Pengfei Chen, Wenzhong Shi, Xiaolin Zhou, Zhewei Liu & Xuandi Fu (2019) STLP-GSM: a method to predict future locations of individuals based on geotagged social media data, International Journal of Geographical Information Science, 33:12, 2337-2362, DOI: 10.1080/13658816.2019.1630630

STLP-GSM: a method to predict future locations of individuals based on geotagged social media data

An increasing number of social media users are becoming used to disseminate activities through geotagged posts. The massive available geotagged posts enable collections of users' footprints over time and offer effective opportunities for mobility prediction. Using geotagged posts for spatio-temporal prediction of future location, however, is challenging. Previous studies either focus on next-place prediction or rely on dense data sources such as GPS data. Introduced in this article is a novel method for future location prediction of individuals based on geotagged social media data. This method employs the hierarchical density-based clustering algorithm with adaptive parameter selection to identify the regions frequently visited by a social media user. A multi-feature weighted Bayesian model is then developed to forecast users' spatio-temporal locations by combining multiple factors affecting human mobility patterns. Further, an updating strategy is designed to efficiently adjust, over time, the proposed model to the dynamics in users' mobility patterns. Based on two real-life datasets, the proposed approach outperforms a state-of-the-art method in prediction accuracy by up to 5.34% and 3.30%. Tests show prediction reliability is high with quality predictions, but low in the identification of erroneous locations.

Keywords: spatio-temporal location, prediction uncertainty, daily trajectory, online footprint, social network

1. Introduction

The popularity of human mobility research arises because of its significant academic value and in particular the value of its application. Understanding human mobility

patterns helps capture the regularity of individual movements and behaviour, thereby supporting the related management and policymaking. Currently, human mobility has attracted the increasing attention of scholars and experts in various scientific fields, such as urban informatics (Gao 2015), geographic information science (Belyi *et al.* 2017), sociology (Cresswell 2010), epidemiology (Meloni *et al.* 2011), economics (Heckman and Mosso 2014). With regard to the study of human mobility in urban system, one of the main issues is human mobility prediction based on individuals' historical mobility data (González *et al.* 2008, Song *et al.* 2010).

Cell phone data has been widely acknowledged as one of the best provision sources of people location information in the area of human mobility analysis (Hidalgo and Rodríguez-Sickert 2008, Song *et al.* 2010). As a piece of what has become indispensable equipment in people's lives, a cell phone plays an important role in providing a record of human daily activities and movements. Traditionally, the location of a cell phone can be obtained either through call detail records (CDRs) or simply via GPS (Lv *et al.* 2017). CDRs, however, are strictly controlled by communication companies and difficult to access due to privacy concerns and security issues (Candia *et al.* 2008, Mosenia *et al.* 2017). Currently, the collection of GPS data requires support from third-party software and remains subject to privacy concerns (Huang 2017). Such disadvantages hamper the application of CDRs and GPS data in the study of human mobility, especially when the population is large scale and hence the data collection even more difficult.

Social networks, such as Twitter, Instagram, Facebook, could be a feasible alternative to GPS data and CDRs, for capturing locations of cell phone users (Huang 2017). In contrast with the restricted collection of GPS data and CDRs data, many social networks provide application programming interfaces (APIs), making the information of users' social media activities available for researches and business analysis (Oleksiak 2014). At present, a geotagged social media post which contains geographic location information, is becoming increasingly popular among social media users when sharing their lives and thoughts. According to a survey (Morstatter *et al.* 2013), Twitter alone can produce roughly 5 to 15 million geotagged posts per day. The large number of geotagged posts offers continuous updates and, in particular, provides long-term sequences of users' spatio-temporal locations for effective use in the study of human mobility prediction.

Individuals' daily activities exhibit a high level of temporal and spatial regularity (Gonzalez *et al.* 2008). This regularity is affected by many mobility features (Noulas *et al.* 2012), such as activity transition, popularity of places, periodical activities and geographic distance. Those mobility features serve as basic elements in the building of a mobility prediction model. Using social media data alone to conduct mobility prediction, however, is still challenging. Social media data, taken virtually as sparse sample points of the users' activities over time, however, will inevitably introduce sampling uncertainty during the generalization of any mobility feature. To reduce that sampling uncertainty, a long sequence of data is always required (Huang 2017). In addition, a combination of multiple mobility features can provide

compensatory effects to improve prediction accuracy (Noulas *et al.* 2012). Nevertheless, given a set of mobility features, different people may have a different degree of predictability for each (Lv *et al.* 2017). Simple uses of the fixed model to combine those mobility features ignore the personal characteristics thus leading to erroneous prediction results.

It is of importance to note that individuals' mobility patterns may change over time due to many factors, such as physical condition, family status, moving, job hopping and road reconstruction. (Sevtsuk and Ratti 2010). The change of mobility pattern can be either temporary or perpetual, hence demanding an updating process of the prediction model. The traditional updating of mobility prediction models accumulates new data (Cho 2016, Lv *et al.* 2017). However, due to the limited amount of new daily social media data from one person in the operation, the accumulation approach could suffer from the time lag of the historical data, thereby undermining the performance of prediction models.

To address the challenges indicated above, the study presented in this paper proposes a novel method for spatio-temporal prediction of individual's future locations based on geotagged social media data (STLP-GSM). The core of STLP-GSM is a prediction model: the multi-feature weighted Bayesian model (MWBM), which integrates multiple mobility features derived from users' historical trajectory data using the Naïve-Bayes algorithm and weighting the mobility features according to their predictability. The objectives of the study presented in this paper include: (1) to identify the activity zones of an individual by analyzing geotagged posts through clustering

techniques; (2) to formulate selected mobility features affecting individuals' mobility pattern; (3) to construct a model of an extensible structure that can include multiple mobility features for mobility prediction; (4) to predict individuals' spatio-temporal location and analyze the reliability of the predictions; (5) to evaluate the proposed approach on several social media data sets and validate the advantages of the approach, by outlining comparisons with state-of-the-art methods. The significance of this research is summarized as follows: (1) filling the research gap, using geotagged social media data for spatio-temporal location prediction; (2) developing a prototype of a probabilistic prediction model, which is theoretically extensible when more mobility features become available; (3) enhancing the understanding of social media users' mobility patterns, to better enable personalized location-based services in practical applications.

This paper is further organized as follows: the related work is reviewed in Section 2; the problem statement and some definitions are given in Section 3; the methodology of the proposed approach is discussed in Section 4; the experiments are demonstrated in Section 5; a discussion and conclusion are given in Section 6.

2. Related work

2.1.Social dynamics analysis using social media data

Over the last decade, social media data have been widely utilized in social dynamics analysis. Cao *et al*. (2015) proposed a spatio-temporal data cube model, based on social media data, which provides a hierarchical structure of cuboids for scalable spatial

analysis. Cranshaw *et al.* (2012) proposed a new methodology, using massive check-in data, to identify the so-called *Livehood* in a city. By considering both spatial and social proximity among venues, Livehood can better explain the activity patterns of residences than the traditional municipal boundary and can yield insights into the impacts from various factors on those patterns. Cesario *et al* (2017) proposed a novel approach for discovering mobility patterns on the basis of geotagged social media data in large-scale public events. This approach can estimate the number of people attending each event, as well as, for given groups of users, those association rules among these events. Altomare *et al.* (2017) established a Cloud platform for the computation of massive trajectory data. Through tests on both real-life and synthetic data, the proposed platform exhibited a significant improvement regarding data throughput, execution time, and speed. These studies have demonstrated the usefulness of social media data and laid a foundation for future research in related fields.

*2.2.*Location prediction

Predicting locations for individuals is an important issue in social dynamics analysis. There are two main approaches in current studies regarding location prediction: namely, the next-location prediction and spatio-temporal prediction. The main related works, including their prediction categories, main data sources, and basic models, are listed in Table 1.

Table 1. List of the main related works

Related work	Prediction category	Main data source	Basic model
Noulas *et al*. (2012)	Next-location	Check-in	M5 model tree
Akoush and Sameh (2007)	Next-location	CDR	Neural network
Ying *et al.* (2011)	Next-location	GPS	Semantic analysis
Gambs et al. (2011, 2012)	Next-location	GPS	Markov chain
Mathew *et al*. (2012)	Next-location	GPS	HMM
Huang (2017)	Next-location	Social media	Markov chain
Lv *et al*. (2017)	Both spatial-temporal and next-location	Cellular data	HMM
Hadachi *et al*. (2014)	Spatio-temporal	CDR	Markov chain
Alvarez-Lozano *et al*. (2015)	Spatio-temporal	GPS	HMM
Liu *et al.* (2016)	Spatio-temporal	Check-in	Neural network
Scellato *et al.* (2011)	Spatio-temporal	GPS	Time series analysis

Next-location prediction: Next-location prediction aims to predict the site where an individual will go next, based on that individual's visited places in the past. Different approaches have been implemented for next-location prediction in existing studies, such as M5 tree model (Noulas *et al.* 2012), neural network (Akoush and Sameh 2007) and in the semantic analysis approach (Ying et al. 2011). Among these approaches, Markov-based approaches, such as the Markov chain model and Hidden Markov model (HMM), appear to be the most popular in the literature. For example, Mathew *et al*. (2012) applied HMM for human mobility prediction with an accuracy of 13.85% on a real-life dataset. Gambs *et al*. (2011) proposed a Mobility Markov chain (MMC) model for next-location prediction by calculating a transition matrix based on the sequence of places visited by an individual. This model was further developed as n-MMC by including n previously visited places in the calculation of the transition matrix. Huang (2017) proposed a sparse mobility Markov chain (SMMC) model for next-location prediction of social media users. According to a test on 52 Twitter users in Washington,

DC, the SMMC model was able to gain about a 2% prediction accuracy enhancement than the common Markov chain model. The best prediction accuracy was reported to be as high as 78.94%. The SMMC model result implies the great potential of the application of geotagged social media data in location prediction.

Spatio-temporal prediction: Spatio-temporal prediction aims to forecast the location of an individual at a specific time in the future. Unlike next-location prediction, spatio-temporal prediction provides temporal information, therefore more mobility features, such as temporal habits of people, need to be involved in any consideration, which may make spatio-temporal models commonly more complex than the models for next-location prediction. For example, considering the impact of living habits on user's mobility patterns, Lv *et al*. (2017) proposed two models for spatial and spatio-temporal location prediction of cell phone users, based on HMM. Hadachi *et al*. (2014) developed four algorithms for mobility prediction, using CDR data, by integrating the behavior rule and temporal rule into classic first- and second-order Markov chain models. Alvarez-Lozano et al. (2015) presented a medium-term prediction model based on HMM and tested on GPS data, from 63 mobile users. The study indicated the possibility of some users having a high change rate in their behavior and emphasized the importance of model updating. Scellato *et al.* (2011) presented a NextPlace approach that used time series analysis to estimate a user's spatio-temporal behavior. Unlike Markov-based methods, the NextPlace approach emphasized the predictability of a user and enabled the estimation of the time duration of a user may stay in relevant

places. Liu *et al.* (2016) introduced the Spatial Temporal Recurrent Neural Networks (ST-RNN) for spatio-temporal prediction, based on users' check-ins on social media. ST-RNN discretized continuous spatial and temporal values through multiple bins and implemented spatial-temporal transition metrics into the recurrent architecture. The latter improved the prediction accuracy of experimental datasets.

Despite the advance of existing models, difficulties remain in practice for models relying on GPS, CDR, or Cellular data, all of which have low accessibility due to privacy, security and commercial issues. The accessibility problem can, however, be "all-or-nothing", which results in researchers either having to pay for data or being authorized to collect. Regarding check-in data, the recorded trajectories may be biased as it is common for people to decline to check in when they visit a less popular venue (Joseph *et al.* 2012). In contrast, geotagged social media data has the advantage of being able to be obtained freely through the given API. However, although social media data are used in the SMMC model for next-location prediction of users (Huang 2017), the SMMC model cannot provide the temporal information of individuals' movements, hence reducing its utility in practical applications. To reflect the spatio-temporal patterns in individuals' movements, it is essential to develop new approaches for spatio-temporal prediction based on geotagged social media data. Of note, is that current prediction models cannot be extended due to their relatively fixed structure. The latter hinders operations to include extra potentially available mobility information. Lastly, updating mechanisms by simply accumulating new data in the current model needs to be improved to cater for adjustments to any changes in individuals' mobility patterns.

Hence, the research presented in this paper aims to develop an extensible model for spatio-temporal location prediction using geotagged social media data. An updating strategy is also designed to account for and accommodate users' dynamic mobility patterns.

2.3.Clustering approaches for regions of interest extraction

Some points of interest (POIs) or regions of interest (ROIs) are quite predictable due to the high level of repetition occurring in people's daily routines (Yuan *et al.* 2017). As POI may reflect unrealistic mobility patterns of social media users (Huang 2017), in this study ROI is used as the basic unit to provide users' daily mobility information. Clustering techniques, including density-based clustering (Zhou *et al.* 2007, Zheng *et al.* 2012, Huang 2017), K-means clustering (Yuan *et al.* 2012), hierarchical clustering (Xu *et al.* 2015), and other algorithms such as G-RoI (Belcastro *et al.* 2018), are commonly employed for ROI identification. The use of the density-based spatial clustering algorithm (DBSCAN) is the most popular, as it is effective in the detection of clusters of varying shapes (Ester *et al.* 1996). The performance of DBSCAN, however, has strong sensitivity to the two parameters: *radius* (*Eps*) *and minimum neighbours* (*MinPts*), both of which are always set fixedly with empirical values (Zhou *et al.* 2007, Huang 2017).

Compared with density-based clustering approaches, to obtain meaningful clusters, hierarchical clustering methods provide a more intuitive way allowing an interactive exploration (Zhao *et al.* 2005). Additionally, given the hierarchical property

of urban areas (Roth *et al.* 2011, Louail *et al.* 2015) and people's movements (Bao *et al.* 2015), hierarchical clustering approaches have a strong theoretical foundation in human mobility analysis. Hierarchical DBSCAN (HDBSCAN), which inherits the merits of both the hierarchical clustering approach and DBSCAN, is prominent among the current clustering algorithms due to its stable performance and well-coded program (Campello *et al.* 2013). In recent studies, HDBSCAN has been also applied to ROI identification (Korakakis *et al.* 2017, Järv *et al.* 2018).

3. Problem statement

This research aims to predict the ROI that a user will visit at a specific future time based on the user's historical geotagged posts. To introduce the proposed method, some basic definitions are clarified as follows:

Definition 1. A *geotagged post* (Po) records the basic information of a user's footprints, which can be expressed as:

$$Po = (u, c, t, l), \tag{1}$$

where u denotes user's ID, c denotes the content of a post, t is the posting time and l indicates posting location (i.e. longitude and latitude).

Definition 2. A *trajectory* (*Tr*) is a user's time-ordered sequence of *Po* on a specific social network (e.g. Twitter), which is defined as:

$$Tr = \left\langle Po_0, Po_1, ..., Po_k \right\rangle \tag{2}$$

where k is the total number of posts.

Definition 3. An ROI (r) is an area that has been visited by a user with a significant high frequency. Since a user can have multiple ROIs, these ROIs form a collection C_{ROI} which is defined as:

$$C_{\mathrm{ROI}} = \{r_0, r_1, ..., r_n\}, \tag{3}$$

where n equals to the number of ROIs.

Definition 4. A *travel sequence item* is the combination of a geotagged post Po and its corresponding ROI r, which is written as:

$$TrSI = (Po, r), \tag{4}$$

Definition 5. A series of *travel sequence items* make up the *travel sequence,* which is written as:

$$TrS = \left\langle TrSI_0, TrSI_1, ..., TrSI_m \right\rangle, \tag{5}$$

where *m* is the length of the *travel sequence*.

Definition 6. A *time slot* Ts is defined as a time interval marked by two time points. With a given slot length T_l (in hours), the day time can be equally divided into *m* end-to-end time slots $Ts_1, Ts_2, \ldots, Ts_m$, where *m* is calculated as $24/T_l$. For example, there will be 24 time slots for a day with $T_l = 1$, in which the first will be 0:00-1:00 and the last will be 23:00-24:00.

4. Methodology

4.1.Framework of the proposed method

Figure 1 is about here.

Figure 1. Framework of STLP-GSM

The proposed STLP-GSM consists of four components as illustrated in Figure 1, including data preparation, model initialization, spatio-temporal prediction, and model updating. The four components are described as follows:

(1) **Data preparation**: Data cleaning will eliminate duplicated data (e.g., successive posts published within a short time). Spatial clustering of a user's geotagged posts will then be applied to identify the ROIs of the user. Subsequently, the temporality of the spatial clusters discloses the user's travel sequence.

(2) **Model initialization**: A prediction model formulates and combines the selected mobility features probability expressions, based on the travel sequence.

(3) **Spatio-temporal prediction**: The initial prediction model predicts the ROI which the user is most likely to visit at a specific time in the future. Subsequent geotagged posts are used to verify the prediction.

(4) **Model updating**: The current travel sequence for constructing a prediction model is updated by deleting outdated data and inserting new data from subsequent geotagged postings. Probabilities, used in the model, will then be recalculated with the updated sequence, and model parameters will be optimized, over time, at a certain frequency. The optimized model will then be used for further prediction.

4.2.ROI identification

The prevalence of duplicate social media posts can lead to some false growth in the local density of users' trajectories (Cuenca-Jara *et al.* 2017). Therefore, directly clustering a user's trajectory will result in false ROI identification. In this case, a data cleaning process is conducted to discard any posts less than a time threshold T_{min} and distance threshold D_{min} from its previous post published by a user.

As a user's geotagged posts, in some regions, may be dense, but dispersive in other, using DBSCAN with empirical and fixed parameters may lead to unrealistic ROI extraction due to the weakness of clustering different density data (McInnes *et al.* 2017). In contrast, HDBSCAN shows a better and more stable performance in handling varying densities in data. Further, its performance in ROI identification of social media data has also been proved in previous studies (Korakakis *et al.* 2017, Järv *et al.* 2018). Therefore, HDBSCAN is herein employed to extract the ROIs of each user.

The performance of HDBSCAN relies on one main parameter m_{pts}, which determines the smallest size that a set of grouping points can be considered as a cluster (McInnes *et al.* 2017). To select a proper m_{pts}, multiple values will be tested, and the one that yields the highest clustered degree (criterion 1) with at least three clusters (criterion 2) will be chosen as the optimal. Under criterion 1, the non-clustered points can be safely considered as noise, and the overall data utilization can be improved since most data are retained; under criterion 2, the model avoids predictions with one or two ROIs, likely to be meaningless and monotonous. Additionally, posts that are not

clustered will be considered as noise and hence discarded. The circumscribed polygons of each cluster will be taken as the ROIs, while the remaining posts will be associated with their corresponding ROI and hence, make up a travel sequence serving as the initial data for further analysis.

4.3.Model initialization

Model construction

Mobility features provide information from multiple dimensions of people's movements (Noulas *et al.* 2012). Predicting human mobility is thus subjective to a given set of *k* mobility features $MFs = \{mf_1, mf_2, \dots, mf_k\}$. Huang (2017) selects a linear model to combine spatial and temporal features by a summation with different weights. The summation of the probabilities of individual mobility features, however, assumes an independence, which may not hold in reality. To effectively include mobility information to the prediction process, the research presented in this paper proposes an extensible prediction model on the basis of the Naïve-Bayes algorithm. The latter has been widely applied in human behavior pattern mining as well as location prediction (Phithakkitnukoon *et al.* 2010, Steiger *et al.* 2015). Let variable R_{T^*} denote the predicted ROI, where T^* refers to the timestamp of R_{T^*}. The general format of the proposed model is written as:

$$P(R_{T^*} \mid mf_1, mf_2, \dots, mf_k) = \frac{P(mf_1, mf_2, \dots, mf_k \mid R_{T^*}) P(R_{T^*})}{P(mf_1, mf_2, \dots, mf_k)}. \quad (6)$$

The denominator on the right side of the equation is a constant as it is independent of R_{T^*}. In addition, based on the Naïve-Bayes algorithm, we have:

$$P(mf_1, mf_2, ..., mf_k \mid R_{T^*})P(R_{T^*}) = P(R_{T^*})\prod_{i=1}^{k} P(mf_i|R_{T^*}) , \quad (7)$$

Due to the predictability of various mobility features, a weight variable w, ranging from 0 to 1, is further assigned as an exponent to each of the conditional probabilities. Therefore, the proposed model can be finally factorized as:

$$P(R_{T^*} \mid mf_1, mf_2, ..., mf_k) \propto P(R_{T^*})\prod_{i=1}^{k} P(mf_i|R_{T^*})^{w_i} . \quad (8)$$

The following discussion introduces comments on the calculation of the conditional probability term $P(mf_i|R_{T^*})$ and the selection of weight w_i.

Mobility features definition

As indicated above, this research focuses on three mobility features, namely, *Previous visit*, *Temporal habit* and *Posting interval*. Of three mobility features *Temporal habit* and *Previous visit* are the most important features to provide the respective temporal and spatial mobility information of a user. *Posting interval* serves as an adjuster to rectify the bias brought by successive posts within the same ROI, especially for a long-term prediction. It is noticeable that more potential mobility features may be extracted from data, such as the semantic meaning of ROIs. However, extracting those potential mobility features relies on many other techniques, such as semantic analysis, which is out of the scope of this study. Thus, in this research, focus is only on the three mobility features.

Previous visit (PV): Previous visit refers to the past places (ROIs) that a user has visited as noted by one's travel sequence. The impact of *Previous visit* on the next location is commonly characterized in a Markov chain (Liu *et al.* 2016, Huang 2017). Let R_{current} denote the current ROI in a user's travel sequence, the probability for $R_{\text{current}} = r_i$ under the condition $R_{T^*} = r_j$ can be written as:

$$P(mf_{\text{PV}}|R_{T^*}) = P(R_{\text{current}} = r_i \mid R_{T^*} = r_j) = \frac{N_{\text{PV}}(r_i, r_j)}{N(r_j)}, \tag{9}$$

where $N_{\text{PV}}(r_i, r_j)$ indicates the number of transitions from r_i to r_j, and $N(r_j)$ is the number of times that the user visits r_j.

Temporal habit (TH): People tend to take time-specific activities (Cho *et al.* 2011, Lv *et al.* 2017). For example, people are more likely to stay at home rather than in a commercial centre between 00:00 to 01:00. *Temporal habit* describes how likely a user will appear in an ROI within a given time slot. The probability for T^* being within the time slot Ts_i under the condition $R_{T^*} = r_j$ can be expressed as:

$$P(mf_{\text{TH}}|R_{T^*}) = P(T^* \in Ts_i \mid R_{T^*} = r_j) = \frac{N_{\text{TH}}(r_j, Ts_i)}{N(r_j)}, \tag{10}$$

where $N_{\text{TH}}(r_j, Ts_i)$ is the number of times that the user visits r_j within Ts_i.

Posting interval (PI): Stay points, which herein indicate successive points within the same ROI, have a significant influence on constructing a Markov chain since they can produce many internal transitions (e.g., transition from a commercial centre to itself) and introduce bias to the prediction model. The stay points are mostly caused by the frequent posting of a user in particular ROIs, such as commercial centre and entertainment venues (Falcone *et al.* 2014). In previous study, stay points were

completely removed from the travel sequence (Huang 2017). The latter seems unrealistic and unreasonable as it completely denies the occurrence of the internal transitions.

To address the problem of stay points, a mobility feature named *Posting interval*, which indicates the time span between two successive posts, is defined. *Posting interval* in those ROIs considered in the stay points problem is expected to be small. Therefore, when a user does not publish any posts for a long period of time since the last posting in an ROI, the user has likely left this ROI. According to previous studies, *Posting interval* for a given ROI can be modelled by an exponential distribution (Kumar *et al.* 2015, Lukasik *et al.* 2015). The parameter λ_{r_j} of the exponential distribution is measured by the average duration in terms of a given time unit *Tu* (e.g., 1 hour) between the post in r_j and its previous one. Therefore, the conditional probability for *Posting interval* under $R_{T^*} = r_j$ can be written as:

$$P(mf_{\mathrm{PI}}|R_{T^*}) = P(\Delta T = m \times Tu | R_{T^*} = r_j) = e^{-(m-1)\lambda_{r_j}} - e^{-m\lambda_{r_j}}, \quad (11)$$

where $\Delta T = T^* - T_{R_{\mathrm{current}}}$, which denotes the time span between R_{T^*} and the current post R_{current}. $T_{R_{\mathrm{current}}}$ corresponds to the timestamp of R_{current}. *m* is the ratio of the time span ΔT to the time unit *Tu*.

Based on the equation (8) to (11), the prediction model in this article can be written as:

$$\begin{aligned} &P(R_{T^*} \mid mf_{\mathrm{PV}}, mf_{\mathrm{TH}}, mf_{\mathrm{PI}}) \\ &\propto P(R_{\mathrm{current}} \mid R_{T^*})^{w_{\mathrm{PV}}} P(Ts_i \mid R_{T^*})^{w_{\mathrm{TH}}} P(T^* - T_{R_{\mathrm{current}}} \mid R_{T^*})^{w_{\mathrm{PI}}} P(R_{T^*}), \end{aligned} \quad (12)$$

where w_{PV}, w_{TH} and w_{PI} are the weights of the three mobility feature discussed above, respectively. The last term $P(R_{T^*})$ can be expressed as the proportion of R_{T^*} in the whole travel sequence:

$$P(R_{T^*}) = \frac{N(R_{T^*})}{\sum_{R_{T^*} \in C_{\mathrm{ROI}}} N(R_{T^*})} . \tag{13}$$

Model parameter selection

The verification accuracy, measuring the goodness of the model parameters (i.e., w_{PV}, w_{TH} and w_{PI}) is defined as the proportion of predictions consistent with the real locations when applying the model to the historical data at hand. To obtain the optimal parameters with the best verification accuracy, an intuitive approach, namely, the grid search, is commonly used to exhaust finite combinations of manually chosen values of weights. The enormous computation burden due to the high dimensions of mobility features, however, challenges the grid search approach (Bergstra and Bengio 2012). The Bayesian Optimization Algorithm (BOA) is a global optimization algorithm that has been widely applied for high dimensional problems and can significantly reduce the computational cost through employing relatively fewer evaluations than other approaches (Snoek *et al.* 2012). Given an initial set of random points at the beginning of the optimization, BOA iteratively calculates the posterior distributions of these points and determines the most possible extremum point for the next exploration (Brochu *et al.* 2010). In addition, the BOA process time can be simply controlled by adjusting the number of iterations. Therefore, BOA is employed in this research for the

selection of model parameters.

As this study focuses on three mobility features, the initial points of BOA are designed to consist of 10 random points and 27 manual points. The 27 manual points correspond to combinations of weights with three optional values: 0, 0.5 and 1. The detail regarding the selection of the number of iterations is discussed in Section 5.3.

4.4. Spatio-temporal prediction

A user's potential location R_{T^*} at a specific time T^* can be predicted as the ROI with the largest probability, mathematically:

$$R_{T^*} = \arg\max_{r \in C_{\mathrm{ROI}}} P(R_{T^*} = r \mid mf_{\mathrm{PV}}, mf_{\mathrm{TH}}, mf_{\mathrm{PI}}), \tag{14}$$

To estimate the reliability of each prediction, Shannon's entropy which reflects the disorderliness and uncertainty within a system is employed (Shannon and Weaver 1949). Let p_i be the short notation for $P(R_{T^*} = r_i | mf_{\mathrm{HV}}, mf_{\mathrm{TH}}, mf_{\mathrm{PI}})$; therefore, each prediction corresponds to a set of probabilities denoted as $S_P = \{p_1, p_2, \dots, p_n\}$. The normalized entropy of the set S_P can be written as:

$$H(S_p) = \frac{-\sum_{i}^{n} p_i \log(p_i)}{\log(n)}. \tag{15}$$

The normalization can make entropy values comparable to users with different numbers of ROIs. $H(S_P)$, reaches its maximum 1 when all the elements in the set S_P are equal (i.e., $p_1 = p_2 = \cdots = p_n$), thus indicating a completely uncertain situation of

the prediction; the minimum value 0 is obtained for $H(S_P)$ when only one element (e.g., p_1) in the set S_P equals to 1.

4.5. Probability recalculation and model optimization

Algorithm 1: probability recalculation

Input: historical travel sequence TrS_o, current probability set S_{prob}, and threshold d_{max}

Output: updated travel sequence TrS_u, and probability set S_{prob_u}

Begin

Receive a new post Po_{new}

$r = \text{find\_cluster}(Po_{new})$ // Find ROI r that the new post belongs to

if $r = \text{Null}$ **do**

Return $TrS_o,\ S_{prob}$ // Remain the current sequence and probability set

if $\text{check\_dup}(Po_{new}) = \text{True}$ **do** // Check if the post is duplicative

Return $TrS_o,\ S_{prob}$

$TrSI_{\text{new}} = (Po_{\text{new}}, r)$ // Create new travel sequence item

$TrS_u = \text{add}(TrS_0,\ TrSI_{\text{new}})$ // Add the new item to current sequence

if $\text{check\_dur}(TrS_u) > d_{max}$ **do** // Check the duration of TrS_u

$\text{trim}(TrS_u)$ // Remove the outdated items in TrS_u

$S_{pro\_u} = \text{rec\_prob}(TrS_u)$ // Recalculate the probability set

Return $TrS_u,\ S_{prob\_u}$

End

The probability recalculation process starts instantly once a new geotagged post is published. The detail of the updating process is demonstrated in Algorithm 1. Let S_{prob} denote the set of the four probability terms needing to be updated, namely: $S_{prob} = \{p(mf_{\mathrm{PV}}|R_{T^*}), p(mf_{\mathrm{TH}}|R_{T^*}), p(mf_{\mathrm{PI}}|R_{T^*}), p(R_{T^*})\}$. The inputs for Algorithm 1 then, include the probability set S_{prob}, the historical travel sequence TrS_o, and a threshold d_{max} which indicates the maximum time span of updated travel sequence. The outputs include the recalculated conditional probability set S_{prob_u} and updated travel sequence TrS_u.

Let Po_{new} denote the new post published by a user, the recalculation process will terminate and return to the current travel sequence TrS_o and probability set S_{prob} if no known ROI can be found for Po_{new} (the function *find_cluster()* has been implemented in HDBSCAN (McInnes *et al.* 2017)) or the post is judged as a duplicate; otherwise, an updated travel sequence TrS_u will be produced by adding the post Po_{new} and its corresponding ROI r as a travel sequence item to the current travel sequence TrS_o. To avoid the effect of outdated data, the earliest items in TrS_u will be discarded until the time span of the data in TrS_u is shorter than the threshold d_{max}. With the updated travel sequence TrS_u, the conditional probability set S_{prob_u} for each mobility feature can be simply recalculated with the function *rec_prob()*.

The model optimization process starts at a certain frequency f represented by the inverse of recalculation times. The optimization process is described in Algorithm 2. The inputs include the updated travel sequence TrS_u and a weight combination set WS, which stores the optimization results over time. BOA is employed to select the

optimal weights. Since changes in the updated travel sequence TrS_u are minor, the optimal weights are expected to be close to previous optimization results. Therefore, to save computation cost, newly obtained weights are added to weight combination set WS and the earliest optimization result removed to control the size of WS below l_{WS}. Finally, a new weight combination w_{opt} is obtained for further location predictions.

Algorithm 2: model optimization

Input: the travel sequence TrS_u, the historical weight combination set WS

Output: Weight combination w_{opt}

Begin

for every 1/f times of 'probability recalculation' **do**

BOA.initial()

BOA.explore_manaul_pts() // Explore manual points

BOA.explore(WS) // Explore items in WS

BOA.iterate(n) // Proceed additional n iterations for BOA

w_{opt} = BOA.optimum // Get the optimal weight combination w_{opt}

if w_{opt} is not in WS **do**

add(WS, w_{opt}) // Add w_{opt} to WS

if WS.size> l_{WS} **do**

clear_old(WS) // Remove the oldest item in WS

end

Return w_{opt} // Return optimal weight combination w_{opt}

End

5. Experiments and results

5.1. Data description and experimental setup

To validate the proposed method, two datasets from Instagram and Twitter were used. The Instagram dataset (HK dataset) was collected from November 2014 to November 2015, containing about 2.1 million geotagged posts from about 58 thousand users in Hong Kong, China. The Twitter dataset (NY dataset) consists of over 1 million geotagged posts from about 100 thousand users in New York, spanning four months from July 2018 to October 2018.

Since people have different habits and preferences in the use of social media, some data should be excluded from building a model for location predictions. Firstly, users whose monthly average number of geotagged posts is less than 100 or more than 600 were excluded, thereby ensuring enough data to support the further modelling and avoiding the noise from automatic posts such as advertisements and broadcast. After this step, 79 and 158 users were left for HK and NY dataset, respectively. Secondly, to ensure that prediction is meaningful, users with less than three ROIs were excluded. Finally, 47 users in the HK dataset and 89 users in the NY dataset were qualified for the experiment.

To avoid bias from outdated mobility information, the threshold d_{max} for controlling the time span of travel sequence, was set to 180 days (approximately six months), over which individual the mobility pattern is believed to be relatively stable (Schneider *et al.* 2013). The time threshold T_{min} for filtering duplicate posts during

the data cleaning was preliminarily set to 300 seconds, and a sensitivity analysis for T_{min} was conducted in the experiment. The distance threshold D_{min} was set to 20 meters, thus close to the uncertainty in GPS readings (Zhou *et al.* 2004). The time slot length T_l was set to 1 hour, enabling a day to be divided into 24 equal time slots for the calculation of *Temporal habit* mobility feature. The time unit Tu for the *Posting interval* was also set to 1 hour**.**

5.2. Clustering method comparison

Figure 2 is about here.

Figure 2. A real case for the ROI identification of a user in Hong Kong

To demonstrate the performance of the proposed method in ROI identification, DBSCAN with different parameters and HDSCAN based on the proposed strategy were applied to the footprints of a selected user in Hong Kong. The heatmap of this user's footprints is shown in Figure 2.(a), and some hot spots can be visually observed. DBSCAN was first tested with a specific parameter setting (i.e., *Eps* = 20, *MinPts* = 4) in Huang's (2017) highly related study, and only three locations with a high level of overlapped points were obtained (Figure 2. (b)), of which the *Eps* seems to be too small to generate meaningful clusters. A similar situation can be observed with *Eps* increasing to 50 as shown in Figure 2.(c). The DBSCAN with *Eps*=100 in Figure 2.(d) and HDBSCAN in Figure 2.(e) show similar performances in this case, and the results seem to be visually reasonable. However, the cluster 7 identified in Figure 2. (e) was merged

to the cluster 1 in the result of HDBSCAN (marked with red boxes), thus matching reality, as both two clusters belong to the commercial zone 'Center' in Hong Kong. In summary, compared with DBSCAN, HDBSCAN with the proposed strategy can significantly reduce the uncertainty in parameter selection and be capable of discovering realistic ROIs.

5.3. Efficiency of BOA

Figure 3 is about here.

Figure 3. Effect of different BOA parameters on the sample data.

To find the appropriate combination set size l_{WS} and iteration number n for BOA, different values were tested on a typical user's trajectory in the HK dataset. The optimization process was conducted 238 times and the results are displayed in Figure 3. From Figure 3.(a), the average optimization time keeps growing when n increases (l_{WS} is fixed to 10), while the average maximum verification accuracy grows slightly after $n = 5$; the similar situation can be observed in Figure 3.(b), in which l_{WS} larger than 10 (n is fixed to 5) does not achieve a significant improvement in accuracy. Therefore, $l_{WS} = 10$ and $n = 5$ were selected for BOA in further experiments. It is notable that the parameters might be set larger as the tradeoff between accuracy and time consumption becomes more budget with a better hardware configuration.

To demonstrate the advantage of the BOA, grid search method and BOA were both applied to the selected sample data. The increment for the grid search method was

set at 0.1 following the setting in the similar optimization process in SMMC. All the cases were run on a desktop computer with an Intel i7-7700U CPU (3.6 GHz) and 16-GB memory. BOA was implemented using an open-source Bayesian Optimization package (Nogueira 2014). The optimization time, as well as the maximum verification accuracy of those two optimization methods, are compared in Figure 4.

Figure 4 is about here.

Figure 4. Performance of BOA and grid search on the sample data.

As shown in Figure 4, the time consumption of BOA is much lower (by approximately half) compared to the grid search. The maximum verification accuracy achieved by BOA is close to that achieved by the grid search, hence implying the similar optimization results (i.e., the weights for MWBM) of these two methods. Therefore, because of the BOA significant advantage on time efficiency, BOA outperforms the traditional grid search method in this case.

5.4. Accuracy evaluation

As weekday human mobility pattern can be different from that at weekends, the geotagged posts published in weekdays were extracted for the accuracy evaluation. To estimate the prediction accuracy of the proposed model, each user's first half posts were selected for training their prediction model and then applied to their rest posts for validation. The prediction accuracy is described as the proportion of the predicted locations consistent with the validation process observation.

Figure 5 is about here.

Figure 5. Prediction accuracy for two datasets with different optimizing frequencies.

The frequency for model optimization affects the time efficiency as well as the prediction accuracy of the proposed approach. To identify an appropriate frequency, different optimizing frequencies were tested on HK and NY datasets. The results are shown in Figure 5. The prediction accuracies are close when the frequency is higher than 1/30 (marked by the dashed line), while the accuracy drops dramatically as the frequency decreases to 1/100. Thus, regarding time efficiency and prediction accuracy, the optimizing frequency is set to 1/30 in the following experiments.

To evaluate the performance of the proposed model and the impact of the updating phase, four control groups, including MWBM without parameter optimization, and probability recalculation, the SMMC model proposed by Huang (2017), and the Baseline model, which permanently picks the commonest ROI over time, were employed for the purpose of comparison. Note that the clustering method used in SMMC is originally DBSCAN; therefore HDBSCAN was implemented for SMMC in this study. In addition, there were no time threshold T_{min} standard values for detecting duplicate posts and slot length T_l for the measurement of the mobility feature *Temporal habit* in previous studies (Alvarez-Lozano *et al.* 2015, Huang and Wong 2015, Chen *et al.* 2016). Different settings for these two parameters will produce different results and introduce the well-known modifiable temporal unit problem (Cheng and Adepeju 2014). To investigate the influence of the two parameters for MWBM on HK and NY datasets, multiple values of T_l ranging from 1 to 4 hour and $T_{\min}$ ranging from 300 to 1500 seconds were tested in the experiment.

Figure 6 is about here.

Figure 6. Comparison of prediction accuracy of different models on two datasets.

The evaluation results are shown in Figure 6 and quantitative details for the comparison between MWBM and SMMC are given in Table 2. According to Figure 6, the prediction accuracies for all four models are all visually much higher than the baseline accuracy, while among other models, the proposed MWBM presents the best over prediction accuracy. A certain amount of decline in terms of prediction accuracy exists for MWBM without the model optimization or probability recalculation process, but the accuracy is still higher than that of SMMC in most cases, thus proving the effectiveness of the whole updating phase. Furthermore, as seen in Table 2, the average prediction accuracy improvement by MWBM against SMMC is more than 3.00% for the HK dataset and 2.59% for the NY dataset. In particular, the best prediction accuracy (54.51%) for the HK dataset is obtained by MWBM with $T_l = 3$ and $T_{\min} = 1500$, hence exhibiting an SMMC enhancement of 5.33%; while the prediction accuracy for the NY dataset, reaches its highest value of 54.34% with $T_l = 4$ and $T_{\min} = 1500$ through MWBM. Hence an improvement of 3.30% compared with SMMC (51.04%) is shown. Thus, regarding prediction accuracy only, $T_l = 3$, $T_{\min} = 1500$ and $T_l = 4$, $T_{\min} = 300$ seem to be the respective relatively good settings for the HK and NY dataset. Taking the results corresponding to the above best parameter settings as an example, the relationship between the number of ROIs and prediction accuracy were plotted in Figure 7. The overall prediction accuracy tends to be higher for the users who

have fewer ROIs. This is reasonable since the transition pattern will get simpler as the number of ROIs decrease; therefore, the prediction accuracy is enhanced.

Table 2. Details of the prediction accuracy on two datasets.

Slot length	T_{min}	HK dataset			NY dataset		
(hour)	(second)	MWBM	SMMC	Improvement	MWBM	SMMC	Improvement
1	300	52.57%	49.93%	2.64%	52.51%	51.04%	1.48%
	600	51.91%	49.63%	2.28%	53.08%	50.14%	2.94%
	900	52.27%	50.37%	1.90%	52.71%	50.50%	2.21%
	1200	51.14%	49.50%	1.65%	50.61%	47.90%	2.71%
	1500	52.00%	49.08%	2.92%	50.21%	48.29%	1.92%
2	300	53.31%	49.93%	3.38%	53.79%	51.04%	2.75%
	600	53.11%	49.63%	3.48%	53.34%	50.14%	3.20%
	900	52.98%	50.37%	2.61%	52.84%	50.50%	2.34%
	1200	52.89%	49.50%	3.39%	50.36%	47.90%	2.46%
	1500	53.23%	49.08%	4.15%	50.33%	48.29%	2.04%
3	300	51.97%	49.93%	2.04%	54.24%	51.04%	3.21%
	600	51.44%	49.63%	1.81%	53.50%	50.14%	3.37%
	900	53.47%	50.37%	3.10%	52.96%	50.50%	2.45%
	1200	52.55%	49.50%	3.05%	50.70%	47.90%	2.80%
	1500	**54.41%**	49.08%	**5.33%**	50.37%	48.29%	2.08%
4	300	52.13%	49.93%	2.21%	**54.34%**	51.04%	**3.30%**
	600	53.40%	49.63%	3.77%	53.28%	50.14%	3.14%
	900	53.50%	50.37%	3.13%	52.85%	50.50%	2.35%
	1200	53.07%	49.50%	3.57%	50.89%	47.90%	2.99%
	1500	52.59%	49.08%	3.51%	50.33%	48.29%	2.04%
Average improvement				3.00%			2.59%

Figure 7 is about here.

Figure 7. Relationship between the number of individual ROIs and prediction accuracy.

Note that the greatest accuracy in these two cases is less than the accuracy claimed in the original SMMC (Huang 2017). Two reasons make clear reduction: 1) The original SMMC did not consider internal transitions, which means the transitions from one ROI to itself. Thus, the baseline accuracy of the original SMMC is expected to be higher than the newly proposed method in this study. For example, if a user has three ROIs and random predictions for the user's next location are made, the average accuracy for original SMMC will be about 50% (two optional ROIs), while that for MWBM will be about 33% (three optional ROIs). 2) As the predictability of people's mobility may differ in different cities, the overall accuracy (i.e., up to 79%) claimed in Huang (2017) is not representative of the general conditions.

5.5.Uncertainty analysis

The density maps of the entropy for each prediction with selected parameters, where the real location is available for validation, are given in Figure 8. It is shown that high entropy values are likely to be associated with the wrong predictions, while the lower entropy values tend to correspond to correct predictions. In addition, the average entropy value (marked by red lines in Figure 8) giving the wrong predictions is always above 0.8, while for correct predictions, the average is less than that number. These tendencies are universal for both HK and NY datasets along with different temporal parameters.

Figure 8 is about here.

Figure 8. Density maps of prediction entropy with different parameters on two datasets. The mean entropy values are marked with red lines.

The relationship between the correctness of the predictions and the corresponding entropy values are further quantified below. The range of entropy value (i.e., 0 to 1) was divided into ten equal intervals, and the proportion of the correct predictions (POC) was calculated for each interval. Here, only the first column in Figure 8 (i.e., $T_l = 1$, $T_{\min} = 300$) was taken as an example for the quantification, and the results are shown in Table 3. POC decreases when the entropy value increases, and the result of the Ordinary Least Squares regression (OLS) demonstrates the linear relation between the prediction entropy (denoted by PE, equals to the middle value of each entropy interval) and POC. Therefore, POC can be an indicator that evaluates the reliability of each prediction.

Table 3. The relationship between the correctness of the predictions and entropy values. PE is equal to the middle value of each entropy interval. All the coefficients of OLS are significant at the 0.05 level.

ID	Interval of entropy	HK dataset			NY dataset		
		Total prediction	Correct prediction	POC	Total prediction	Correct prediction	POC
1	0 to 0.1	10	10	100.00%	246	245	99.59%
2	0.1 to 0.2	52	45	86.54%	181	167	92.27%
3	0.2 to 0.3	68	51	75.00%	214	178	83.18%
4	0.3 to 0.4	205	160	78.05%	543	471	86.74%

5	0.4 to 0.5	301	229	76.08%	580	469	80.86%
6	0.5 to 0.6	295	199	67.46%	782	547	69.95%
7	0.6 to 0.7	529	314	59.36%	1455	819	56.29%
8	0.7 to 0.8	742	382	51.48%	2507	1334	53.21%
9	0.8 to 0.9	933	421	45.12%	3347	1451	43.38%
10	0.9 to 1	1029	378	36.73%	3834	1508	39.33%
OLS		$POC = (0.99 - 0.63 * PE) * 100\%$			$POC = (1.05 - 0.69 * PE)$ *100%		

The overall reliability for the spatio-temporal prediction of each user's daily movements can be further evaluated, based on the OSL model. A user from the HK dataset with six ROIs was selected as an example to demonstrate the evaluation. The mean entropy value of the predictions was firstly counted for over one day, for each time slot (i.e., an hour). This was followed by the overall prediction reliability of this user calculated using the given OSL model. As shown in Figure 9, the predictions for the selected user's locations are relatively reliable in the early morning (6 a.m. to 8 a.m.), the afternoon (2 p.m. to 3 p.m.), and the night (00:00 to 1 a.m.). However, the predictions of the user's locations are quite uncertain in the late morning (10 a.m. to 11 a.m.) and evening (5 p.m. to 8 p.m.).

Figure 9 is about here.

Figure 9. An example of a user's daily prediction reliability.

6. Discussion and conclusion

The human mobility prediction is one major issue in the study on human dynamics and social science. Predicting spatio-temporal location based on geotagged social media

posts remains challenging in existing models for the next-location prediction or subject to the dependence on receiver-based location data, such as GPS and CDR data. In this article, a spatio-temporal location prediction method called STLP-GSM was proposed and validated on two real-life social media datasets. The ROIs of a user were extracted on the basis of the clusters generated by HDBSCAN with self-adaptive parameter selection. An MWBM was trained using a user's past travel footprints represented by the extracted ROIs. In MWBM, multiple mobility features, which may affect the spatial and temporal regularity in people's daily movements, were combined on the basis of Naïve-Bayesian theory. To avoid redundancy caused by outdated data and deal with the changes in users' mobility patterns, an updating strategy, including a probability recalculation process and a model optimization process, was designed and tested. Furthermore, the reliability of the prediction results was analyzed and demonstrated based on Shannon's entropy and linear regression.

A comparing with the state-of-the-art prediction model SMMC indicated that the proposed method outperformed it, in terms of prediction accuracy, with an improvement by up to 5.33% on one dataset and 3.30% on the other, based on selected parameters. In addition, the prediction accuracy of the proposed method declines significantly when excluding the probability recalculation or model optimization, which demonstrates the necessity and effectiveness of these two processes in the updating phase.

One limitation of this study is that the number of qualified social media users for prediction is small. This appears to be a general problem in previous studies based

on social media data. For example, in the SMMC model original paper, the number of qualified users from Washington, D.C. is 52 (Huang 2017). However, two reasons cause the authors of this paper to believe that the limitation can be mitigated: To achieve this firstly, as long as more data can be collected over a longer period, more sophisticated filters can be constructed to improve the ratio of data usage. For example, with long-term observation, focus can be on the active periods of a user, and days with few posts can be excluded. In this way, the constraint from 'total number of geotagged posts' can be relieved, and the qualified users will increase. Secondly, there are huge growth spaces for geotagged social media data. According to the study of Huang (2017), most Twitter users add geotags to only 20% of their posts. Together with the development of location-based services and the increasing popularity of social networks, the density of geotagged social media data can be expected to increase in the future.

This work could be strengthened in the future. The proposed method provides an intuitive way to combine multiple mobility features to enhance the accuracy of spatio-temporal location prediction. The extensible structure allows the inclusion of more mobility features, such as the user's profile, semantic meanings of ROIs. The global mobility patterns over the area may also be included if available. Future research may assess the derivation of the conditional probability from those mobility features and their integration into the current MWBM.

Acknowledgement

This study was supported by Ministry of Science and Technology of the People's Republic of China (2017YFB0503604), the Innovation and Technology Fund of the Hong Kong Government (No. ITP/053/16LP), and Hong Kong Polytechnic University (1-ZVF2, 1-ZEAB). We are grateful to the editor Dr May Yuan and the anonymous reviewers who provided insightful comments on improving this article. We also thank Mrs. Elaine Anson for her valuable advises on the improvement of English writing.

Reference

Akoush, S. and Sameh, A., 2007. Mobile user movement prediction using bayesian learning for neural networks. *In*: *Proceedings of the 2007 international conference on Wireless communications and mobile computing*. ACM, 191–196.

Altomare, A., Cesario, E., Comito, C., Marozzo, F., and Talia, D., 2017. Trajectory Pattern Mining for Urban Computing in the Cloud. *IEEE Transactions on Parallel and Distributed Systems*, 28 (2), 586–599.

Alvarez-Lozano, J., García-Macías, J.A., and Chávez, E., 2015. Crowd location forecasting at points of interest. *International Journal of Ad Hoc and Ubiquitous Computing*, 18 (4), 191–204.

Bao, J., Zheng, Y., Wilkie, D., and Mokbel, M., 2015. Recommendations in location-based social networks: a survey. *GeoInformatica*, 19 (3), 525–565.

Belcastro, L., Marozzo, F., Talia, D., and Trunfio, P., 2018. G-RoI: Automatic Region-of-Interest Detection Driven by Geotagged Social Media Data. *ACM Transactions on Knowledge Discovery from Data*, 20 (3), 1–22.

Belyi, A., Bojic, I., Sobolevsky, S., Sitko, I., Hawelka, B., Rudikova, L., Kurbatski, A., and Ratti, C., 2017. Global multi-layer network of human mobility. *International Journal of Geographical Information Science*, 31 (7), 1381–1402.

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13 (Feb), 281–305.

Brochu, E., Cora, V.M., and De Freitas, N., 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Campello, R., Moulavi, D., and Sander, J., 2013. Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery …*, 160–172.

Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical*, 41 (22), 224015.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., and Soltani, K., 2015. A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, 51, 70–82.

Cesario, E., Marozzo, F., Talia, D., and Trunfio, P., 2017. SMA4TD: A social media analysis methodology for trajectory discovery in large-scale events. *Online Social Networks and Media*, 3–4, 49–62.

Chen, X., Shi, D., Zhao, B., and Liu, F., 2016. Periodic Pattern Mining Based on GPS Trajectories. Atlantis Press.

Cheng, T. and Adepeju, M., 2014. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PLoS ONE*, 9 (6), 1–10.

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

Cho, S.-B., 2016. Exploiting machine learning techniques for location recognition and prediction with smartphone logs. *Neurocomputing*, 176, 98–106.

Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N., 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. *In*: *Sixth International AAAI Conference on Weblogs and Social Media*.

Cresswell, T., 2010. Towards a politics of mobility. *Environment and planning D: society and space*, 28 (1), 17–31.

Cuenca-Jara, J., Terroso-Saenz, F., Valdes-Vela, M., and Skarmeta, A.F., 2017. Fuzzy Modelling for Human Dynamics Based on Online Social Networks. *Sensors (Basel, Switzerland)*, 17 (9).

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Kdd*, 2, 635–654.

Falcone, D., Mascolo, C., Comito, C., Talia, D., and Crowcroft, J., 2014. What is this place? Inferring place categories through user patterns identification in geo-tagged tweets. *In*: *Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on*. IEEE, 10–19.

Gambs, S., Killijian, M.-O., and del Prado Cortez, M.N., 2011. Show me how you move and I will tell you who you are. *Transactions on Data Privacy*, 4 (2), 103–126.

Gambs, S., Killijian, M.-O., and del Prado Cortez, M.N., 2012. Next place prediction using mobility Markov chains. *In*: *Proceedings of the First Workshop on Measurement, Privacy, and Mobility - MPM '12*. Presented at the the First Workshop, Bern, Switzerland: ACM Press, 1–6.

Gao, S., 2015. Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation*, 15 (2), 86–114.

González, M.C., Hidalgo, C.A., and Barabási, A.-L., 2008. Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782.

Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.-L., 2008. Understanding individual human mobility patterns. *nature*, 453 (7196), 779.

Hadachi, A., Batrashev, O., Lind, A., Singer, G., and Vainikko, E., 2014. Cell phone subscribers mobility prediction using enhanced Markov Chain algorithm. *In*: *2014 IEEE Intelligent Vehicles Symposium Proceedings*. Presented at the 2014 IEEE Intelligent Vehicles Symposium Proceedings, 1049–1054.

Heckman, J.J. and Mosso, S., 2014. The economics of human development and social mobility. *Annu. Rev. Econ.*, 6 (1), 689–733.

Hidalgo, C.A. and Rodríguez-Sickert, C., 2008. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387 (12), 3017–3024.

Huang, Q., 2017. Mining online footprints to predict user's next location. *International Journal of Geographical Information Science*, 31 (3), 523–541.

Huang, Q. and Wong, D.W.S., 2015. Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. *Annals of the Association of American Geographers*, 105 (6), 1179–1197.

Järv, P., Tammet, T., and Tall, M., 2018. Hierarchical Regions of Interest. *In*: *2018 19th IEEE International Conference on Mobile Data Management (MDM)*. Presented at the 2018 19th IEEE International Conference on Mobile Data Management (MDM), 86–95.

Joseph, K., Tan, C.H., and Carley, K.M., 2012. Beyond local, categories and friends: clustering foursquare users with latent topics. *In*: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, 919–926.

Korakakis, M., Spyrou, E., Mylonas, P., and Perantonis, S.J., 2017. Exploiting social media information toward a context-aware recommendation system. *Social Network Analysis and Mining*, 7 (1), 42.

Kumar, S., Liu, H., Mehta, S., and Subramaniam, L.V., 2015. Exploring a Scalable Solution to Identifying Events in Noisy Twitter Streams. *In*: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*. Presented at the the 2015 IEEE/ACM International Conference, Paris, France: ACM Press, 496–499.

Liu, Q., Wu, S., Wang, L., and Tan, T., 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. *In*: *AAAI*. 194–200.

Louail, T., Lenormand, M., Picornell, M., Cantú, O.G., Herranz, R., Frias-Martinez, E., Ramasco, J.J., and Barthelemy, M., 2015. Uncovering the spatial structure of mobility networks. *Nature Communications*, 6, 6007.

Lukasik, M., Srijith, P.K., Cohn, T., and Bontcheva, K., 2015. Modeling Tweet Arrival Times using Log-Gaussian Cox Processes. *In*: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 250–255.

Lv, Q., Qiao, Y., Ansari, N., Liu, J., and Yang, J., 2017. Big Data Driven Hidden Markov Model Based Individual Mobility Prediction at Points of Interest. *IEEE Transactions on Vehicular Technology*, 66 (6), 5204–5216.

Mathew, W., Raposo, R., and Martins, B., 2012. Predicting future locations with hidden Markov models. *In*: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, 911–918.

McInnes, L., Healy, J., and Astels, S., 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2 (11), 205.

Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., and Vespignani, A., 2011. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports*, 1, 62.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K.M., 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *In*: *ICWSM*.

Mosenia, A., Dai, X., Mittal, P., and Jha, N., 2017. PinMe: Tracking a Smartphone User around the World. *IEEE Transactions on Multi-Scale Computing Systems*.

Nogueira, F., 2014. *Bayesian Optimization: Open source constrained global optimization tool for Python*.

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C., 2012. Mining User Mobility Features for Next Place Prediction in Location-Based Services. *In*: *2012 IEEE 12th International Conference on Data Mining*. Presented at the 2012 IEEE 12th International Conference on Data Mining (ICDM), Brussels, Belgium: IEEE, 1038–1043.

Oleksiak, P., 2014. Analysing and processing of geotagged social media. *Information Systems in Management*, 3, 11.

Phithakkitnukoon, S., Veloso, M., Bento, C., Biderman, A., and Ratti, C., 2010. Taxi-aware map: Identifying and predicting vacant taxis in the city. *In*: *International Joint Conference on Ambient Intelligence*. Springer, 86–95.

Roth, C., Kang, S.M., Batty, M., and Barthélemy, M., 2011. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLOS ONE*, 6 (1), e15923.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V., and Campbell, A.T., 2011. Nextplace: a spatio-temporal prediction framework for pervasive systems. *In*: *International Conference on Pervasive Computing*. Springer, 152–169.

Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., and González, M.C., 2013. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10 (84), 20130246.

Sevtsuk, A. and Ratti, C., 2010. Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks. *Journal of Urban Technology*, 17 (1), 41–60.

Shannon, C.E. and Weaver, W., 1949. The Mathematical Theory of Communication. *University of Illinois Press*.

Snoek, J., Larochelle, H., and Adams, R.P., 2012. Practical bayesian optimization of machine learning algorithms. *In*: *Advances in neural information processing systems*. 2951–2959.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L., 2010. Limits of Predictability in Human Mobility. *Science*, 327 (5968), 1018–1021.

Steiger, E., De Albuquerque, J.P., and Zipf, A., 2015. An Advanced Systematic Literature Review on Spatiotemporal Analyses of T witter Data. *Transactions in GIS*, 19 (6), 809–834.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., and Li, Q., 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation*, 42 (4), 625–646.

Ying, J.J.-C., Lee, W.-C., Weng, T.-C., and Tseng, V.S., 2011. Semantic trajectory mining for location prediction. *In*: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 34–43.

Yuan, J., Zheng, Y., and Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. *In*: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. Presented at the the 18th ACM SIGKDD international conference, Beijing, China: ACM Press, 186.

Yuan, Q., Zhang, W., Zhang, C., Geng, X., Cong, G., and Han, J., 2017. PRED: Periodic Region Detection for Mobility Modeling of Social Media Users. *In*: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 263–272.

Zhao, Y., Karypis, G., and Fayyad, U., 2005. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10 (2), 141–168.

Zheng, Y.-T., Zha, Z.-J., and Chua, T.-S., 2012. Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3 (3), 56.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L., 2004. Discovering personal gazetteers: an interactive clustering approach. *In*: *Proceedings of the 12th annual ACM international workshop on Geographic information systems*. ACM, 266–273.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L., 2007. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, 25 (3), 12.