Point-of-Interest Demand Modeling with Human Mobility Patterns

Yanchi Liu†
Rutgers University
yanchi.liu@rutgers.edu

Chuanren Liu†
Drexel University
chuanren.liu@drexel.edu

Xinjiang Lu
Northwestern Polytechnical University, China
xjlu@mail.nwpu.edu.cn

Mingfei Teng
Rutgers University
mingfei.teng@rutgers.edu

Hengshu Zhu
Baidu Talent Intelligence Center
zhuhengshu@baidu.com

Hui Xiong*
Rutgers University
hxiong@rutgers.edu

ABSTRACT

Point-of-Interest (POI) demand modeling in urban regions is critical for many applications such as business site selection and real estate investment. While some efforts have been made for the demand analysis of some specific POI categories, such as restaurants, it lacks systematic means to support POI demand modeling. To this end, in this paper, we develop a systematic POI demand modeling framework, named Region POI Demand Identification (RPDI), to model POI demands by exploiting the daily needs of people identified from their large-scale mobility data. Specifically, we first partition the urban space into spatially differentiated neighborhood regions formed by many small local communities. Then, the daily activity patterns of people traveling in the city will be extracted from human mobility data. Since the trip activities, even aggregated, are sparse and insufficient to directly identify the POI demands, especially for underdeveloped regions, we develop a latent factor model that integrates human mobility data, POI profiles, and demographic data to robustly model the POI demand of urban regions in a holistic way. In this model, POI preferences and supplies are used together with demographic features to estimate the POI demands simultaneously for all the urban regions interconnected in the city. Moreover, we also design efficient algorithms to optimize the latent model for large-scale data. Finally, experimental results on real-world data in New York City (NYC) show that our method is effective for identifying POI demands for different regions.

CCS CONCEPTS

• **Information systems** → **Location based services**; *Geographic information systems*; • **Computing methodologies** → **Factor analysis**; • **Human-centered computing** → *Mobile computing*;

KEYWORDS

Region demand; Point-of-Interest; human mobility

1 INTRODUCTION

Identification of POI demands in spatially differentiated regions is fundamental for governments, also it is critical for the survival of local businesses. In some cases, failing to estimate demands properly is enough to force a company to go out of business. The demand analysis helps the entrepreneur and the authority in making decisions for the efficient allocation of limited resources, such as business site selection, real estate investment and land use planning [31]. Take site selection for example, a business will have high chance to success if it is placed in a highly-demanded region, otherwise it may result in serious business risk and even failure. In this paper, we aim to provide a systematic POI demand analysis for urban regions with the help of large-scale human mobility data.

Currently, local businesses and governments largely rely on labor-intensive surveys to inform their decision-making. For example, to understand region demands, a series of research has been done based on survey data [21, 22]. However, the information obtained through the surveys may not be sufficient and timely enough. Recently, the wide availability of Information Communications Technology (ICT) has enabled unprecedented opportunities to collect large-scale human mobility data, e.g., taxi GPS traces, which is able to cover the whole urban area with fine-grained time and location information. It reflects the underlying dynamics of residents in the city, which is much more detailed, and has larger scale than survey data. While more and more efforts have been put into analyzing the large-scale human mobility data to understand urban dynamics, few of them provide a systematic POI demand modeling but focus on specific POI categories such as restaurant and gas station [11, 20]. Actually, *people vote with their feet*, which is an important economic logic [24], indicates that people have the ability to choose what they need by traveling. For example, as illustrated in Figure 1, if people from one region frequently travel to other regions for restaurants and doctors, it is much likely people need new or improved restaurants and health services in this region. Therefore, the human mobility patterns can be utilized to identify daily needs of people, and provide governments and local businesses with opportunities to better understand POI demand and formulate future planning.

†Both authors contributed equally to this manuscript. *Contact author.

KDD '17, August 13-17, 2017, Halifax, NS, Canada

ACM ISBN 978-1-4503-4887-4/17/08…$15.00
https://doi.org/10.1145/3097983.3098168

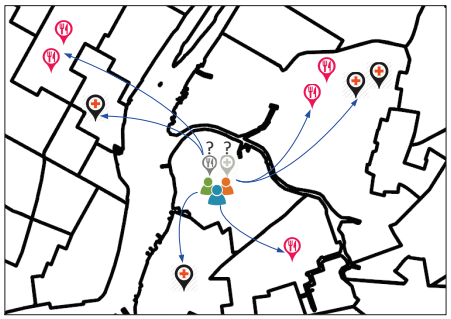


Figure 1: Identification of Region POI Demand

Indeed, in this paper, we investigate how to identify POI demands for urban regions by modeling region POI preferences, region POI supplies, and region demographic features. To the best of our knowledge, this is the first attempt to identify POI demands over different categories in the city scale. The main challenges of this work include: 1) Human mobility data is highly skewed between different regions, even no human mobility data collected for underdeveloped regions which in fact need more attention. Take NYC as an example, Staten Island is one borough of NYC but separated from other boroughs by New York Bay, which makes it difficult to travel by taxi to downtown NYC but by ferry. As shown in Figure 3 (a), the number of taxi trips in Staten Island is much less than other boroughs. 2) How to integrate region POI profiles and demographic data together with human mobility data to better model the POI demand of urban regions in a holistic way. 3) How to learn the demands simultaneously for all the regions in a city with community opinions considered.

To this end, in this paper, we develop a systematic POI demand modeling framework, named Region POI Demand Identification (RPDI), to model POI demands by exploiting the daily needs of people identified from their large-scale mobility data. Specifically, in our proposed framework, we first partition the urban space into spatially differentiated neighborhood regions formed by many small local communities. Then, a Bayesian model [7] considering context POI information like distance, rating, and popularity is exploited to infer trip activities. Besides, we aggregate the trips that origin from the same region and identify the underlying demands of all the regions simultaneously using a latent factor model, which integrates region preferences, POI supplies, and demographic features. Furthermore, we apply the identified demands for region POI demand ranking. Finally, the main contributions of this paper can be summarized as follows:

- A systematic framework, named RPDI, is developed to identify POI demands for urban regions in a city. RPDI is able to identify POI demands over different categories in the city scale with large-scale human mobility data.
- A latent factor model is proposed to integrate region preferences, POI profiles, and demographic features for POI demand modeling. The model helps to identify a set of latent factors, which in turn can be leveraged for learning region demands in a coherent way.
- The RPDI framework has been evaluated on large-scale real-world data for region POI demand identification. The experimental results show that our method outperforms baseline methods in terms of multiple metrics such as Normalized Root-Mean-Square Error (NRMSE), F-measure, and Normalized Discounted Cumulative Gain (NDCG).

2 THE REGION POI DEMAND IDENTIFICATION FRAMEWORK

In this section, we first formally introduce the problem of region POI demand identification, and then provide an overview of our proposed Region POI Demand Identification framework.

2.1 Problem Statement

Assume that we have M POIs denoted by the set $\mathcal{P}$ and N regions in the set $\mathcal{R}$. For simplicity, we let $\mathcal{P} = \{1, 2, \cdots, M\}$, i.e., we use integers to represent the POIs. For the n-th region in time slot $t \in \mathcal{T}$, where $n = 1, 2, \cdots, N$, and $t = 1, 2, \cdots, T$, we have its activity records represented as a vector of visiting probabilities to POIs $x_t^n = (x_{1t}^n, x_{2t}^n, \cdots, x_{Mt}^n)$, which indicates needs of people. Thus all the activities for all the regions in all time slots form a region activity cube $\mathbf{X} = \{x_{pt}^n\}$. However, $\mathbf{X}$ is very sparse since not all the POIs are visited, and moreover, not all the activities are recorded for certain regions. Formally, our idea of demand identification is to recover the visiting probabilities of all the regions using community opinions with region preferences and region supplies considered. With recovered $\hat{x}_{pt}^n$ from the demand inference model, we further derive the region demand d^n at POI level and category level. Then we apply the identified demand for region POI demand ranking, which is one of many applications leveraging POI demand. Table 1 lists some notations used in this paper.

Table 1: Mathematical Notations

Notation	Description
N, M, C, T, D	The number of regions, POIs, POI categories, time slots, dimensions of region features, respectively.
$\mathcal{R}$, $\mathcal{P}$, $\mathcal{C}$, $\mathcal{T}$	The set of regions, POIs, POI categories, and time slots, respectively.
$\mathbf{F}$	The matrix of region features, including POI profiles and demographic features.
$\mathbf{X}$, $\mathbf{Y}$	The region activity cube at POI level and category level, respectively.
K	The dimension of latent factors.
α, υ	The latent demand patterns at POI category level and POI level, respectively.
β	The region coefficients for features.
u	The latent region pattern coefficients.

2.2 Framework Overview

Figure 2 shows the proposed Region POI Demand Identification framework. This framework first segment an urban area into regions which are shared by local communities. Then with the help of collected geographic and demographic data, we are able to extract POI profiles and demographic features for each region. In the mean

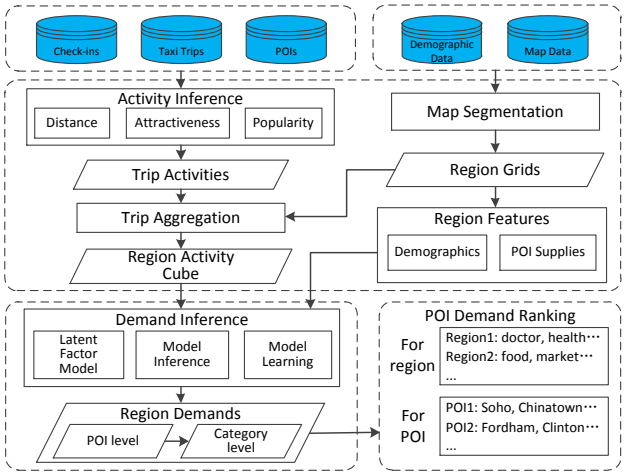


Figure 2: An overview of RPDI framework

time, we infer trip activities from taxi GPS traces with context information considered, then propagate trips to the region level. We treat the activities performed outside the region as the underlying POI demand of this region. With the region activity cube and region features obtained, we model the region demand with a latent factor model, which integrates region preferences, region supplies and demographics. With the proposed demand model, we aim to infer region activities with partial activity information given. We also develop efficient algorithms to optimize the latent model with large-scale data. Finally, we derive the region demands from the learned region activities and further apply them for POI demand ranking, which is useful for various applications such as business site selection and real estate investment.

3 PRELIMINARIES

In this section, we first introduce the region partition for urban area, followed by the details of demographics and POI profiles in regions. At last we introduce how to extract human mobility patterns from taxi GPS traces and POIs.

3.1 Region Partition

The urban area can be partitioned into regions with different methods, e.g., grid-based, road network-based [29]. However, these methods do not take the socioeconomic factor into account. Instead, we use Neighborhood Tabulation Areas (NTAs)[1] provided by New York City government as our region partition method, which is shown in Figure 3. NTAs are created to project populations at a small area level for the long-term sustainability plan for New York City. NTAs are a valuable summary level for use with both the 2010 Census and the American Community Survey (ACS). These geographic areas offer a good compromise between the very detailed data for census tracts (2,168) and the broad strokes provided by community districts (59). As a result, we obtain 195 neighborhood regions which are shared by local communities.

[1]https://www.nyc.gov

3.2 Region Demographics

The POI demand indeed is the demand of people. The different distributions of people in a region will affect the demands significantly. Therefore, the demographic information could be a complement to human mobility patterns to estimate region demands. To this end, we integrate the demographic information collected from the US Census data[1] into our proposed model.

For each neighborhood, population density, sex, age, composition are used to depict population information in that region. Moreover, household income, household type, housing occupancy, and housing tenure are used to depict household information. In total we have 25 attributes to depict the demographic features $df^n = (df_1^n, df_2^n, \cdots, df_{25}^n)$ for region r_n, and some typical ones are shown in Figure 3. Note that we describe population composition using eight attributes with one for each race type (e.g., white, black, asian, etc.), but in Figure 3 (c) we only show the entropy of these eight attributes for visualization, where a higher value means a more mixed population compostion.

3.3 Region POI Profiles

Existing POIs in one region indicate the POI supply of this region provided, which is on the other side of region demand. As long as the supply and demand can be balanced, there is no need to add new POIs to increase the supply. From this point of view, what we try to estimate in this paper is the region POI demand cannot be fulfilled locally.

For the n-th region r_n, the number of POIs in each POI category can be counted. The frequency density of POI category c in region r_n is calculated by:

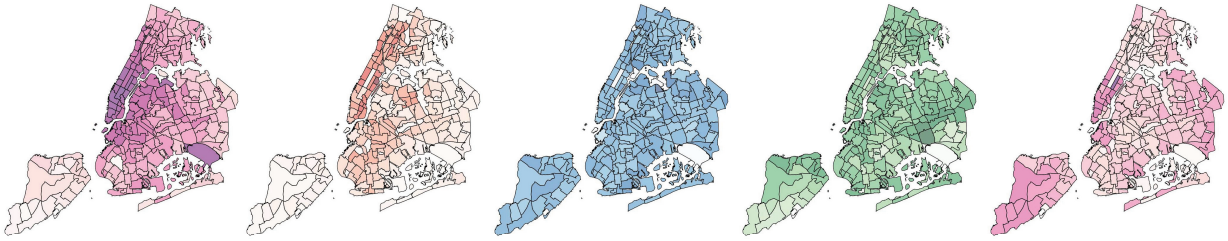
$$pf_c^n = \frac{|\{p|p \in r_n, c\}|}{Area\ of\ r_n}$$

and the region POI feature vector of region r_n is denoted by $pf^n = (pf_1^n, pf_2^n, \cdots, pf_C^n)$, where C is the number of POI categories. The region feature vector $f_n = (df_n, pf_n) \in \mathbf{F}$, consists of demographic and POI features, is regarded as the metadata of each region.

3.4 Trip Activity Inference

The original human mobility from taxi GPS traces is described by trip origin and destination, where the activities participated for the trips are unknown. Fortunately, efforts have been made to infer the activities involved for each trip [4, 12]. Here, we leverage the trip context information, i.e., POIs around destination, to describe the trip activities instead of one destination point. Specifically, we utilize the Bayesian activity inference model proposed by Gong et al. [7], which take both spatial and temporal constraints into consideration, to estimate the probabilities of possible destination POIs for each trip. After we get the probabilities of POIs for each trip, we aggregate the trips for regions. Gong et al. [7] show this model can effectively infer trip activity at the aggregation level.

Given trip $tr = (tr.ori, tr.dest, tr.time)$, the inference process first chooses a set of candidate POIs $\mathcal{PC} \subset \mathcal{P}$ within the walking distance δ of the trip destination, then assigns different visiting probabilities by considering influences of 1) distance decay, further the distance from POI to destination smaller the chance of visiting;



(a) Trip Number (b) Population Density (c) Median Age (d) Population Composition (e) Household Income

Figure 3: Features of regions. Note that darker color stands for a higher value in that region.

2) popular time, POI categories have different popularities at different time of day; and 3) attractiveness, POIs with higher ratings can attract more people to visit.

As we know, the region demands may vary over time of day, e.g., more people visiting bar in the evening than in the morning. We thus segment time into multiple segments in terms of T defined time slots. For example, we first segment days by weekday and weekend, then segment each day into 24 slots with each hour as a slot, and finally we get 48 time slots. The daily average number of trips for each time slot is shown in Figure 4.

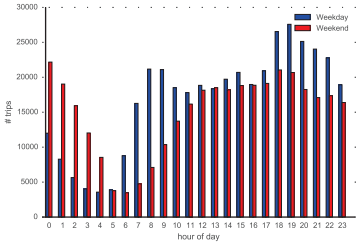


Figure 4: Number of trips at different hours of a day

Specifically, in this model, the probability of user choosing a POI $p \in \mathcal{P}C$ for a trip $tr, tr.t \in t$ is formulated as follows.

$$\Pr(p|tr) = \frac{A_p \cdot dist(tr.d, p)^{-\lambda} \cdot pop(p|t)}{\sum_{q \in \mathcal{P}C} A_q \cdot dist(tr.d, q)^{-\lambda} \cdot pop(q|t)} \quad (1)$$

where $dist(a, b)$ is the distance between two points a and b, and $dist \leq \delta$, A_p is the attractiveness of p, which is represented by the rating from Google POI data, $pop(p|t)$ is the popularity of p at time slot t which can be derived from Foursquare check-in data with

$$pop(p|t) = \frac{1}{|\{p \in c\}|} \cdot \frac{|\{checkin \in c, t\}|}{\sum_{c \in C} |\{checkin \in c, t\}|}.$$

$Pr(p|tr)$ ranges from 0 to 1, and the sum of the visiting probabilities of all the candidate POIs for one trip equals to 1. According to [7], we choose $\delta = 200m$, $\lambda = 1.5$ as the parameters. Finally, for each trip tr, we extract the *activity* tuple: $act(tr) = (tr.ori, tr.time, tr.pois)$.

4 DEMAND INFERENCE

In this section, we introduce how to derive the POI demands for all the regions by integrating human mobility records, POI and region profiles. We start with the semantic aggregation of the mobile activities of each region.

4.1 Region Activity Aggregation

After we infer the trip activities, next we aggregate the trips from the same region to obtain region activities. Specifically, for the n-th region r_n and time slot t, we aggregate the intentions of activities originating from r_n as follows:

$$pr^n_{pt} = \sum_{tr: \substack{tr.ori=r_n \\ tr.time=t}} \Pr(p|tr),$$

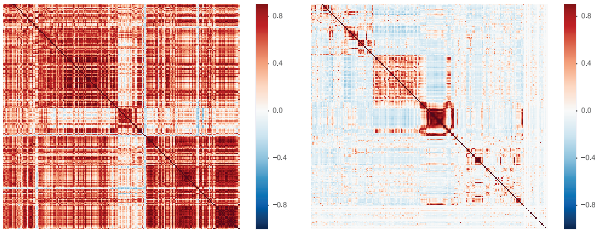
where $p \in \mathcal{P}$ is the index of the POIs. Moreover, we aggregate and normalize the probability score as

$$x^n_{pt} = \frac{pr^n_{pt}}{\sum_{q \in \mathcal{P}} pr^n_{qt}},$$

and obtain the region activity cube $\mathbf{X} = \{x^n_{pt}\}$ at the POI level.

Other than that, the region activities can also be aggregated to POI category level, which is commonly used in literature [4]. Specifically, with the probabilities of POIs visited by people from region r_n inferred, we can further summarize the visiting probability of those POIs per category $c \in C$ and obtain the category-level aggregated visit probability as $y^n_{ct} = \sum_{p \in c} x^n_{pt}$. In this way, we construct the representation of POI visit as an aggregated visiting probability vector over different POI categories. Similarly, we obtain the region activity cube $\mathbf{Y} = \{y^n_{ct}\}$ at the POI category level.

Figure 5 shows the correlations of extracted features and human mobility patterns between regions, from which we can see several region clusters formed indicating that we may learn from peers. Please note our latent factor model can be applied at both the POI level and the POI category level to infer the POI demand.



(a) (b)

Figure 5: Correlation Map (a) features of regions, (b) human mobility patterns of regions.

4.2 Latent Factor Model

As aforementioned, human mobility data can tape the "foot voting", i.e., where people go is for what they need but cannot be fulfilled locally. Therefore, one straightforward way to infer the region demand is to directly aggregate the probabilities of the destination POIs. However, the distribution of the mobility data is highly skewed, thus some regions may not have enough or even no observations to recover the demand by activity aggregation. Moreover, one trip may visit multiple POIs and the POIs can compete with each other, but the simple aggregation cannot take these factors into account. In the following, we develop a latent factor model, which considers profiles of regions and POIs, to learn the region demand with skewed mobility records. In other words, the regions with enough human mobility data can help the regions with few observations in the modeling process.

Intuitively, a person starting an activity from a region first needs to decide which demand category (e.g., shopping, eating, recreation) to be fulfilled. If the demand cannot be fulfilled locally, which cost the least amount of time and energy, then the person needs to decide which POIs in which regions to go. Along this line, we model the *demand patterns* at both POI category level (α) and POI level (v). Given a time slot t, in each column vector $\alpha_{ct} \in \mathbb{R}^K$ and $v_{pt} \in \mathbb{R}^K$ are the *pattern coefficients* for one category (c) and one POI (p), respectively. Similarly, the latent variables in the matrix u encode the *pattern coefficients* of the regional POI demands. The column vector $u_n \in \mathbb{R}^K$ is for the n-th region. In this way, the observed activity can be modeled as:

$$x_{pt}^n \sim u_n' v_{pt}.$$

The structure of the proposed region demand inference model is shown in Figure 6. Note that, the region demands may vary over time of the day, e.g., more people visiting bar in the evening than in the morning. We thus segment time everyday into multiple time slots indexed by $t = 1, 2, \cdots, T$ and learn the demand patterns for each time slot.

As shown in the graphical model (Figure 6), we also use side-information (e.g., POI category, regional POI supply and demographic data) to enhance the model. Specifically, suppose we use K latent demand patterns to model the demand portfolio of the region. The pattern coefficients $u_n \in \mathbb{R}^K$ and the demographic features $f_n \in \mathbb{R}^D$ of the n-th region is modeled as:

$$u_n \sim \beta' f_n,$$

where the matrix $\beta \in \mathbb{R}^{D \times K}$ will be learned in the modeling process. Similarly, we have:

$$v_{pt} \sim \alpha_{c(p)t},$$

if the p-th POI is in the $c(p)$-th POI category.

We put the above modeling processes in an unified probabilistic framework, with the following distribution specifications: $x_{pt}^n \sim \mathcal{N}(u_n' v_{pt}, \sigma) \in \mathbb{R}$, $u_n \sim \mathcal{N}(\beta' f_n, \sigma_u I_K) \in \mathbb{R}^K$, and $v_{pt} \sim \mathcal{N}(\alpha_{c(p)t}, \sigma_v I_K) \in \mathbb{R}^K$, where σ, σ_u, and σ_v are standard deviations of the normal distributions, respectively. Now, we use the negative log-likelihood of the model as the objective function (Equation 2) to

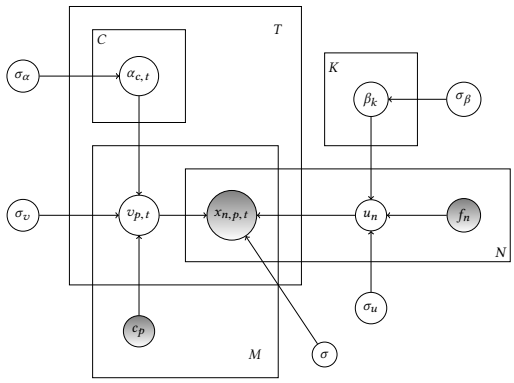


Figure 6: The Demand Inference Model

optimize the model parameters (α, β, u, and v):

$$\begin{aligned}\mathcal{L}(\alpha, \beta, u, v | x, \sigma) =& \frac{1}{2\sigma^2} \sum_{n,p,t} (x_{pt}^n - u_n' v_{pt})^2 \\ &+ \frac{1}{2\sigma_u^2} \sum_n \|u_n - \beta' f_n\|^2 + \frac{1}{2\sigma_v^2} \sum_{p,t} \|v_{pt} - \alpha_{c(p)t}\|^2 \\ &+ \frac{1}{2\sigma_\alpha^2} \sum_{c,t} \|\alpha_{ct}\|^2 + \frac{1}{2\sigma_\beta^2} \sum_k \|\beta_k\|^2\end{aligned} \tag{2}$$

The last two terms are added to reduce the generalization error with the following priors on the latent variables α and β: $\alpha_{ct} \sim \mathcal{N}(0, \sigma_\alpha I_K) \in \mathbb{R}^K$, and $\beta_k \sim \mathcal{N}(0, \sigma_\beta I_D) \in \mathbb{R}^D$.

4.3 Learning Algorithm

In this section we introduce an efficient algorithms to optimize the latent factor model with large-scale data. In the proposed model, we have parameters in: 1) α, v for time-aware demand patterns for POI categories and POIs respectively; 2) u for latent demand preferences for individual regions; 3) finally β for regression coefficients between the region features and the demand preferences of each region. We iteratively update these parameters to optimize the objective function in Equation 2.

Specifically, to optimize α with other parameters fixed, the problem is equivalent to minimize:

$$\frac{1}{2\sigma_v^2} \sum_{p,t} \|v_{p,t} - \alpha_{c(p),t}\|^2 + \frac{1}{2\sigma_\alpha^2} \sum_{c,t} \|\alpha_{c,t}\|^2 \tag{3}$$

Therefore, for each POI category c and time index t, to compute $\alpha_{c,t}$, we minimize:

$$\frac{1}{2\sigma_v^2} \sum_{p:c(p)=c} \|v_{p,t} - \alpha_{c,t}\|^2 + \frac{1}{2\sigma_\alpha^2} \|\alpha_{c,t}\|^2 \tag{4}$$

For this problem, we have closed form solution:

$$\alpha_{c,t} = \frac{\sum_{p:c(p)=c} v_{p,t}}{M_c + \sigma_v^2/\sigma_\alpha^2} \tag{5}$$

where $M_c = |\{p \mid c(p) = c\}|$ is the number of POIs in category c.

The problem to optimize β is the so-called ridge regression which is to minimize:

$$\frac{1}{2\sigma_u^2} \sum_n \|u_n - \beta' f_n\|^2 + \frac{1}{2\sigma_\beta^2} \sum_k \|\beta_k\|^2$$

Since $\mathbf{F}\mathbf{F}' + \sigma_u^2/\sigma_\beta^2 I_D$ is not singular, we also have closed form solution as:

$$\beta = (\mathbf{F}\mathbf{F}' + \sigma_u^2/\sigma_\beta^2 I_D)^{-1}\mathbf{F}u' \in \mathbb{R}^{D\times K}$$

where $\mathbf{F} \in \mathbb{R}^{D\times N}$, $u \in \mathbb{R}^{K\times N}$ and $I_D \in \mathbb{R}^{D\times D}$ is the identity matrix.

For updating u and v, we use gradient descent optimization. To this end, we have

$$\frac{\partial \mathcal{L}}{\partial u_n} = -\frac{1}{\sigma^2}\sum_{p,t}(x_{pt}^n - \langle u_n, v_{pt}\rangle)v_{pt} + \frac{1}{\sigma_u^2}(u_n - \beta' f_n)$$

$$\frac{\partial \mathcal{L}}{\partial v_{pt}} = -\frac{1}{\sigma^2}\sum_{n}(x_{pt}^n - \langle u_n, v_{pt}\rangle)u_n + \frac{1}{\sigma_v^2}(v_{pt} - \alpha_{c(p),t})$$

4.4 Variations and Extensions

There are several variations of our modeling process. For example, we can use the the following objective function to directly identify the POI demand of urban regions at the POI category level with region activity cube **Y**:

$$\begin{aligned}\mathcal{L}(\alpha,\beta,u|x,\sigma) =& \frac{1}{2\sigma^2}\sum_{n,c,t}(y_{ct}^n - u_n'\alpha_{ct})^2 \\ &+ \frac{1}{2\sigma_u^2}\sum_n \|u_n - \beta' f_n\|^2 \\ &+ \frac{1}{2\sigma_\alpha^2}\sum_{c,t}\|\alpha_{ct}\|^2 + \frac{1}{2\sigma_\beta^2}\sum_k \|\beta_k\|^2\end{aligned} \tag{6}$$

In this way, we can estimate:

$$\hat{y}_{ct}^n \sim u_n'\alpha_{ct}.$$

This might be different with the simple aggregation of the results at the POI level, e.g.,

$$\hat{x}_{pt}^n \sim u_n' v_{pt},$$
$$\hat{y}_{ct}^n \sim \sum_{p\in c}\hat{x}_{pt}^n,$$

where u and v are from Equation 2. We named the model variation in Equation 6 as RPDI_c, which will be investigated in our empirical studies in section 5.

5 EXPERIMENTAL RESULTS

In this section we first introduce the data and settings of our experiments. Then we evaluate the performances of the proposed region demand inference model. Finally we show the results of our model applying to POI demand ranking.

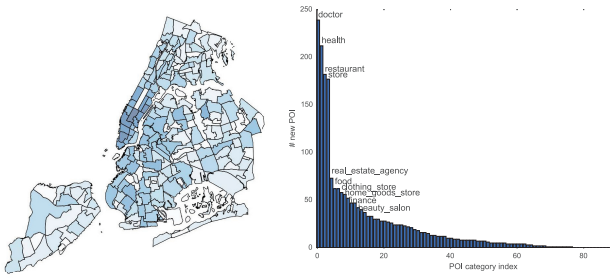
5.1 Experimental Data

All the experiments were performed on real-world datasets including one taxi trip dataset from the New York City (NYC)[2], one POI dataset collected from Google Map[3], one Location-based Social Network (LBSN) dataset collected from Foursquare[4], and one demographic dataset as introduced in subsection 3.2.

The taxi trip dataset is generated by about 50,000 taxis in New York City from January to June, 2016, in total we have around 72 million trips collected. Each trip is associated with pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Here we focus on the time, origin, and destination information related to taxi trips.

The POI dataset is collected from Google Places API and have a flat category structure, which contains 97 fine categories, such as restaurant, store, park, etc. Due to space limit, we cannot list all the categories here, please refer to Google Places[5] for the whole list of categories. This dataset can be split into two sets, one contains 297,078 POIs created before June, 2016, which are employed to estimate the human activities. Another set contains 3,817 POIs, which are created after June, 2016, is used as our validation set of demand ranking. The distribution of newly created POIs are shown in Figure 7 (a) and (b) in terms of regions and categories, respectively. Here we assume that people are rational and the new POIs are created to meet the demands.

The Foursquare dataset includes 504,152 check-in observations for 55,717 POIs. Each check-in contains the user ID, check-in time, venue ID and the venue's geo-coordinates. Note the Foursquare dataset has 418 POI categories which is more detailed than Google POI, and we manually match the 418 POI categories to the Google POI categories to make them consistent.



(a) Regions (b) Categories

Figure 7: The distribution of new POIs in NYC. Note that in (b) only top 10 most POI categories are labeled due to space.

5.2 Evaluation Metrics

In our experiments, we first evaluate the performances of our proposed latent factor model by comparing learned POI demands with observed demands. Then we apply our model for POI demand ranking, and the performances are presented.

Model Performance. We removed a uniform random subset of 10% of the entries in **X** (or **Y**) as a test set and trained on the remaining 90%. We chose to remove random entries in **X** (or **Y**) as opposed to random trips so as to avoid the obvious bias that regions will tend to revisit the same POIs. The goal of demand identification is to estimate the demands for a region which may not be observed. So, by limiting the test set to new region-POI pairs we are able to define an evaluation metric more inline with the problem we are solving. We learn the models and obtain the estimated POI demands for each region, then compare this estimation with testing data. The evaluation metric used for the comparison is Normalized Root-Mean-Square Error (NRMSE).

[2]http://www.nyc.gov/html/tlc/html/home/home.shtml
[3]https://developers.google.com/places/
[4]https://www.foursquare.com/
[5]https://developers.google.com/places/supported_types

Normalized Root-Mean-Square Error (NRMSE). The root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. Normalizing the RMSE with respect to the standard deviation (or mean) facilitates the comparison between datasets or models with different scales.

$$RMSE = \sqrt{\frac{\sum_i^n (\hat{x}_i - x_i)^2}{n}}, NRMSE = \frac{RMSE}{std(x)} \tag{7}$$

POI Demand Ranking. We rank our identified POI demands to see what's the most needed POI categories in a region, and which regions are with the most demand for certain POI category. Therefore, our experiments are conducted in two-folds: (1) Given a region, rank the demands for different POI categories; (2) Given a POI category, rank the demands of different regions.

For practical usage, we train a model for each time slot. To provide a unified region demand to users, the output of these models can be aggregated as $d_p^n = \sum_t w_t^n \cdot \hat{x}_{pt}^n$, where $w_t^n = \frac{|\{tr \in r_n, t\}|}{\sum_t |\{tr \in r_n, t\}|}$, d_p^n stands for demand of region r_n for POI p. And further normalized by category with supply information considered, $\bar{d}_c^n = \frac{1}{|\{p \in r_n, c\}|+1} \cdot \sum_{p \in c} d_p^n$. Here $\bar{d}_c^n$ is the marginal demand for one POI in category c, which measures the potential demand can be delivered to a new POI. Our ranking result is given by ranking $\bar{d}_c^n$ in descending order. To evaluate the ranking list given, we use the newly created POIs after June, 2016 as our groundtruth of demand, which is ranked in descending order by the ratio of increased new POIs: $\frac{|\{q \in r_n, c\}|}{|\{p \in r_n, c\}|+1}$, where $q \in \mathcal{NP}, p \in \mathcal{P}$, and $\mathcal{NP}$ is the new POI set.

F-measure. F-measure combines precision and recall together with a harmonic mean, which is defined as

$$F_1@\,top\text{-}k = 2 \cdot \frac{Precision@k \times Recall@k}{Precision@k + Recall@k}. \tag{8}$$

Given a top-k ranking list S_{rank} sorted in a descending order based on the estimated demands, precision and recall can be obtained as follows: $Precision@\,top\text{-}k = \frac{S_{rank} \bigcap S_{new}}{k}$, and $Recall@\,top\text{-}k = \frac{S_{rank} \bigcap S_{new}}{S_{new}}$, where S_{new} are the POI categories newly created in the groundtruth data. The F-measure for the entire city are computed by averaging all the F-measure values of all the regions (or categories).

Normalized Discounted Cumulative Gain (NDCG). Given a top-k ranking list sorted in a descending order of the estimated demands, NDCG [9] is defined as

$$NDCG@\,top\text{-}k = \frac{1}{IDCG} \times \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log(i+1)}, \tag{9}$$

where IDCG is the maximum possible DCG for a given set of ranking list, and rel_i is 1 if the ranked POI category at position i is newly created and 0 otherwise. NDCG measures the ranking quality of the recommender system based on a graded relevance scale of recommendations. The NDCG for the entire city are computed by averaging all the NDCG values of all the regions (or categories).

5.3 Baselines

We compare the proposed method (RPDI) with five baselines, which are introduced as follows.

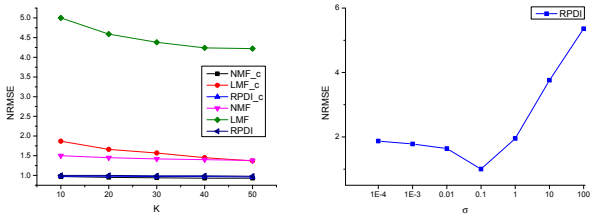
Non-Negative Matrix Factorization (NMF) is a matrix factorization model, which factorize a matrix into (usually) two matrices, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect.

Logistic Matrix Factorization (LMF) [10] is a factorization model for the implicit case in which it models the probability of a user choosing an item by a logistic function.

Moreover, NMF_c, LMF_c, RPDI_c are the methods we apply NMF, LMF, RPDI to the aggregated region activity cube $\mathbf{Y}$ at POI category level, respectively.

5.4 Results of Model Performances

We evaluate our proposed model and baseline models using the NRMSE evaluation metric for a differing number of latent factors K ranging from 10 to 50. As shown in Figure 8 (a), we can see RPDI achieves best performance among all the models, with relatively small NRMSE obtained. While LMF performs worst among all the methods. For the models applied to $\mathbf{Y}$, they are able to obtain relatively small NRMSE, because $\mathbf{Y}$ is much smaller than $\mathbf{X}$ and easier to be factorized. With the increasing of K, the NRMSE decreases slowly. We also find that increasing the number of the latent factors beyond 50 did not improve performance on our dataset. Moreover, the priors of latent variables $(\sigma, \sigma_\alpha, \sigma_\beta, \sigma_u, \sigma_v)$ are validated with values (0.0001, 0.001, 0.01, 0.1, 1, 10, 100), and we find 0.1 achieves the best performance (as shown in Figure 8 (b)).



(a) Dimension of latent factors (b) Priors

Figure 8: NRMSE of models

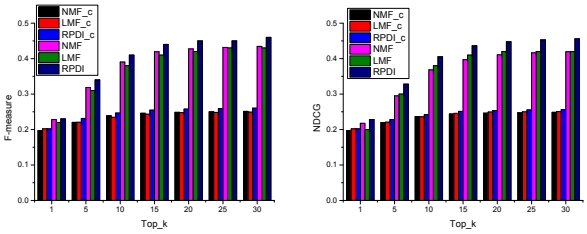
5.5 Results of POI Demand Ranking

In the next, we investigate the performances of our methods on ranking region POI demand in two-folds: rank the POI demand for every region, and rank the region demand for every POI category.

First, given a region, we rank the demands for POI categories. And aggregate the results for all the regions as our final result. The performances in terms of F-measure and NDCG with respect to top-k categories are shown in Figure 9. In the figures, we can see that RPDI achieves better overall performances than the others, outperforming the second best model NMF by 8.5%. The performances of models using category visiting probabilities are not as good as the models using POI visiting probabilities, probably due to the aggregation of category cannot reveal people's choices of POIs when going out. Among the ranking lists of all the regions, the top 10 most needed POIs are as follows: restaurant, bar, lodging, health, doctor, dentist, school, clothing store, beauty salon, cafe. We can see most of these POIs are quite related to local businesses and can bring much convenience to local residents if demands can be satisfied. Moreover, to better illustrate the ranking results for regions, we show several examples of the top-10 identified demands of regions in Table 2.

Table 2: Identified POI demands for regions.

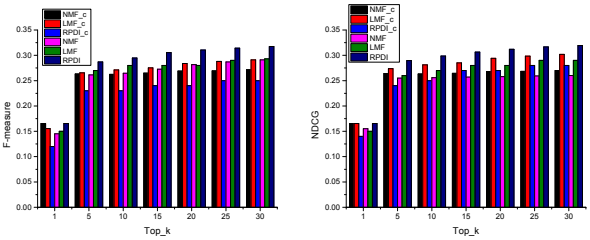
Region Name	Identified Demands@top-10	Groundtruth
Homecrest	'**restaurant**' 'bar' 'school' 'lodging' '**beauty salon**' '**store**' 'cafe' '**clothing store**' 'bakery' '**finance**'	'beauty salon' 'store' 'restaurant' 'clothing store' 'car repair' 'health' 'finance' 'jewelry store' 'doctor' 'hair care'
Central Harlem South	'**beauty salon**' 'dentist' '**store**' 'night club' '**health**' 'clothing store' 'electronics store' '**health**' 'general contractor' 'bank'	'school' 'beauty salon' 'store' 'gym' 'health' 'laundry' 'mosque' 'liquor store' 'doctor'
Clinton Hill	'**restaurant**' '**bar**' 'school' '**store**' '**beauty salon**' 'lodging' 'cafe' 'dentist' '**clothing store**' '**bakery**'	'beauty salon' 'store' 'bar' 'pharmacy' 'restaurant' 'bakery' 'grocery or supermarket' 'clothing store' 'car repair' 'liquor store'
Bedford	'**restaurant**' '**beauty salon**' '**school**' 'bar' 'lodging' 'cafe' '**clothing store**' 'night club' 'health' '**food**'	'school' 'beauty salon' 'food' 'store' 'restaurant' 'real estate agency' 'grocery or supermarket' 'clothing store' 'health' 'laundry'
Fordham North	'**beauty salon**' '**health**' '**clothing store**' '**restaurant**' 'moving company' 'car repair' 'finance' 'car wash' 'doctor' 'general contractor'	'beauty salon' 'restaurant' 'clothing store' 'health'



(a) F-measure (b) NDCG

Figure 9: Rank categories for regions with different top-k

Then, given a POI category, we rank the POI demands of regions. And aggregate the results for all the POI categories as our final result. The performances in terms of F-measure and NDCG with respect to top-k regions are shown in Figure 10. In the figures, we can see that RPDI is still able to obtain better overall performances than the others. However, the ranking results are not as good as ranking for regions, since we have more regions than POI categories which makes it a harder problem. Similar to the ranking for regions, the performances of models using category visiting probabilities are not as good as the models using POI visiting probabilities.



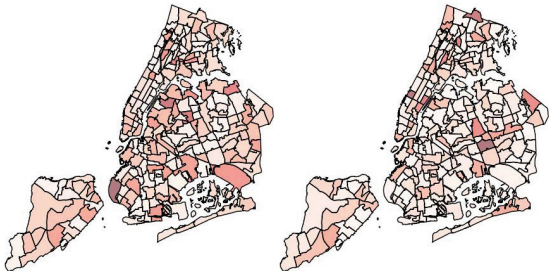
(a) F-measure (b) NDCG

Figure 10: Rank regions for categories with different top-k

To better illustrate the ranking results for POI categories, we show the estimated demands of two typical kinds of POIs, restaurants and health services. As shown in Figure 11, it is notable to see that the estimated demands of restaurants (shown in (a)) in Manhattan do not rank high among all the regions. Although Manhattan is the central area of NYC and the market of restaurants is huge, there is not much demand for new restaurants since the supply is also high. As for health services, there are higher demands in areas with lower household income in Brooklyn, Bronx, and Queens. From this point of view, the government should make efforts to allocate more health services in these areas.

6 RELATED WORK

POI Recommendation. POI recommendation, targeting at recommending the right POIs to the target users [14, 15, 17], can be



(a) Restaurants (b) Health Services

Figure 11: Identified POI demands for categories. Note that darker color stands for a higher demand in that region.

seen as discovering POI demands for users. Previous studies often used collaborative filtering (CF) algorithm to fuse the check-in information, e.g., user interest preferences, social influence, temporal influence and geographical influence [5, 28]. In [26], Ye et al. considered the social influence under the framework of a user-based CF model, and modeled the geographical influence by a model-based method (a Bayesian CF algorithm). Moreover, Yuan et al. [28] and Gao et al. [5] introduced temporal preference to enhance the algorithm efficiency and effectiveness. The authors separated a day into different time slots and user preferences were learned for each slot, thus POIs can be recommended according to different times of a day. Cheng et al. [3] considered more comprehensive information, such as the multi-center of user check-in patterns, and the skewed user check-in frequency. Moreover, Liu et al. [16] proposed a geographical probabilistic factor analysis framework to analyze the joint effects of multiple factors by considering user preference for locations as a multiplication of interest in the locations, location popularity and distance.

Different from the above works, we consider the POI demand problem at the region level. Moreover, we integrate the region supply information and demographic information into the proposed model to learn region demands more accurately.

Site Selection. Traditionally, the site selection problem has been studied by researchers from land economy community using spatial interaction models, and multiple regression discriminant analysis [2]. In recent years, location-based services have been widely used to tackle this problem [11]. Karamshuk et al. selected optimal retail store location from a list of locations by using supervised learning with features mined from Foursquare check-in data. Li et al. [13] studied ambulance stations site selection by using real traffic information so as to minimize the average travel time to reach the emergency requests. Niu et al. [20] extracted discriminative features of gas stations from heterogeneous mobile data and then formalized

a gas station ranking predictor to select gas station location. Xu et al. [25] proposed a framework to combine the spatial distribution of user demands with the popularity and economic attributes for optimal location selection.

Different from the above works, we consider a more general framework to identify region POI demands. In this framework, it learns demand of regions for all POIs simultaneously instead of focuses on specific POI categories like restaurant or gas station.

Human Mobility Pattern Mining. Understanding human mobility in urban environments is central to traffic forecasting, location-based services, and urban planning. A significant number of papers on human mobility analysis have been published in recent years thanks to the widely available mobility data, such as GPS data, cellular network data, and transportation data [1, 8, 18, 23].

To the best of our knowledge, we are the first to work on the problem of discovering region POI demand by leveraging human mobility patterns. Although there is no existing work on the exact application we are working on, there are many existing work on making use of human mobility patterns for different novel applications. Giannotti et al. [6, 19] developed trajectory pattern mining, and applied it to predict the next location at a certain level of accuracy by using GPS data. Zheng et al. [30] detected flawed designs in current road network with a frequent graph method on taxi GPS traces. Yuan et al. [27] proposed a topic-based inference model that discovers regions of different functions, such as educational areas and business districts, using both human mobility data and POIs.

7 CONCLUSION

In this paper, we investigated how to exploit human mobility patterns, geographic data, and demographic data for identifying region POI demands. Along this line, we first proposed a framework, named Region POI Demand Identification (RPDI), to model POI demands with the daily needs identified from their large-scale mobility data. Specifically, in this framework, an urban space was first partitioned into spatially differentiated neighborhood regions formed by local communities. Then, the daily activity patterns of people traveling in the city were extracted from human mobility data. However, the trip activities, even aggregated, were sparse and insufficient to directly identify the POI demands, especially for underdeveloped regions. Therefore, with a proposed demand inference model considering POI preferences and supplies together with demographic features, we estimated the POI demands of all the regions simultaneously. As shown in the experimental results on real-world data, the proposed RPDI framework could provide effective POI demand identification for different regions.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation grant IIS-1648664. Also, it was supported in part by the Natural Science Foundation of China (71329201, 71531001).

REFERENCES

[1] Tengfei Bao, Huanhuan Cao, Enhong Chen, Jilei Tian, and Hui Xiong. 2012. An unsupervised approach to modeling personalized contexts of mobile users. *Knowledge and Information Systems* 31, 2 (2012), 345–370.
[2] Oded Berman and Dmitry Krass. 2002. The generalized maximal covering location problem. *Computers & Operations Research* 29, 6 (2002), 563–581.
[3] Chen Cheng, Haiqin Yang, Irwin King, and Michael R Lyu. 2012. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*.
[4] Yanjie Fu, Hui Xiong, Yong Ge, Yu Zheng, Zijun Yao, and Zhi-Hua Zhou. 2016. Modeling of Geographic Dependencies for Real Estate Ranking. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 1 (2016), 11.
[5] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2013. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of RecSys*. ACM, 93–100.
[6] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *KDD*. ACM, 330–339.
[7] Li Gong, Xi Liu, Lun Wu, and Yu Liu. 2016. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science* 43, 2 (2016), 103–114.
[8] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.
[9] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM TOIS* 20, 4 (2002), 422–446.
[10] Christopher C Johnson. 2014. Logistic matrix factorization for implicit feedback data. *NIPS* 27 (2014).
[11] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of KDD*. ACM, 793–801.
[12] Ravi Kumar, Mohammad Mahdian, Bo Pang, Andrew Tomkins, and Sergei Vassilvitskii. 2015. Driven by food: Modeling geographic choice. In *Proceedings of WSDM*. ACM, 213–222.
[13] Yuhong Li, Yu Zheng, Shenggong Ji, Wenjun Wang, Zhiguo Gong, et al. 2015. Location selection for ambulance stations: a data-driven approach. In *Proceedings of GIS*. ACM, 85–94.
[14] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. 2014. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *KDD*. 831–840.
[15] Bin Liu and Hui Xiong. 2013. Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness.. In *Proceedings of SDM*, Vol. 13. 396–404.
[16] Bin Liu, Hui Xiong, Spiros Papadimitriou, Yanjie Fu, and Zijun Yao. 2015. A General Geographical Probabilistic Factor Model for Point of Interest Recommendation. *TKDE* 27, 5 (2015), 1167–1179.
[17] Yanchi Liu, Chuanren Liu, Bin Liu, Meng Qu, and Hui Xiong. 2016. Unified Point-of-Interest Recommendation with Temporal Interval Assessment. In *Proceedings of KDD*. ACM, 1015–1024.
[18] Yanchi Liu, Chuanren Liu, Nicholas Jing Yuan, Lian Duan, Yanjie Fu, Hui Xiong, Songhua Xu, and Junjie Wu. 2014. Exploiting heterogeneous human mobility patterns for intelligent bus routing. In *Proceedings of ICDM*. IEEE, 360–369.
[19] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. 2009. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of KDD*. ACM, 637–646.
[20] Hongting Niu, Junming Liu, Yanjie Fu, Yanchi Liu, and Bo Lang. 2016. Exploiting Human Mobility Patterns for Gas Station Site Selection. In *Proceedings of DASFAA*. 242–257.
[21] Vaida Pilinkienė. 2008. Market demand forecasting models and their elements in the context of competitive market. *Engineering Economics* 60, 5 (2008).
[22] Vaida Pilinkienė. 2008. Selection of market demand forecast methods: Criteria and application. *Engineering Economics* 58, 3 (2008).
[23] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.
[24] Charles M Tiebout. 1956. A pure theory of local expenditures. *The journal of political economy* (1956), 416–424.
[25] Mengwen Xu, Tianyi Wang, Zhengwei Wu, Jingbo Zhou, Jian Li, and Haishan Wu. 2016. Demand driven store site selection via multiple spatial-temporal data. In *Proceedings of GIS*. ACM, 40–49.
[26] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of SIGIR*. 325–334.
[27] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2015. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2015), 712–725.
[28] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *SIGIR*. 363–372.
[29] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. 2014. Diagnosing New York city's noises with ubiquitous data. In *Proceedings of UbiComp*. ACM, 715–725.
[30] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. 2011. Urban computing with taxicabs. In *Proceedings of UbiComp*. ACM, 89–98.
[31] Hengshu Zhu, Hui Xiong, Fangshuang Tang, Qi Liu, Yong Ge, Enhong Chen, and Yanjie Fu. 2016. Days on market: Measuring liquidity in real estate markets. In *Proceedings of KDD*. 393–402.