Neurocomputing 403 (2020) 153–166

Contents lists available at ScienceDirect

Neurocomputing

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

A hierarchical temporal attention-based LSTM encoder-decoder model for individual mobility prediction

Fa Li[a,b,c,d], Zhipeng Gui[a,b,c,e,*], Zhaoyu Zhang[a,f], Dehua Peng[a,b,c], Siyu Tian[a,e], Kunxiaojia Yuan[b,d], Yunzeng Sun[a], Huayi Wu[b,c], Jianya Gong[a,c], Yichen Lei[a,g]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[c] Collaborative Innovation Center of Geospatial Technology, Wuhan, China
[d] Lawrence Berkeley National Laboratory, Berkeley, United States
[e] Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China
[f] School of Artificial Intelligence, Nanjing University, Nanjing, China
[g] Department of Urban Spatial Analytics, University of Pennsylvania, Philadelphia, United States

ARTICLE INFO

Article history:
Received 2 November 2018
Revised 22 March 2020
Accepted 24 March 2020
Available online 1 May 2020

Communicated by Prof. Zidong Wang

Keywords:
Human mobility
Mobility prediction
Temporal attention
Sequence prediction
Travel regularity
LSTM encoder-decoder model

ABSTRACT

Prediction of individual mobility is crucial in human mobility related applications. Whereas, existing research on individual mobility prediction mainly focuses on next location prediction and short-term dependencies between traveling locations. Long-term location sequence prediction is of great importance for long-time traffic planning and location advertising, and long-term dependencies exist as individual mobility regularity typically occurs daily and weekly. This paper proposes a novel hierarchical temporal attention-based LSTM encoder-decoder model for individual location sequence prediction. The proposed hierarchical attention mechanism captures both long-term and short-term dependencies underlying in individual longitudinal trajectories, and uncovers frequential and periodical mobility patterns in an interpretable manner by incorporating the calendar cycle of individual travel regularities into location prediction. More specifically, the hierarchical attention consists of local temporal attention to identify highly related locations in each day, and global temporal attention to discern important travel regularities over a week. Experiments on individual trajectory datasets with varying degree of traveling uncertainty demonstrate that our method outperforms four baseline methods on three evaluation metrics. In addition, we explore the interpretability of the proposed model in understanding individual daily, and weekly mobility patterns by visualizing the temporal attention weights and frequent traveling patterns associated with locations.

1. Introduction

Understanding and prediction of human mobility is significant in urban planning [15], ubiquitous computing [34], contextual advertisement [1], as well as intelligent transportation systems [9,10]. With the advancement of data collection technology, abundance of emerging trajectory data has been recorded, and supports quantitative analysis and prediction of human mobility [42]. For example, GPS data, mobile phone data, and transit smart card data record where people go. Social media data, credit card data, and mobile online payment data (e.g., Alipay) not only record locations of where people go, but also what people do at these locations [8]. Based on these data, there is a rising demand for individual mobility prediction as such prediction is a critical enabler for various human mobility related applications, such as intelligent urban transportation systems [45], and location based advertising [32]. In studies of individual mobility prediction, location prediction is one of the most notable branches. Numerous insightful works have been done on this area, however, the problem of predicting individual locations remains challenging.

- Long-term and short-term dependencies commonly exist in individual mobility patterns. Individual mobility regularity typically occurs daily and weekly, and capturing dependencies between different temporal locations in a longitudinal travel sequence is essential for prediction [31,37].
- The dependencies between a location and its context may change over time. A travel event often does not occur in isola-

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.
E-mail address: zhipeng.gui@whu.edu.cn (Z. Gui).

https://doi.org/10.1016/j.neucom.2020.03.080
0925-2312/

tion, and should be considered as a part of context with multiple travel events. Regular travel events orderly occur over time, and often periodically and frequently repeated with their surrounding contexts [10,16,20].
- External factors, such as weather, emotion, and the interaction with other individuals, may exert influences on travel decision-making of an individual. The model needs to be evaluated on different datasets with varying degrees of traveling uncertainty to show its effectiveness [16,37,45].

Currently, the majority of research focus on next location prediction, while the most commonly used next location prediction models discard long-term dependencies on the past mobility patterns [23]. Markov Chain and its variants are often used in current location prediction tasks [12,37,45]. However, they are limited to look back in time because of their inherent assumptions that the current state only depends on the states of previously limited time steps [23]. Meanwhile, next location prediction may work well in instantaneous applications, while long-term prediction is also needed to achieve a better long-term planning. Long Short-Term Memory (LSTM) have been used for individual mobility prediction, however it may fail to capture long-term dependencies when the length of input sequence increases [33]. In addition, LSTM cannot uncover the mobility regularities hidden in the black-box framework, which is useful for understanding of travel behaviors, travel preference analytics, and targeted demand management [10,16]. A solution to capture short-term as well as long-term dependencies by paying more attention to regular travel patterns underlying in the historical location sequences dynamically, may largely improve the accuracy and interpretability of individual location prediction.

To handle aforementioned challenges and fill the research gap, we propose a hierarchical temporal attention-based LSTM encoder-decoder model which is capable to predict day-long, and week-long trajectories where an individual will go. The advantages of our method are verified through the comparison with four baseline methods on real-world datasets with varying degrees of traveling uncertainty. Meanwhile, the interpretability of the model is revealed by visualizing the hierarchical temporal attention mechanisms.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes the temporal attention-based individual location sequence prediction method. Experiments in Section 4 demonstrate the advantages of proposed method. Section 5 further analyzes and discusses the effectiveness, interpretability, and limitation of the proposed method. Section 6 concludes this article and points out future research.

2. Literature review

In studies of location prediction, individual mobility is represented as a series of time-stamped locations, and the prediction problem is commonly framed as that of predicting an individual's next location [12,37,45]. To solve the problem of next location prediction, a plethora of methods have been proposed. Most used methods are based on Markov Chain (MC) by modeling sequential patterns of individual location histories. These methods predict individual mobility by applying each MC model to each individual person, and have demonstrated the ability to achieve high prediction performance [21,28]. However, individually fitted MC models are prone to overfitting, and unable to predict locations that users have never visited before. To address these issues, individuals with similar mobility characteristics are firstly clustered before applying a MC-based method [4,6]. In addition to MC-based methods, Bayes network models [3], n-gram model [45], as well as artificial neural networks models [12], also have been applied to next location prediction. Previous research of location prediction mainly focusses on the individual next location prediction problem instead of prediction of a whole location sequence that consists of multiple ordered locations within a long period. Short term prediction (e.g., next location prediction) may perform well for near real-time applications with known of previous locations. However, for applications that need to beforehand know where an individual is going during a relative long period, long-term prediction may be required. To simultaneously achieve short-term and long-term prediction, we treat location prediction as a sequence prediction problem by using the historical location sequence to predict the future location sequence in a certain period.

Location sequence prediction is more challenging than next location prediction. To predict a location sequence, methods of individual's next location prediction need to iteratively generate the next location by regarding the newly predicted location as the previous known location. This requires a higher accuracy in each location prediction to alleviate the problem of error propagation caused by prediction error of previous locations [30]. However, the mainly used MC model as well as its variants in next location prediction may be not competent because of its limitation in capturing long-time dependencies [23]. The mostly used one-order MC assumes that the next status only depends on its previous status, meanwhile, existing research also shows that higher order MC suffers complex computation issues, and cannot significantly improve the prediction accuracy [23,45]. However, individual travel behaviors often demonstrate daily, weekly, or specific repetition regularities during a long period [16]. For example, a person will go to the cinema every Friday night. To predict this location, the long-term travel regularity plays a decisive role instead of its previous location. A sequence model that can capture long-term and short-term dependencies as well as individual travel regularities (e.g., daily and weekly) is highly-desired.

Neural network sequence models provide a promising path for individual mobility prediction. Long Short-Term Memory (LSTM) is a notable variant of recurrent neural network (RNN) that has been widely used in many applications of sequence data [17]. Unlike MC-based models, LSTM has the advantage of having a continuous space memory which theoretically allows it to use arbitrarily length of past observations for sequence prediction. Except for the basic LSTM, the LSTM based encoder-decoder model also has shown excellent performance for Seq2Seq tasks, like machine translation [29], vehicle trajectory prediction [30], as well as time series prediction [26,33]. It uses one LSTM as an encoder to process the input sequence, and another LSTM as a decoder to generate the output sequence. Nevertheless, one problem with encoder-decoder network is that their performance will deteriorate rapidly as the length of input sequence increases [33]. This imperfection limits the sequence length in location prediction when we expect to make predictions based upon a relatively long input series. In addition, such a black-box framework cannot intuitively tell us the frequential and periodic individual mobility patterns captured by the model, while these patterns are useful for travel preference inference and recommendation [16]. Temporal attention-based encoder-decoder network resolves this issue by employing an attention-mechanism to select parts of hidden states across all time steps of encoder, and shows the importance of each time step in a sequence [26,33,43]. However, research that specifically designs, or applies the LSTM encoder-decoder model with the attention-mechanism is insufficient in individual mobility prediction. For studies of human mobility using deep learning methods, Alahi et al. [2] proposed an LSTM model, and predicted the trajectories of pedestrians to avoid collisions in autonomous navigation. Krishna et al. [23] developed two LSTM-based models to forecast human activity sequence (viz. eating, commuting, etc.) with associated durations. Deep learning is also utilized to reveal the

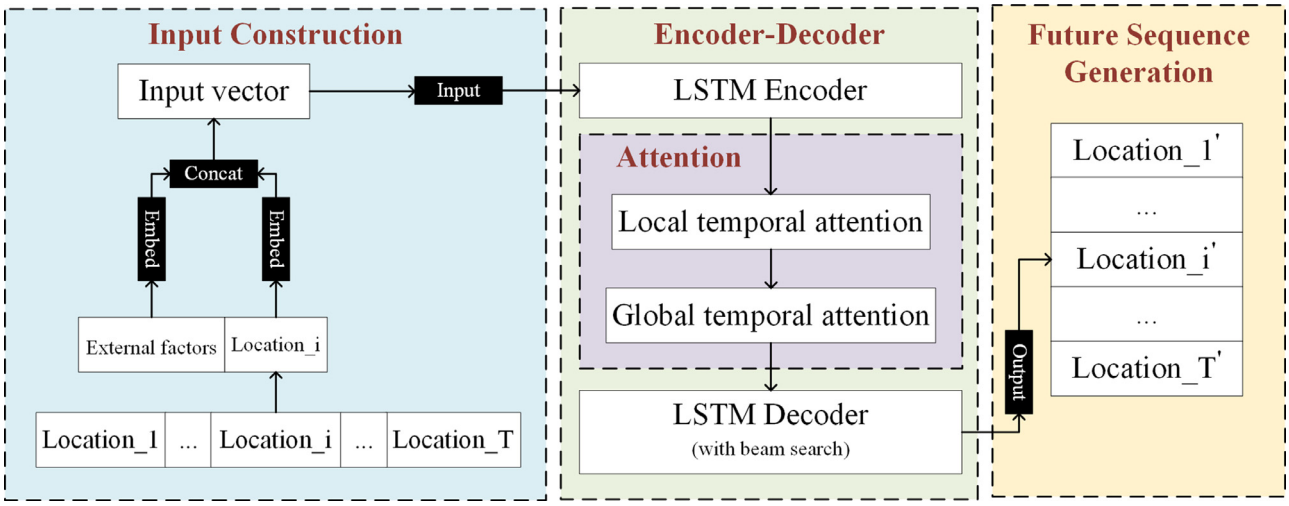


Fig. 1. Proposed individual location sequence prediction framework.

relationship between human mobility and personality information [22]. Further study is needed to increase the performance as well as model interpretability of individual mobility prediction.

In this paper, we propose a novel hierarchical temporal attention-based LSTM encoder-decoder model for individual location sequence prediction. To achieve short-term and long-term location prediction, the next location prediction problem is treated as a sequence-to-sequence (Seq2Seq) problem, and a LSTM encoder-decoder framework [38] with a beam search algorithm is designed for predicting location sequence of where an individual is going. To capture dynamical dependencies between traveling locations and uncover frequential and periodical mobility patterns hidden in the black-box of deep learning models in an interpretable manner, we integrate the calendar cycles of individual mobility patterns [16] into our model architecture and develop a hierarchical temporal attention mechanism, consisting of local and global temporal attention. During each location prediction, local temporal attention adaptively extracts related sub-location-sequence within a day, while global temporal attention captures travel regularities across a week. Experiments demonstrate that the proposed method outperforms four baseline methods by a substantial margin. We also visually analyze the hierarchical attention mechanisms to explore the interpretability of our method in uncovering individual underlying daily and weekly mobility regularities.

3. Methods

In this paper, the individual location prediction is treated as a Seq2Seq problem. Given a sequence of time-stamped locations denoted as $X=(x_1, x_2 \dots x_T)$ where an individual orderly visited during a time period, the sequence prediction model aims to learn a nonlinear mapping to the location sequence $Y=(y_{1'}, y_{2'} \dots y_{T'})$ during next time period that the individual will orderly visited, where x_i $(1 \le i \le T)$ represents the i-th visited location of totally visited T locations, while $y_{i'}$ is the i'-th $(1' \le i' \le T')$ location that the individual will visit in order. The problem is formulated as $Y = (y_{1'}, y_{2'} \dots y_{T'}) = F(X) = F(x_1, x_2 \dots x_T)$, where $F(*)$ is the nonlinear function the model aims to learn.

The overall architecture of individual location sequence prediction is demonstrated in Fig. 1. As shown in Fig. 1, three parts are included: input, temporal attention-based encoder-decoder LSTM model, and output. During input construction, locations and external affect factors (e.g., time stamp) that have influence on location sequence prediction are embedded, and concatenated into vectors as the input of encoder-decoder model. In the encoder-decoder architecture, we employ two separate LSTMs, one to encode the input-sequences during last period, and another one with a beam search method to predict the top K probable output location-sequences during next period. More specifically, our framework is composed of hierarchical temporal attention mechanisms, namely local and global temporal attentions, to respectively capture regular mobility patterns of an individual within a locally short period as well as a long period. After training of this model, individual location sequence in next period can be iteratively predicted. Details of the framework are described in the following sub-sections.

3.1. Input construction

The input of LSTM encoder-decoder model is a vector, concatenated by two parts, location and other attributes that affect location prediction. To represent each location, we use the occupancy grid map (OGM) that has been widely used in robotics and location prediction for the object localization [10,30]. The OGM divides the study area into equal-sized grids and each grid has a unique ID to identify which grid an individual is in. We linearly assigned the grid IDs, which range from one to the number of grids. A location sequence therefore is represented as a string of grid IDs. However, the grid IDs do not represent the spatiotemporal dependencies between grids and cannot be fed to neural networks directly due to its data type. Therefore, to capture dependencies between grids and make grid ID readily used for machine learning, we transform each grid ID into a finite-dimensional real-valued vector using the embedding method, which is capable to embed the enumerative values representing similar patterns into the close locations in embedding space [13,41]. Specially, the embedding method maps each categorial value v∈[V] to a real space $R^{E \times 1}$ by multiplying a parameter matrix $W \in R^{E \times V}$. For location sequence prediction, V represents the number of locations, and E represents the dimension of the real space. Parameters of W are obtainable by training the whole prediction model (described in Section 3.4). Through embedding, each location is represented as a vector with E dimensions. We use P_i to denote location i $(P_i \in R^{E \times 1})$.

Besides location itself, time periodicity also matters in location sequence prediction. The day and the week are most conventional calendar cycles that regular travel events repeat [16]. To capture such daily and weekly periodicity as well as avoid data sparseness problem, we organize location sequences by day. Locations where an individual orderly visited in each day is represented as a lo-

cation sequence, and we input location sequences of last week to predict the location sequences of next week. In addition to location information, time stamp information, such as day of week, is also helpful for human mobility inferring [27] as mobility patterns on different days (e.g., weekday and weekends) may be different. We apparently embed, and concatenate the day of the week as a part of the input. We use day-ID, an enumeration value ranging from one to seven, to denote the day of a week. Similarly, the embedding method [13] is used to transform each day-ID to a vector. We use D_d to denote the vector of day-ID d ($D_d \in R^{M \times 1}$), where M is the dimension of the day-ID vector and $1 \le d \le 7$. As shown in formula (1), two day-IDs are used, one to denote which day the location belongs to and another one to denote which day to be predicted in next week. After embedding the location and day-IDs, we concatenate the embedded vectors of each location and each day-ID using (1). As formulated in (1), the concatenation x_i is the final representing vector of location i considering time periodicity.

$$x_i = \left[P_i; D_k(lastweek); D_j(nextweek)\right] \tag{1}$$

where $x_i \in R^{Q \times 1}$, $Q = E + 2M$, and $1 \le k,j \le 7$.

Through such construction, the input of LSTM encoder-decoder is a sequence of vectors $(x_1, x_2 \dots x_T)$, that represents the location-sequence of last week. The goal of the model is to learn regular travel patterns from the input location sequence, and predict the location-sequence of next week, noted as $(y_{1'}, y_{2'} \dots y_{T'})$ where $y_{t'}$ is the label of a predefined location at time step t' ($y_{t'} \in [V]$).

3.2. LSTM encoder-decoder and beam search

LSTM is a variant of RNN that overcomes the vanishing gradient issue of RNN by introducing gating mechanism [17]. The LSTM consists of hidden state and cell memory, that respectively stores the summary of the past input sequence, and controls the information flow between the input and output through gating mechanism [17]. The following recursive equations demonstrate how the LSTM works.

$$\begin{aligned} f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\ i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{2}$$

where f_t, i_t, and o_t are gating vectors, that respectively control how much information for the cell memory to forget, update, and output. c_t and h_t respectively are cell memory state vector and hidden state vector (c_t and $h_t \in R^{n \times 1}$). In these equations, $\sigma = \frac{1}{1+e^{-x}}$ is the sigmoid function (element-wise), $\odot$ is element wise product, and x_t is the input vector. W_{xf}, W_{hf}, W_{xi}, W_{hi}, W_{xo}, W_{ho}, W_{xc}, and W_{hc} are linear transformation matrices whose parameters need to be learned, while b_f, b_i, b_o, and b_c are corresponding bias vectors. We simplify the LSTM representations in (2) as the Eq. (3) shows.

$$h_t,\ c_t = LSTM(x_t, h_{t-1},\ c_{t-1}) \tag{3}$$

The LSTM encoder-decoder architecture is based on LSTM, and now has been applied as the state-of-the-art sequence prediction architecture. As shown in Fig. 2, two LSTM networks, called encoder and decoder, respectively reads and generates variant-length sequences. The encoder recursively inputs the sequence $x_1, \dots, x_T$ of length T and updates the cell memory state vector c_t and hidden state vector h_t at each time step t through Eq. (3). After T time steps, the encoder summarizes the whole input sequence into the final vectors c_T and h_T.

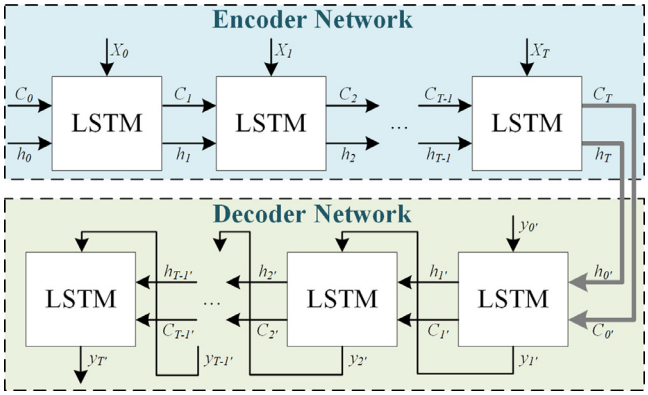


Fig. 2. LSTM encoder-decoder framework.

$$\begin{aligned} Y_{t'} &= W_l h_{t'},\ \text{where } W_l \in R^{V \times n},\ Y_{t'} \in R^{V \times 1}, Y_{t'} = [y_1 \dots y_V]^T \\ y_{t'} &= \frac{e^{y_{t'}}}{\sum_{j=1}^{V} e^{y_j}} \\ y_{t'} &= \max(y_{t'}) \end{aligned} \tag{4}$$

The decoder uses c_T and h_T passed from the encoder as its initial cell memory state vector ($c_{0'} = c_T$) and initial hidden state vector ($h_{0'} = h_T$) for T'-length sequence generation. During sequence generation, the decoder firstly uses a dummy input y_0, and the initial vectors $c_{0'}$ and $h_{0'}$, to obtain $c_{1'}$ and $h_{1'}$ through Eq. (3). Equations in (4) are subsequently used to compute $y_{1'}$ that represents the location where an individual will go. For simplification, we use formula (5) to represent equations in (4). Similarly, by feeding $c_{t-1'}$, $h_{t-1'}$, and $y_{t-1'}$ to Eqs. (3) and (5), the output sequence $y_{1'}, \dots, y_{T'}$ are recursively generated. Note that, the LSTM decoder produces $y_{t'}$ for given $y_{t-1'}$. If $y_{t-1'}$ is wrongly estimated, estimation of subsequent values may be affected.

$$y_{t'} = f(h_{t'}) \tag{5}$$

To alleviate error propagation, beam search algorithm is introduced into decoder process. As formulated in (4), the way to determine $y_{t'}$ is the greedy search strategy that simply picks the value $y_{t'}$ that maximizes the conditional probability $p(y_{t'}|y_{t-1'},\ c_{t-1'}, h_{t-1'})$. Unfortunately, such greedy strategy suffers from the error propagation since wrong decision made at the current time step would be propagated to the subsequent time steps. The basic idea of the beam search is to choose K most probable hypotheses according to $p(y_{t'}|y_{t-1'},\ c_{t-1'}, h_{t-1'})$ at each iteration [29]. After T' iterations, the decoder generates K T'-length sequences as the most K probable result sequences. Note that the beam search with $K = 1$ degenerate into the greedy search. Details of beam search algorithm can be referred in [29].

We apply aforementioned LSTM based encoder-decoder framework for individual location sequence prediction. LSTM of encoder iteratively processes each location vector x_t of the input sequence constructed in Section 3.1 as a hidden vector h_t through the Eq. (3); LSTM of decoder with a beam search algorithm forecasts the most probable K location sequences where an individual will go in next week using the Eq. (5). In spite of beam search algorithm, the performance deterioration problem of this framework exists especially when the length of input sequence increases [43]. As mentioned above, the encoder treats the final cell memory state vector c_T and hidden state vector h_T as the summarization and output of the input sequence, and passes these two vectors to the decoder as the initial input of decoder. This strategy may lead to information loss of input sequence as all information is summarized to the final cell memory state and hidden state at time step T [26,33]. To resolve this issue, we employ an attention-based

encoder-decoder network with the temporal attention mechanism to select parts of hidden states that are highly related to target location across all time steps of encoder, instead of the final states.

3.3. Temporal attention mechanisms

3.3.1. Temporal attention

The temporal attention mechanism extracts regular travel behaviors by adaptively paying more attention to relevant hidden states of the encoder during future sequence generation [26,33]. The temporal attention mechanism is essentially the weighted sum of sequence $\{h_t,\ 1 \le t \le T\}$ as shown in Eq. (6). The weighted sum $H_{t'}$ is used to predict the location $y_{t+1'}$ at time $t+1'$ through decoder. For the prediction of location $y_{t+1'}$, some locations may be highly correlated. For example, an individual will regularly visit the location $y_{t+1'}$ after orderly visiting some other specific locations. The temporal attention mechanism adaptively assigns greater weights to locations with higher correlations for the target location prediction. Travel regularities underlying in location sequences therefore, can be quantitatively analyzed by comparing the values of these weights, making the encoder-decoder model more interpretable.

$$H_{t'} = \sum_{t=1}^{T} u_{t'}^{t} \cdot h_t \tag{6}$$

Where $H_{t'}$ represents the summarization of encoder, and is used for prediction of $y_{t+1'}$ in decoder, $u_{t'}^{t}$ is the weight of the t-th hidden state vector of encoder, computed as defined in Eq. (7).

$$u_{t'}^{t} = V_a^T \tanh\left(W_a'[h_{t-1'};\ c_{t-1'}] + W_a h_t + b_a\right)$$

$$u_{t'}^{t} = \frac{e^{u_{t'}^{t}}}{\sum_{j=1}^{T} e^{u_{t'}^{j}}} \tag{7}$$

Where $W_a \in R^{m \times n}$, $W_a' \in R^{m \times 2n}$, V_a and $b_a \in R^{m \times 1}$ are parameter matrixes that need to be learned through model training process.

In addition to the input locations of encoder, the previous location of decoder may also exert affects in current location prediction. The weighted sum $H_{t'}$ represents the summarization of the input location sequence, and $y_{t'}$ is the previous location of $y_{t+1'}$. We obtain the combination of $H_{t'}$ and $y_{t'}$ using (8), and treat $\widehat{y_{t'}}$ as the input of decoder for $y_{t+1'}$ prediction following equations in (9).

$$\widehat{y_{t'}} = \widehat{W_c}[y_{t'}; H_{t'}] \tag{8}$$

$$h_{t'} = LSTM(\widehat{y_{t'}},\ h_{t-1'},\ c_{t-1'})$$
$$y_{t+1'} = f(h_{t'}) \tag{9}$$

where $W_c \in R^{Q \times (Q+n)}$.

Integrated with the temporal attention mechanism, LSTM encoder-decoder model more efficiently captures regular travel behaviors. Travel regularity indicates the degree to which subsequences of travel events are repeated [16]. For a location sequence prediction, different locations may be correlated to different sub-sequences. The temporal attention mechanism adaptively assigns greater weights to those higher-related sub-sequences during each location prediction. However, locations of regular subsequences may be along with irregular locations. For example, a person regularly goes to work place from his home, and goes home from his work place during every workday. Along with these regular travel sub-sequences, this person may occasionally go to some places that do not belong to regular travel behaviors (e.g., occasionally go to the cinema). Irregular locations contained in sequences may affect the performance of the model. Especially when the length of sequence increases, the temporal attention may be distracted by the increased number of irregular locations [11,40]. Inspired by theories of human attention that behavioral results are best modeled by two-stage attention mechanism, as well as some dual-stage attention-based research for time series prediction [26,33], we propose a novel hierarchical temporal attention mechanism for individual location sequence prediction.

3.3.2. Hierarchical temporal attention

The hierarchical temporal attention networks for location sequence prediction are demonstrated in Fig. 3. Different from aforementioned one-layer temporal attention, the hierarchical temporal attention consists of two layers, local temporal attention and global temporal attention. Local temporal attention measures the relative importance of different locations within each day. Global temporal attention pays attention to the relative importance of different days within a week. Instead of directly assigning weights to T-length locations of a week in one-layer temporal attention, hierarchical temporal attention computes the weights in two stages.

Local temporal attention: Given the location sequence of the dth day $(x_{1_d},\ x_{2_d} \ldots\ x_{L_d})$, the local temporal attention obtains the weighted summation vector of this day. Not all locations of a day equally contribute to the prediction of the target location. Hence, we introduce attention mechanism to extract locations that are highly correlated to the target location, and aggregate representations of those related locations to form a summarization vector that represents regularity in this day. Different from one-layer temporal attention, the local temporal attention is weighted sum of the sequence $\{h_{t_i}, 1 \le t_i \le L_d\}$ of each day, where L_d is the number of locations in the dth day, $1 \le d \le 7$, and h_{t_i} is the corresponding hidden state vector of location x_{t_i}. As shown in formula (10), $H_{t_d'}$ represents the regularity vector of the dth day. The corresponding weights are computed in Eq. (10).

$$H_{t_d'} = \sum_{t_i=1}^{L_d} u_{t_d'}^{t_i} \cdot h_{t_i}$$

$$u_{t_d'}^{t_i} = V_a^T \tanh\left(W_a'[h_{t-1'};\ c_{t-1'}] + W_a h_{t_i} + b_a\right)$$

$$u_{t_d'}^{t_i} = \frac{e^{u_{t_d'}^{t_i}}}{\sum_{t_i=1}^{L_i} e^{u_{t_d'}^{t_i}}} \tag{10}$$

Local temporal attention obtains the regularity vector sequence $\{H_{t_1'},\ H_{t_2'} \ldots\ H_{t_7'}\}$ of a week. To predict the target location of $y_{t+1'}$ on dth day of next week $(1 \le d \le 7)$, not all days equally exert influences. For daily and weekly periodic travel regularities, the specific days of last week may be more correlated to the corresponding days in next week. For example, travel patterns of weekday during last week may be more similar to weekday patterns of next week, instead of patterns of weekends. To reward days that are clues to correctly predict the target location in next week, we proposed the global temporal attention, a week-level weight assignment mechanism.

Global temporal attention: The global temporal attention obtains the weighted summation vector of a week. As shown in (11), $\widehat{H_{t'}}$ is the week-level vector that summarizes all correlated information of last week for target location prediction. The corresponding weights are computed using equations in (11). Similar to one-layer temporal attention, we combine $\widehat{H_{t'}}$ and $y_{t'}$ using (12), and use the combined value $\widehat{y_{t'}}$ for $y_{t+1'}$ prediction through (9).

$$\widehat{H_{t'}} = \sum_{d=1}^{N_d} u_{t_d'}^{d} \cdot H_{t_d'}$$

$$u_{t_d'}^{d} = V_a^T \tanh\left(W_a'[h_{t-1'};\ c_{t-1'}] + W_a H_{t_d'} + b_a\right) \tag{11}$$

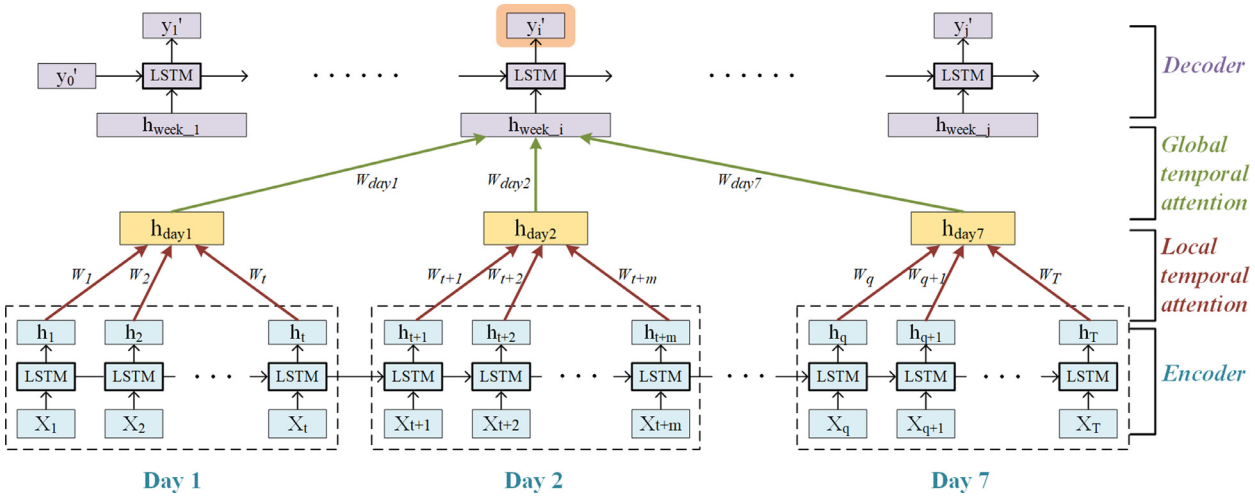


Fig. 3. Hierarchical temporal attention networks for location sequence prediction.

$$u^d_{t_d'} = \frac{e^{u^d_{t_d'}}}{\sum_{d=1}^{N_d} e^{u^d_{t_d'}}}$$
$$\widehat{y_{t'}} = \widehat{W_c}\left[y_{t'}; \widehat{H_{t'}}\right] \quad (12)$$

where N_d is the number of days used for global attention computation. In our case, N_d equals to 7.

3.4. Model training

The parameters of the temporal attention-based LSTM encoder-decoder (including the embedding matrices) are trained referring to a multi-class classification problem [30]. Since each location is denoted as a grid ID, prediction of each location is essentially to classify which grid the location belongs to. The grid containing the target location should be selected among all V grids. We use the one hot vector $O_i = [0, \ldots 0,\ 1, 0, \ldots 0]$ to indicate the target location at time step i. The index of the entry in O_i that equals to one is the true-value grid ID of this location. For location sequence prediction, the normalized vector Y_i computed through (4) indicates the probability distribution of each grid. Following studies of classification as well as previous trajectory prediction study [30], we use cross entropy as the loss function (13) to train the model via back-propagation algorithm [36].

$$\mathrm{L}(\Omega) = -\sum_{i=1}^{T'}\sum_{j=1}^{V} o_{i,j} ln y_{i,j} \quad (13)$$

where Ω denotes the set of parameters in our model. T' is the total number of locations in output sequence, and V is the total number of different locations that an individual tends to visit. $o_{i,j}$ is the jth element of O_i, and $y_{i,j}$ is the jth element of Y_i associated with the label $o_{i,j}$.

4. Experiments

In this section, we present the performance of the proposed method based on long-observed individual's GPS trajectory data. Four baseline methods are selected and three indicators are designed to evaluate the effectiveness of our method.

4.1. Data collection and preprocessing

A dataset consisting of individual GPS trajectories of private cars is used for experiments. The dataset totally records 37,854 trajectories of 49 individuals from March, 2017 to October, 2018 through GPS equipment loaded in vehicles. Each trajectory is represented by a sequence of time-stamped points, recording where an individual goes from and to. Each point contains the information of latitude, longitude, time, and individual ID which is a unique string for identifying the person who generates the trajectory. The Origin and Destination (OD) of each trajectory is extracted, and all OD pairs within a day are chronologically organized as the individual traveled location sequence of the day. Note that each location is primitively represented as a coordinate with continuous values, the OGM in Section 3.1 is used to merge points that are within the same grid. Considering that individuals may not select driving as a traffic mode when the travel distance is less than 1 km in China [18,44], we set the grid size as 1 km × 1km. Hereto, the visited location sequence of an individual is represented by a sequence of grid-IDs. In following studies, a location is equivalent to a grid.

To make a comprehensive experiment on individuals with varying degrees of traveling uncertainty, the entropy of stay points of each individual in the dataset is calculated. In information theory, entropy measures the level of randomness or unpredictability of a process. In human mobility studies, entropy has been used to measure the variation or regularities of individual travel behaviors [16,45]. Higher entropy indicates higher uncertainty in individual travel patterns, and is generally more difficult to predict, vice versa [37]. We divide the individuals with distinct levels of traveling uncertainty into groups according to the entropy of stay points. The individuals with continuous long-term recordings and maximum entropy of stay points in each group are selected as the representatives of the groups. Finally, the representatives in three groups are selected for experiments. Statistical results of the three datasets are depicted in Table 1.

As shown in Table 1, the entropy of stay points increases from Data_1 to Data_3, showing an increasing traveling uncertainty on the three datasets. The trajectories, OD interactions, and activity hotspots of three exemplary individuals in three datasets are visualized in Fig. 4. In the figure, travel activities in Data_1 are more densely distributed in fewer hotspots compared to the other two datasets, indicating a lower traveling uncertainty with higher mobility regularities. Frequential travel activities in Data_2 and Data_3 are dispersed in more hotspots, and the travel interactions between hotspots are more diverse than that in Data_1, implying a higher uncertainty and lower predictability of Data_2 and Data_3. In experiments, we equally and randomly partition the data into non-overlapped ten subsamples for each dataset. A subsample is randomly selected as a test set, and the other nine subsamples are selected as the training set.

Table 1
Details of selected three datasets.

Datasets	Number of different locations per year	Number of location sequences per year	Average length of location sequences per week	Yearly entropy of stay points	Average entropy of stay points per week
Data_1	45	1635	55	3.2	2.6
Data_2	82	694	95	3.8	3.2
Data_3	317	1138	60	5.8	3.5

Table 2
The model performance over five-times repeating experiments, including best performance, mean, and variance.

	Methods	Data_1			Data_2			Data_3		
		MRE	MA	MR	MRE	MA	MR	MRE	MA	MR
Best result	MC	77.3%	23.0%	57.2%	70.1%	34.6%	57.2%	79.7%	21.3%	45.5%
	LSTM	18.0%	87.3%	80.3%	29.7%	79.4%	80.1%	26.6%	80.7%	82.2%
	ED	12.7%	92.6%	90.8%	29.7%	80.4%	80.1%	17.6%	88.0%	87.3%
	TAED	7.2%	96.2%	94.9%	21.3%	86.1%	85.1%	14.4%	89.9%	89.0%
	HTAED	**4.1%**	**97.5%**	**97.0%**	**9.8%**	**92.4%**	**93.9%**	**9.0%**	**93.3%**	**94.4%**
Mean	MC	78.0%	22.2%	54.3%	72.9%	29.8%	52.3%	84.6%	16.1%	41.5%
	LSTM	19.3%	84.6%	80.3%	30.3%	78.5%	79.4%	28.2%	78.4%	80.4%
	ED	13.7%	91.1%	88.9%	30.3%	79.2%	79.4%	20.7%	85.6%	84.6%
	TAED	7.5%	95.3%	94.2%	23.6%	83.9%	84.0%	16.9%	87.7%	87.1%
	HTAED	**5.9%**	**96.5%**	**95.6%**	**16.1%**	**88.3%**	**87.4%**	**14.3%**	**89.5%**	**89.4%**
Standard deviation	MC	0.005	**0.005**	0.018	0.032	0.045	0.050	0.029	0.031	0.027
	LSTM	0.014	0.033	**0.006**	**0.006**	**0.009**	**0.008**	**0.019**	0.028	**0.023**
	ED	0.010	0.020	0.032	0.010	0.013	**0.008**	0.040	0.030	0.034
	TAED	**0.004**	0.010	0.009	0.024	0.019	0.014	0.030	**0.021**	**0.023**
	HTAED	0.013	**0.009**	0.010	0.038	0.024	0.037	0.037	0.027	0.033

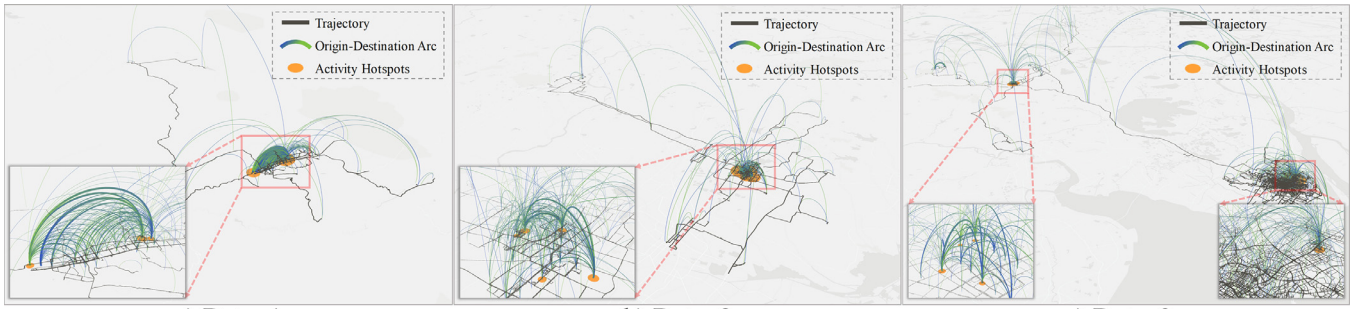


Fig. 4. Trajectories, Origin-Destination arcs and activity hotspots of the exemplary representatives in three datasets.

4.2. Settings

4.2.1. Evaluation metrics

Multiple criteria are used to evaluate our model, including mean relative error (MRE), mean accuracy (MA), and mean recall (MR). In the calculation of MRE, the error between the predicted location sequence and the target location sequence is measured using the Levenshtein distance (also called edit distance). Levenshtein distance has been widely used to measure the similarity between two sequences [35]. It is the minimum number of insertions, deletions, and substitutions required to transform one sequence into the other, therefore the orders of locations in the sequence are considered [16]. The formula of MRE is defined as follows:

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{Edit\left(Seq_{predicted\_i},\ Seq_{target\_i}\right)}{\max\left(length\left(Seq_{predicted\_i},\ Seq_{target\_i}\right)\right)} \tag{14}$$

Where n is the number of sequences, $Edit\left(Seq_{predicted\_i},\ Seq_{target\_i}\right)$ is the Levenshtein distance of ith predicted sequence compared with its corresponding target sequence. The denominator normalizes the error, making MRE value in [0,1] [24].

In addition to MRE, the other two indicators, namely MA and MR, that have been widely used in classification tasks, are used to evaluate the performance of the model regardless of the location orders in a sequence. The formula of MA and MR are shown in (15) and (16). In the formulas, $N_{correct\_i}$ is the number of correctly predicted locations in the ith predicted sequence. Note that when a location both exists in the predicted sequence as well as in the target sequence, and the times of appearance in predicted sequence is not more than that in target sequence, we will regard it as a correctly predicted location and one will be added to $N_{correct\_i}$.

$$\mathrm{MA} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_{correct\_i}}{length\left(Seq_{predicted\_i}\right)} \tag{15}$$

$$\mathrm{MR} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_{correct\_i}}{length\left(Seq_{target\_i}\right)} \tag{16}$$

4.2.2. Baselines

We compare our model with four baselines as follows:

MC: Many studies of individual location prediction are based on Markov Chain (MC), and the literature has shown that first-order MC, can approach the limit of predictability for next location prediction problem, and increasing the order does not necessarily improve the prediction performance [28]. We therefore use first-order MC as a benchmark method, and iteratively predict individual location sequence by treating the predicted location as a previous known location.

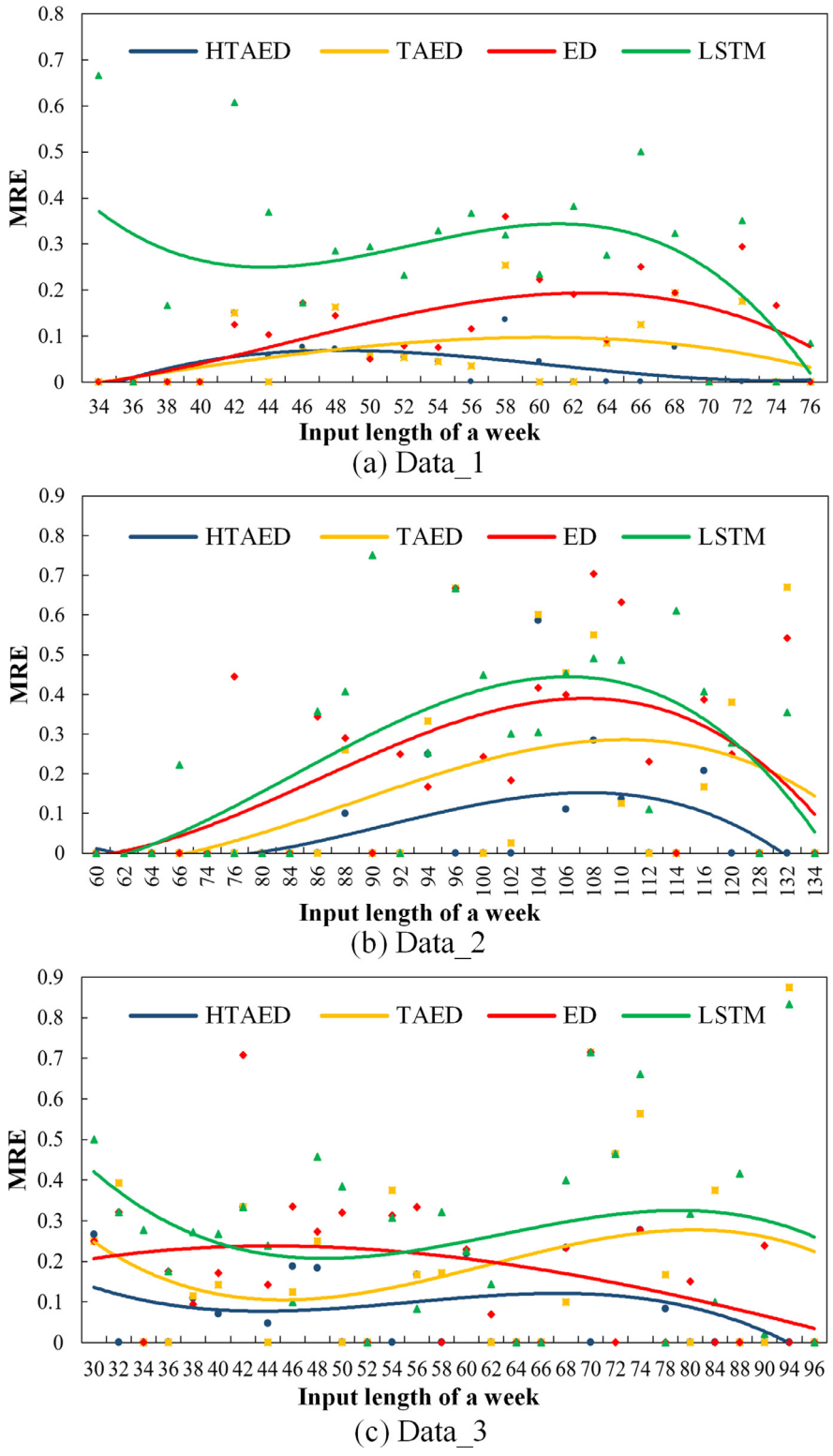


Fig. 5. Model performance comparison under different weekly input sequence lengths and the performance trends are plotted by fitting cubic polynomial curves.

LSTM: LSTM has demonstrated its advantage over MC in human activity sequence prediction and vehicle trajectory prediction [23].

LSTM **Encoder-Decoder (ED)**: It uses a LSTM to encode the input sequence and another LSTM to iteratively predict the future sequence, that has outperformed basic LSTM in trajectory prediction [30].

Temporal Attention based Encoder-Decoder (TAED): Temporal attention based encoder-decoder model has shown performance improvement than basic encoder-decoder model in time series prediction and document classification [33,43]. There is no existing temporal attention-based encoder-decoder model directly for individual trajectory prediction. We design and develop a temporal attention-based encoder-decoder following our sequence predic-

tion framework, to compare with the proposed hierarchical temporal attention-based encoder-decoder model (HTAED).

4.2.3. Parameter settings

In experiments, we set parameters of HTAED model as follows. In input construction step, we compared different embedding dimensionality for location and day over {64,128,256,512} and {5,10,15} respectively, and finally embedded location ID to R^{256}, and embedded day ID to R^{10}. During the training phase, Stochastic Gradient Descent with momentum is used as the optimizer, and we evaluated our model for different momentum values from 0 to 0.9. The learning rate was set 0.01 after testing 0.1, 0.01, and 0.001. For simplicity, we set the same dimensionality to hidden state vectors in encoder and decoder. The dimensionality was finally set to 256 as it outperformed over the other values. The K value in beam search algorithm was 3 because a larger value not significantly increased the performance of the model. We trained the model for 100 epochs, and repeated each experiment for five-times. Following the evaluation rule used in [41] and [26], we use our best model performance to compare the best performance of each baseline method under different parameter settings. For LSTM-related baselines, the settings, including dimensionality of location ID, day ID, and hidden state, the optimizer, learning rate, as well as training process, are the same to that of HTAED.

4.3. Model comparison

The comparison results between the proposed method and the four baselines on three datasets, including best performance, mean, and variance over five-times repeating experiments are listed in Table 2. As shown in Table 2, the best performance and mean performance of our method outperform all the baselines on three datasets and three metrics. The proposed algorithm shows significant improvements in accuracy both considering the orders of locations and regardless of the orders. This advantage is especially obvious on the second dataset, whose average sequence length of a week is larger than the other two datasets. The standard deviation of our model is larger than that of LSTM, ED, and TAED in general as our model involves more parameters, and introduce more uncertainty in parameter initialization process. However, the mean performance of our model is superior to the best performance of the other four methods except MA on Data_3 which is lower than TADE, showing its effectiveness on performance improvement. In summary, LSTM based methods achieve better performance than MC, demonstrating the ability of LSTM on capturing long time dependency. LSTM based encoder-decoder models are superior to LSTM due to the positive effects of the decoder component. In addition, the TAED as well as HTAED bring significant improvement in performance, that illustrates the power of temporal attention mechanism in Seq2Seq problems.

Table 2 shows that travel activities with higher entropy of stay points (Data_2 and Data_3) are overall more difficult to be predicted compared with lower entropy (Data_1), indicating that uncertainty is also a challenge for individual mobility prediction. Meanwhile, the results also show that entropy has its limitations in measuring travel uncertainty or predictability. The entropy of stay points of Data_2 is lower than that of Data_3, however, traveling activities in Data_2 are more difficult to be predicted than that of Data_3 according to results in Table 2. Entropy of stay points only measures the probability distribution of different locations, while the orders and the dependencies between different locations in the location sequences are not considered. The average length of location sequences in Data_2 are much larger than that of Data_3 as shown in Table 1, which may be one of the reasons why Data_2 is less predictable as it may contain more complex and long-distant dependencies in traveling activities.

5. Analysis and discussion

5.1. Model effectiveness for different length input and output sequences

As the length of input and output sequence may affect the model, we evaluate the performance of different models under variant-length input and output sequences. Because of the limit of MC on predicting the length of a location sequence, we removed it in this evaluation. The MRE of different methods are represented as points with different symbol styles and colors in Figs. 5 and 6, and the performance trends of different models are plotted by fitting polynomial curves.

As depicted in Fig. 5, the performances of HTAED under different length of input sequences are better than that of the other three models as it captures more meaningful dependencies by incorporating the daily-weekly mobility structure into the model architecture. In the figure, the performance of the four models shows similar trends as the input sequence length changes in general, indicating that the predictability of individual mobility patterns changes with input sequence length. However, the performance of these models does not show a significant decline as the input length increases over the three datasets, implying that the predictability is not simply controlled by input sequence length. In fact, entropy, dependency strength, and dependency distance between different locations may all exert influences on predictability [16]. For individual location sequence prediction, longer-term dependencies exist in longer length of sequences, but the longer length of a sequence does not necessarily contain longer-term dependencies. For example, if predicting the next location relies on its previous 100 locations, the length of the input sequence is at least 100. While if predicting the next location only relies on its dependency on its previous location, the 100 locations become unnecessary. To furtherly explore the dependencies or regularities contained in each sequence, more advanced metric considering frequential patterns as well as orders needs to be studied. Overall, the HTAED outperforms the other methods by a considerable margin, illustrating the power of the proposed hierarchical temporal attention mechanism in enhancing the long-term predictive performance.

Furthermore, we explore the performance of different models on prediction under different output sequence lengths. In the figure, the proposed method outperforms the other methods with lower MREs on prediction of different length location sequences over three datasets. Meanwhile, the performance of different models does not significantly deteriorate when predicting on more distant future over the three datasets. Longer-term prediction does not necessarily mean less predictability, which is actually determined by multiple factors, such as entropy and dependencies between different locations [16]. Different length of location sequences may represent different travel patterns. Travel patterns with a longer location sequence of a day consisting of more different locations may be different from travel patterns with few locations. Good performance on generating location sequences with variant lengths indicates the effectiveness of capturing different patterns. Therefore, the proposed hierarchical temporal attention mechanism improves the performance in forecasting different length of travel sequences against the baselines.

5.2. Visualization and analysis of local and global temporal attention weights

To further investigate our model, we visualize, and analyze the local and global temporal attention weights. Previous research has demonstrated the interpretability characteristics of temporal attention mechanisms [26,33]. For example, in document classification,

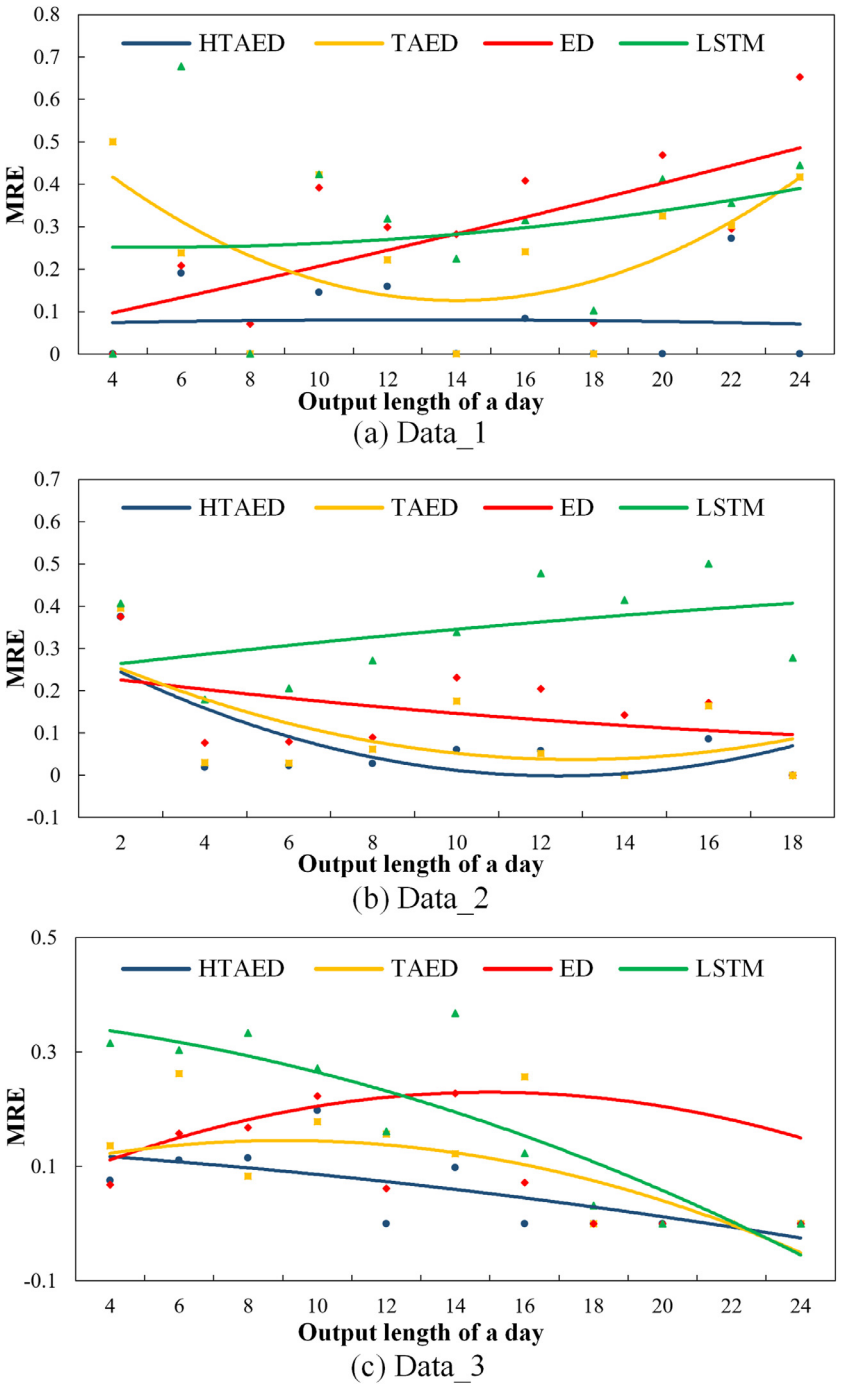


Fig. 6. Model performance comparison under different daily output sequence lengths and the performance trends are plotted by fitting quadratic polynomial curves.

a larger weight assigned to a word indicates that this word plays a more important semantic role in deciding the semantic category of a document [43]. This characteristic makes the model more interpretable than traditional black-box deep learning methods, such as basic LSTM and LSTM encoder-decoder models [5]. For individual location sequence prediction, a location with a larger weight indicates the higher importance of the location in future sequence prediction. So, we perform a case study and show how the proposed method captures travel regularities through our hierarchical temporal attention mechanisms.

As shown in Fig. A in Appendix, the hierarchical temporal attention weights of four predictions with different destination locations and days for an individual are visualized. Recall that the local and global attention weights represent the relative importance of different locations within a day and different days over a week respectively. Blue lattices in Fig. A denotes the location weights within a day, and pink lattices denotes the day weights of a week. The deeper the colors are, the greater the weights are, and different numbers marked in the grids represent different locations. In the figure, we can find that, when predicting a target location, the same location but with different orders in the input sequence, exerts different weights of influences, revealing that the order in which an individual completes trips and activities is an integral component of the structure in their travel routines [16]. We can also find that daily-scale, and weekly-scale long-term dependencies commonly exist in individual location sequence prediction.

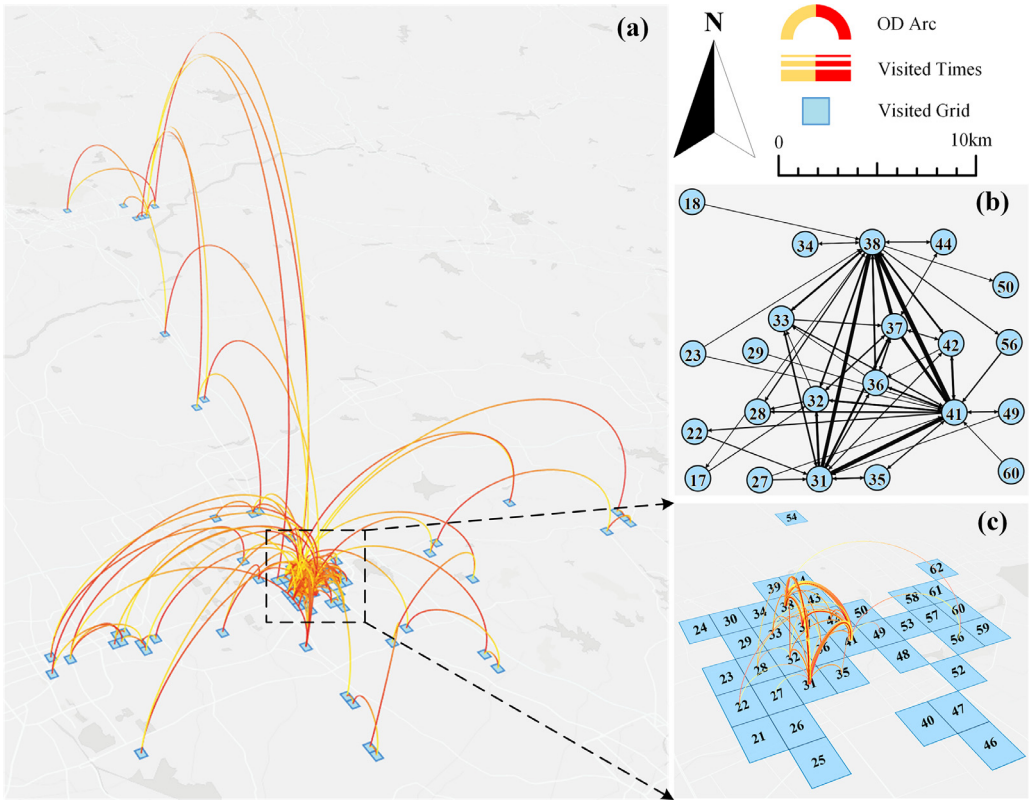
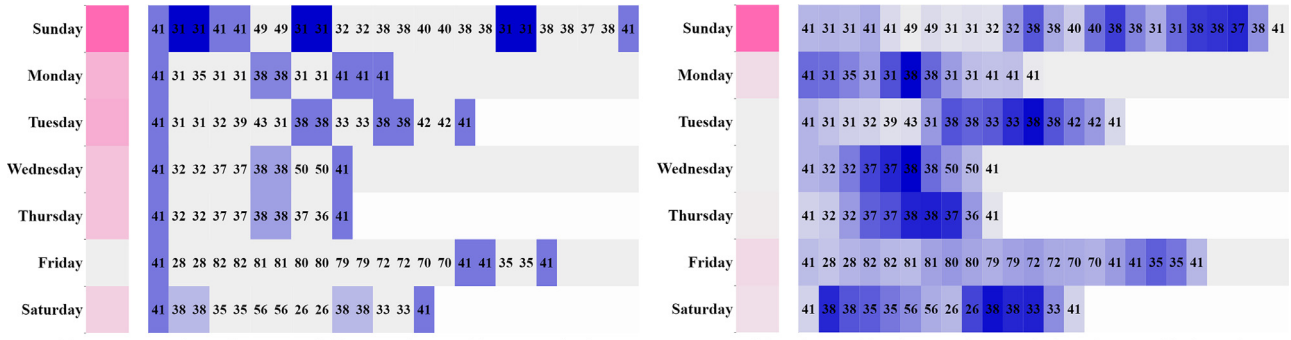


Fig. 7. Individual frequently visited locations. (a) global visited grids and Origin-Destination (OD) arcs, (b) network of the individual visited grids, and (c) Highlight of the most frequently visited grids and their interactions through OD arcs. The arcs in (b) and (c) are OD pairs whose frequency is greater than two.



(a) Frequent locations on different days of last week that associate with the No. 37 location on this Sunday

(b) Hierarchical attention weights for predicting the No. 37 location in next week

Fig. 8. Frequent traveling patterns and hierarchical attention weights associated with the target No.37 location on Sunday. Each row of the subfigures represents the location sequence where an individual orderly visited per day of last week in the form of Origin-Destination location ID pairs.

Through such an attention mechanism, we may more intuitively understand how a location matters in each day, and how each day affects the location prediction of an individual.

To further explore the travel patterns captured by the model, we take an individual as an example, and analyze the travel patterns associated with traveling locations in physical space. The traveling activities and locations that the individual frequently visited are shown in Fig. 7(a). The interactions and their frequency between locations in traveling are visualized in Fig. 7(b). The trips with frequency lower than two are removed to make the frequently visited locations and their associations represented more clearly. From Fig. 7(b), we can find that the locations, including No.31, No.38, and No.41, are the most correlated locations with the No. 37 location. The most frequently visited grids and their interaction are highlighted in Fig. 7(c). From Fig. 8, we can find that our model captures such overall and generalized mobility patterns. In Fig. 8(b), all these locations are highlighted by the hierarchical attention weights, while locations that are not highly relevant to the No.37 location are weakened.

In addition, we use FP-growth [19], one of the mostly used frequent-pattern-mining algorithm, to explore the ability of the proposed model for obtaining frequently time-periodic mobility patterns. For example, when the individual goes to the No. 41 location and No. 31 location on last Sunday, the individual will go to the No.37 location with a 90% probability in next Sunday. We count the number of frequent patterns that each location is associated with the target No.37 location on Sunday. The statistics of frequent patterns is shown in Fig. 8(a). In the figure, the value of pink color

indicates the total number of frequent patterns of each day, and the value of blue color indicates the number of frequent patterns involved by each location. We can find that most of frequent patterns occur on Sunday, and the corresponding global weight value in Fig. 8(b) is also the highest. Most of the blue-highlighted locations in Fig. 8(a) are also highlighted in Fig. 8(b). However, differences do exist in these two figures. The reason is that the latter one considers the order of each location, and captures varying dependencies within a week, while the former one not. Overall, the temporal attention mechanism helps us explore and understand complex travel regularities underlying in individual travel histories.

5.3. Importance and limitations of hierarchical temporal attention

The reason why hierarchical temporal attention matters is that it incorporates structural knowledge into individual location sequence prediction, and alleviates performance degradation for sequences with long-term dependencies. Our hierarchical temporal attention includes daily-level, and weekly-level attention networks. The two levels are the most conventional calendar cycles that regular travel events repeat [16]. We integrate the background knowledge of calendar cycle regularity into model structure, which actually tells the model more spatiotemporal boundary and hierarchy information [26,33,43]. As shown in Fig. 8, frequently periodic mobility patterns do exist in physical space, and our model is capable to highlight such structural mobility patterns. In addition to structure knowledge, the temporal attention also improves the performance of location sequence prediction with long-term dependencies. For individual mobility prediction, most used MC-based methods only rely on limited number of previous locations for the next location. However, daily and weekly long-term dependencies are an integral part of individual mobility regularities. Our model adaptively selects highly relevant hidden state of the encoder during decoding, improving the performance of encoder-decoder architecture for long-term sequence prediction. Dual-staged attention learns structured individual mobility patterns, and as a result, we may more intuitively understand how a traveling location matters in each day, and how traveling regularities occur over a week.

In spite of the advantages, our model is just in its infancy for individual mobility prediction in real world, and there are several limitations as well as remaining challenges. Firstly, the current model requires long-term trajectory records to learn traveling regularities of each individual. Data sparsity problem may be encountered for individuals with short historical trajectories. Secondly, we follow previously numerous studies in individual mobility [12,37,45], and also treat the model as a multi-class classification problem. Therefore, the targets are limited in individual historically visited places. The model needs to be extended if an individual visits a historically unvisited place. For example, an individual changes his or her job, and moves to another place. Thirdly, our model only involves limited external factors for individual mobility prediction. Other factors, such as weather [27], spatial factors (e.g., land use and road network) [25], specific events, and the interactions between individuals, may also affect individual traveling. For example, the outbreak of COVID-19 significantly changes individual mobility patterns [7]. Fourthly, our model only focuses on location sequence prediction. While, human activities along with time-stamped location sequences may be more generic and logical for human decision-making as well as prediction, such as lunching out at a restaurant at 12:30 pm., going back to the company for working at 1:30 pm. Finally, the proposed model is based on LSTM that inherently precludes parallelization because of its recurrent computation process. Advanced sequential models could be adopted to reduce computation complexity.

6. Conclusion and future work

We propose a novel hierarchical temporal attention-based LSTM encoder-decoder model for individual location sequence prediction. The hierarchical temporal attention networks consist of location temporal attention, and global temporal attention, to respectively capture travel regularities with daily, and weekly long-time dependencies. We evaluate our model on three levels of predictable datasets, and experiments show that the proposed model achieves the best performance against four baseline methods (one commonly used next-location prediction method, and three advanced sequence prediction methods) in terms of three evaluation metrics. We find that (1) the proposed method largely enhances the performance of location sequence prediction, both on variant length input sequences, as well as generation of different length output sequences; (2) the temporal attention mechanism can more interpretably uncover, and illustrate the daily and weekly underlying travel regularities compared with basic LSTM or LSTM encoder-decoder model.

Further investigations can be conducted from the following aspects. To reduce data sparsity problem and make the model applicable for predicting unvisited locations, we will extend our model to a group of people with similar travel patterns [4,45], and integrate multiple data sources such as point of interest (POI) and event dataset, to achieve semantic-level human activities prediction along with durations and locations at different spatiotemporal resolutions [16,23]. To improve computational efficiency, we will leverage more advanced sequential models, e.g., self-attention [39] and convolutional Seq2Seq models [14], to capture short-term and long-term dependencies while remaining interpretability.

Declaration of Competing Interest

We would like to declare that the work described is original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All authors are aware of the submission of this manuscript for publication and there is no potential conflict of competing interests.

CRediT authorship contribution statement

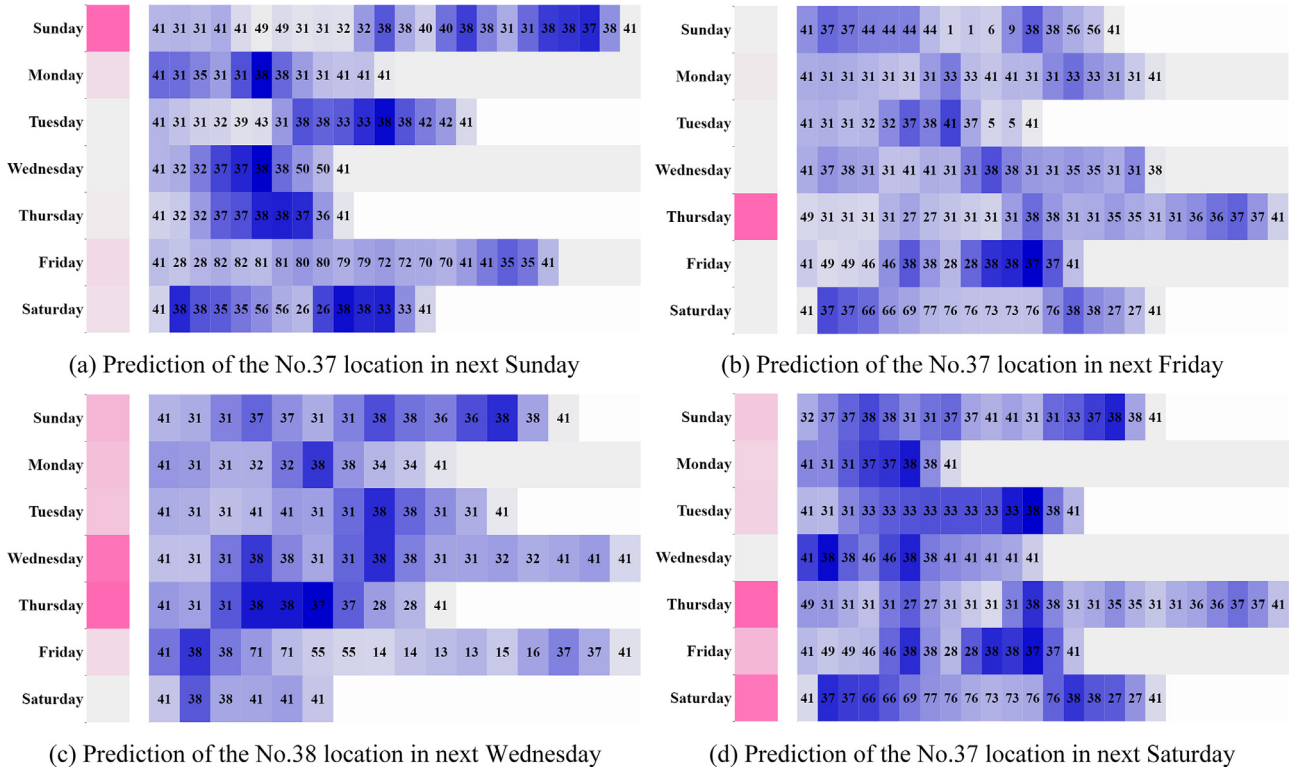
Fa Li: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Zhipeng Gui:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Zhaoyu Zhang:** Software, Visualization, Investigation. **Dehua Peng:** Visualization, Investigation, Validation. **Siyu Tian:** Conceptualization, Methodology. **Kunxiaojia Yuan:** Writing - review & editing. **Yunzeng Sun:** Visualization, Validation. **Huayi Wu:** Supervision. **Jianya Gong:** Supervision. **Yichen Lei:** Validation.

Acknowledgments

This paper is supported by National Key R and D Program of China (Nos. 2017YFB0503704, 2017YFB0503802, and 2018YFC0809806) and National Natural Science Foundation of China (Nos. 41971349, 41930107, and 41501434). Thanks to Zhenxu Zhai for providing helps in visualization, and data preprocessing. Thanks to Stephen C. McClure for language assistance.

Appendix

Fig. A.



(a) Prediction of the No.37 location in next Sunday

(b) Prediction of the No.37 location in next Friday

(c) Prediction of the No.38 location in next Wednesday

(d) Prediction of the No.37 location in next Saturday

Fig. A. Visualization of local and global attention weights in prediction of two locations at different calendar days of random selected weeks. Each row of the subfigures represents the location sequence where an individual orderly visited per day of last week in the form of Origin-Destination location ID pairs.

References

[1] L. Aalto, et al., Bluetooth and WAP push based location-aware mobile advertising system, in: Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services, ACM, Boston, MA, USA, 2004, pp. 49–58.

[2] A. Alahi, et al., Social LSTM: human trajectory prediction in crowded spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 961–971.

[3] F. Alhasoun, M. Alhazzani, M.C. GonzaĹez, City scale next place prediction from sparse data through similar strangers, in: Proceedings of ACM KDD Workshop, Halifax, Canada, 2017.

[4] A. Asahara, et al., Pedestrian-movement prediction based on mixed Markov-chain model, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2011, pp. 25–33.

[5] A. Brown, et al. Recurrent neural network attention mechanisms for interpretable system log anomaly detection, (2018). arXiv:1803.04967.

[6] F. Calabrese, G. Di Lorenzo, C. Ratti, Human mobility prediction based on individual and collective geographical preferences, in: Proceedings of the International IEEE Conference on Intelligent Transportation Systems, IEEE, 2010, pp. 312–317.

[7] M. Chinazzi, et al., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak, Science (2020).

[8] Y. Choi, L. Sun, Reuse intention of third-party online payments: a focus on the sustainable factors of alipay, Sustainability 8 (2) (2016) 147.

[9] S. Çolak, A. Lima, M.C. González, Understanding congested travel in urban areas, Nat. Commun. 7 (1) (2016) 1–8.

[10] A. Cuttone, S. Lehmann, M.C. González, Understanding predictability and exploration in human mobility, EPJ Data Sci. 7 (1) (2018) 1–17.

[11] R.W. Engle, Working memory capacity as executive attention, Curr. Dir. Psychol. Sci. 11 (1) (2002) 19–23.

[12] V. Etter, et al., Where to go from here? Mobility prediction from instantaneous information, Pervasive Mob. Comput. 9 (6) (2013) 784–797.

[13] Y. Gal, Z. Ghahramani, A theoretically grounded application of dropout in recurrent neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 1019–1027.

[14] J. Gehring, et al., Convolutional sequence to sequence learning, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017, pp. 1243–1252.

[15] M.C. González, C.A. Hidalgo, A.L. BarabÃ si, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.

[16] G. Goulet-Langlois, et al., Measuring regularity of individual travel patterns, IEEE Trans. Intell. Transp. Syst. 19 (5) (2017) 1583–1592.

[17] A. Graves, Long Short-Term Memory, Springer, Berlin Heidelberg, 2012.

[18] H. Guan, S. Huang, Study on traffic access mode choice of urban railway system in Beijing, Researcher 1 (2009) 5–57.

[19] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: Proceedings of the ACM SIGMOD Record, 2000, pp. 1–12.

[20] J.O. Huff, S. Hanson, Repetition and variability in urban travel, Geogr. Anal. 18 (2) (1986) 97–114.

[21] M.O. Killijian, Next place prediction using mobility Markov chains, in: Proceedings of the Workshop on Measurement, 2012, p. 3.

[22] D.Y. Kim, H.Y. Song, Method of predicting human mobility patterns using deep learning, Neurocomputing 280 (2018) 56–64.

[23] K. Krishna, et al., An LSTM based system for prediction of human activities with durations, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (4) (2017) 1–31.

[24] Y. Li, B. Liu, A Normalized Levenshtein Distance Metric, IEEE Computer Society, 2007.

[25] Y. Liang, Z. Jiang, Y. Zheng, Inferring traffic cascading patterns, in: Proceedings of the 25th ACM Sigspatial International Conference on Advances in Geographic Information Systems, 2017, pp. 1–10.

[26] Y. Liang, et al., GeoMAN: multi-level attention networks for geo-sensory time series prediction, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 3428–3434.

[27] Y. Liang, et al., UrbanFM: inferring Fine-Grained Urban Flows, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 3132–3142.

[28] X. Lu, et al., Approaching the limit of predictability in human mobility, Sci. Rep. 3 (10) (2013) 2923.

[29] Neubig, G. 2017. Neural machine translation and sequence-to-sequence models: a tutorial. arXiv:1703.01619.

[30] Park, S., et al. 2018. Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture. arXiv:1802.06338.

[31] E.I. Pas, F.S. Koppelman, An examination of the determinants of day-to-day variability in individuals' urban travel behavior, Transportation (AMST) 14 (1) (1987) 3–20.

[32] Y. Qiao, et al., A hybrid Markov-based model for human mobility prediction, Neurocomputing 278 (2018) 99–109.

[33] Qin, Y., et al. 2017. A dual-stage attention-based recurrent neural network for time series prediction. 2627–2633.

[34] D. Quercia, et al., Recommending social events from mobile phone location data, in: Proceedings of the IEEE International Conference on Data Mining, 2010, pp. 971–976. 13-17 Dec. 2010.

[35] E.S. Ristad, P.N. Yianilos, Learning string-edit distance, IEEE Trans. Pattern Anal. Mach. Intell. 20 (5) (1998) 522–532.
[36] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors. 323(6088) (1986) 399–421.
[37] C. Song, et al., Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.
[38] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks. 4, (2014) 3104–3112.
[39] A. Vaswani, et al., Attention is all you need, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[40] P.E. Wais, et al., Neural mechanisms underlying the impact of visual distraction on retrieval of long-term memory, J. Neurosci. 30 (25) (2010) 8541–8550.
[41] D. Wang, et al., When will you arrive? estimating travel time based on deep neural networks, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[42] Y. Yang, et al., Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies, Travel Behav. Soc. 1 (2) (2014) 69–78.
[43] Z. Yang, et al., Hierarchical attention networks for document classification, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
[44] H. Yin, et al., Study of Urban Resident Travel Mode Choice Behavior, in: Proceedings of the Tenth International Conference of Chinese Transportation Professionals, 2010, pp. 1807–1815.
[45] Z. Zhao, H.N. Koutsopoulos, J. Zhao, Individual mobility prediction using transit smart card data, Transp. Res. Part C Emerg. Technol. 89 (2018) 19–34.

Fa Li is a Ph.D. candidate in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University. He is also an affiliate researcher of Lawrence Berkeley National Laboratory. His research interests include theory and applications of machine learning and causal inference on social and environmental science.

Zhipeng Gui is an Associate Professor of Geographic Information Science in the School of Remote Sensing and Information Engineering, Wuhan University. His research interest is high-performance spatiotemporal data mining, geovisual analytics and Distributed Geographic Information Processing (DGIP), especially on (1) Spatiotemporal point pattern analysis and GeoAI; (2) High-performance geocomputation and spatial cloud computing; (3) Geospatial service chain modeling and optimization; (4) QoGIS-aware monitoring and evaluation of geospatial web services. He now serves as the Co-chair of International Society for Photogrammetry and Remote Sensing (ISPRS) Working Group V/4: Web-based Resource Sharing for Education and Research.

ZhaoYu Zhang is a master student in the School of Artificial Intelligence, Nanjing University. His research interests include machine learning and data mining, especially time series.

Dehua Peng is a master student in the School of Remote Sensing and Information Engineering, Wuhan University. He will become a doctoral student in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing in Sep., 2020. His research interests include clustering algorithms and point pattern mining.

Siyu Tian holds a master's degree in Geo-information Science from CUHK and bachelor's degree in remote sensing from WHU. His current research interests include geospatial intelligence and geo-statistics.

Kunxiaojia Yuan is a Ph.D. candidate in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. She is also an affiliate researcher of Lawrence Berkeley National Laboratory. Her current research interests include theory and applications of machine learning, causal inference, and spatial statistics.

Yunzeng Sun is a master student in the school of remote sensing and information engineering, Wuhan University. His research interests include spatiotemporal analysis and trajectory data mining.

Huayi Wu is now a full professor in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include high-performance geospatial computing and intelligent geospatial web services.

Jianya Gong is now a full professor in the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include theory and applications of geographic information science as well as remote sensing.

Yichen Lei is a master student at the Department of Urban Spatial Analytics, University of Pennsylvania. He received his B.E. degree from Wuhan University. His research fields are spatial analysis and big data mining.