Computers, Environment and Urban Systems 62 (2017) 74–85

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A novel methodology for prediction of spatial-temporal activities using latent features

QiuLei Guo *, Hassan A. Karimi

Geoinformatics Laboratory, School of Information Sciences, University of Pittsburgh, United States

ARTICLE INFO

Article history:
Received 4 July 2016
Received in revised form 8 October 2016
Accepted 9 October 2016
Available online 8 November 2016

ABSTRACT

In today's era of big data, huge amounts of spatial-temporal data are generated daily from all kinds of citywide infrastructures. Understanding and predicting accurately such a large amount of data could benefit many real-world applications. In this paper, we propose a novel methodology for prediction of spatial-temporal activities such as human mobility, especially the inflow and outflow of people in urban environments based on existing large-scale mobility datasets. Our methodology first identifies and quantifies the latent characteristics of different spatial environments and temporal factors through tensor factorization. Our hypothesis is that the patterns of spatial-temporal activities are highly dependent on or caused by these latent spatial-temporal features. We model this hidden dependent relationship as a Gaussian process, which can be viewed as a distribution over the possible functions to predict human mobility. We tested our proposed methodology through experiments conducted on a case study of New York City's taxi trips and focused on the mobility patterns of spatial-temporal inflow and outflow across different spatial areas and temporal time periods. The results of the experiments verify our hypothesis and show that our prediction methodology achieves a much higher accuracy than other existing methodologies.

1. Introduction

Today's city infrastructures are retrofitted with GPS and RFID, among other sensors, and with the popularity and widespread use of location-based social networks (LBSN), a huge amount of spatial-temporal data is generated and accumulated every day. This Big Data trend brings great opportunities for tackling many real-world challenges. In this paper, we propose a novel methodology for prediction of spatial-temporal activities using latent spatial and temporal factors extracted from existing large-scale mobility datasets at a city level. Of spatial-temporal activities, we are interested in human mobility, especially the inflow and outflow of people in neighborhoods/areas during certain time periods. Understanding the inflow/outflow of people in urban environments spatially and temporally and predicting them correctly are essential to solve many real-world problems. Example applications that can benefit from the proposed methodology include: a taxi/Uber company can predict the number of people who would leave or enter some neighborhoods by taking taxis/Uber vehicles during the next hour so that they can allocate their vehicles optimally; traffic agencies can estimate the number of people heading to certain neighborhoods and infer the corresponding traffic situations in the near future; transportation managers can dynamically optimize schedules of their public bus and subway services based on people's real-time needs, among other factors.

Given the importance of understanding and predicting human's spatial-temporal mobility patterns correctly, many methodologies and techniques have emerged. However, a review of existing methodologies and techniques reveals that they have limitations. For example, many existing studies employ techniques such as k-means to cluster, interpret, and predict human's mobility patterns based on spatial closeness (Kaltenbrunner, Meza, Grivolla, Codina, & Banchs, 2010; Froehlich, Neumann, & Oliver, 2009; Cranshaw, Schwartz, Hong, & Sadeh, 2012; Gao, Liu, Wang, & Ma, 2013; Guo, Zhu, Jin, Gao, & Andris, 2012). There are a few problems with such methodologies. One problem is that there is no definition of how close two neighborhoods should be in order to share a similar pattern. Another problem is that close neighborhoods do not necessarily share a similar pattern. Existing works also have problems with temporal characteristics because they usually simply use latest historical data to predict future activities, ignoring the fact that the relationship between future human's mobility patterns and the previous ones might vary over time.

To fill this research gap, we proposed a new methodology for prediction of spatial-temporal activities by identifying and using latent spatial and temporal features. One major motivation behind our methodology is that we think the patterns of many spatial-temporal activities, such as human mobility, are dependent on or even caused by hidden characteristics of spatial areas, different time periods, and some other factors. Take the spatial-temporal inflow and outflow of people in different

* Corresponding author.
E-mail addresses: qig6@pitt.edu (Q. Guo), hkarimi@pitt.edu (H.A. Karimi).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.006
0198-9715/© 2016 Elsevier Ltd. All rights reserved.

neighborhoods in a city as an example. It can be seen that residential neighborhoods and office districts have high volumes of outflow and inflow in the morning and in the evening, respectively. While this is an interesting observation analyzed qualitatively, it is not sufficient to make any prediction, like the number of people who would be leaving/entering a residential neighborhood during certain time periods. With our proposed methodology, we can use this simple and initial qualitative information to predict various spatial-temporal activities. Our methodology comprised three major steps, as described below.

In the first step, we used a 3D tensor to model human spatial-temporal movements in a geographic area like a city and extract the latent spatial and temporal features of different neighborhoods and time periods through tensor factorization. In the tensor, one dimension (mode) is for origin neighborhood, one dimension is for destination neighborhood, and one dimension is for time interval (e.g., one hour). Each entry of the tensor stores the average number of movements from a particular origin neighborhood to a particular destination neighborhood at a particular time. The 3D tensor is used to extract the latent features of origin neighborhood, destination neighborhood, and time interval through tensor factorization, respectively.

In the second step, we explored and verified the dependent relationship between mobility patterns and extracted latent spatial and temporal features. In particular, we hypothesized that the mobility patterns in different neighborhoods and time periods are highly dependent on the latent spatial and temporal features, among other factors. Then we inferred the causal structure that would generate those mobility patterns through the PC algorithm (Spirtes, Glymour, & Scheines, 2000).

Once the dependent relationship between the mobility patterns and the latent features was verified, in the third step, we mathematically modeled this relationship and used it for the prediction of spatial-temporal activities. Note that instead of relating this dependent relationship to some specific models such as linear, quadratic, cubic, or even non-polynomial, which could have numerous possibilities, we modeled this relationship as the free-form Gaussian process. From the function standpoint, the Gaussian process can be thought of as defining a distribution over functions where inferencing is directly made in the space of functions (Rasmussen, 2006). One major reason for using the Gaussian process to represent this relationship between mobility and latent features is due to the observation that spatial-temporal activities, such as outflow/inflow of people, are usually generated from an underlying process that is smooth and continuous. This process has typical amplitude and variations and takes place over spatial and temporal characteristics, among other things.

One major advantage of our methodology is that it inherently considers both spatial and temporal data characteristics. More specifically, through modeling the characteristics of different spatial area, different time periods, and their relationship with mobility patterns mathematically as a Gaussian process, predictions can be made using the data from not only one specific spatial area or temporal time period of interest, but also from other areas and time periods with similar patterns. To validate our proposed methodology, we conducted experiments on a case study of taxi trips in New York City. The results of the experiments showed that our prediction methodology achieved a higher accuracy than the existing methodologies (around 20% error rate reduction).

The contribution of the paper is a new methodology for prediction of spatial-temporal activities using latent spatial-temporal features. We systematically verified the causality relationship between spatial-temporal activities and latent spatial-temporal features, and modeled this relationship as a Gaussian process. Through our designed covariance function, the prediction methodology inherently considers both the spatial and temporal characteristics of the data. We applied our methodology to a large-scale real-world dataset to demonstrate the advantages of our methodology compared to other existing methodologies.

The rest of the paper is organized as follows. Section 2 reviews related work in the literature. Section 3 describes the tensor model of human mobility and the extraction of latent spatial and temporal features. Section 4 discusses the verification of the dependent/causality relationship between the extracted latent features and human mobility. Section 5 presents the prediction methodology through modeling the relationship between latent features and human mobility as a Gaussian process. Section 6 demonstrates the performance of our methodology through a series of experiments with the taxi trips in New York. Section 7 concludes the paper and discusses future research direction.

2. Related work

Given the importance of gaining a deeper understanding of many spatial-temporal activities, like human mobility, and predicting them accurately, related work in this area has been published in various fields, such as computer science, urban planning, sociology, and other areas. In this section, some of the works relevant to different aspects of mobility patterns are overviewed.

2.1. Human mobility and land use study

Liu, Wang, Xiao, and Gao (2012) derived the urban land use information by classifying the study area into six types of “source-sink” areas with the pick-ups and drop-offs taxi data in Shanghai. Cranshaw et al. (2012) introduced a clustering model and research methodology for studying the structure and composition of a city on a large scale based on the social media generated by its residents. Pei et al. (2014) applied a semi-supervised fuzzy clustering approach to infer the land use types from the mobile phone data. Liu, Kang, Gong, and Liu (2016) proposed a novel unsupervised land use classification method with a new type of place signature that incorporates the spatial interaction patterns of mobility data. Zhang and Pelechrinis (2016) investigated how certain street fairs would affect human's mobility and location business. Noulas and Mascolo (2013) inferred the functions of each neighborhood in the city by using the Foursquare POIs and cellular data. Liu et al. (2015) provided a comprehensive overview of how to use the social sensing data in geographical areas to better understand the socioeconomic environments.

2.2. Prediction of human mobility

Froehlich et al. (2009) provided a spatial-temporal analysis of bicycle station usage in Barcelona and compared the experimental results from four simple predictive models. Kaltenbrunner et al. (2010) also provided the spatial-temporal analysis for the bicycle usage in Barcelona and adopted the autoregressive-moving-average (ARMA) model to predict the number of available bikes and docks for each bike station. Guo, Luo, Li, Wang, and Geroliminis (2013) and Matthias et al. (2008) predicted the popularity/density of the routes in a road network by estimating the distribution of moving vehicles' future trajectories. Different from the mobility prediction for the city level, Scellato, Musolesi, Mascolo, Latora, and Campbell (2011) made the spatial-temporal location prediction for a single user based on his/her own historical trajectories. Zhang, Lin, and Pelechrinis (2016) introduced EigenTransitions, a spectrum-based, generic framework for analyzing mobility datasets and predicting individual user's mobility like next visit area.

2.3. Tensor/matrix factorization

Tensor, also known as multidimensional matrix, has shown its usage in applications such as recommendation systems (Baltrunas, Ludwig, & Ricci, 2011) and has gained the attention of researchers in urban computing. Among many applications, one typical usage of tensor is to recover the missing/sparse data through tensor factorization. Zhang, Wilkie, Zheng, and Xie (2013) used the tensor factorization to infer urban refueling behavior by incorporating POI data, traffic features, and gas stations as contextual features together. Zheng et al. (2014) modeled the noise situation of NYC with a 3D tensor, then recovered the missing noise situations in NYC through a context-aware tensor decomposition. Another usage of tensor is to isolate and analyze the patterns hidden in a dataset. Fan, Song, and Shibasaki (2014) represented people's different life patterns in the city as a linear combination of basic tensors and conducted series of spatial-temporal analyses on them. Kang and Qin (2016) identified a set of high-level statistical features of the taxicabs' operation behaviors through collecting the digital traces of 6000 + taxicabs in Wuhan, China, and conducting the nonnegative matrix factorization. Zhi et al. (2016) identified a series of latent spatial-temporal activity structures from the social media check-in data through the matrix factorization and clustered the studied areas into five significant types of regions. Peng, Jin, Wong, Shi, and Liò (2012) analyzed the passengers' traffic patterns with the taxi trips by employing the nonnegative matrix factorization and identified three main daily travel purposes.

2.4. Causation

Spirtes et al. (2000) presented the fundamental principles of causation and several algorithms to search the causal structure from the given datasets. Hoyer, Shimizu, Kerminen, and Palviainen (2008) took advantage of the non-Gaussian data and made causal inference possible even in the presence of hidden variables. Zhang and Hyvärinen (2009) considered the causal models in which the relationship among the variables is nonlinear while disturbances have linear effects. In causal structure inference, one important procedure is the independence test. Gretton et al. (2007) provided a novel test of the independence hypothesis for one particular kernel independence measure, the Hilbert-Schmidt independence criterion (HSIC). Zhang, Peters, Janzing, and Schölkopf (2012) proposed a Kernel-based Conditional Independence test (KCI-test) for continuous variables with high dimensionality.

2.5. Gaussian process regression

A Gaussian process is a generalization of the Gaussian probability distribution. While a probability distribution describes random variables that are scalars or vectors (for multivariate distributions), a stochastic process governs the properties of functions (Rasmussen, 2006). Rasmussen (2006) provided an introduction to the Gaussian process, and Roberts et al. (2013) offered an introduction to the Gaussian process for time series data analysis. Kim, Lee, and Essa (2011) used Gaussian process regression to model the trajectory in video, allowing for incrementally predicting possible paths and detecting anomalous events from online trajectories.

Different from existing works, which usually used the explicit features like the data from nearby neighborhoods or previous time series records for the spatial-temporal prediction, we (a) proposed to extract and quantify the latent spatial and temporal characteristics of the data, (b) systematically explored the causality relationship between these latent features and human spatial-temporal activities, and (c) mathematically modeled the relationship between them as Gaussian process for future prediction. Our validation experiments showed that our methodology has higher prediction accuracy than the existing ones.

3. Modeling human spatial-temporal movements

In this section, we provide a detailed description of using a 3D tensor to model human spatial-temporal movements between different neighborhoods and the extraction of latent spatial-temporal features.

Fig. 1 shows our model of human's fluxes between different neighborhoods with a tensor $\mathcal{H}\in\mathcal{R}^{N\times N\times L}$. The 3D tensor $\mathcal{H}$ denotes N origin neighborhoods, N destination neighborhoods, and L time slots, respectively. Each entry of the tensor $\mathcal{H}(i,j,k)$ stores the average number of trips starting from neighborhood i to neighborhood j during time period k.

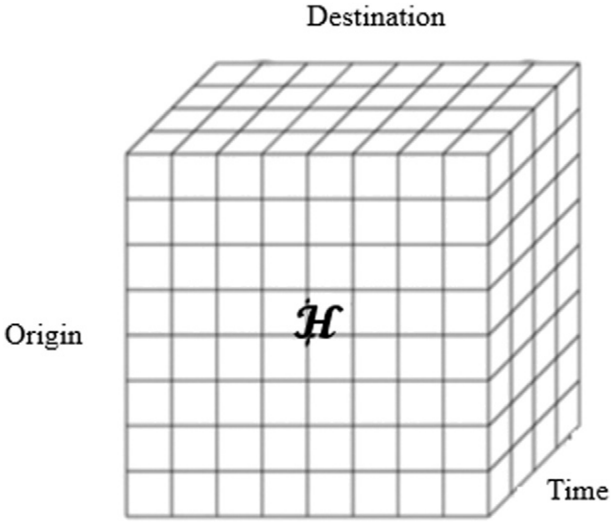


Fig. 1. Tensor model of human spatial-temporal movements.

With this tensor model, we can extract the latent features of each origin neighborhood, destination neighborhood, and time slot by decomposing the tensor $\mathcal{H}$ into three matrixes $\mathcal{S}_o{}^{N\times Q}$, $\mathcal{S}_d{}^{N\times P}$, $\mathcal{T}^{K\times R}$, and a core tensor $G^{Q\times P\times R}$, respectively, as shown in Fig. 2, with the tucker factorization (Kolda & Bader, 2009).

$$\mathcal{H} \approx G\times_1\mathcal{S}_o\times_2\mathcal{S}_d\times_3\mathcal{T} \tag{1}$$

To achieve this decomposition, we can define the loss function with regularization by using Eq. (2):

$$\begin{aligned} L(\mathcal{H},\mathrm{G},\mathcal{S}_o,\mathcal{S}_d,\mathcal{T}) = {} & \frac{1}{2}\|\mathcal{H}-G\times_1\mathcal{S}_o\times_2\mathcal{S}_d\times_3\mathcal{T}\|^2 \\ & +\frac{\lambda}{2}\left(\|G\|^2+\|\mathcal{S}_o\|^2+\|\mathcal{S}_d\|^2+\|\mathcal{T}\|^2\right) \end{aligned} \tag{2}$$

This optimization problem can be solved in many ways, for example, by using the gradient-descent method, but discussion of these algorithms is beyond the scope of this paper.

The motivation behind using the tensor factorization is due to the existence of some latent features and interactions among them that usually determine how people in one neighborhood (origin) move to another neighborhood (destination) during certain time periods. For example, two residential neighborhoods would both have a high volume of outflow (to an office district) in the morning. Similarly, two nightlife districts would both attract a high volume of inflow in the evening. This is a simple qualitative analysis and difficult to extend to general cases since most regions are not monofunctional and people's flow is usually a mix of a variety of life patterns. However, by discovering the

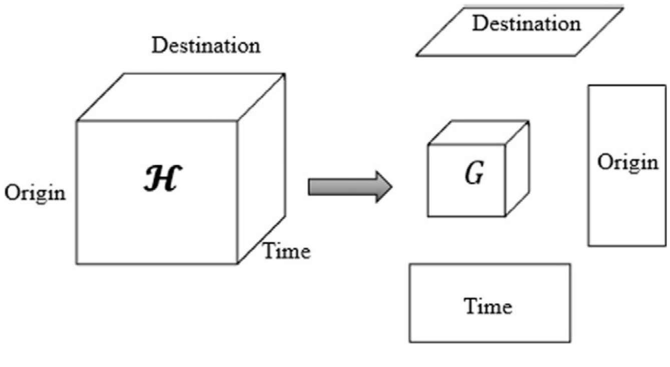


Fig. 2. Tensor factorization.

latent features and the interactions among them, we can predict the number of outflow/inflow of people with respect to a certain neighborhood during certain time periods. This is somewhat similar to the recommendation system like the one Netflix uses where a multidimensional tensor represents how different users rate different movies under various contexts such as different times. For example, two users would give a high rating to a certain movie if they both liked the actors/actresses in the movie, or if the movie was a romantic movie, which was preferred by both users in the previous couple of weeks. Hence, if we can discover these latent features, we should be able to predict a rating with respect to a certain user and a certain item under specific contexts. Similarly, given the extracted latent features of origin neighborhoods (like users), destination neighborhoods (like movies), the specific time period, and some other features, we could predict the human flow.

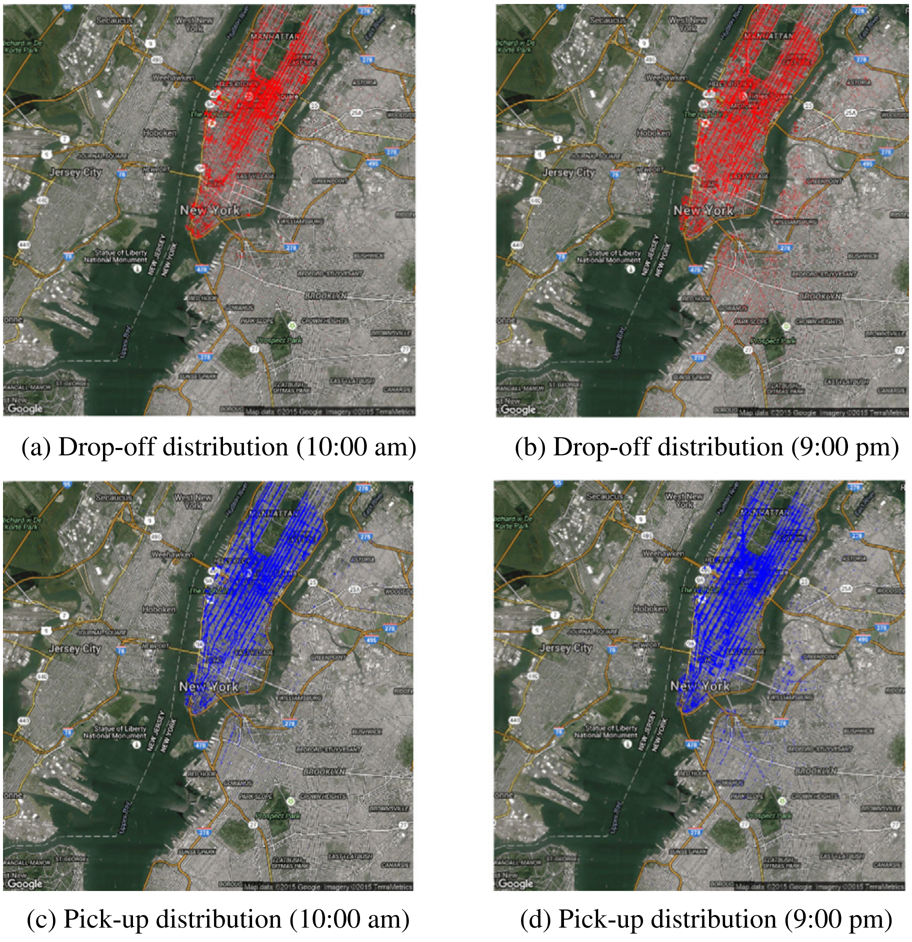
After the tensor factorization, the row i of matrix $\mathcal{S}_o$, $\mathcal{S}_{oi}$, is the feature vector indicating the characteristics of origin neighborhood i. Similarly, the row j of matrix $\mathcal{S}_d$, $\mathcal{S}_{dj}$, is the feature vector indicating the characteristics of destination neighborhood j. $\mathcal{T}_k$ is the feature vector indicating the characteristics of the corresponding time slot k. Each entry of the core tensor G indicates the level of interaction among different components of $\mathcal{S}_o$, $\mathcal{S}_d$, and $\mathcal{T}$, respectively. We will discuss the relationship between the patterns of inflow/outflow of people and use the identified latent features for prediction in the next sections.

4. Causal analysis of human's mobility patterns

As discussed in Section 1, one major motivation behind our prediction methodology is that we think the patterns of many spatial-temporal activities, like human mobility, are dependent on and/or caused by the corresponding characteristics of spatial areas, temporal time periods, and some other factors. Therefore, before we used the latent features for prediction, we verified our hypothesis that the inflow/outflow of people in different spaces and times are dependent on and/or caused by latent spatial-temporal features. In particular, we argued that the volumes of outflow $x_{oi,k}$ and inflow $x_{\iota i,k}$ of a neighborhood i at time slot k are dependent on their latent spatial features $\mathcal{S}_{oi}$, $\mathcal{S}_{di}$, latent temporal features $\mathcal{T}_k$, and their previous values, $x_{oi,k-1}$, $x_{\iota i,k-1}$, respectively. For this verification, we used the PC algorithm (Spirtes et al., 2000) whose steps are: (1) create a complete undirected graph of all p variables where each variable is a graph node/vertex; (2) for each set of variables X and Y, if $X \perp\!\!\!\perp Y$ (which means X is independent from Y) remove the edge between the nodes representing them; (3) for each set of variables X and Y which are still connected, if $X \perp\!\!\!\perp Y \mid$ other l variables (which means X and Y are independent conditioning on other l variables), remove the edge between them (X and Y); (4) repeat Step 3 until l is larger than $p-2$; and (5) orient all the edges, e.g., through colliders.

One key part of implementing the PC algorithm is the independence test and conditional independence test. Given $\mathcal{S}_{oi}$, $\mathcal{S}_{di}$, $\mathcal{T}_k$ all are high-dimensional continuous feature vectors, the conventional chi-square test for categorical variables does not work. Instead, we used the kernel-based independence and conditional independence tests (Gretton et al., 2007; Zhang et al., 2012). We performed the independence test for $\mathcal{S}_{oi}$ and $\mathcal{T}_k$ to give an example. If $\mathcal{S}_{oi} \perp\!\!\!\perp \mathcal{T}_k$, we can express the null hypothesis as:

$$H_0 : P_{\mathcal{S}_{oi},\mathcal{T}_k} = P_{\mathcal{S}_{oi}} \times P_{\mathcal{T}_k}$$


MANHATTAN
Hoboken
Jersey City
New York
Statue of Liberty National Monument
BROOKLYN
Google
Map data ©2015 Google Imagery ©2015 TerraMetrics

(a) Drop-off distribution (10:00 am)

(b) Drop-off distribution (9:00 pm)

(c) Pick-up distribution (10:00 am)

(d) Pick-up distribution (9:00 pm)

Fig. 3. Pick-up and drop-off distributions in a single day.

and the alternative hypothesis as:

$$H_1 : P_{\mathcal{S}_{oi},\mathcal{T}_k} \neq P_{\mathcal{S}_{oi}} \times P_{\mathcal{T}_k}.$$

Then we defined the sample set as $Z ≔ \{(\mathcal{S}_{oi1}, \mathcal{T}_{k1}), (\mathcal{S}_{oi2}, \mathcal{T}_{k2}), \ldots, (\mathcal{S}_{oim}, \mathcal{T}_{km})\}$. If the squared Hilbert-Schmidt norm $HSIC(Z)$ is smaller than a given threshold, we will accept the null hypothesis. To compute the threshold, as suggested in Gretton et al. (2007) and Johnson, Kotz, and Balakrishnan (1994), we can make the approximation of $mHSIC_b(Z)$ as the Gamma distribution:

$$mHSIC_b(Z) \sim \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \tag{3}$$

$$\alpha = \frac{(E(HSIC_b(Z)))^2}{var(HSIC_b(Z))} \tag{4}$$

$$\beta = \frac{m * var(HSIC_b(Z))}{E(HSIC_b(Z))} \tag{5}$$

With these, we can determine the significance of independence given a significant level.

5. Prediction using Gaussian process regression

After we verified that the patterns of people mobility (outflow/inflow) are dependent on and/or caused by the characteristics of corresponding spatial areas, temporal factors, and past values, we mathematically modeled this relationship and used it for prediction. For this, we assumed that people mobility is generated from a smooth and continuous process. This process has typical amplitude and variations in the function which takes place over spatial, temporal, and other characteristics. Generally, this relationship can be expressed as:

$$x_{oi,k} = f(\mathcal{S}_{oi}, \mathcal{T}_k, x_{oi,k-1}, \ldots) \tag{6}$$

$$x_{\iota i,k} = f(\mathcal{S}_{di}, \mathcal{T}_k, x_{\iota i,k-1}, \ldots) \tag{7}$$

where $x_{oi,k}$ is the volume of outflow in the neighborhood i during time period k and $x_{\iota i,k}$ is the volume of inflow in the neighborhood i during time period k.

Note that instead of relating this relationship to some specific models such as linear, quadratic, cubic, or even non-polynomial, which may have numerous possibilities, we modeled this relationship as the free-form Gaussian process. One reason for using the Gaussian process is that for any spatial-temporal activity y (e.g., $x_{oi,k}$) to be predicted, it will likely be generated by the same process and have similar values as the historical ones sharing similar latent spatial-temporal features. We can take advantage of this relationship and use it for prediction. Formally, the Gaussian process can be represented as (Rasmussen, 2006):

$$y \sim f(X) \sim GP(m(X), K(X, X)) \tag{8}$$

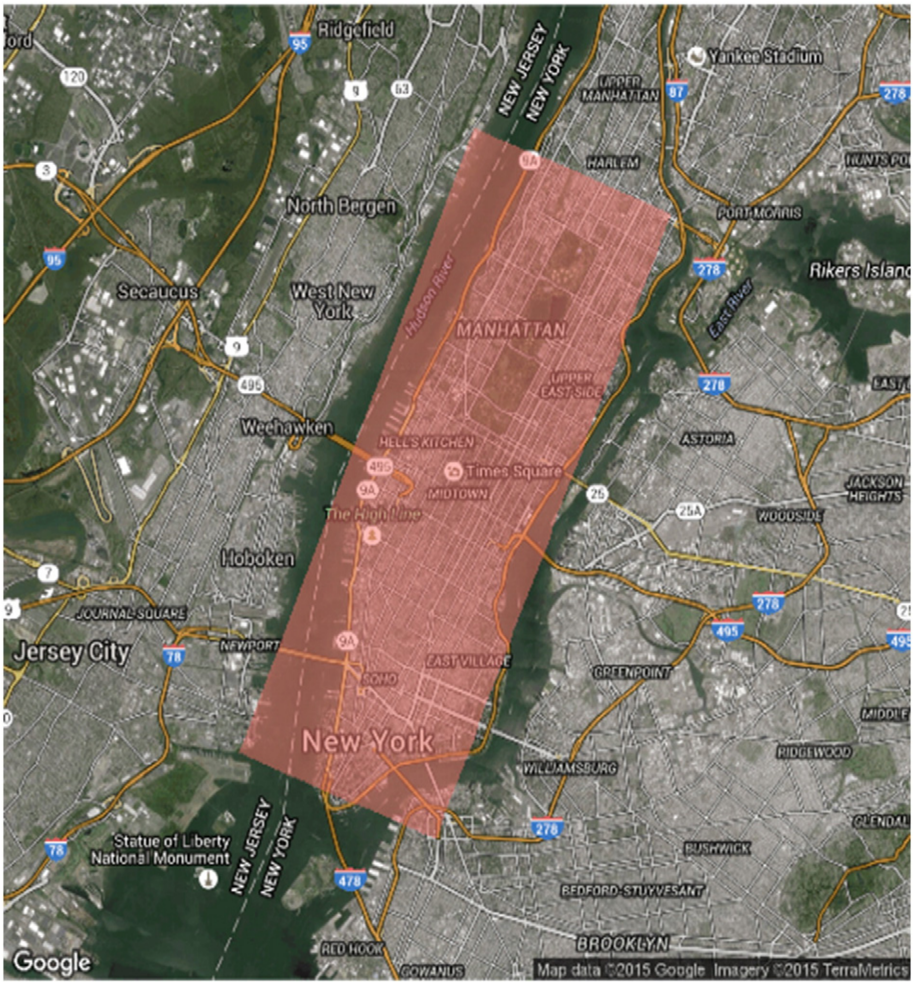


Fig. 4. The study area.

where y is a vector containing series of human activities $\{y_1, y_2, \ldots, y_n\}$, X is the features matrix of y (here for an activity $x_{oi,k}$, the corresponding feature in X would be $(\mathcal{S}_{oi}, \mathcal{T}_k,\ x_{oi,k-1},\ \ldots)$); $m(X)$ is the expected value of the generating process $f(X)$; and $K(X,X)$ is the covariance matrix where its element $k_{i,j}$ measures the similarity between the input features of activity y_i and y_j. We can also represent the relationship above as:

$$p(y(X)) \sim \mathcal{N}(m(X), K(X,X)) \tag{9}$$

For a future activity y^* to be predicted, we have:

$$p\left(\frac{y}{y^*}\right) \sim \mathcal{N}\left(\left(\frac{m(X)}{m(X^*)}\right), \begin{bmatrix} K & K^{*T} \\ K^* & K^{**} \end{bmatrix}\right) \tag{10}$$

where K, K^*, and K^{**} are the abbreviations of covariance matrix $K(X,X)$, $K(X^*,X)$, and $K(X^*,X^*)$, respectively, and T indicates matrix transpose. The key idea in Eq. (9) and Eq. (10) is that we assumed the future data are generated from the same process as the existing data. In other words, the future data and existing data have the same distribution. This is a reasonable assumption since the characteristic of a neighborhood and resulting mobility pattern there are usually stable and won't change significantly in a short time period.

Since we already had the historical datasets, we were more interested in the conditional probability of $p(y^*|y)$ that given the exiting datasets, what is the probability distribution of an unknown value y^*. Based on the transformations given in Rasmussen (2006), this conditional probability distribution is:

$$y^*|y \sim \mathcal{N}\left(m(X^*) + K^*K^{-1}(y - m(X)), K^{**} - K^*K^{-1}K^{*T}\right) \tag{11}$$

So the best estimate for y^* is the mean value of this distribution:

$$y^* = m(X^*) + K^*K^{-1}(y - m(X)) \tag{12}$$

In many applications, the general assumption is that the mean function $m(X)$ is a constant value, e.g., 0. Here we assumed $m(X)$ is a constant $\mathbb{C}_o$. Hence in our problem, the prediction for $x_{oi,k}$ became (similar for $x_{\iota i,k}$):

$$x_{oi,k} = \mathbb{C}_o + K^*K^{-1}(x_o - \mathbb{C}_o) \tag{13}$$

Note that in the input features, we have past values $x_{oi,k-1}, \ldots$, here we only considered one step backwards $x_{oi,k-1}$. One problem is that the input feature $(\mathcal{S}_{oi}, \mathcal{T}_k,\ x_{oi,k-1})$ of $x_{oi,k}$ contains three variables, spatial latent feature $\mathcal{S}_{oi}$, temporal latent feature $\mathcal{T}_k$,and past outflow volume $x_{oi,k-1}$, each having different meanings, amplitudes, and dimensions. To consider together the spatial factors, temporal factors, and flow volume,

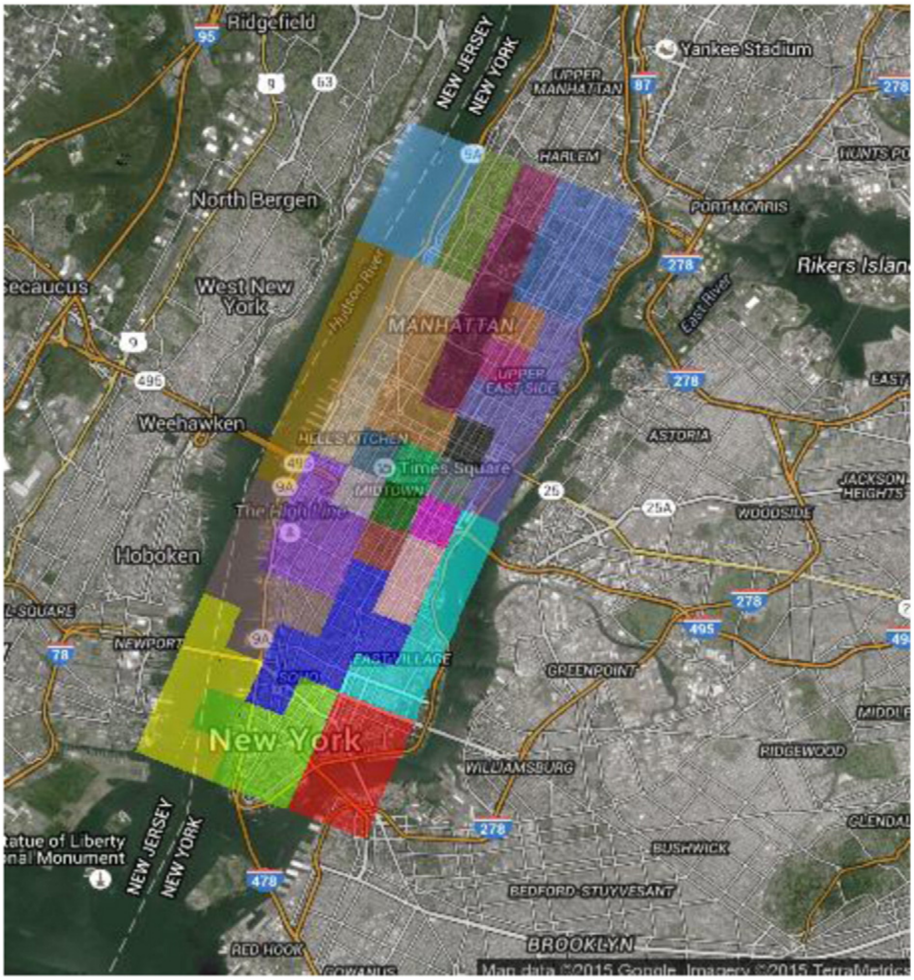


Fig. 5. The clustered neighborhoods.

we designed a new covariance function:

$$k\left(\left(\mathcal{S}_{oi1},\mathcal{T}_{k1},\ x_{oi_1,k_1-1}\right),\left(\mathcal{S}_{oi2},\mathcal{T}_{k2},\ x_{oi_2,k_2-1}\right)\right)$$
$$=\sigma_s^2\exp\left(-\frac{1}{2l_s^2}\left|\mathcal{S}_{oi1}-\mathcal{S}_{oi2}\right|^2\right)+\sigma_t^2\exp\left(-\frac{1}{2l_t^2}\left|\mathcal{T}_{k1}-\mathcal{T}_{k2}\right|^2\right)$$
$$+\sigma_p^2\exp\left(-\frac{1}{2l_p^2}\left|\,x_{oi_1,k_1-1}-x_{oi_2,k_2-1}{}^2\right.\right) \qquad (14)$$

where σ_s, σ_t, σ_p, l_s, l_t, l_p are all hyper parameters to be inferred, $|\mathcal{S}_{oi1}-\mathcal{S}_{oi2}|$, $|\mathcal{T}_{k1}-\mathcal{T}_{k2}|$, and $|x_{oi_1,k_1-1}-x_{oi_2,k_2-1}|$ are Euclidean distance between latent spatial features, temporal features, and past outflows, respectively. Eq. (14) computes the differences between spatial features, temporal features, and mobility in isolated infinity dimensional spaces and merges them together. Therefore, by defining the covariance function like this, the predictions made through Eq. (13) are based on the historical datasets of different (but similar) spatial areas, temporal time periods, and mobility trends, instead of just one specific neighborhood and time period of interest.

6. Experiment

In this section, we present a case study, for experimenting with the proposed methodology, based on the mobility data collected from New York City's taxi trips. First, we introduce and analyze the collected dataset and explore the characteristics of spatial areas that have similar latent features. Then we conduct causality analysis to mathematically verify the relationship between mobility patterns and the extracted latent features. Finally, we test the prediction accuracy of our methodology across different areas and time periods, and compared it with existing methodologies.

6.1. Dataset

Taxis play a very important transportation role in many metropolitan areas. Given the popularity and the importance of taxis, many previous works view them as the ubiquitous mobile sensors constantly probing a city's rhythm and pulse, such as traffic flows on road surfaces and citywide travel patterns of people (Zheng, Liu, Yuan, & Xie, 2011). In New York City, each day almost 13,000 taxis carry over one million passengers and make, on average, 500,000 trips—totaling over 170 million trips a year (Ferreira, Poco, Vo, Freire, & Silva, 2013). A taxicab company records each taxi trip with the pick-up time, pick-up location, drop-off time, and drop-off location. Predicting how people move around through taxis not only help optimize the taxi operation itself, but also reveals the cultural and geographic aspects of the city and detects abnormal events, among other things. The New York government has an open data project that provides data to the public including millions of taxi trip records. For our experiments, we collected the taxi trip data spanning from September 1, 2014, to October 31, 2014, where there are approximately 29 million distinct trip records totally. Given there are not enough weekend data in our dataset, our analysis will be focused on weekdays. We also exclude trips on Fridays because mobility patterns on Fridays are somewhat different from the other weekdays.

We first visualized the pick-ups and drop-offs distribution in the morning (10:00 am) and at night (09:00 pm) in a randomly selected

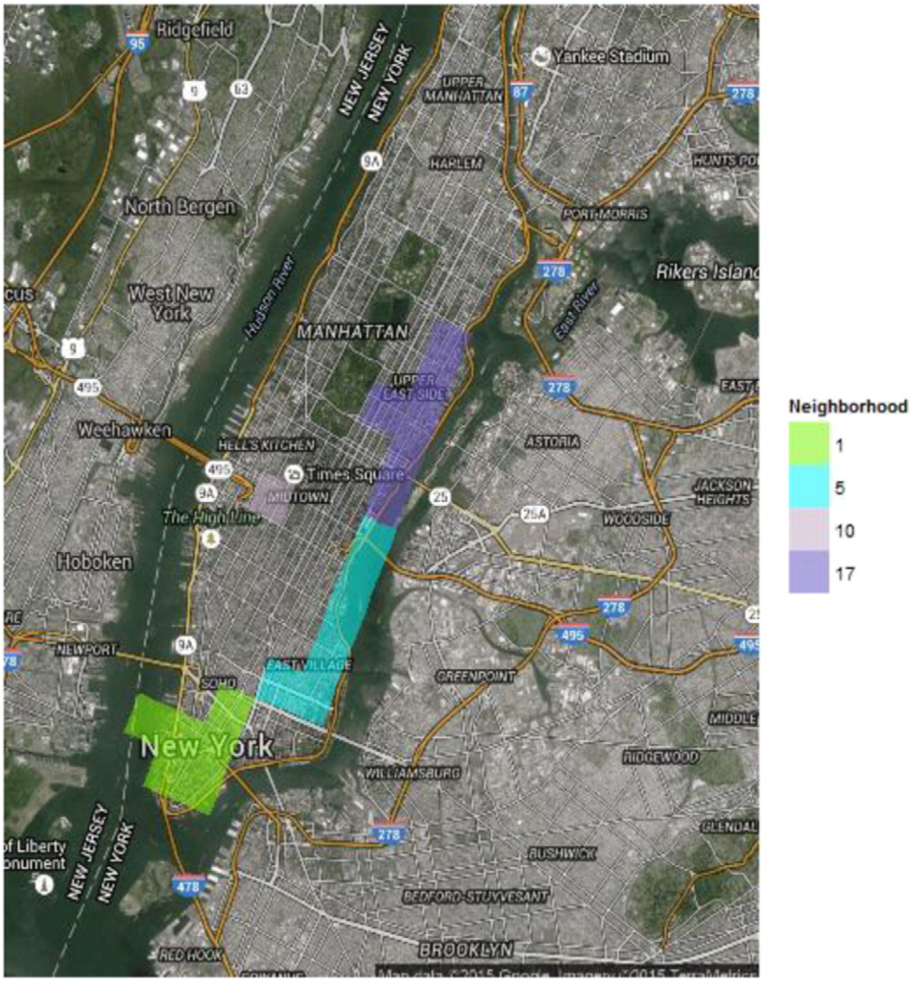


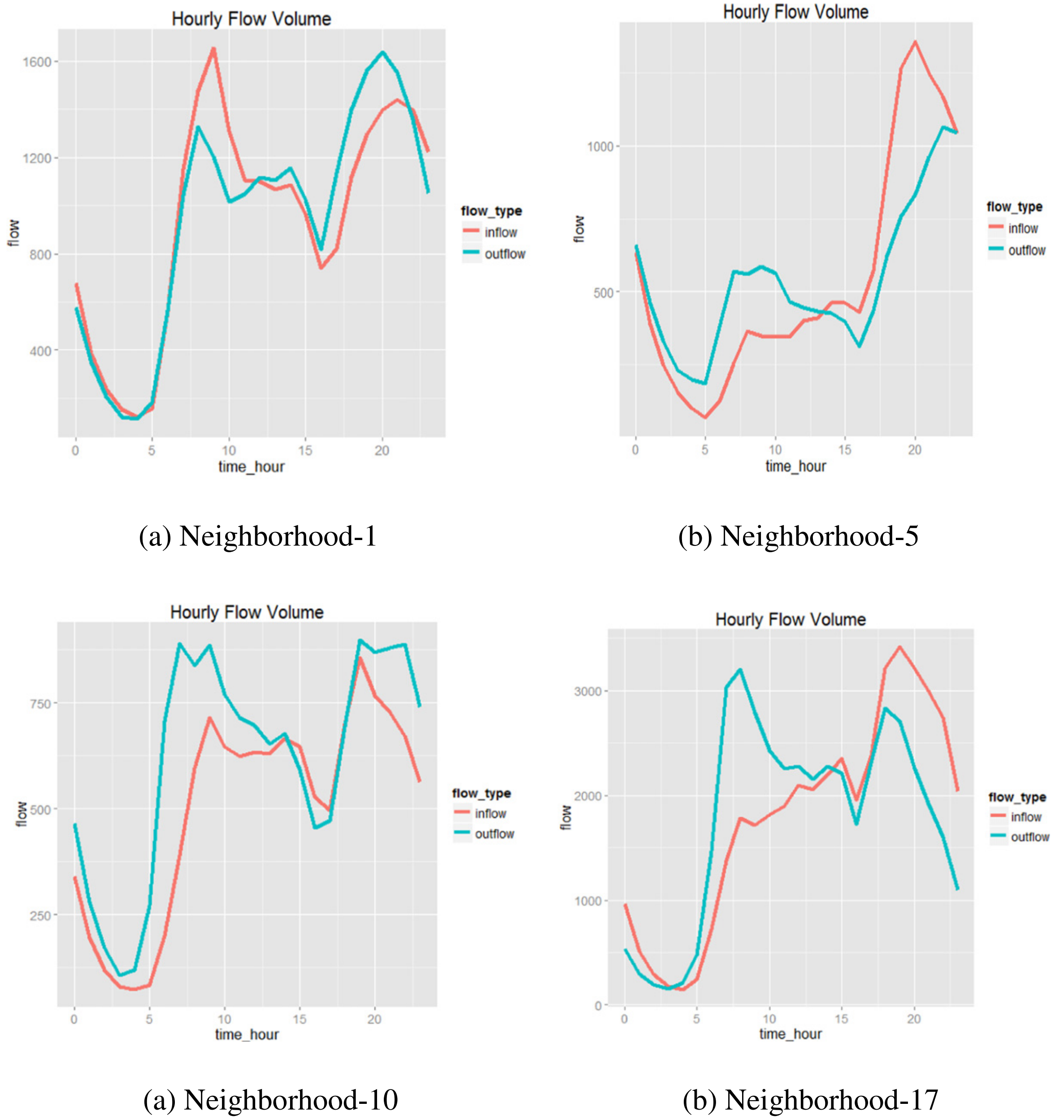
Fig. 6. Neighborhoods used in the study.

day (Fig. 3). From these visualizations we noticed most of the taxi activities happened within the Manhattan district although there were some pick-ups and drop-offs outside the Manhattan at night. Among all the neighborhoods within Manhattan district, the districts near Times Square generally have the most pick-ups and drop-offs. This phenomenon is reasonable since Times Square is a highly commercial district, with many people working there, and a tourist attraction. Another observable interesting phenomenon is that in the lower east district, there are significantly more pick-ups and drop-offs at night compared with the daytime, a sign of night life district. The spatial clustering result in the next subsection based on the extracted latent features will also confirm this.

Since most of the taxi pick-up and drop-off activities happen in Manhattan district, we will focus our analysis on that district (shown in Fig. 4). We partitioned the district into small parallelogram grids, each with approximately 0.8 km on each side. As discussed in Liu et al. (2015), while exploring human's spatial-temporal activities with social sensing data, discretizing the studied areas into spatial units with area between 0.25 km^2 and 1 km^2 would be appropriate and has been adopted by many previous works (Liu et al., 2012; Reades, Calabrese, & Ratti, 2009; Toole, Ulm, González, & Bauer, 2012). So the resolution we used (0.64 km^2 per unit) is reasonable and fine enough to demonstrate the accuracy of our prediction methodology in small areas where human's mobility patterns might have high variances. We used one hour as the time unit. We then constructed the mobility tensor as discussed in Section 3 and conducted the tensor factorization to extract the latent spatial features of each grid as the origin and destination and the latent temporal features in each hour.

6.2. Neighborhood characteristics

With the extracted latent features, we explored how the mobility patterns of neighborhoods with different latent spatial features would differ. In particular, we clustered the grids with similar latent spatial features. Since each grid can be either an origin or a destination, we defined



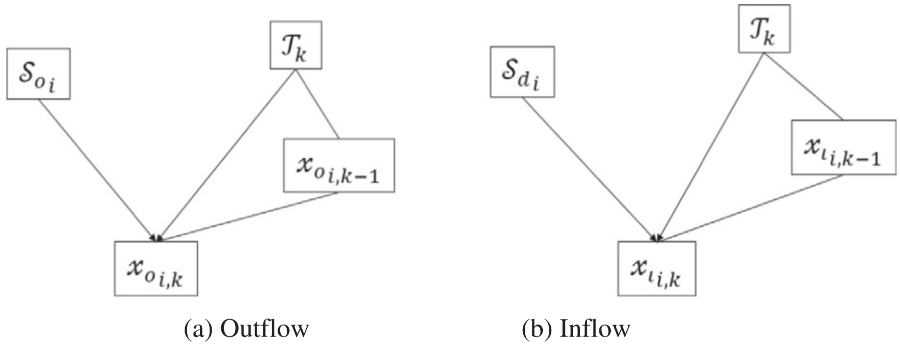
(a) Neighborhood-1

(b) Neighborhood-5

(a) Neighborhood-10

(b) Neighborhood-17

Fig. 7. Average hourly inflow/outflow of selected neighborhoods.



(a) Outflow (b) Inflow

Fig. 8. The causal structures.

the mobility feature vector of grid *i* as:

$$\mathcal{S}_i = (\mathcal{S}_{oi}, \mathcal{S}_{di}) \tag{15}$$

and the distance between the two grids *i* and *j* as:

$$s_{ij} = |\mathcal{S}_i - \mathcal{S}_j * \frac{\mathcal{S}_i * \mathcal{S}_j}{|\mathcal{S}_i| * |\mathcal{S}_j|}| \tag{16}$$

Note that $\frac{\mathcal{S}_i * \mathcal{S}_j}{|\mathcal{S}_i| * |\mathcal{S}_j|}$ is the cosine between two spatial vectors. This distance function takes both direction and magnitude of the latent spatial features into account.

To cluster the grids with similar spatial latent features in neighborhoods, we adapted a bottom-up spatial hierarchical clustering approach. Specifically, in the beginning we assumed every grid is a neighborhood. Then we iteratively searched the pair of adjacent neighborhoods that have the smallest complete-linkage and merged them together. We repeated this merging procedure until certain criteria are met; for example, the smallest complete-linkage is larger than a given threshold. The clustered result is shown in Fig. 5. Note that the size of clusters varies as some central areas (e.g., the ones close to Times Square) would form smaller clusters while the other suburban areas tend to form larger clusters. This phenomenon is reasonable and similar to some of the previous spatial clustering works (Zhi et al., 2016) because there are more taxi activities in those central areas than other places, and the clustering result should reflect this.

With the clustered neighborhoods, we can explore mobility patterns between them. For our analysis, we chose four representative neighborhoods: 1, 5, 10, and 17 (see Fig. 6). We plotted their average volume of inflow and outflow in a day (see Fig. 7). One notable common pattern among all four neighborhoods (but unrelated to neighborhood characteristics) is the drop of outflow volume between 3:00 pm and 4:00 pm that is caused by the shift switch of taxi drivers. We also observed that these four neighborhoods have very unique mobility patterns. The inflow volume of neighborhood 1 has the highest peak in the morning at around 9:00 am while its outflow volume has the highest peak at around 8:00 pm, which indicates neighborhood 1 is an office district; in fact, neighborhood 1 is mainly composed of financial district, one of the busiest business and tourist areas in New York City. On the other side, the highest peaks of neighborhood 17's outflow and inflow are in the morning and evening, respectively, which implies it is mainly a residential district mixed with some other functions. As for neighborhood 10, it has significant inflows and outflows in almost all times starting from 10:00 am to midnight; neighborhood 10 is Times Square, one of New York City's most popular tourist attractions. Different from all areas, neighborhood 5 has relatively low inflow and outflow volume in daytime but both activities increase significantly in the evening, a sign of nightlife district. From these examples we can infer that our extracted latent features generally distinguish between different neighborhoods with diverse unique characteristics.

6.3. Causality analysis

After exploration of the characteristics of neighborhoods with different latent spatial features, we mathematically verified the relationship among them. More specifically, given a specific neighborhood *i*, we explored the causality relationship between its outflow $x_{oi,k}$ at time period *k*, its previous record $x_{oi,k-1}$, its latent spatial feature $\mathcal{S}_{oi}$, and the latent temporal feature $\mathcal{T}_k$ at time period *k*. For this, we used the PC algorithm (Spirtes et al., 2000) and performed the independence test and conditional independence test based on the method described in Gretton et al. (2007) and Zhang et al. (2012), and we studied the outflow/inflow independently. The inputs to the algorithms are the hourly outflow/inflow records of each small parallelogram grids and their corresponding spatial and temporal latent features. The outputted causal graphs are plotted in Fig. 8, which shows both outflow and inflow having similar causal structures. Take the outflow $x_{oi,k}$ as an example. This spatial-temporal value $x_{oi,k}$ is generally decided by the corresponding latent spatial features $\mathcal{S}_{oi}$ and latent temporal feature $\mathcal{T}_k$ and is dependent on its previous record $x_{oi,k-1}$. This result verified our hypothesis.

6.4. Prediction

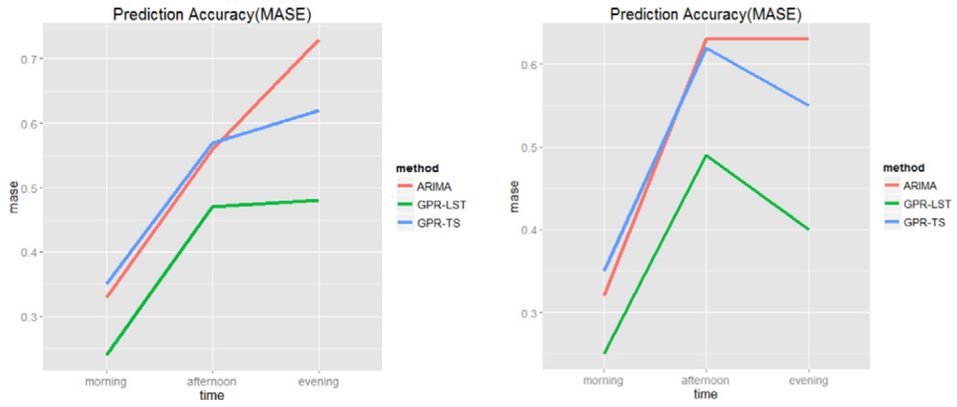
We tested the prediction accuracy of our methodology and compared it with existing ones. We named our methodology (Gaussian process regression with latent spatial and temporal features) as GPR-LST for short and compared it with two existing methodologies. One methodology is the parametric seasonal ARIMA model where we take each grid as a fixed point and build seasonal ARIMA models for its time-series outflow and inflow, respectively. Another methodology is the non-parametric

Table 1
Prediction accuracy of outflow.

	MSE	MASE
GPR-LST	1919	0.38
ARIMA	3419	0.53
GPR-Naive	3276	0.50

Table 2
Prediction accuracy of inflow.

	MSE	MASE
GPR-LST	1173	0.36
ARIMA	2455	0.50
GPR-Naive	2011	0.48

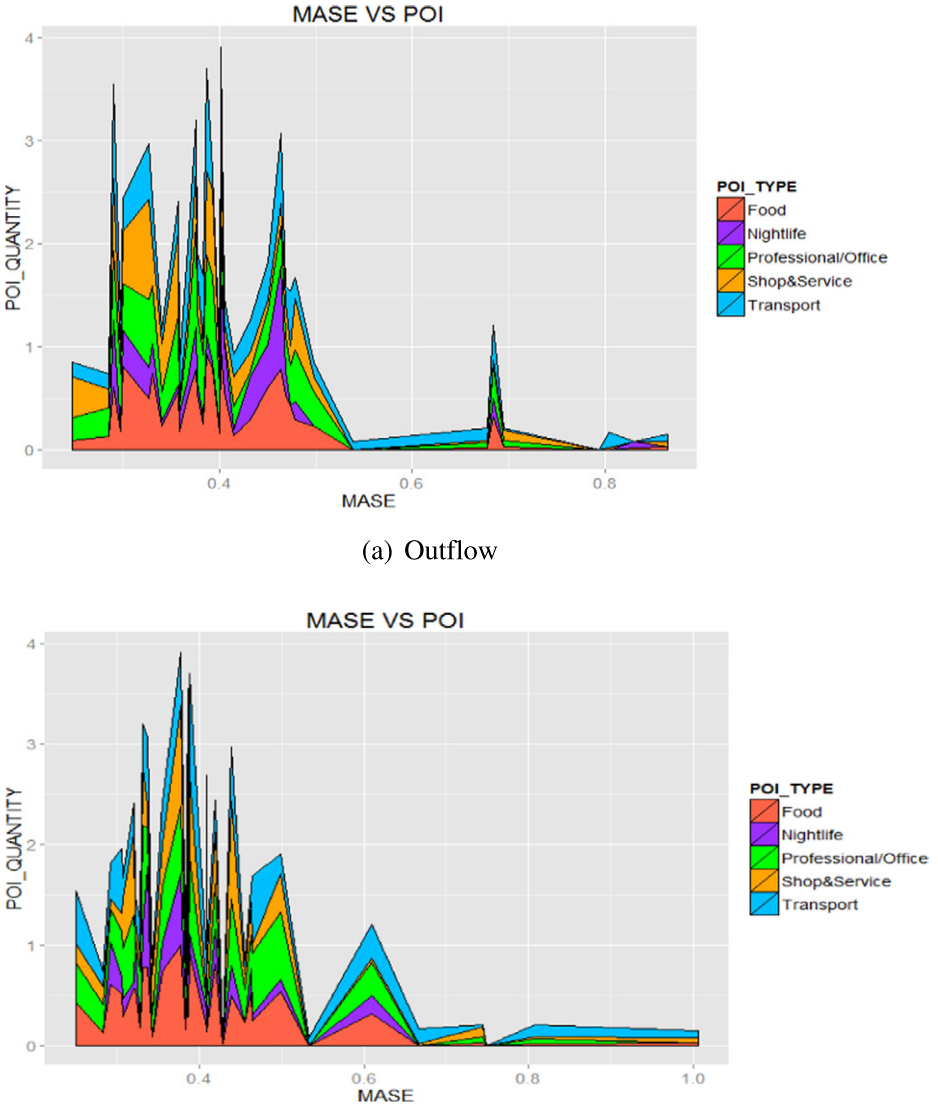


(a) Outflow

(b) Inflow

Fig. 9. Prediction accuracy (MASE) at different time periods.

model, naive Gaussian Process regression (GPR), which uses the explicit previous time-serious records like ($x_{oi,k-1}$, $x_{oi,k-2}$, $x_{oi,k-3}$, ...,) as the input features and the squared exponential kernel with a separate length scale per predictor as the covariance function. We named this methodology (Naive Gaussian process regression for time series records) as GPR-Naive for short. We have one GPR-Naive model for outflow and one



(a) Outflow

(b) Inflow

Fig. 10. The prediction accuracy (MASE) at different spatial units.

GPR-Naive model for inflow. We used the first 6 weeks data in our dataset as the training set and the remaining 2 weeks data for the test and performed all the methodologies on each parallelogram grid and predicted each parallelogram grid's inflow and outflow in the next hour sequentially.

For prediction accuracy measurements, we used the mean squared error (MSE) and mean absolute scaled error (MASE) (Franses, 2016) instead of absolute percentage error (MAPE). This is because many areas usually have very different scales of hourly inflow/outflow within a day (or even have zero records at times like late midnight), which makes the mean absolute percentage error (MAPE) inappropriate. Eq. (17) shows how MASE is calculated.

$$MASE = \frac{1}{T}\sum_{t=1}^{T}\frac{|e_t|}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|} \tag{17}$$

where e_t is the error of prediction at time t. The general idea of MASE is to compare the prediction methodology with the naive one-step forecast methodology that makes predictions based on the previous value, e.g., to predict human's outflow $x_{oi,k}$ at time period k; the one-step forecast methodology uses the value of $x_{oi,k-1}$ directly.

The prediction errors of different methodologies are shown in Table 1 and Table 2. From these two tables we can see that our proposed methodology is the most accurate one. For example, for outflow, our methodology has a much lower MASE (0.38) than the seasonal ARIMA (0.53).

We further analyzed the prediction accuracies of different methodologies at different time periods. We separated a day into three main time periods, morning (6:00 am–11:59 am), afternoon (12:00 pm–17:59 pm), and evening (18:00 pm–23:59 pm) and plotted the prediction accuracies of different methodologies in Fig. 9. From these plots, we can see that our proposed methodology (GPR-LST) performs best at any time period. Apart from this, there are also some other interesting observations worth mentioning. The first one is that all methodologies are most accurate in the morning and are generally less accurate in the afternoon and evening. One explanation for this could be that people's mobility patterns in the morning are quite simple since most people probably would head to work places in the morning. However, people's mobility patterns are more complicated in the afternoon and evening since they might go to restaurants, places for business, shopping mall, night clubs, etc., which makes the prediction more difficult. The second observation is that for inflow, the prediction accuracies in the evening are generally higher than the accuracies in the afternoon. One reason for this could be that people's inflow patterns in the evening become simpler compared with the one in the afternoon (although not as simple as the morning) since people are more likely to go to residential districts, night life neighborhoods, and little chance to office districts in the evening, which would make the prediction easier.

From the experiments above, we can see that our proposed methodology performed best, compared to some of the existing methodologies, and reduced the prediction error significantly. Furthermore, we assessed how our prediction methodology performed across different regions. More specifically, for each partitioned grid, we explored the relationship between the prediction accuracy of our methodology and the POI (point of interest) distribution. We collected POI data from the OpenStreetMap (OpenStreetMap) and focused on 5 types of POIs: food, nightlife, professional/office, shop & service, transport. We do not consider the residential data here because the residential data in OpenStreetMap is very sparse and incomplete. Note that the size of different POI types varies, e.g., in an office area, there could be more restaurants than actual offices. Hence, it is difficult to judge the function of a region based on the absolute number of POIs. To address this, we normalize the scale of each POI type in each partitioned grid into [0,1] with:

$$P'_{i,k} = \frac{P_{i,k} - \min_i(P_{i,k})}{\max_i(P_{i,k}) - \min_i(P_{i,k})} \tag{18}$$

where $P_{i,k}$ is the number of POI of type k within grid i and $P'_{i,k}$ is the normalized $P_{i,k}$. We plot the prediction accuracy (MASE) and the normalized POI values of each grid in Fig. 10. It is a stacked area plot where the x-axis indicates the MASE of our prediction methodology for different grids and the y-axis indicates the normalized value of different POIs in the corresponding grid. From the plot, we can see when there are certain amounts of POIs (the sum of normalized POI values is larger than a threshold, like 0.8) in an area, our prediction methodology generally achieves a high accuracy (the MASE is less than 0.5). This makes sense since in the urban areas with more POIs and more people's activities, the pattern of taxis' pick-ups and drop-offs tend to be more regular compared to suburban areas where people would take taxi less frequently and more randomly. But this relationship does not change smoothly. In other words, there is no strict increase/decrease function and some exceptions do exist. One reason for this is the inherent complication of human's mobility pattern, and many people usually do not take taxi frequently and regularly. Another reason could be that our collected POI data is not very complete, e.g., lack of residential data and the scale/popular of each POI is also not considered here, e.g., a big office POI like New York City Hall would definitely have a larger impact on the taxi

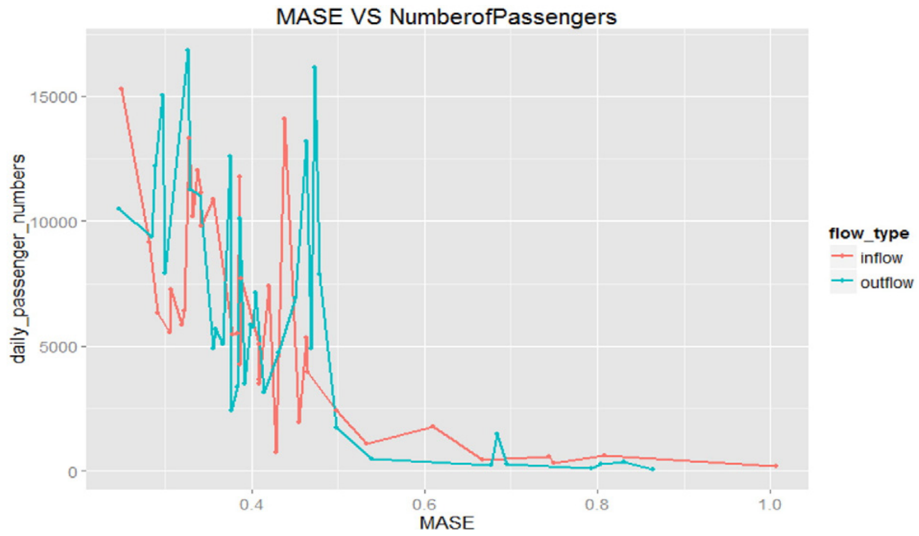


Fig. 11. The prediction accuracy (MASE) vs. number of pick-ups and drop-offs.

demand than a POI of small company. Lastly, our sample is relatively small, with less than hundred grids in a city.

Besides the number of POIs, we also explored the relationship between prediction accuracy and the number of passengers in each area. The result is plotted in Fig. 11, from which we can see when more people took taxis in an area (more than 2500 pick-ups/drop-offs a day), our prediction methodology achieved quite high prediction accuracies (with MASE less than 0.5), confirming one of our hypotheses that when there are more human activities, it is easier to predict the number of pick-ups and drop-offs. But this relationship is also not a strict increase/decrease function.

7. Conclusion and future research

In this work, we proposed a new methodology for the prediction of spatial-temporal activities like human mobility, especially the inflow and outflow of people in neighborhoods/areas during certain time periods. Our methodology comprised three steps: (1) use of a 3D tensor to model human mobility and extract latent spatial and temporal features of different neighborhoods and time periods through tensor factorization; (2) explored and verified the dependent relationship between mobility patterns and the extracted latent spatial and temporal features; and (3) modeled this relationship as a Gaussian process for prediction of human mobility.

For validation of the proposed methodology, we experimented with New York City's taxi trips. The results showed that our extracted latent features successfully distinguish between neighborhoods with diverse unique characteristics such as nightlife areas, office districts, and residential neighborhoods, among other characteristics. Through the PC algorithm, we further verified our hypothesis that there is a causality relationship between human mobility and the extracted latent spatial and temporal features. The results also showed that our methodology achieved a much higher accuracy than existing ones (MASE error is reduced by 20%+). Lastly, we also explored how our methodology performed across different areas and times.

With regard to future research directions, we would (1) test our proposed methodology with the data in other sources; (2) detect the abnormal events; and (3) incorporate multisource data (e.g., weather, citi-bike system, public bus, and subways) into the tensor model, and design corresponding new mean and covariance functions of the Gaussian process to address the relationship among multisources in order to achieve a higher prediction accuracy.

References

Baltrunas, L., Ludwig, B., & Ricci, F. (2011). Matrix factorization techniques for context aware recommendation. *Proceedings of the fifth ACM conference on recommender systems* (pp. 301–304).

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. *International AAAI conference on weblogs and social media* (pp. 58).

Fan, Z., Song, X., & Shibasaki, R. (2014). Cityspectrum: A non-negative tensor factorization approach. *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 213–223).

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, *19*, 2149–2158.

Franses, P. H. (2016). A note on the mean absolute scaled error. *International Journal of Forecasting*, *32*, 20–22.

Froehlich, J., Neumann, J., & Oliver, N. (2009). Sensing and predicting the pulse of the city through shared bicycling. *IJCAI* (pp. 1420–1426).

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, *17*, 463–481.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2007). A kernel statistical test of independence. *Advances in neural information processing systems* (pp. 585–592).

Guo, D., Zhu, X., Jin, H., Gao, P., & Andris, C. (2012). Discovering spatial patterns in origin-destination mobility data. *Transactions in GIS*, *16*, 411–429.

Guo, Q., Luo, J., Li, G., Wang, X., & Geroliminis, N. (2013). A data-driven approach for convergence prediction on road network. *Web and wireless geographical information systems* (pp. 41–53). Springer.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., & Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, *49*, 362–378.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions. 1–2*. John Wiley & Sons: New York.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., & Banchs, R. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, *6*, 455–466.

Kang, C., & Qin, K. (2016). Understanding operation behaviors of taxicabs in cities by matrix factorization. *Computers, Environment and Urban Systems*, *60*, 79–88.

Kim, K., Lee, D., & Essa, I. (2011). Gaussian process regression flow for analysis of motion trajectories. *Computer vision (ICCV), 2011 IEEE international conference on* (pp. 1164–1171).

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*, 455–500.

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning*, *106*, 73–87.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., et al. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*, 512–530.

Liu, X., Kang, C., Gong, L., & Liu, Y. (2016). Incorporating spatial interaction patterns in classifying and understanding urban land use. *International Journal of Geographical Information Science*, *30*, 334–350.

Matthias, H. -P. K. M. R., & Zuefle, S. A. (2008). *Statistical density prediction in traffic networks*.

Noulas, A., & Mascolo, C. (2013). Exploiting foursquare and cellular data to infer user activity in urban environments. *Mobile Data Management (MDM), 2013 IEEE 14th international conference on* (pp. 167–176).

OpenStreetMap (d). Available https://www.openstreetmap.org

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. -L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, *28*, 1988–2007.

Peng, C., Jin, X., Wong, K. -C., Shi, M., & Liò, P. (2012). Collective human mobility pattern from taxi trips in urban area. *PloS One*, *7*, e34487.

Rasmussen, C. E. (2006). *Gaussian processes for machine learning*.

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: Analysing cities using the space–time structure of the mobile phone network. *Environment and Planning. B, Planning & Design*, *36*, 824–836.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *371*, 20110550.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V., & Campbell, A. T. (2011). NextPlace: A spatio-temporal prediction framework for pervasive systems. *Pervasive computing* (pp. 152–169). Springer.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search. 81*.MIT press.

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring land use from mobile phone activity. *Proceedings of the ACM SIGKDD international workshop on urban computing* (pp. 1–8).

Zhang, K., & Hyvärinen, A. (2009). Causality discovery with additive disturbances: An information-theoretical perspective. *Machine learning and knowledge discovery in databases* (pp. 570–585). Springer.

Zhang, K., & Pelechrinis, K. (2016). Do street fairs boost local businesses? A quasi-experimental analysis using social network data. *Joint European conference on machine learning and knowledge discovery in databases* (pp. 161–176).

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv*, *1202*(3775).

Zhang, F., Wilkie, D., Zheng, Y., & Xie, X. (2013). Sensing the pulse of urban refueling behavior. *Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing* (pp. 13–22).

Zhang, K., Lin, Y. -R., & Pelechrinis, K. (2016). EigenTransitions with hypothesis testing: The anatomy of urban mobility. *Tenth international AAAI conference on web and social media*.

Zheng, Y., Liu, Y., Yuan, J., & Xie, X. (2011). Urban computing with taxicabs. *Proceedings of the 13th international conference on ubiquitous computing* (pp. 89–98).

Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., & Chang, E. (2014). Diagnosing New York city's noises with ubiquitous data. *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 715–725).

Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., et al. (2016). Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science*, 1–12.