

# Urban Human Mobility Data Mining: An Overview

Kai Zhao<sup>1</sup>, Sasu Tarkoma<sup>2</sup>, Siyuan Liu<sup>3,4</sup> and Huy Vo<sup>1,5</sup>

<sup>1</sup>Center for Urban Science and Progress, New York University

<sup>2</sup>Department of Computer Science, University of Helsinki

<sup>3</sup>Smeal College of Business, Penn State University

<sup>4</sup>Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

<sup>5</sup>Department of Computer Science, the City College of New York

{kai.zhao, huy.vo}@nyu.edu

**Abstract**—Understanding urban human mobility is crucial for epidemic control, urban planning, traffic forecasting systems and, more recently, various mobile and network applications. Nowadays, a variety of urban human mobility data have been gathered and published. Pervasive GPS data can be collected by mobile phones. A mobile operator can track people’s movement in cities based on their cellular network location. This urban human mobility data contains rich knowledge about locations and can help in addressing many urban challenges such as traffic congestion or air pollution problems. In this article, we survey recent literature on urban human mobility from a data mining view: from the data collection and cleaning, to the mobility models and the applications. First, we summarize recent public urban human mobility data sets and how to clean and preprocess such data. Second, we describe recent urban human mobility models and predictors, e.g., the deep learning predictor, for predicting urban human mobility. Third, we describe how to evaluate the models and predictors. We conclude by considering how applications can utilize the mobility models and predictive tools for addressing city challenges.

**Index Terms**—human mobility; spatio-temporal data mining; machine learning; smart city

## I. INTRODUCTION

Urban human mobility pertains to how people move in cities, for example, characterizing mobility patterns such as walking home, driving to working places or utilizing public transportation. Understanding human mobility is crucial for epidemic control [1], [2] urban planning [3], [4], [5], traffic forecasting systems [6], [7], [8] and, more recently, various mobile and network applications [9], [10], [11]. Nowadays, a variety of urban human mobility data have been gathered and published. The pervasive GPS data can be collected by mobile phones. A mobile operator can track people’s movement in cities based on their cellular network location. This urban human mobility data contains rich knowledge about locations and can help in addressing many urban challenges such as traffic congestion or air pollution problems.

In this article, we survey recent literature on urban human mobility from a **data mining** view (see Fig. 1). According to Fayyad et al. [12], a knowledge discovery process based on data mining or machine learning methods includes:

- 1) **Finding target data.** We summarize recent public urban human mobility data sets in Section II.
- 2) **Data cleaning and preprocessing.** We show how to clean and preprocess the mobility data in Section III.
- 3) **Exploratory model selection.** We describe recent urban human mobility models in Section IV.
- 4) **Searching for patterns of interests using data mining or machine learning methods.** We describe and classify recently proposed predictive algorithms for urban human mobility, e.g., the Long-Short-Term-Memory (LSTM) predictor [13], in Section V.
- 5) **Evaluating the mined patterns.** We show how to evaluate the models and the predictors in Section VI.
- 6) **Acting on discovered knowledge.** We examine applications using the mobility models and predictive tools for addressing urban challenges in Section VII.

To the best of our knowledge, this is the first article summarizing the overall urban human mobility data mining process, from the data collection and cleaning, to the mobility models, and applications.

Human mobility has been studied for a very long time. In 1885, the publication of The Laws of Migration [14] in the Journal of the Royal Statistical Society can be considered as the first modern attempt to understand human mobility. Due to the significant growth of mobile phones, the study of human mobility has significantly changed. Mobile phones utilize cell tower information and the Global Positioning System (GPS) for fine-grained location tracking. Billions of people carry their phone every day, which provides a large amount of data on human movement. The growing volumes of urban mobility data being collected and made available open up new opportunities for modeling and predicting the urban human mobility more accurately. We summarize recent public urban human mobility data sets in Section II and the methods for cleaning and preprocessing such data sets in Section III.

The ability to model and predict urban human mobility is a fundamental problem in mobile computing and wireless networks. An accurate location predictor can improve the performance of many mobile applications as well as the infrastructure. For example, accurate location prediction is

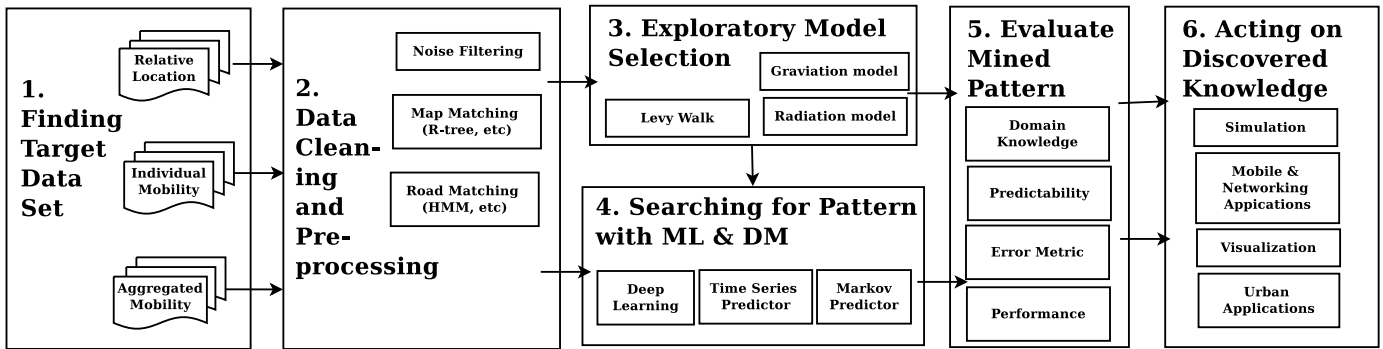


Fig. 1: The process of urban human mobility data mining.

vital for enabling autonomous vehicles and making them safe, and optimizing wireless base station performance. Another example is that knowledge on where the people would visit in a city can be advantageous to both taxi drivers and taxi companies. Taxi drivers can drive to areas where there is a big demand of taxi services if the urban human mobility can be correctly predicted. The taxi companies (e.g, Uber) may re-allocate their taxis in advance to meet the service demand of passengers in a region. The growing volumes of urban human mobility data set can help us modeling and predicting the human mobility more accurately.

An urban human mobility model captures the basic properties and features of human movement using mathematical or physical models, such as Lévy Walks, for the simulation and prediction of human mobility in cities. One of the first large scale human mobility modeling studies based on big urban data set was published in Nature in 2008 [15]. By studying cell phone user’s locations it was shown that trajectories in human mobility have statistically similar features to Lévy Walks. Other urban human mobility models that are not based on Lévy Walks (e.g., the radiation model [16]) have been proposed recently. We summarize recent urban human mobility models in Section IV.

Urban human mobility prediction pertains to the estimation of the next location that a person will visit in a city. It has been observed that urban human mobility exhibits strong regularities [15]. For example, people usually go to work during daytime on weekdays, and go shopping after work. Each person has a significant probability to return to a few highly frequented locations such as home or working places. Many of recent urban human mobility predictors try to capture such spatial-temporal regularities. We summarize recent human mobility predictive algorithms in Section V.

The evaluation of the proposed mobility models and predictors usually involve multiple factors: the error metrics, the computing performance or the feedback from domain experts. The error metrics such as symmetric Mean Absolute Percentage Error (sMAPE [17]) are often used to evaluate the errors of the predictive algorithms to the ground truth data. We show how to evaluate the proposed mobility models and predictors in Section VI.

The urban human mobility models and predictive tools can help us in many applications in addressing the urban problems. For example, while building Mobile Ad Hoc Networks (MANET), previous research usually use a synthetic model such as the Random Way Point (RWP) model. Later studies find that human mobility actually follow the Lévy Walk model. The routing performance in a RWP model in MANET studies tends to be overestimated compared to Lévy Walk based models [18]. We summarize recent urban applications based on the latest urban mobility models and predictive algorithms in Section VII.

The contribution of this article is mainly threefold:

- We summarize recent studies on urban human mobility from a data mining view: from the data collection and cleaning, to the mobility models, and the applications. To the best of our knowledge, this is the first article summarizing recent urban human mobility studies from the overall data mining process.
- We describe the urban human mobility models from both the complex network (Physics) and machine learning (Computer Science) view. Current surveys mainly classify and describe mobility models or predictive tools either from the physicist’s view, or from the computer scientist view. Our paper provides a unified view to the topic.
- We classify and describe recent public urban human mobility data sets, and how we can use such data sets for building mobility models or validating the prediction results. In addition, we provide a mathematical synthesis for the well-known models and predictors. There is still significant potential for high impact research in the area.

## II. URBAN MOBILITY DATA SETS

A variety of urban human mobility data have been gathered and published, due to the significant growth of sensing technologies and large-scale computing infrastructures. This urban human mobility data contains rich knowledge about locations and can help in addressing many urban challenges. For example, understanding human movements inside a city can help forecasting of the traffic [8]. Another example is that we can identify the functions of locations by the means of the transitions between these locations [4], [10], e.g., people

Experimental data set	Relative Location			Individual Mobility		Aggregated Mobility			
	INFOCOM06	Reality	UCSD	Geolife	Nokia MDC	Rome	San Francisco	T-Drive	NYC Yellow Taxi
Location	Barcelona	Boston	San Diego	Beijing	Geneva	Rome	San Francisco	Beijing	New York
Context settings	Conference	Campus	Campus	City	City	City	City	City	City
Number of samples	24 M	11 M	11 M	21 M	17 M	11 M	21 M	17 M	868 M
Localization Technology	Bluetooth	GSM	WiFi	GPS	GPS	GPS	GPS	GPS	GPS
Participants	78	97	274	182	200	536	316	10,357	13,237
Duration of trace	4 days	9 months	2 months	5 years	1.5 year	1 month	1 month	1 week	7 year
Sampling interval	120s	300s	120s	1-5s	10s	64s	9s	177s	1s

TABLE I: Urban Human Mobility Data Sets

usually go to work during daytime on weekdays, and visit shopping centers after work.

The main human mobility data sets are recorded according to: 1. relevant location with access points (Bluetooth or WiFi direct access points, Cellular Tower, etc.), 2. GPS information by individual devices, 3: Aggregated GPS points recorded by vehicles such as taxis or buses. In this article, we mainly summarize the available public data sets falling in these three categories (see Table I).

#### A. Relative Location Data Sets

The relative location data sets collect the proximity information of a mobile device (e.g., mobile phone) carried by a person to the access points (cellular towers or a WiFi access points). If we know the location of the access point, we can infer the individual relative location. The advantage of the access point data sets is that they contain information pertaining to the social networks of a person. For example, two persons can be considered as friends or acquainted if they stay at the same location for a long time [19]. The disadvantage is that the granularity of the access point data sets is usually low, a cellular tower location accuracy is usually 2-3 kilometers, much larger compared to the GPS location. It should also be noted that mobile operators are typically not willing to share their data sets due to privacy issues. Privacy-enhancing technologies aim to solve such problem [20]. Mir et al. [21] propose a method for generating synthetic Call Detail Records (CDRs), to capture the mobility patterns of real metropolitan populations while preserving privacy. The accuracy of their method has been validated against billions of relative location samples for hundreds of thousands of cell phones in the New York and Los Angeles metropolitan areas.

Here we briefly introduce three real-world access point mobility data traces: (i) The Infocom06 data set [19] contains opportunistic Bluetooth contacts between 98 mobile devices in a conference in Barcelona, 78 of which were distributed to the participants and 20 of which were static. The relative location of the 78 participants to the 20 static devices were recorded, from which we can infer the location of each participant. (ii) The MIT Reality trace [22] comprises 95 participants carrying GSM enabled cell-phones over a period of 9 months. The cellular tower location were also provided by the data set. (iii) In the UCSD data set [23], 274 WiFi-enabled PDAs were respectively used by 274 freshmen to log nearby Access Points for an 11-week period between Sep 22, 2002 and Dec 8, 2002.

#### B. Individual Mobility Data Sets

Instead of logging the relative locations to the access points, the individual mobility data sets record the GPS position of each participant. This is usually the best data set for modeling the individual mobility. However, due to privacy issues, it is hard to collect such data sets at large scale. Individual mobility data can reveal the everyday behavior of the people: where they live, where they work, where they have dinners, and so forth. All this information is related with the private personal life and could be potentially lead to undesirable and unlawful consequences. Many privacy-enhancing technologies for the individual mobility data have been proposed, see the survey from Calabrese et al. [20] for an overview of recent methods.

Here we introduce two public individual mobility data sets: (i) Geolife [24] is a public data set with 182 users' GPS trajectories over five years (from April 2007 to August 2012) gathered mainly in Beijing, China. This data set contains over 24 million GPS samples with a total distance of 1,292,951 kilometers and a total of 50,176 hours. It includes not only daily life routines such as going to work and back home in Beijing, but also some leisure and sports activities, such as sightseeing, and walking in other cities. The transportation mode information in this data set is manually logged by the participants. (ii) The Nokia MDC data set [25] is a public data set from Nokia Research Switzerland that aims to study smart-phone user behavior. The data set contains extensive the smartphone data of two hundred volunteers in the Lake Geneva region over one and a half years (from September 2009 to April 2011). This data set contains 11 million data points and the corresponding transportation modes.

#### C. Aggregated Mobility Data Sets

Public transportation data set, such as bus data, taxi data or subway data represents the aggregated human mobility. Take the taxi data set as an example, it usually contains the following information: taxi id, timestamp and taxi position (longitude, latitude). In the taxi mobility patterns, the drivers typically either move to pick up or drop off customers, or stay in parking areas while waiting for new customers. Thus the pick-up location and drop-off location can be considered as the trip origin and destination for one person.

Here we introduce four public aggregated mobility data sets: (i) The San Francisco data set [26] is a public data set from the Exploratorium project that aims to study the invisible economic, social, and cultural trends of the city. The data

set contains extensive GPS data of five hundred Yellow Cab vehicles in the San Francisco region over one month (from 17th May 2008 to 10th June 2008). This data set contains 11 million data points and the corresponding timestamps. (ii) The Rome data set [27] is a public data set containing mobility traces of 316 taxi cabs in Rome over 30 days. Each taxi driver had a tablet that was set to retrieve the GPS position every 7 seconds after which the position was sent to a central server. (iii) The Beijing data set [28] is a public data set gathered by Microsoft Research Asia. It records the GPS trajectories of 10,357 taxis in Beijing from Feb.2 to Feb.8, 2008. There are about 15 million GPS points in this data set, and the total distance for each trajectory reaches up to 9 million kilometers. (iiii) The New York Taxi & Limousine Commission (TLC) captures the detailed information about each trip through the meters installed in each vehicle, and store them in the public yellow taxi data set [29]. Every day there are over 500,000 taxi trips serving roughly 600,000 people in New York City. Each trip consists of two spatial attributes (pickup and dropoff locations), two temporal attributes (pickup and dropoff times), and additional attributes including taxi identifier, distance traveled, fare, medallion code, and tip amount.

### III. URBAN MOBILITY DATA CLEANING AND PREPROCESSING

The collected urban mobility data sets are not always accurate. For example, the GPS samples collected are heavily influenced by the tall buildings in cities and thus can be inaccurate. The quality of the GPS receiver algorithm might also lead to inaccurate GPS positions. Fig. 2 displays such errors. We plot the New York yellow taxi GPS samples and find that many of the taxi GPS samples are in rivers, in the ocean and even outside North America. In this section, we introduce the data cleaning and preprocessing methods for urban human mobility data.

#### A. Data Cleaning

Freire et al. survey the challenges and solutions while cleaning urban mobility data sets [29]. Visualization tools is an effective mechanism to identify the GPS errors. To remove GPS inconsistencies, one common method is to use the geographical boundaries to clear all the GPS samples that out of the boundary. Besides the spatial errors, another common problem is that the temporal recorded might also contain some errors. For example, while analyzing the taxi data in New York City, there is a large number of overlapping trips for the same taxi. That is, for a given taxi, a new trip starts before the previous trip has ended. The reason behind this error is unclear: some trips may overlap due to a device error, or simply because the taxi driver forgot to log the end of a trip after dropping off passengers. Nevertheless, they certainly affect further analysis on the data. Such inconsistencies must be removed before using the data set.

#### B. Data Preprocessing

After data cleaning, the next step is to preprocess the data for the specific usage. For example, if we want to identify the

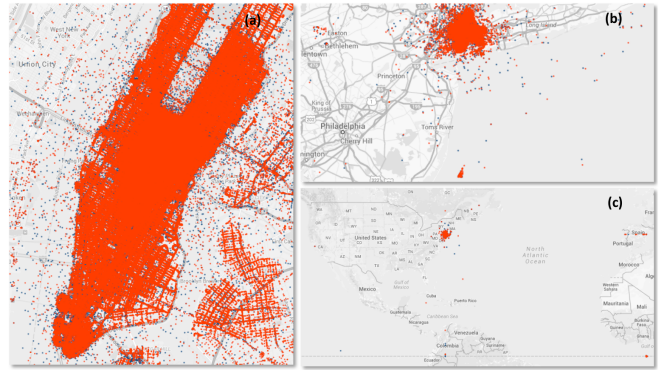


Fig. 2: Inaccurate New York taxi GPS points (a) in rivers, (b) in the ocean, and (c) outside North America [29].

number of people leaving a building block, we need to map the taxi pick-up samples with the associated building blocks. R-tree [30] is often used for mapping the GPS point with the shape file to identify the associated building blocks. Zheng et al. [31] give an overview of the urban data preprocessing algorithms such as Hidden Markov Models [32].

Currently the mapping of human mobility data to geospatial features such as building blocks, roads or neighborhoods, requires a lot of processing given the volume of the data set. E.g., in New York City, every day there are over 500,000 taxi trips serving roughly 600,000 people [29]. The big data processing platform such as Spark and Hadoop are commonly used for dealing with such huge data sets. It has been found that it takes about ten minutes for a R-tree based algorithm to map matching the 14 million GPS samples with the associated building blocks [33].

### IV. URBAN MOBILITY MODELING

The growing volumes of urban human mobility data sets can help us modeling and predicting the human mobility more accurately. In this section we mainly summarize and compare recent urban human mobility models (see Table II). Random Way Points (RWP) [34], Lévy Walks [18] Gravity Model [6] and Radiation Model [16] are the most commonly used mobility models.

#### A. Synthetic Mobility Model

Synthetic mobility models are created without the use of observation, based only assumptions about certain properties of movement, such as changes in direction or changes in flight length. A flight is defined as a trip of a person from one location to another without pause.

Here we introduce two synthetic mobility models: In RWP model [34], the mobile nodes move randomly and freely without any restrictions. The destination, speed and direction are all chosen randomly and independently of the other nodes. In Brownian Motion (BM) [35], the mobile nodes move with a mean flight and a mean pause time between flights. A flight is defined as a trip of a person from one location to another without pause. In BM, the flights are normally distributed.

Publications	Data Set	Human Mobility Model (Flight distributions)	Transportation Mode	Explanation
RWP [34]	None	Random	None	None
BM [35]	None	Normal	None	None
LEVY [36][18]	Aggregated Mobility	Power-Law	None	Levy Walk
GM [15]	Relative Location, 2-3 km	Power-Law and Gaussian model	None	Temporal and spatial regularity
TRAF [37][38]	Aggregated Mobility , 5-10 m	Power-Law	None	Hierarchical Traffic systems
AGI [39]	Individual Mobility	Power-Law	Yes	Aggregated individual mobility
TRANS [40]	Individual Mobility	Power-Law and log-normal	Yes, multiple transportation modes	Lévy Walks decomposed by transportation modes
EXP [41], [42]	Aggregated Mobility	Exponential	Yes, taxi	Population Density
Gravity [6]	Aggregated Mobility	Gravity Model	Yes	Population
Radiation [16]	Relative Location	Radiation Model	Yes	Population and opportunities
coMobile [43]	Individual Mobility	Multi-view Learning	Yes	None

TABLE II: Comparison of recent urban human mobility models

Fig. 3 (a) and (c) shows the sample trajectory of RWP and BM respectively.

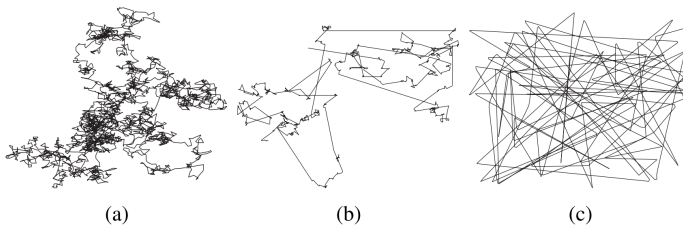


Fig. 3: Sample trajectory of (a) BM, (b) Lévy Walks and (c) RWP [44].

### B. Lévy Walks

Both the RWP and BM are not based on the real human mobility studies so that they do not reflect how people move in real life. Recent data-driven research has shown that trajectories in human mobility have statistically similar features as Lévy Walks by studying the tracing of bank notes [36], cell phone users' locations [15] and GPS traces [18], [39], [37], [38]. According to the Lévy Walks model, human movement contains many short flights and some long flights (see Fig. 3 (b)). The flight length  $l$  follows a power-law distribution,

$$P(l) \sim l^{-(1+\beta)} \quad (1)$$

where the displacement exponent  $\beta < 2$ .

Although recently human mobility has been empirically observed to exhibit Lévy flight characteristics and behaviour with power-law distributed jump size [39], [37], [38], the fundamental mechanisms behind this behavior has not yet been fully explained. Later studies propose explanations for the emergence of the Lévy Walks pattern:

1) *Gaussian Model*: Gonzalez et al. [15] model human mobility as a stochastic process centered around a single location. They indicate the the power-law jump size distribution is due to the convolution between the statistics of the motion of individuals and the population heterogeneity. That is, each individual mobility follows the power-law distribution and there is also a population-based heterogeneity coexists between individuals.

2) *Hierarchy Traffic System*: The hierarchy of traffic networks [37] or road networks [38] are also possible reasons behind the Lévy Walks. Han et al. [37] model the human mobility as a random walk process in hierarchical Euclidean networks and such system can reproduce the statistics of Lévy Walks pattern. Each node in the hierarchical network represents a city such as first-layer city or second-layer city. The edges represent the connection between the cities. Their model implies that the human mobility are strongly affected by the geographical structure of traffic systems. Similar results have also been found by Jiang et al. [38] while examining the human mobility on the street networks.

3) *Aggregation of Individual Mobility*: In [39] Yan et al. observe that the individual human mobility patterns do not follow Lévy Walks and Lévy Walks are due to the aggregation of individual mobility patterns. The aggregated displacement distribution can be explained by the mixed nature of human mobility under the maximum entropy principle. The maximum entropy principle also predicts that the human mobility with the single transportation mode follows the exponential distribution, which is consistent with other findings [41], [42].

4) *Decomposition by Transportation Modes*: Intuitively, these long and short flights in the Lévy walk model reflect different transportation modalities. The short flights might be associated with walking or bicycling mode, whereas the long flights might be associated with the subway or train trips. Zhao et al. [40] propose to explain the Lévy walk behaviour by decomposing the trips into different classes according to different transportation modes, such as Walk/Run, Bike, Train/Subway or Car/Taxi/Bus. They observe that human mobility can be modelled as a mixture of different transportation modes, and these single transportation movement patterns can be approximated by a lognormal distribution, rather than a power-law distribution. They demonstrate that the mixture of the decomposed lognormal flight distributions associated with each modality is a power-law distribution, providing an explanation of the Lévy walk human mobility.

### C. The Exponential-scaling Human Mobility Model

Recent research results [41], [42] investigate the urban human mobility of a single transportation mode such as taxi

and they found that the scaling of human flights is exponential:

$$P(l) \sim e^{-\lambda l} \quad (2)$$

In [41] Liang et al. propose that this is possibly because few people tend to travel long distances by taxi due to economic considerations. In [42] they explain the exponential law of urban human mobility as a result of the exponential decrease in average population density in urban areas. They find that the empirical and analytical results indicating the same exponential decaying rate between the flight length and the population density.

#### D. Gravity Model

Jung et al. [6] investigate the traffic flows of the Korean highway system for 30 selected cities. They find that there is a positive correlation between the traffic flow and the population of two cities. The traffic flows between city  $i$  and  $j$ ,  $T_{ij}$ , form a Gravity model:

$$T_{i,j} = \frac{P_i P_j}{r_{ij}^2} \quad (3)$$

where  $P_i$  and  $P_j$  are the population of city  $i$  and  $j$  and  $r_{ij}$  is the distance between  $i$  and  $j$ .

#### E. Radiation Model

One big flaw with the Gravity model is that it can not describe the number of individual flows in both directions between two locations. To address this problem, Simini et al. [16] propose the radiation human mobility model. In this model, the number of trips  $T_{ij}$  from location  $i$  to  $j$  is:

$$T_{i,j} = T_i \frac{P_i P_j}{(P_i + P_{ij})(P_i + P_j + P_{ij})} \quad (4)$$

Here  $P_i$  and  $P_j$  are the population of location  $i$  and  $j$ .  $T_i$  is the total number of trips starting from  $i$ .  $P_{ij}$  is the total population of locations (other than  $i$  and  $j$ ), from which the distance are less than or equal to the distance  $d_{ij}$  between  $i$  and  $j$ . They observe that the Radiation model can significantly improve the accuracy of predictive tools in a wide range of phenomena, from long-term migration patterns to communication volume between different regions.

#### F. Multi-view Learning Model

Existing urban human mobility are mostly driven by data from a single view, e.g., data from a single transportation view [41] such as taxi, bus, subway or a cellphone view [15] such as call records. To address this issue, Zhang et al. [43] propose a new human mobility model based on a multi-view learning framework. They find that the new multi-view human mobility model outperforms a single-view model by 51% on average.

They improve the performance of single-view model based on tensor decomposition with correlated context. Take the call record data as an example, they construct a three dimensional tensor, an entry in this tensor represents [user id, time, location] (see Fig. 4). Due to the sparsity of the call records, they use the tensor decomposition methods to decompose the tensor

into a core tensor, with small latent factors. Then they try to optimize the decomposition problem by reducing the errors and using the regularization function to avoid over-fitting. The obtained approximate tensor can improve the accuracy of each single-view model.

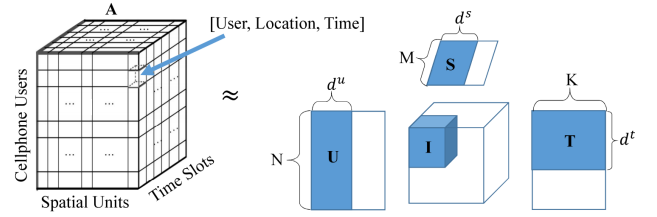


Fig. 4: Tensor decomposition for single-view mobility modeling [43].

Then they integrate these improved single-view human mobility models together for multi-view learning to iteratively obtain mutually-reinforce knowledge. A human mobility graph is formed, which is a combination of many single-view human mobility graphs such as call records or transportation data. Then they use the iterative multi-view learning method to obtain the ground truth of the edges in the human mobility graph, that is the volume of passengers traveling from one place to another. After that, a human mobility graph is formed with high accuracy for modeling urban human mobility.

## V. URBAN MOBILITY PREDICTION

Urban human mobility prediction pertains to the estimation of the next location that a person will visit in a city. Urban human mobility exhibit strong temporal regularities, e.g., people usually go to work during daytime on weekdays, and go shopping after work. Marta et al. observe that the trajectories in urban human mobility exhibit strong regularities by studying cell phone user's locations [15]. Each person has a significant probability to return to a few highly frequented locations such as home or working places.

In Fig. 5 we show the directions of taxi flow from the lower Manhattan to other regions for three time steps. The taxi flow indicates the aggregated human mobility. At 8 am we observe that the probability of taxis moving beyond Midtown is low. However, after 4 pm the probability of taxis moving towards Upper Manhattan is high. This is mainly due to the fact that the lower Manhattan is mainly a working place and Upper Manhattan (e.g., Upper East) is mainly a residential place. People tend to go home after work and the probability that the taxi moves from Lower to Upper Manhattan increases.

Such spatial-temporal regularities can be utilized for predicting urban human mobility using data mining methods. Here we give a formal definition of the urban mobility prediction problem: suppose  $d_j$  represent the person's location at time  $j$  ( $1 \leq j \leq n$ ). For this person, we have the historical location visits as a sequence  $D_n = d_1 d_2 d_3 \dots d_n$ . Given  $D_n = d_1 d_2 d_3 \dots d_n$ , our goal is to predict the person's next location  $d_{n+1}$  at time  $n + 1$ . Different predictors have been

Publications	Data Set	Features	Methodology	Approaching error bound	Computation Time
Markov [45]	Relative Location	Location, time	Markov	Yes	Fast
LZW [45]	Relative Location	Location, time	Compression-based	No	Fast
ARIMA [33]	Aggregated Mobility	Location, time	Time-series	Unknown	Relatively Fast
LSTM [45]	Individual Mobility	Location,time,transportation mode	deep learning	Unknown	Slow

TABLE III: Comparison of recent urban human mobility predictors.

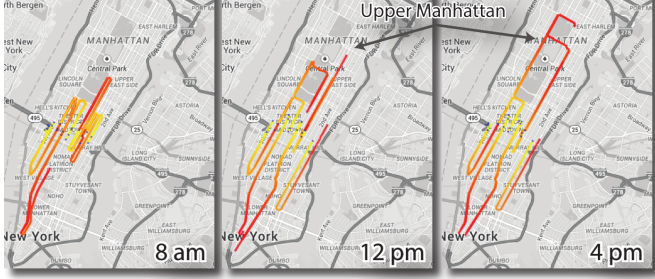


Fig. 5: Directions of taxi flow at different time of the day. Note that the taxis tend to cross into Upper Manhattan only in the latter part of the day, e.g., after 4 pm. The direction is color encoded from red to yellow [46].

proposed for predicting the human mobility, in this paper we mainly survey the Markov-based predictors, compression-based predictor, time-series-based predictor and deep learning predictor (see Table III). The evaluation of the predictors can be found in Section VI.

#### A. Markov Predictor

In this subsection, we discuss the Markov-based predictor. The order- $k$   $O(k)$  Markov predictor can be used for predicting the future location of a user from the  $k$  most recent location history sequence  $d_{n-k+1}, d_{n-k+2}, \dots, d_n$  [45]. The location that the person visits during time  $j$  can be viewed as a random variable  $X_j$ . Let  $X_{j,k}$  donates the sequences of random variable  $X_j, X_{j+1}, X_{j+2}, \dots, X_k$  for  $1 \leq j \leq k \leq n$  representing the person's past locations. Considering the person's visiting history  $D_n = d_1 d_2 d_3 \dots d_n$  and  $N$  as the set of all possible locations that the person can visit, following Markov assumption we have

$$P(X_{n+1} = \beta | X_n = D_n) \quad (5)$$

$$= P(X_{n+1} = \beta | X_{n-k+1, n} = c) \quad (6)$$

$$= P(X_{j+k+1} = \beta | X_{j+1, j+k} = c). \quad (7)$$

Here  $P(X_{n+1} = \beta | X_n = D_n)$  means that the probability of the person is at the location  $\beta$  during the time interval  $n+1$ .  $c$  is the sub-sequence of the previous human mobility history where  $d_{n-k+1} d_{n-k+2} \dots d_n = d_{j+1} d_{j+2} \dots d_{j+k} = c$ .

#### B. Compression-based Predictor

The LZW predictor is based on the Lempel-Ziv-Welch compression algorithm (LZW) [47], [45]. Given a person's visit historical sequence  $D_n$ , LZW algorithm partitions  $D_n$  into distinct subsequence  $s_0, s_1, s_2, s_3, \dots, s_m$ , where  $s_j$  represents

the shortest subsequence starting at the time  $j$  which does not appear from 1 to  $j-1$ . We have the LZW predictor:

$$P(X_{n+1} = \beta | D_n) = \frac{N^{LZ}(s_m \beta, D_n)}{N^{LZ}(s_m, D_n)} \quad (8)$$

Here  $P(X_{n+1} = \beta | D_n)$  represents the probability of the person that is at the location  $\beta$  during the time interval  $n+1$ .  $\frac{N^{LZ}(s_m \beta, D_n)}{N^{LZ}(s_m, D_n)}$  represents the probability of the subsequence  $s_m \beta$  occurs in the mobility sequence  $s_m$ .

#### C. Time-series Predictor

Li et. al [48] investigate human mobility patterns in an urban taxi transportation system. They propose an improved Auto-Regressive Integrated Moving Average (ARIMA) based predictive algorithm to forecasting the spatial-temporal variation of passengers in hotspots in a city. ARIMA is a classical approach for time series analysis. The ARIMA predictor first build a mathematical model with the historical human mobility data for representing the regular pattern of a time series. Then it use this model and the historical values forecasting the future value. We give a basic ARIMA predictor below. Given the urban human mobility data  $d_1 d_2 d_3 \dots d_{n-1}$ , the ARIMA is going to predict  $d_n = \beta$  by solving the following equations:

$$\phi(B) \nabla^d \beta = \theta(B) a_i \quad (9)$$

Here  $B$  is the lag operator,  $\phi(B)$  is the auto-regressive process,  $\nabla^d$  is the differencing operator,  $\theta(B)$  is moving average process, and  $a_i$  is a random walk process. The predicted value  $\beta$  can be obtained from historical values  $d_1 d_2 d_3 \dots d_{n-1}$ .

#### D. Deep Learning Predictor

Song et al. building a deep LSTM learning architecture (see Fig. 6) for predicting the urban human mobility [13]. Recurrent Neural Network (RNN) is able to capture the temporal and spatial evolution of human mobility patterns. However, it has been shown that the traditional RNN fail to capture the long temporal dependency for the input sequence [49]. LSTM-a special RNN architecture is developed for sequence prediction tasks, which can learn the time series with long time spans and determine the optimal time lags automatically. Given a person's observed mobility history data, they find that the deep LSTM is able to predict his future movements and transportation mode with over 80% accuracy.

## VI. EVALUATING THE MODELS

The evaluation of the proposed mobility models and predictors usually involve multiple factors: the error metrics, the computing performance or the feedback from domain experts.

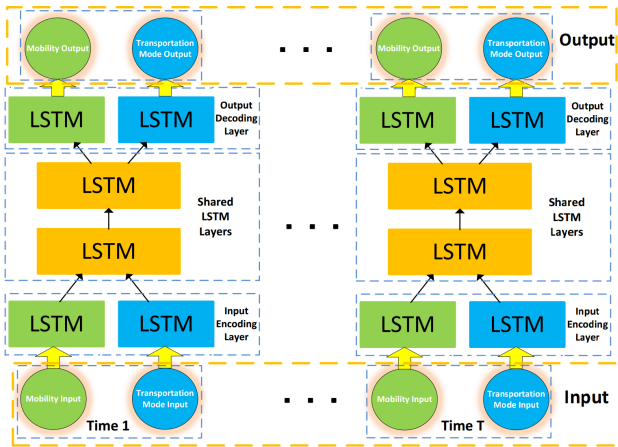


Fig. 6: Deep LSTM Learning Architecture for predicting urban human mobility [13].

We summarize the common evaluation metrics here and also apply some metrics to evaluate the predictors in Table III.

#### A. Error Metric

The error metrics such as sMAPE are often used for evaluating the performance of different mobility models and predictive algorithms. For example, Song et al. [13] compare the prediction accuracy of the LSTM predictor (see Section V-D) and Gaussian Model (see Section IV-B1) and they observe that the LSTM predictor achieve better performance than the Gaussian Model using sMAPE.

1) *Lower Bound of Predictive Errors*: Song et al. [50] introduced the limits of predictability  $\Pi^{max}$  in urban human mobility. They define the limits of predictability  $\Pi^{max}$  as the highest potential accuracy (lower bound of errors) that a predictive algorithm can reach for predicting human mobility. The limits of predictability  $\Pi^{max}$  is obtained by measuring the entropy of the human mobility sequence considering both the randomness and the temporal correlation of human movement. They analyze 50,000 users mobility and find that there is a potential 93% predictability.

The limits of predictability captures the degree of the temporal correlation in human mobility [51]. For most people, their mobility patterns are governed by a certain amount of randomness (e.g., unexpected events) and some degree of regularity (e.g., weekly patterns), which can be exploited for prediction. For a person with  $\Pi^{max} = 0.3$ , that means that at least 70% of the time the his location appears to be random, and only in the remaining 30% of the time can we hope to predict the location that he appears. In other words, no matter how good the predictive algorithm is, we cannot predict with better than 30% accuracy the future location of a person with  $\Pi^{max} = 0.3$ .  $\Pi^{max}$  represents the fundamental limit for prediction accuracy of the human mobility.

2) *Approaching the Lower Bound of Errors*: Lu et al. find that the limits of predictability (the lower bound of errors) is not only a fundamental theoretical limit for the potential predictive algorithm, but also an approachable target for actual

prediction accuracy [51]. They implement a set of the Markov predictors to predict the actual location visited by each user. Results show that the order  $O(1)$  Markov predictor can approach the limits. The higher order Markov predictor does not generate improved prediction accuracy when compared to a  $O(1)$  Markov predictor (see Fig. 7).

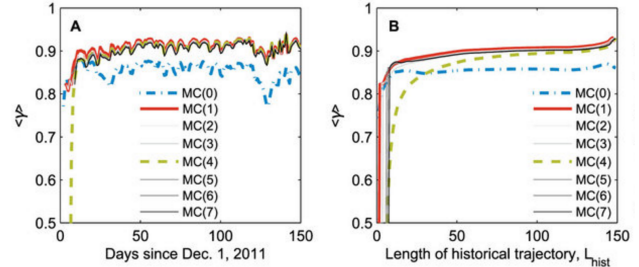


Fig. 7: (a) The accuracy of predictive algorithms for each day based on historical data, (b) the accuracy increases with the length of historical trajectories [51].

#### B. Performance

Beside the prediction errors, the computation performance should also be considered while evaluating the models and predictors. First, the mobility data is growing significantly and so is the computation time for processing such big data. Second, some of the urban mobility based applications such as the traffic forecasting system require the real-time output. Scalability and computation time [45] are the two performance metrics often to be examined during the evaluation. For example, it has been found that [33] the Markov predictor computation time is about 0.03% of the deep learning method. Choosing the algorithms that is able to scale up and compute efficiently while maintaining prediction accuracy will save a lot of computation time.

#### C. Domain Knowledge

Urban human mobility relates to many urban applications, such as urban traffic analysis. To validate the accuracy of the model, it is often required to have the domain experts in the loop. For example, as shown in Fig. 5, we observe that the human mobility flow from the Lower Manhattan to Upper Manhattan only occurs after the working time. Domain experts from Department of Transportation or TLC can verify such mobility patterns based their previous experience.

## VII. URBAN MOBILITY APPLICATION

The urban human mobility models and predictor summarized above can help in addressing many urban problems. In this section, we give four examples using the urban mobility models for solving urban problems:

1) *Traffic Forecasting*: Urban traffic anomalies are usually caused by accidents, control, protests, sport events, celebrations, disasters and other events. Pan et al. [8] propose a method for detecting and describing such anomalies by analysing human mobility patterns. They evaluate their



mobility-based system with a GPS trajectory data set generated by over 30,000 taxicabs in Beijing. The evaluation results show significant advantages over the traditional traffic volume-based anomaly detection methods regarding accuracy and computation performance.

2) *Air Pollution Detection*: Zheng et al. [9] observe that there is a positive correlation between the concentration of  $PM_{10}$  in a region and the number of people arriving at and departing from that region. While there are limited air-quality-monitor-stations in a city, the urban human mobility model can be an very important feature inferring the real-time and fine grained air quality information.

3) *Functional Region Detection*: There are different functional regions in a city, e.g., residential areas, business districts and educational areas. The functions of a region have a strong correlation with the urban human mobility. In the workdays people usually go the working places in the morning and return to residential places in the afternoon. Jing et al. [4] use a topic-based inference model for inferring the functions of each region with urban human mobility patterns.

4) *Mobile Ad Hoc Networks*: In Mobile Ad Hoc Networks (MANET), whenever mobile devices (vehicles, phones, etc.) encounter each other, they can exchange content via short-range communications (e.g., Bluetooth or WiFi) for increasing the network throughput [52]. Since people carry their mobile devices everywhere everyday, human mobility model plays an important role in such network. The choice of the mobility model has a significant impact on the behaviour and performance of a MANET algorithm. Lévy Walks provide a more accurate mobility model compared to other existing models. The heavy-tail tendencies of the Lévy Walks model induce heavy-tail routing delays and throughput in MANET [18].

## VIII. RELATED SURVEYS

Several surveys have been presented regarding urban human mobility in the past few years. Camp et al. [53] give an overview of several synthetic mobility models such as RWP and BM back in 2002. Aschenbruck et al. [54] review and discuss several publicly available mobility data sets. Musolesi et al. [55] survey the mobility models that utilize information from social networks. Gonçalves et al. [56] review the urban human mobility models from the physicist's view. Hess et al. [57] provide a data-driven human mobility model survey for mobile networking applications. They take an engineering approach and discuss the steps of model creation and validation.

The difference between our survey and previous works is that, many of them focus on a single process, e.g., finding human mobility data sets [54] or human mobility models [57]. No existing articles summarize urban human mobility studies from the overall data mining process. We summarize urban human mobility studies from a data mining view: from the data collection and cleaning, to the mobility models, and the applications. In addition, we describe the urban human mobility models from both the complex network (Physics) and machine learning (Computer Science) view. Current surveys mainly classify and describe mobility models or predictive

tools either from the physicist's view [56], or from the computer scientist view [55]. Our paper provides a unified view to the topic.

## IX. CONCLUSION

Today, 50% of the world's population lives in cities, rising to 70% by 2050; North America is already 80% in cities and the number will be 90% by 2050 [58]. Understanding urban human mobility is crucial for epidemic control, urban planning, traffic forecasting systems and, more recently, various mobile and network applications in cities. The growing volumes of urban mobility data being collected and made available open up new opportunities for modeling and predicting the urban human mobility more accurately. In this paper, we survey recent studies on urban human mobility from a data mining view: from the data collection and cleaning, to the mobility models, and the applications.

Due to page limits, some aspects are not covered in this paper, such as the summary of the mobility data privacy protection technologies [21] or the recent map matching algorithms [32]. In the future work, we plan to extend the current work by adding those missing components. Besides, a benchmark across the surveyed models and predictors will also be implemented in future works.

## ACKNOWLEDGMENT

This work was supported in part by a CUNY IRG Award and the NYU Center for Urban Science and Progress. Siyuan Liu's work was supported by Basic Research Program of Shenzhen: JCYJ20140610152828686, and Natural Science Foundation of China: 61572488.

## REFERENCES

- [1] S. Ni and W. Weng, "Impact of travel patterns on epidemic dynamics in heterogeneous spatial metapopulation networks," *Physical Review E*, vol. 79, no. 1, p. 016111, Jan. 2009.
- [2] V. Belik, T. Geisel, and D. Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases," *Phys. Rev. X*, vol. 1, Aug 2011.
- [3] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *UbiComp*, 2011, pp. 89–98.
- [4] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *SIGKDD'12, Beijing, China, August 12-16, 2012*, 2012, pp. 186–194.
- [5] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang, "Measuring social functions of city regions from large-scale taxi behaviors," in *IEEE PerCom 2011, 21-25 March 2011, Seattle, WA, USA, Workshop Proceedings*, 2011, pp. 384–388.
- [6] W.-S. Jung, F. Wang, and H. E. Stanley, "Gravity model in the korean highway," *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48005, 2008.
- [7] S. Goh, K. Lee, J. S. Park, and M. Y. Choi, "Modification of the gravity model and application to the metropolitan seoul subway system," *Phys. Rev. E*, vol. 86, p. 026102, Aug 2012.
- [8] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2013, pp. 344–353.
- [9] Y. Zheng, F. Liu, and H. Hsieh, "U-air: when urban air quality inference meets big data," in *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, 2013, pp. 1436–1444.
- [10] K. Zhao, M. P. Chinnasamy, and S. Tarkoma, "Automatic city region analysis for urban routing," in *IEEE ICDMW 2015, Atlantic City, NJ, USA, November 14-17, 2015*, 2015, pp. 1136–1142.

- [11] W. Rao, K. Zhao, E. Lagerspetz, P. Hui, and S. Tarkoma, "Energy-aware keyword search on mobile phones," in *Proceedings of MCC@SIGCOMM 2012, Helsinki, Finland*, 2012, pp. 59–64.
- [12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.
- [13] X. Song, H. Kanasugi, and R. Shibasaki, "Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level," in *IJCAI 2016, New York, NY, USA, 9-15 July 2016*, July 2016.
- [14] E. G. Ravenstein, "The laws of migration," *Journal of the Royal Statistical Society*, 1885.
- [15] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.
- [16] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.
- [17] S. Makridakis and M. Hibon, "The m3-competition: results, conclusions and implications," *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476, 00 2000.
- [18] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, 2011.
- [19] A. Chaintreau, A. Múbaa, L. Massoulié, and C. Diot, "The diameter of opportunistic mobile networks," in *CoNEXT*, 2007, p. 12.
- [20] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: A survey of research," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 25:1–25:20, 2014.
- [21] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, and R. N. Wright, "DP-WHERE: differentially private modeling of human mobility," in *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, 2013, pp. 580–588.
- [22] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.
- [23] M. McNett and G. M. Voelker, "Access and mobility of wireless PDA users," *Mobile Computing and Communications Review*, vol. 9, no. 2, pp. 40–55, 2005.
- [24] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.
- [25] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, and L. J., "Towards rich mobile phone datasets: Lausanne data collection campaign," in *Proc. ACM Int. Conf. on Pervasive Services (ICPS)*, Berlin., 7 2010.
- [26] M. Piorowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "A parsimonious model of mobile partitioned networks with clustering," in *Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International*, Jan 2009, pp. 1–10.
- [27] R. Amici, M. Bonola, L. Bracciale, A. Rabuffi, P. Loreti, and G. Bianchi, "Performance assessment of an epidemic protocol in vanet using real traces," vol. 40, no. 0, 2014, pp. 92 – 99, fourth International Conference on Selected Topics in Mobile Wireless Networking (MoWNet2014).
- [28] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in *18th ACM SIGSPATIAL GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*, 2010, pp. 99–108.
- [29] J. Freire, A. Bessa, F. Chirigati, H. T. Vo, and K. Zhao, "Exploring what not to clean in urban data: A study using new york city taxi trips," *IEEE Data Eng. Bull.*, vol. 39, no. 2, pp. 63–77, 2016.
- [30] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, June 18-21, 1984*, 1984, pp. 47–57.
- [31] Y. Zheng and X. Zhou, *Computing with Spatial Trajectories*, 1st ed. Springer Publishing Company, Incorporated, 2011.
- [32] P. Newson and J. Krumm, "Hidden markov map matching through noise and sparseness," in *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2009, pp. 336–343.
- [33] K. Zhao, D. Khryashchev, J. Freire, C. Silva, and H. Vo, "Predicting taxi demand at high spatial resolution: Approaching the limit of predictability," in *IEEE International Conference on BigData*, 2016.
- [34] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Trans. Mob. Comput.*, vol. 2, no. 3, pp. 257–269, 2003.
- [35] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 483–502, 2002.
- [36] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006.
- [37] X.-P. Han, Q. Hao, B.-H. Wang, and T. Zhou, "Origin of the scaling law in human mobility: Hierarchy of traffic systems," *Phys. Rev. E*, vol. 83, p. 036117, Mar 2011.
- [38] B. Jiang, J. Yin, and S. Zhao, "Characterizing the human mobility pattern in a large street network," *Phys. Rev. E*, vol. 80, p. 021136, Aug 2009.
- [39] X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou, "Diversity of individual mobility patterns and emergence of aggregated scaling laws," *Sci. Rep.*, vol. 3, Sep 2013.
- [40] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Nature Scientific Reports*, vol. 5, no. 9136, March 2015.
- [41] X. Liang, J. Zhao, L. Dong, and K. Xu, "Unraveling the origin of exponential law in intra-urban human mobility," *Sci. Rep.*, vol. 3, Oct 2013.
- [42] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu, "The scaling of human mobility by taxis is exponential," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 5, pp. 2135–2144, 2012.
- [43] D. Zhang, J. Zhao, F. Zhang, and T. He, "comobile: real-time human mobility modeling at urban scale using multi-view learning," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, November 3-6, 2015*, 2015, pp. 40:1–40:10.
- [44] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy-walk nature of human mobility," in *INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13-18 April 2008, Phoenix, AZ, USA*, 2008, pp. 924–932.
- [45] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location predictors with extensive wi-fi mobility data," in *Proceedings IEEE INFOCOM 2004, Hong Kong, China, March 7-11, 2004*, 2004.
- [46] J. Poco, H. Doraiswamy, H. T. Vo, J. L. D. Comba, J. Freire, and C. T. Silva, "Exploring traffic dynamics in urban environments using vector-valued functions," *Comput. Graph. Forum*, vol. 34, no. 3, pp. 161–170, 2015.
- [47] C. T. Cheng, R. Jain, and E. van den Berg, "Mobile wireless systems: Location prediction algorithms," in *Encyclopedia of Wireless and Mobile Communications*, 2008.
- [48] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers of Computer Science in China*, vol. 6, no. 1, pp. 111–121, 2012.
- [49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [50] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
- [51] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the Limit of Predictability in Human Mobility," *Scientific Reports*, vol. 3, Oct. 2013.
- [52] W. Rao, K. Zhao, P. Hui, Y. Zhang, and S. Tarkoma, "Maximizing timely content delivery in delay tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 4, pp. 755–769, 2015.
- [53] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 483–502, 2002.
- [54] N. Aschenbruck, E. Gerhards-Padilla, and P. Martini, "A survey on mobility models for performance analysis in tactical mobile networks," *JTIT*, pp. 54–61, 2008.
- [55] M. Musolesi and C. Mascolo, "Mobility models for systems evaluation," in *Middleware for Network Eccentric and Mobile Applications*. Springer, 2009, pp. 43–62.
- [56] B. Gonçalves and N. Perra, *Social phenomena: From data analysis to models*. Springer, 2015.
- [57] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven human mobility modeling: A survey and engineering guidance for mobile networking," *ACM Comput. Surv.*, vol. 48, no. 3, p. 38, 2016.
- [58] "United nations world urbanization prospects: The 2011 revision," [http://www.un.org/en/development/desa/population/publications/pdf/urbanization/WUP2011\\_Report.pdf](http://www.un.org/en/development/desa/population/publications/pdf/urbanization/WUP2011_Report.pdf).