Discovering Urban Travel Demands Through Dynamic Zone Correlation in Location-Based Social Networks

Wangsu Hu[1], Zijun Yao[2], Sen Yang[1], Shuhong Chen[1], and Peter J. Jin[1(✉)]

[1] Rutgers University, New Brunswick, USA
{wh251,sy358,sc1624,peter.j.jin}@rutgers.edu
[2] IBM Thomas J. Watson Research Center, Yorktown, USA
zijun.yao@ibm.com

Abstract. Location-Based Social Networks (LBSN), which enable mobile users to announce their locations by checking-in to Points-of-Interests (POI), has accumulated a huge amount of user-POI interaction data. Compared to traditional sensor data, check-in data provides the much-needed information about trip purpose, which is critical to motivate human mobility but was not available for travel demand studies. In this paper, we aim to exploit the rich check-in data to model dynamic travel demands in urban areas, which can support a wide variety of mobile business solutions. Specifically, we first profile the functionality of city zones using the categorical density of POIs. Second, we use a Hawkes Process-based State-Space formulation to model the dynamic trip arrival patterns based on check-in arrival patterns. Third, we developed a joint model that integrates Pearson Product-Moment Correlation (PPMC) analysis into zone gravity modeling to perform dynamic Origin-Destination (OD) prediction. Last, we validated our methods using real-world LBSN and transportation data of New York City. The experimental results demonstrate the effectiveness of the proposed method for modeling dynamic urban travel demands. Our method achieves a significant improvement on OD prediction compared to baselines. Code related to this paper is available at: https://github.com/nicholasadam/PKDD2018-dynamic-zone-correlation.

Keywords: Origin-Destination (OD) analysis
Travel demand prediction · Location-Based Social Networks

1 Introduction

Due to the ubiquity of smartphone and the pervasiveness of social media, there have been rapid developments in Location-Based Social Networks (LBSN) research. Using LBSN, mobile users are able to check-in to Points-of-Interest (POI) for sharing their experiences and enjoying a variety of location-based services such as POI recommendation. This type of check-in data shows the presence

M. Berlingerio et al. (Eds.): ECML PKDD 2018, LNAI 11052, pp. 88–104, 2019.
https://doi.org/10.1007/978-3-030-10928-8_6

of users at different locations, and so can reveal large-scale human mobility in urban areas over time. With proper analysis, check-in data can be a rich source of intelligence for supporting real-time decision making in smart city applications.

While literature has shown the promising effectiveness of analyzing urban area check-in data in many applications (such as POI recommendation, business demand forecasting, human mobility analysis, and users' behavior prediction [1–4]), there are limited studies exploiting check-in data for modeling urban travel demands across city zones. Traditional transportation studies mainly use sensor-based approaches for travel demand analysis [5]. However, sensor data usually does not contain trip purpose information which is the fundamental motivation for traveling. Recent studies have shown that check-in data can be an alternative source to sense both static and time-of-day Origin-Destination (OD) flow patterns [6]. This motivates investigation into exploiting check-in data to model dynamic travel demands in urban areas.

The work presented in this study is firstly motivated by the need to fully utilize the trip purpose information revealed by check-in data. Trip purpose greatly influences a traveler's decisions, such as whether to travel, the departure time, and the destination choice [7]. Since mobile users check-in to POIs of different tags (multiple-level categorical information - e.g., nightlife spot), their check-ins can consistently expose their trip purposes (e.g., recreation). By aggregating all the individual-level check-ins at the level of zones and correlating trip purpose portfolios, we can discover the OD flow patterns between zones. Second, the proposed method is motivated by the temporal effects of travel demands. OD flow patterns between zones changes across time, but most of the existing work only provides static results. We instead propose to learn a time-aware correlation between zones by modeling dynamic OD flow patterns. To achieve this, we incorporate transportation models (i.e., zone gravity modeling and radiation modeling) into our dynamic OD analysis framework. Lastly, this work is motivated by the uniqueness of zonal analysis applications. Zonal OD flow patterns study the destination choices at an aggregated level (such as census tracts, neighborhoods, or regions) for supporting urban planning, traffic operation, and transportation management [5]. However, most check-in modeling approaches focus on individual-level prediction; therefore, they are difficult to directly apply to zonal analysis applications. To overcome this challenge, we propose to use the dynamic variations of check-in activities at different zones and different times to study zonal OD flow patterns. For example, we may see a decrease of check-in activities at residential areas followed by an increase at workplaces area in the morning period. From this observation, we may infer the AM Peak commuting trip from residential zones to commercial zones.

Alone this line, in this paper, we develop a dynamic framework to predict the time-aware travel demands across city zones through dynamic zone correlation modeling using massive mobile check-ins. Depart from prior works [3] which applied check-in data to estimate zonal trip arrivals, the proposed models move one-step forward to generate the dynamic OD flow patterns among zonal check-in activities. Specifically, we first introduce a profiling method to infer the functionality of city zones based on the POI categorical distribution and

density. Second, we adopt a Hawkes Process-based State-Space (HPSS) formulation to infer the dynamic trip arrivals using check-in arrivals. Moreover, in order to capture the travel demands for any pair of zones, we develope a dynamic OD flow predictive framework called Pearson Product-Moment Correlation Gravity Model (PPMC-GM) which consists of dynamic zone correlation identification and zone gravity optimization. Last, we conduct experiments to validate the performance of PPMC-GM using a real-world dataset collected from Foursquare. The experimental results indicate that PPMC-GM outperforms other state-of-the-art baselines at predicting OD flow patterns. In addition, we reported qualitative results which study the pairwise time-of-day OD flow patterns between zones of different functionality.

2 Related Work

There are two areas related to this study: OD estimation models and check-in data-based spatial-temporal correlation models.

2.1 OD Estimation Model

OD estimation models are crucial for understanding the mechanisms of human mobility. There are three major approaches for predicting the OD flow patterns at an aggregated zonal level: the gravity model, the intervening opportunities model, and the regression model. The gravity model, inspired by the Newton's law of gravitation, assumes that the amount of trips between two locations is proportional to their populations and decays by a function of the distance [8]. The gravity model was originally applied to model large-scale migration patterns, then has been widely used for predicting destination choices of network users [6,9,10]. These models were able to replicate the OD patterns on static scenarios to reveal insightful temporal and spatial patterns. Jin *et al.* [6] predicted the OD flows over a target time period to reveal a static urban travel demand pattern. The model was developed for offline planning purposes and the prediction required recalibration to satisfy different scenarios. In contrast to the gravity model, Stouffer [11] proposed the intervening opportunities model, which assumes that individuals are more attracted to locations with higher interest (i.e. employment opportunities, venue capacities) rather than closer distance. Since then, extensive studies have been conducted to model human mobility patterns [12–14]. These models were applied to either model static zonal OD flow patterns or recommend individual-level locations. The third approach deploys regression. Regression requires survey-based partial OD data [15–17], which includes spatial-temporal information of participating travelers (i.e., GPS survey data, cellular data, and location-based services data). The real-time requirements are met by an auto-regressive process [17].

2.2 Spatial-Temporal 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 [18–20]. 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 [21–24] 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 [25–27] 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*. [28] proposed a time-aware Gaussian Mixture model combining periodic short-range movements and sporadic long-distance travels. Wang *et al*. [29] provided a Regularity Conformity Heterogeneous (RCH) model to predict user location at specific times, considering both regularity and conformity. Lian *et al*. [30] incorporated temporal regularity into a Hidden Markov framework to predict regular user locations. Finally, taking advantage of sequential patterns in human movement [28], various sequential mining techniques [4,29,31] 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.

3 Methodology Overview

In this section, we first define some basic concepts in our method. Then we formulate the problem of zonal travel demand modeling. Finally we show the overview of the proposed dynamic OD flow pattern matching based framework.

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 Problem Statement

Let $X = \{x_1, x_2, ..., x_n\}$ be a set of check-ins where each check-in has a location (e.g., latitude and longitude), a timestamp and a POI category. Let $Z = \{z_1, z_2, ..., z_m\}$ be a set of zones. We aggregate check-ins by each zones and store them as a 3-dimensional tensor V indexed by zone, category, and timeslot. Each entry of V indicates the number of observed check-ins. Our goal is to predict the 3-dimensional OD flow tensor F indexed by origin zone, destination zone, and timeslot. Each value in F indicates the number of trips from a origin zone to a destination zone during a timeslot.

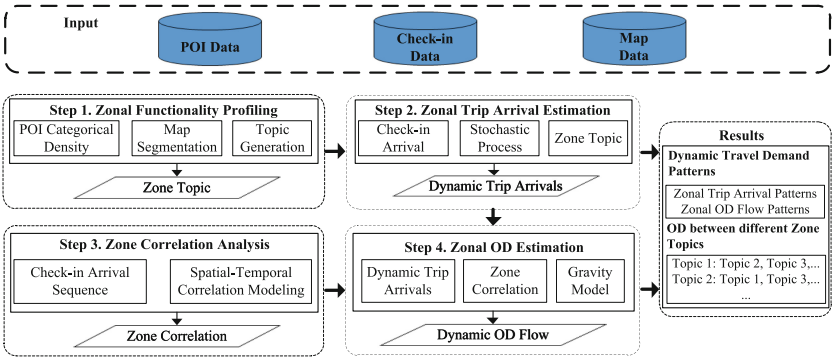

Fig. 1. An overview of our dynamic travel demand estimation framework.

3.3 General Framework

We propose a procedure for estimating the dynamic OD flow patterns (Fig. 1), including four major parts: zonal functionality profiling, zonal trip arrival estimation, zone correlation analysis, and zonal OD flow pattern estimation.

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.

For zonal trip arrival estimation, we applied HPSS formulation to model the trip arrival patterns based on the check-in arrival patterns and the discovered zone topics. The predicted trip arrivals are a form of dynamic urban travel demand patterns. The result is then applied as the input to the dynamic zonal OD estimation model.

For zone correlation analysis, we calculated the PPMC matrix to represent the pairwise zone correlation based on check-in arrival sequence. The result is then used in zonal OD estimation.

Finally, given the gravity model framework, we proposed a joint PPMC-GM model to predict the dynamic OD flow patterns. The zone topics will be

combined with the predicted OD flow patterns to discover time-of-day variations between pairwise zones of different zonal functionality.

4 A Dynamic Travel Demand Model

In this section, we describe the four components of the proposed dynamic travel demand estimation framework.

4.1 Zonal Functionality Profiling

To infer the zonal functionality, we introduced M zone topics 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:

$$d_{i,c} = \frac{numbers\ of\ POI_{c \in C}}{area\ of\ zone\ i}. \tag{1}$$

Using Latent Dirichlet allocation (LDA) method, we treat the zone functionalities $m \in \{1, ..., M\}$ as the document topics, the zones as documents (each POI distribution D_i as one documentation), and POI categorical density $d_{i,c}$ as words in the documentation. Then the zone functionality can be uncovered from the POIs. We construct a hierarchical topic model as following:

$$\theta_i \sim Dir(\eta_1); Multinomial(z_{i,d_{i,c}}, \theta_i); \varphi_m \sim Dir(\eta_2); Multinomial(\omega_{i,d_{i,c}}, \varphi_m), \tag{2}$$

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, $z_{i,d_{i,c}}$ and $\omega_{i,d_{i,c}}$ are the chosen topic and word for zone i. Then the probability of POIs within one zone being covered by zone topic m is:

$$P(m|D_i) = \prod_m P(\varphi_m|\eta_2) \prod_i P(\theta_i|\eta_1) \prod_c P(z_{i,d_{i,c}}|\theta_i) P(\omega_{i,d_{i,c}}|\varphi_{1:M}, z_{i,d_{i,c}}). \tag{3}$$

4.2 Zonal Trip Arrival Estimation

Efforts have been made to model zonal trip arrival based on check-in data [3]. Here, we leverage the HPSS formulation to predict zonal trip arrival patterns based on the discovered zone topics and the check-in arrival patterns. The method includes a state equation and an observation equation. The state equation represented in Eq. 4 introduces every check-in arrival as an event occurrence following the Hawkes point process. Then an observation equation represented in Eq. 5 generates zonal trip arrivals, which are compared with the observed check-in arrivals. Through the state-space model framework, check-in arrivals

can then be coupled with the Hawkes process model to estimate dynamic zonal trip arrivals. The zonal trip arrival modeling can be represented as follows:

$$\lambda_{i,m}^{t} = \frac{\tilde{A}_i^{t-\Delta t}}{\Delta t} + \alpha_{i,m}^{t} \sum_{t:t_1<t_2} exp(-\beta_{i,m}^{t}(t_1 - t_2)) \tag{4}$$

$$\tilde{A}_i^{t} = \delta_{i,m}^{t}(\int_{t}^{t+\Delta t} \lambda_{i,m}^{t} dn) + (1 - \delta_{i,m}^{t})\gamma_{i,m}^{t} x_i^{t}, \tag{5}$$

where $\lambda_{i,m}^{t}$ represents a counting process for trip arrivals at zone i at time slot t for zone topic m, $\tilde{A}_i^{t-\Delta t}$ is the predicted zonal trip arrivals at zone i at previous time slot $t - \Delta t$, $\tilde{A}_i^{t}$ is the predicted zonal trip arrivals at zone i at time slot t, x_i^{t} is check-in arrivals at zone i at time slot t, n is the predicted sequence of trip arrival occurrence, and the set of parameters $(\alpha, \beta, \gamma, \delta)$ to be calibrated. Given the ground truth data of zonal trip arrival A_i and the time-of-day arrival A^t, we applied the following objective function for parameter estimation:

$$\min_{\alpha,\beta,\gamma,\delta} (\sum_{i} abs((\sum_{t} \tilde{A}_i^{t}) - A_i)) + \sum_{t} abs((\sum_{i} \tilde{A}_i^{t}) - A^t). \tag{6}$$

4.3 Zone Correlation Analysis

We used PPMC to measure the zone correlation. Let s_i^t represent check-in arrival sequences that consist of a set of check-in arrivals $\{x_i^t, x_i^{t+1}, \ldots, x_i^{t+w}\}$. Then we normalized s_i^t as follows:

$$f_i^t(w) = \frac{s_i^t(w) - \mu}{\sigma}, \tag{7}$$

where w is the selected sequence length to be calibrated, μ and σ is the mean and standard deviation of sequence, respectively. The zone correlation was quantified as follows:

$$RCC_{ij}^{td} = f_i^t(w) f_j^d(w), \tag{8}$$

where RCC_{ij}^{td} is a vector that contains $(2w - 1)$ elements and $d = t + \tau$ indicates the time slot after a POI-category-dependent time delay τ. Each element of RCC_{ij}^{td} has a value between $+1$ and -1 inclusive, where 1 indicates linear correlation, and 0 indicates no correlation. We selected element as the correlation coefficient rcc_{ij}^{td} based on:

$$rcc_{ij}^{td} = \max_{k \in 2w-1} abs(\{RCC_{ij}^{td}\}) \; s.t. RCC_{ij}^{td}(k) >= r_{threshold}, \tag{9}$$

where $r_{threshold}$ is the threshold to be calibrated, $I_{[clause]}$ is an indicator function for a logic clause. The PPMC procedure's purpose is to generate the rcc_{ij}^{td} followed Algorithm 1.

Algorithm 1. Zone Correlation Identification Learning

Require: Identification of zone spatial-temporal correlation
Input: $x_{i,c}^t$ # Check-in arrivals at zone i at time slot t.
Output: rcc_{ij}^{td} # PPMC coefficient for zone correlation between zone i at time slot t and zone j at time slot d.

Initialization $\forall rcc_{ij}^{td} = 0$
for $t \in T$ **do** # Let T be a set of time slots.
  for $i \in N$ **do** # Let N be a set of zones
    $s_i^t = \{x_i^t, x_i^{t+1}, \ldots, x_i^{t+w}\}$
    for $j \in N$ **do**
      $d = t + \tau$
      $s_j^d = \{x_j^d, x_j^{d+1}, \ldots, x_j^{d+w}\}$
      Generate correlation coefficient vector RCC_{ij}^{td} with the normalized s_i^t, s_j^d
      Select one element of RCC_{ij}^{td} as rcc_{ij}^{td}
    end for
  end for
end for
return all rcc_{ij}^{td}

4.4 Zonal OD Estimation

We propose a joint PPMC-Gravity Model (PPMC-GM) to predict the dynamic OD flows. The gravity model assumes knowledge of travel cost when calculating relative zone attractiveness. Normally, such travel cost is a function of travel distance or travel time. When integrated with the inflow and outflow of travelers, the GM model predicts the traffic flows from one zone to another. Considering both the HPSS formulation and PPMC coefficients, 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 follows:

$$P_{ij}^{td} = \frac{A_i^t A_j^d g(rcc_{ij}^{td})}{\sum_j A_j^d g(rcc_{ij}^{td})}, \tag{10}$$

where P_{ij}^{td} stands for the probability of a trip 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 using the PPMC coefficient. Given the ground truth OD flow matrix $F = \{f_{ij}\}$, we sample $N = \sum_{i,j} f_{ij}$ trips following the predicted probability p_{ij}^{td} and additional constraints to generate the OD flow matrix $\tilde{F} = \{\tilde{f}_{ij}\}$. We consider four different types of constraints on the proposed joint PPMC-GM model:

Unconstrained model (UM). The only constraint on UM is that 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 ground truth data. The $\tilde{N}$ trips are randomly sampled from the multinomial distribution:

$$Multinomial(N, (P_{ij}^{td})). \tag{11}$$

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

$$Multinomial(O_i, \frac{\sum_{t,d} P_{ij}^{td}}{\sum_{j,t,d} P_{ij}^{td}}). \tag{12}$$

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

$$Multinomial(D_j, \frac{\sum_{t,d} P_{ij}^{td}}{\sum_{i,t,d} P_{ij}^{td}}). \tag{13}$$

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

$$\tilde{f}_{ij}^{td} = B_i B_j P_{ij}^{td}; \sum_{j,t,d} \tilde{f}_{ij}^{td} = O_i; \sum_{i,t,d} \tilde{f}_{ij}^{td} = D_j; Multinomial(N, \frac{\sum_{t,d} \tilde{f}_{ij}^{td}}{\sum_{i,j,t,d} \tilde{f}_{ij}^{td}}), \tag{14}$$

where B are balancing factors from Iterative Proportional Fitting procedure [32].

5 Evaluation

5.1 Experiment Setting

Experimental Datasets. Manhattan Island of New York City was selected as the study area. We used the following datasets to evaluate our approach.

POI data. The POI dataset obatined from Foursquare covers 96,263 POIs of Manhattan. Each POI was recorded with location information (i.e., latitude and longitude) and category. The POI category is given in a comprehensive multiple-level classification provided by Foursquare [33]. We chose the first level, as the second level had more than 100 categories, making each category too trivial for our analysis. The selected 9 POI categories were: Nightlife Spot, Food, Shop & Service, College & University, Arts & Entertainment, Travel & Transport, Professional & Other Places, Outdoors & Recreation, and Residence.

Check-in data. We extracted check-in arrivals and their sequences from 1,168, 073 anonymous records collected in Manhanttan between August 1st, 2016 and March 31st, 2017 to extract check-in arrivals and check-in arrival sequences. Each check-in data contains spatial-temporal information demonstrating where and when it was generated.

Map segmentation. The study area can be partitioned into zones with different methods, e.g., grid-based or road network-based [18]. Based on the spatial resolution of the ground truth OD data, we selected census tracts as our zone partitions. As a result, we obtained 318 zones in the study area.

Travel demand observation data. The ground truth data includes weekday zonal OD matrices and time-of-day arrival of year 2017 provided by the New York Metropolitan Transportation Council.

We seperate the check-in data and ground truth data into two datasets, one for model training and the other for testing. The training set contains a random 50 out of the 318 zones in Manhattan. The learned parameters were then evaluated on the testing set. All 318 zones are included to ensure complete visualization and analysis of the full mobility pattern in the study area.

Baseline Methods. 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 [34], the probability of a trip between two zones 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 $cost_{ij}$, which is modeled with an exponential distance decay function:

$$P_{ij} = \frac{O_i D_j g(cost_{ij})}{\sum_j D_j g(cost_{ij})}; g(cost_{ij}) = exp(-\beta distance_{ij}). \tag{15}$$

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

$$P_{ij} = \frac{O_i D_j g(cost_{ij})}{\sum_j D_j g(cost_{ij})}; g(cost_{ij}) = (distance_{ij})^{-\beta}. \tag{16}$$

Schneider Intervening Opportunity Model (Schneider). In this model, the probability of a trip between two zones is proportional 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 [12]:

$$P_{ij} = O_i \frac{exp(-\beta S_{ij}) - exp(-\beta(S_{ij} + O_i))}{\sum_j exp(-\beta S_{ij}) - exp(-\beta(S_{ij} + O_i))}. \tag{17}$$

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

$$P(1|O_i D_j, S_{ij}) = \frac{O_i D_j}{(O_i + S_{ij})(O_i + D_j + S_{ij})}. \tag{18}$$

Evaluation Measurement. We used Mean Absolute Error (MAE), Normalized Root Mean Square Error (NRMSE), and Coincidence Ratios (CR) [36] as metrics to evaluate the performance of zonal OD estimation:

$$MAE = \frac{\sum_{i,j} abs(f_{ij} - \tilde{f}_{ij})}{\sum_{i,j} 1}; NRMSE = \frac{\sum_{i,j} abs(f_{ij} - \tilde{f}_{ij})^2}{\sum_{i,j} f_{ij}} \tag{19}$$

$$CR = \frac{\sum_k min(\tilde{tl}_{distance_k}, tl_{distance_k})}{\sum_k min(\tilde{tl}_{distance_k}, tl_{distance_k})}, \tag{20}$$

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.

5.2 Experimental Results

Zonal Functionality Profiling. To uncover the zonal functionality, we generated 5 latent zone topics based on the POI distribution under 9 POI categories shown in Table 1. We found that the POIs such as "Shop & Service","Food", and "Art & Entertainment" 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. Therefore, we considered the POI categories containing not only the highest but also the 2nd or 3rd probability to determine the zonal functionality. Five land use types guided by the New York City zoning and land use data [37] were selected as zone topic labels: "Commercial-Retail", "Commercial-Work", "Residence", "Transportation", and"Open Space". We mapped out in Fig. 2 the distribution of POIs and zone topics. We observed that POIs were most densely distributed in the Midtown and lower Manhattan area, while few were observed in the ring area of Central Park and Upper Manhattan. The zone topics discovered from the POI data indeed resemble the functional diversity of Manhattan's census tracts: commercial-work area for "Financial District", open space area "Central Park", and residential area in "Upper West Side".

Table 1. Zonal topic profiling.

Topic 1	Prob.	Topic 2	Prob.	Topic 3	Prob.	Topic 4	Prob.	Topic 5	Prob.
S	0.395	S	0.298	S	0.197	F	0.191	S	0.167
N	0.170	T	0.184	F	0.154	A	0.145	P	0.161
T	0.119	P	0.157	R	0.092	O	0.133	F	0.139
F	0.102	S	0.146	T	0.067	N	0.108	T	0.113
P	0.098	O	0.125	O	0.063	C	0.071	A	0.109
R	0.094	N	0.124	P	0.027	T	0.066	O	0.100
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.

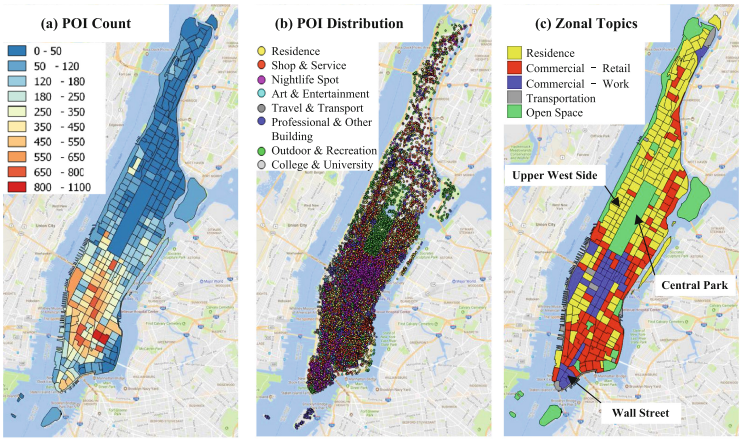

Fig. 2. Spatial distribution of the collected POIs and the generated zonal topics. (a) A heat map representing the POI categorical density. (b) We colored each POI based on the POI categories. (c) The zonal topics distribution generated by LDA. (Color figure online)

Zonal Trip Arrival Estimation. Regarding the mimicking of the trip arrival patterns in the ground truth data, the predicted trip arrivals from HPSS formulation were aggregated hourly to generate the trip arrival patterns in Fig. 3. 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. The regression line has a slope of 0.66, and the $R^2 = 0.80$ under the statistically significant level $p - value < 0.01$.

Zonal OD Estimation. We evaluated the generated daily OD flow patterns against four baseline models using three indicators: MAE, NRMSE, and CR. 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 curve between the ground truth and predicted OD flow patterns. The performance of five OD estimation models is presented under four different constraints as shown in Fig. 4. A total 20 model-constraint combinations were explored. 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

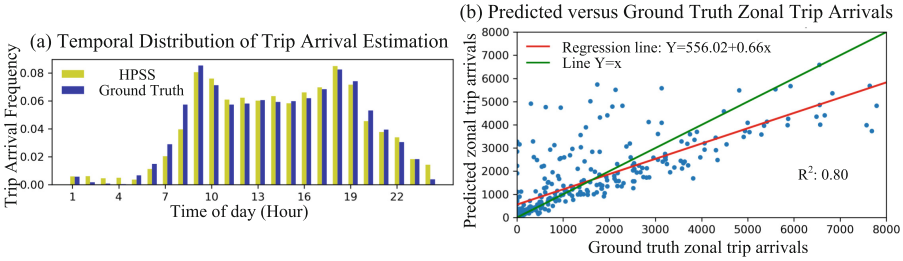

Fig. 3. Model Performance of trip arrival estimation. (a) blue bar indicates the ground truth temporal distribution of trip arrival over the study area; yellow bar indicates the predicted one. The model performance for specific zones can be found in Fig. 5; (b) Regression analysis for the zonal trip arrival prediction over the entire day. Each dot represents the reference value as x coordinate and the predicted value as y coordinate. (Color figure online)

distribution. Globally, the result obtained with the proposed PPMC-GM model achieves the lowest MAE/NRMSE value and the highest CR.

OD Flow Patterns Under Zone Topics. Since the travel demand observation data only has uniform sets of time-of-day arrival for each trip purpose, it cannot provide the time-of-day pattern for different land use type. The proposed model can report the time-of-day OD flow patterns under different zone topics. There are a total of 5 discovered zone topics shown in Fig. 5 including residential area (R), transportation hub area (TH), open space area (OS), commercial-retail area (CR), and commercial-workplace area (CW). Given the six selected zones representing five different zone topics, we evaluate different outflow patterns.

Residential area. Zone 81 is selected as one typical residential (R) area to explore its outflow pattern to other zones during the weekday. A morning peak period can be observed in the outflow of the commercial-workplace (CW) and the transportation hub (TH) area; this may reflect morning commuting activities. We notice that there is a time delay between the AM peak of the R-CW trips compared with that of the R-TH trips. This is consistent with the transit stop during commuters' trip from their homes to the office. Meanwhile, the outflows to CR area and OS area increase into the day starting from the late morning period. This is consistent with typical starting times of trips attempting to avoid rush hours. Finally, for R-CR trips, another two fluctuations can be observed in mid-afternoon and evening for late afternoon shoppers and dinnertime activities.

Commercial-Retail area. Zone 110 contains the major landmark "Time Square" in a commercial area. Trip patterns originating from this zone are analyzed across different destinations. For CR-R trips, the model did capture peaks indicating home-returning activities before and after dinnertime. Meanwhile, travelers leaving the nightlife spots and other recreational attractions within the targeted zone also generated significant late-night outflow trips to the residential area. For CR-TH trips, the afternoon peak to the transportation hub area

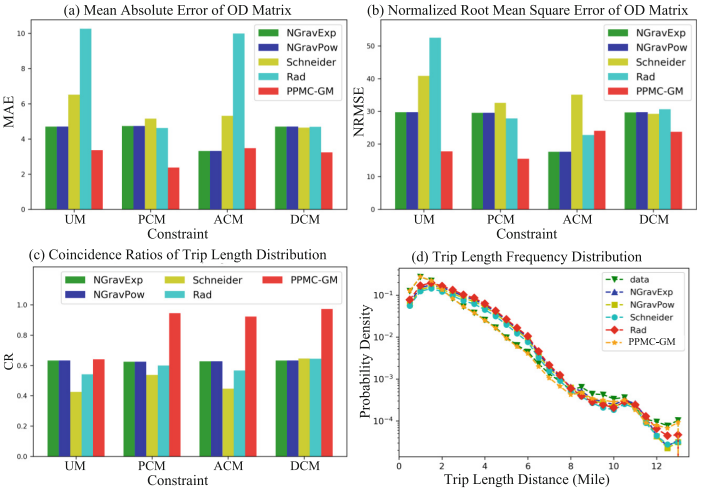

Fig. 4. Performance of the unconstrained model (UM), the production constrained model (PCM), the attraction constrained model (ACM) and the doubly constrained model (DCM) according to 4 baseline models and the proposed model. (a) Average MAE. (b) Average NRMSE. (c) Average CR. (d) Average trip length distribution curve. 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 in (d).

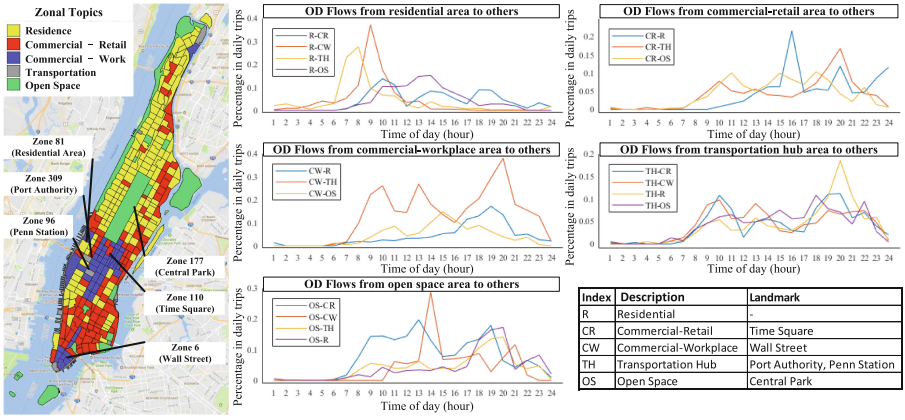

Fig. 5. Sample distribution of time-of-day OD flow patterns. (a) the selected zones and landmark involved to represent different zone topics. (b) the outflow from one zone under one zone topic to others under different zone topics.

coincides with afternoon commuting activities. CR-OS trips maintain an average rate during the daytime, then decrease when entering the night.

Commercial-Workplace area. Zone 5 is located in the Financial District of Manhattan. It contains typical workplaces along "Wall Street". The outflow patterns clearly show that 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 lunchtime peak can be seen for CW-OS trips. This may reflect people resting in the nearby open space area and recreational area during lunch breaks.

Zones containing Transportation Hubs. Zone 309 and Zone 96 contain "Port Authority Bus Terminal" and "Penn Station", respectively. Both are representative transportation hubs. For TH-CW trips, an AM peak is captured by the model indicating 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. Trends are noticeably not aligned with the morning and afternoon commuting rush hours, since the users here consist of mostly 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.

Open Space area. Zone 177 is fully occupied by "Central Park", and 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 home-returning activities. The OS-CR trips exhibit high flow during most of the morning and early afternoon and another peak around dinnertime. This partially reflects the touring trip chains such as visiting retail shops after visiting central park.

6 Conclusion

In this study, we presented a joint PPMC-GM model to generate dynamic OD flow patterns based on check-in data. The model explored adopting the spatial-temporal correlation coefficient into the traditional gravity model to evolve a dynamic OD estimation. We applied check-in data collected from the Foursquare platform in Manhattan Island area. The evaluation showed promising results with low MAE/NRMSE values and high CR values compared to the baseline models. Furthermore, several empirical insights were obtained by analyzing the dynamic OD patterns between zones of different functionality.

References

1. Liu, B., Xiong, H.: Point-of-interest recommendation in location based social networks with topic and location awareness. In: SDM, pp. 396–404. SIAM (2013)
2. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. In: ICWSM, pp. 81–88 (2011)
3. Hu, W., Jin, P.J.: An adaptive hawkes process formulation for estimating zonal trip arrivals with LBSN data. TRP-C: Emerg. Technol. 79, 136–155 (2017)
4. Chong, W.-H., Dai, B.-T., Lim, E.-P.: Prediction of venues in foursquare using flipped topic models. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 623–634. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_69
5. de Grange, L., Fernández, E., de Cea, J.: A consolidated model of trip distribution. TRP Part E **46**(1), 61–75 (2010)
6. Jin, P., Cebelak, M., Yang, F., Zhang, J., Walton, C., Ran, B.: Exploration into use of doubly constrained gravity model for OD estimation. TRR **2430**(1), 72–82 (2014)
7. Elldér, E.: Residential location and daily travel distances: the influence of trip purpose. J. Transp. Geogr. **34**, 121–130 (2014)
8. Carey, H.C.: Principles of Social Science, vol. 3. JB Lippincott & Company, Philadelphia (1867)
9. Medina, A., Taft, N., Salamatian, K., Bhattacharyya, S., Diot, C.: Traffic matrix estimation: existing techniques and new directions. In: SIGCOMM, vol. 32, no. 4, pp. 161–174 (2002)
10. Zhang, J.D., Chow, C.Y.: Spatiotemporal sequential influence modeling for location recommendations: a gravity-based approach. ACM TIST **7**(1), 11 (2015)
11. Stouffer, S.A.: Intervening opportunities: a theory relating mobility and distance. Am. Sociol. Rev. **5**(6), 845–867 (1940)
12. Schneider, M.: Gravity models and trip distribution theory. Reg. Sci. **5**(1), 51–56 (1959)
13. McArdle, G., Lawlor, A., Furey, E., Pozdnoukhov, A.: City-scale traffic simulation from digital footprints. In: SIGKDD UrbComp, pp. 47–54. ACM (2012)
14. Tarasov, A., Kling, F., Pozdnoukhov, A.: Prediction of user location using the radiation model and social check-ins. In: SIGKDD UrbComp, p. 8. ACM (2013)
15. Lee, J.H., Gao, S., Goulias, K.G.: Can Twitter data be used to validate travel demand models. In: IATBR (2015)
16. Bierlaire, M., Crittin, F.: An efficient algorithm for real-time estimation and prediction of dynamic OD tables. Oper. Res. **52**(1), 116–127 (2004)
17. Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-time urban monitoring using cell phones: a case study in Rome. IEEE Trans. Intell. Transp. Syst. **12**(1), 141–151 (2011)
18. Liu, B., Fu, Y., Yao, Z., Xiong, H.: Learning geographical preferences for point-of-interest recommendation. In: SIGKDD, pp. 1043–1051. ACM (2013)
19. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and POIs. In: SIGKDD, pp. 186–194. ACM (2012)
20. Shi, Y., Serdyukov, P., Hanjalic, A., Larson, M.: Nontrivial landmark recommendation using geotagged photos. ACM TIST **4**(3), 47 (2013)
21. Bao, J., Zheng, Y., Mokbel, M.F.: Location-based and preference-aware recommendation using geosocial network data. In: SIGSPATIAL, pp. 199–208. ACM (2012)

22. Ference, G., Ye, M., Lee, W.C.: Location recommendation for out-of-town users in location-based social networks. In: CIKM, pp. 721–726, ACM (2013)
23. Wang, H., Terrovitis, M., Mamoulis, N.: Location recommendation in location-based social networks using user check-in data. In: SIGSPATIAL, pp. 374–383. ACM (2013)
24. Ying, J.J.C., Kuo, W.N., Tseng, V.S., Lu, E.H.C.: Mining user check-in behavior with a random walk for urban POI recommendations. ACM TIST **5**(3), 40 (2014)
25. Gao, H., Tang, J., Hu, X., Liu, H.: Exploring temporal effects for location recommendation on location-based social networks. In: RecSys, pp. 93–100. ACM (2013)
26. Yuan, Q., Cong, G., Ma, Z., Sun, A., Thalmann, N.M.: Time-aware point-of-interest recommendation. In: SIGIR, pp. 363–372. ACM (2013)
27. Yuan, Q., Cong, G., Sun, A.: Graph-based point-of-interest recommendation with geographical and temporal influences. In: CIKM, pp. 659–668. ACM (2014)
28. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: SIGKDD, pp. 1082–1090. ACM (2011)
29. Wang, Y., et al.: Location prediction using heterogeneous mobility data. In: SIGKDD, pp. 1275–1284. ACM (2015)
30. Lian, D., Xie, X., Zheng, V.W., Yuan, N.J., Zhang, F., Chen, E.: A collaborative exploration and returning model for location prediction. ACM TIST **6**(1), 8 (2015)
31. Li, X., Lian, D., Xie, X., Sun, G.: Lifting the predictability of human mobility on activity trajectories. In: ICDMW, pp. 1063–1069. IEEE (2015)
32. Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Stat. **11**(4), 427–444 (1940)
33. Foursquare. https://developer.foursquare.com/docs/resources/categories
34. Barthélemy, M.: Spatial networks. Phys. Rep. **499**(1–3), 1–101 (2011)
35. Simini, F., González, M.C., Maritan, A., Barabási, A.L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96 (2012)
36. Martin, W.A., McGuckin, N.A.: Travel Estimation Techniques for Urban Planning, vol. 365. National Academy Press, Washington (1998)
37. Zola. http://maps.nyc.gov/doitt/nycitymap/template?applicationName=ZOLA