Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 3463330, 14 pages
https://doi.org/10.1155/2023/3463330

*Research Article*

# A New Individual Mobility Prediction Model Applicable to Both Ordinary Conditions and Large Crowding Events

**Bao Guo**,[1] **Kaipeng Wang**,[1] **Hu Yang**,[1] **Fan Zhang**,[2] **and Pu Wang**[1]

[1]*School of Traffic and Transportation Engineering, Rail Data Research and Application Key Laboratory of Hunan Province, Central South University, Changsha 410000, China*
[2]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China*

Correspondence should be addressed to Pu Wang; wangpu@csu.edu.cn

Received 13 December 2022; Revised 29 May 2023; Accepted 19 June 2023; Published 27 June 2023

Academic Editor: Tomio Miwa



Accurate prediction of individual mobility is crucial for developing intelligent transportation systems. However, while previous models usually focused on predicting individual mobility under ordinary conditions, the models that are applicable to large crowding events are still lacking. Here, we employ the smart card data of 6.5 million subway passengers of the Shenzhen Metro to develop a Markov chain-based individual mobility prediction model (i.e., SCMM) applicable to both ordinary and anomalous passenger flow situations. The proposed SCMM model improves the Markov chain model by incorporating the station-level anomalous passenger flow index and the collective mobility patterns of similar passengers. Compared with the benchmark models, the SCMM model achieves the highest prediction accuracy in both ordinary conditions and large crowding events. Our results highlight the importance of combining an individual's own historical mobility data with collective mobility data and suggest the appropriate weights of individual and collective information considered in individual mobility modeling.

## 1. Introduction

An in-depth understanding of individual human mobility is of significant importance for urban planning [1], transportation management [2], and the development of intelligent transportation systems [3, 4]. With increasingly abundant big data recording individuals' temporal and spatial information, human mobility research has experienced rapid development over the last 15 years [1]. Various types of big data, from banknote circulation data [5], mobile phone data [6], to social media data [7], and individual GPS trajectory data [8], were employed to uncover the hidden laws of human travel. Moreover, human mobility was discovered to be highly predictable [9, 10], and many pioneering models were proposed to reproduce human mobility laws or predict individual or collective human movements [11–14]. In recent years, increasing attention has been paid to human mobility under anomalous conditions, for instance, during special events [15, 16], natural disasters [17], extreme weather [18], and epidemic spreading [19, 20]. Yet, individual mobility prediction models applicable to anomalous conditions are still lacking. In this study, we develop a new individual mobility prediction model applicable to both ordinary conditions and large crowding events. The developed model can provide useful information for crowd safety management [21, 22] and crowd disaster prevention [23], which facilitates the development of smart cities.

The existing individual mobility prediction models mainly include location-based models and trip-based models. Location-based models predict the location that an individual will visit [24, 25], whereas trip-based models predict an individual's location in the next time interval [26, 27], or simultaneously predict the departure time, the origin, and the destination of his/her next trip [28]. Although many individual mobility prediction models have been proposed, most of these models are not applicable to anomalous mobility conditions, for instance, in large crowding events [29]. The main challenge is that individual mobility shows dramatically different patterns in large crowding events, and such patterns were not captured by historical data [23, 30]. Nevertheless, this inspires us to combine real-time anomalous passenger flow information from large crowding events with individuals' historical mobility habits. Therefore, in what follows, we first review previous works on individual mobility prediction, clustering of travelers, and prediction of anomalous mobility patterns.

In the recent decade, hidden Markov model (HMM) [25], principal component analysis (PCA) [31], Bayesian network [32], and deep learning methods [33] were employed to model and predict individual mobility. For examples, Al-Molegi et al. [34] proposed a recurrent neural network (RNN) model to predict individuals' next location; Li et al. [33] used the long short-term memory (LSTM) model to capture the daily and weekly travel regularities of individuals. Given the importance of the temporal property of human mobility, some researchers investigated methods for predicting both the location and the time of the next trip. For examples, Hsieh et al. [35] extracted a location time distribution (LTD) for each location and a transition time distribution (TTD) for each pair of locations, and individuals' next location is predicted based on how well the location sequence matches the LTD and TTD; Zhao et al. [28] employed large-scale smart card data of subway passengers to simultaneously predict three attributes of a passenger trip (i.e., the departure time, the origin, and the destination of the trip); and Mo et al. [36] proposed an input-output hidden Markov model (IOHMM) to predict the time and the location of individuals' next trip. In the area of individual mobility prediction, a number of advanced methods and techniques have been applied or developed; however, individual mobility under anomalous conditions has not been sufficiently investigated, probably due to the complex dynamics of human mobility during rare events.

It is difficult to predict an individual's location if the location has never been visited by the individual. To solve this, some researchers clustered individuals into groups based on their temporal-spatial mobility similarities and proposed hybrid models that integrated individual mobility data and collective mobility data of similar individuals to improve the prediction accuracy [37]. Based on similar ideas, a number of individual mobility prediction models have been proposed. Asahara et al. [38] split individuals into different groups using an expectation-maximization (EM) algorithm and proposed the mixed Markov model (MMM) to predict the mobility patterns of each group of individuals. Mathew et al. [25] clustered historical individual locations based on the time period that each location was recorded and trained a hidden Markov model (HMM) for each cluster of individuals. Alhasoun et al. [27] identified the "similar strangers" of each individual and predicted individual mobility by integrating the individual's historical mobility information and his/her similar strangers' collective mobility information in a dynamic Bayesian network model. Yang et al. [39] grouped subway passengers based on their trip frequency in each time slot and their visited locations, and the future movements of passengers in each group were predicted using the Markov chain model and the hidden Markov model. Taken together, we find the increasing use of individual clustering techniques in mobility prediction; however, existing models are mostly applicable to ordinary mobility conditions. We are still lacking individual mobility prediction models applicable to large crowding events [40, 41].

Given their significant importance in scientific crowd management and crowd disaster prevention, researchers have investigated methods for predicting collective human mobility patterns at large events. For examples, Pereira et al. [42] considered the time of the next event and event type to develop an artificial neural network (ANN) for predicting passenger flows at bus stops or subway stations during large events; Rodrigues et al. [43] used the time of the event, event topics, and venues to generate a Bayesian additive model for predicting the volume of subway trips heading to the event area; Ni et al. [44] discovered that the passenger flow at a subway station is positively correlated with the social media post rate, and the discovered correlation was used for predicting the station passenger flow during sports events. Anomalous mobility conditions may also emerge when no event information is released on the Internet [45]. To identify and predict anomalous collective mobility, Huang et al. [23] developed the anomalous mobility network approach to capture anomalous passenger flows and anticipate large crowding events. Zheng et al. [46] proposed a hybrid model to predict anomalous passenger flow in an urban metro, where the complex network index $k_{in}$ was used to determine the time for implementing online learning. In addition, Cheng et al. [47] analyzed the causal relationship between returning flow and incoming demand, which improves the prediction accuracy of passenger flows during special events. Reviewing recent works in this area, we find that a few collective mobility prediction approaches applicable to large crowding events have been proposed; however, individual mobility prediction approaches are still lacking.

In this study, we develop an improved Markov chain-based individual mobility prediction model (i.e., SCMM) applicable to both ordinary and anomalous passenger flow situations (i.e., at large crowding events). Specifically, we propose an anomalous mobility index derived from historical and real-time station-level passenger flow, which can capture the anomalous passenger mobility patterns during large crowding events. In addition, we incorporate the collective mobility patterns of similar passengers into the SCMM model, where the K-means algorithm is employed to classify the individuals based on their temporal mobility patterns, and the collective mobility probability of each group of passengers is calculated using the improved Markov chain model. Moreover, a weight index is used to balance the weights of individual and collective mobility information considered in the SCMM model. The developed SCMM model is validated using the smart card data of subway passengers in the Shenzhen Metro. Compared with the benchmark models, the proposed SCMM model achieves the highest prediction accuracy in both ordinary conditions and large crowding event scenarios, which could be employed to prevent crowd disasters and develop smart cities.

The remainder of this paper is organized as follows: Section 2 introduces the data used in this study; in Section 3, the developed SCMM model for predicting individual passenger mobility under ordinary and anomalous passenger flow situations is presented; in Section 4, the proposed SCMM model is validated using the large-scale smart card data of subway passengers; and Section 5 concludes the findings of this work. The limitations of the research and future research directions are also discussed.

## 2. Data

The data used in the present study were provided by the Shenzhen Transportation Authority. The geographic information system (GIS) data for the Shenzhen Metro were collected in 2014. During the data collection period, there were 5 lines and 118 stations in the studied subway network (Figure 1(a)). The smart card data were also collected in 2014 (from November 1 to December 31). Each time a subway passenger entered or exited a subway station, the unique ID of the anonymous subway passenger, the station ID, and the time when the passenger swiped the card were recorded. During the data collection period, 163,238,950 smart card records were generated by 6,500,941 passengers. The temporal pattern of the smart card records is shown in Figure 1(b). There were data missing on November 20, December 1–8 and December 18–20, 2014 (colored in grey in Figure 1(b)). Only the passenger trips collected in the remaining 49 days were used.

During the two-month data collection period, nine large crowding events occurred near five subway stations (Table 1). Here, the five subway stations are denoted as the crowding stations. The out-passenger flow $f_{\text{out}}$ of a subway station is calculated by aggregating the trips with destinations at the subway station. As shown in Figure 2, the out-passenger flow $f_{\text{out}}$ at the Sea World station (a crowding station) increased prominently during the crowding events.

Given that sufficient historical mobility records are needed for predicting individual mobility, only the passengers with at least 41 trips recorded (at least 1 trip per day averagely) in the training data (439,560 passengers in total) are selected for training and validating the individual mobility prediction model. The developed individual mobility prediction model is trained using the smart card records collected from November 1 to December 23, 2014 (84.6% of all 30,832,570 trips) and tested using the smart card records collected from December 24 to December 31, 2014 (15.4% of all trips).

## 3. Model

*3.1. The Modeling Framework.* In this study, we developed the SCMM model to predict individual passenger mobility under ordinary and anomalous passenger flow situations. The proposed SCMM model is based on the Markov chain model and incorporates the station-level passenger flow information and the collective mobility patterns of similar passengers to further improve prediction accuracy. The SCMM model mainly consists of four modules, as illustrated in Figure 3.

*3.1.1. Inferring Individual Location Sequences.* Although the smart card data recorded the location information of subway passengers, for each passenger, there were no smart card records in the majority of time slots. Here, we use consecutive trip records of each passenger to infer the passenger's location in time slots when there were no smart card records. The inferred individual location sequences are used to train and validate the SCMM model.

*3.1.2. Analyzing Station-Level Anomalous Passenger Flow.* Anomalous passenger mobility may cause a prominent fluctuation (increase) in out-passenger flow $f_{\text{out}}$ at subway stations. Here, we propose an anomalous mobility index to measure the fluctuation of out-passenger flow at a subway station, which is used to evaluate the attractiveness of the subway station to passengers.

*3.1.3. Clustering Subway Passengers.* Passengers who have had similar mobility patterns in history are likely to have similar mobility patterns in the future. We split passengers into different groups according to their temporal mobility patterns. The collective mobility patterns of similar passengers are integrated into the SCMM model.

*3.1.4. Predicting Passenger Locations.* Combining station-level passenger flow information and collective mobility patterns of similar passengers, we train the SCMM model using the training data and validate the effectiveness of the model using the test data.

*3.2. Inferring Individual Location Sequences.* We use individual location sequences to capture the mobility patterns of each subway passenger. Here, a day is divided into 24 one-hour time slots. An individual location sequence $L = \{l_1^1, l_2^1, \ldots, l_t^d, \ldots, l_{24}^D\}$ contains $n = D \cdot 24$ locations, where $D$ is the number of days in the observation period ($D = 49$ for this study), and $l_t^d$ is the location of the passenger in time slot $t$ of day $d$. The number of time slots, $n$, is equal to the number of days studied times the number of time slots (i.e., $n$ = 1,176). Most subway passengers only had smart card records in a few time slots. Hence, we need to infer a passenger's location in time slots when there are no smart card data recorded [48].

Specifically, let trip $i$ represent the $i^{\text{th}}$ trip of the passenger, $t_i$ represents the departure time of trip $i$, $o_i$ represents the origin of trip $i$, and $d_i$ represents the destination of trip $i$. A passenger's location sequence is inferred as follows:

Step 1: for time slots in which a passenger has trips, the origin of the first trip in the time slot is inferred as the location of the passenger in this time slot.

Step 2: for each day, if trip $i$ is the first trip of the day, $o_i$ is inferred as the location of the passenger in the time slots before $t_i$ and if trip $i$ is the last trip of the day, $d_i$ is inferred as the location of the passenger in the time slots after $t_i$.

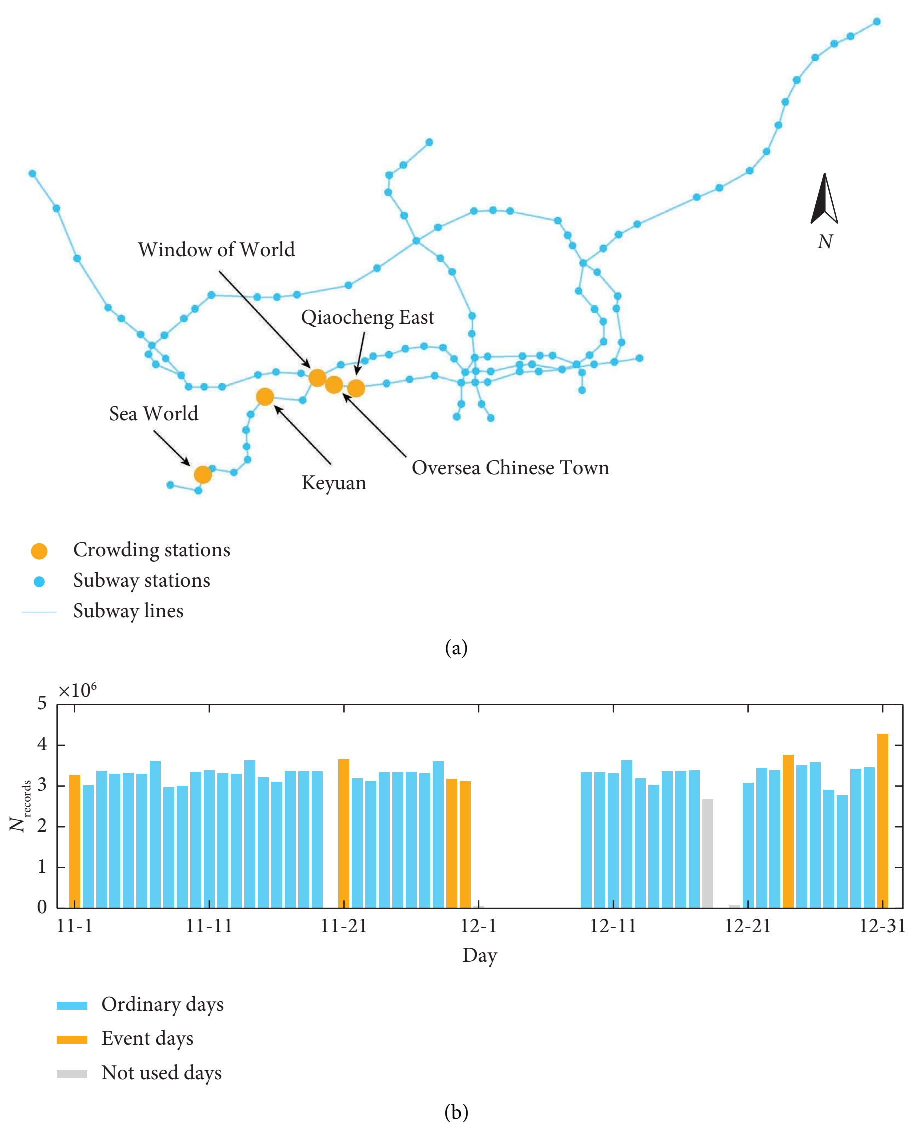


FIGURE 1: (a) Illustration of the subway network of Shenzhen (Shenzhen Metro). The crowding stations are highlighted in orange. (b) The number of smart card records $N_{\text{records}}$ in each hour during the observation period.

TABLE 1: Crowding events occurred in Shenzhen during November and December 2014.

| ID | Date | Crowding station | Event |
|---|---|---|---|
| 1 | Nov. 1 | Window of World | Halloween recreational activity |
| 2 | Nov. 1 | Oversea Chinese Town | Halloween recreational activity |
| 3 | Nov. 21 | Keyuan | Thanksgiving star concert |
| 4 | Nov. 29 | Keyuan | Simple K-pop tour in China |
| 5 | Nov. 30 | Qiaocheng East | Open day of Shenzhen police |
| 6 | Dec. 24 | Sea World | Christmas Eve activity |
| 7 | Dec. 31 | Sea World | New Year's Eve activity |
| 8 | Dec. 31 | Window of World | New Year's Eve activity |
| 9 | Dec. 31 | Qiaocheng East | New Year's Eve activity |

Step 3: if trip $i$ is neither the first trip nor the last trip of the day, and there are time slots between trip $i-1$ and trip $i$, we need to infer the passenger's locations during these time slots. Specifically, if $d_{i-1}$ and $o_i$ are the same, $o_i$ is inferred as the locations of the passenger in the time slots between trip $i-1$ and trip $i$. Otherwise, the passenger's locations during these time slots are marked as unknown.

Step 4: if there are no trips in a day, the location of the passenger in each time slot on this day is marked as

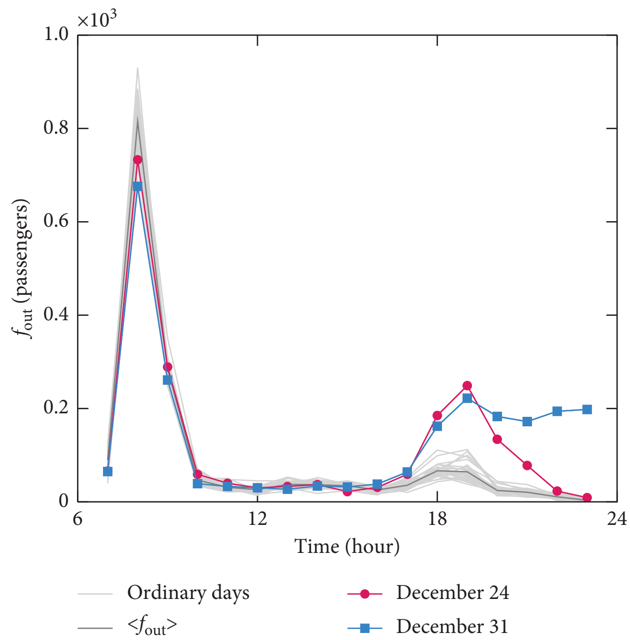


FIGURE 2: The out-passenger flow $f_{out}$ at the Sea World station on ordinary weekdays versus the out-passenger flow $f_{out}$ at the Sea World station during the large crowding events. The bold grey line represents the average out-passenger flow $\langle f_{out} \rangle$ at the Sea World station.

unknown. In addition, if a passenger takes the subway during a whole time slot, the passenger's location is also marked as unknown. For example, if a passenger entered a subway station at 7:50 a.m. and exited a subway station at 9:05 a.m., the passenger's location in the time slot from 8:00 a.m. to 9:00 a.m. is marked as unknown.

Using the method mentioned above, the individual location sequence of each passenger is obtained.

*3.3. Analyzing Station-Level Anomalous Passenger Flow.* When a crowding event occurs, the out-passenger flow $f_{out}$ at the crowding station will increase greatly. However, previous individual mobility models did not make full use of this essential real-time passenger flow information. To incorporate this essential information into the individual mobility prediction model, we propose an anomalous mobility index $\delta_{t,s}$ as follows:

$$\delta_{t,s} = \frac{f_{t,s} - \langle f_{t,s} \rangle}{\sigma(f_{t,s})}, \tag{1}$$

where $f_{t,s}$ is the out-passenger flow at station $s$ in time slot $t$, $\langle f_{t,s} \rangle$ is the mean of out-passenger flow $f_{t,s}$ in time slot $t$, and $\sigma(f_{t,s})$ is the standard deviation of out-passenger flow $f_{t,s}$ in time slot $t$. We calculate $\langle f_{t,s} \rangle$ and $\sigma(f_{t,s})$ for weekdays and weekends using the training data, respectively. On ordinary days, out-passenger flow $f_{t,s}$ is close to $\langle f_{t,s} \rangle$, and the anomalous mobility index $\delta_{t,s}$ is close to 0. When a large crowding event occurs, the out-passenger flow $f_{out}$ at the crowding station increases dramatically, and accordingly, $\delta_{t,s}$ increases prominently. The anomalous mobility index of a subway station captures the real-time attractiveness of the subway station to passengers. A station is in a anomalous passenger flow situation when the anomalous mobility index of the station $\delta_{t,s} > 3$ on weekdays or $\delta_{t,s} > 2.6$ on weekends [49]. Otherwise, the station is in an ordinary passenger flow situation.

Anomalous mobility index $\delta_{t,s}$ is first calculated for each time slot covered by the training data. For the time slots covered by the test data, we inferred the anomalous mobility index $\delta_{t,s}$ in time slot $t$ based on the anomalous mobility index in time slot $t-1$, $\delta_{t-1,s}$ as follows:

$$\delta_{t,s} = \begin{cases} \delta_{t-1,s}, & \delta_{t-1,s} > 1, \\ 1, & \text{otherwise}. \end{cases} \tag{2}$$

If $\delta_{t-1,s} > 1$, there might be anomalous passenger flow at station $s$, and the anomalous mobility index $\delta_{t,s}$ is set to $\delta_{t-1,s}$ to denote the increased traffic demand at station $s$. Otherwise, $\delta_{t,s}$ is set to 1, and the mobility patterns of subway passengers are the same as the mobility patterns on ordinary days.

When predicting the individual location sequence of a passenger, we use $l_t$ to indicate the subway station $s$ where the passenger visits in time slot $t$. Thus, $\delta_{t,s}$ is expressed as $\delta_{t,l_t}$ in the following text.

*3.4. Clustering Subway Passengers.* We cluster the subway passengers with similar mobility patterns into the same group and calculate the collective mobility probability of passengers in each group to calibrate the synthetic mobility probability of a passenger. Here, subway passengers are clustered based on their temporal mobility patterns [50]. The detailed method is given below.

For each passenger, a time series $H = \{H_1^1, H_2^1, \ldots, H_t^d, \ldots, H_{24}^{D'}\}$ is first generated to extract the passenger's temporal mobility pattern (Figure 4), where $H_t^d = 1$ when the passenger swiped his/her smart card in time slot $t$ on day $d$; otherwise, $H_t^d = 0$, and $D' = 41$ are the number of days covered by the training data. The operation hours of the Shenzhen Metro were from 6:00 a.m. to 12:00 a.m. Thus, $H_t^d = 0$ for the time slots of 0:00 a.m. to 6:00 a.m. Next, we generate the *overlapped slots* $S = \{S_1, S_2, \ldots, S_t, \ldots, S_{22}\}$. The length of each overlapped slot is set to 3 hours. For instance, overlapped slot $S_1$ denotes 0:00 a.m. to 2:59 a.m.; overlapped slot $S_2$ denotes 1:00 a.m. to 3:59 a.m.; and there are 22 overlapped slots, i.e., $t \in [1, 2, \ldots, 22]$. Note that a subway trip could have multiple overlapped slots. Finally, time series $H$ is used to calculate the attributes of each overlapped slot $S_t$, which include the proportion of active days with trips $D_t$ and the average number of trips $F_t$, both of which are calculated using the training data:

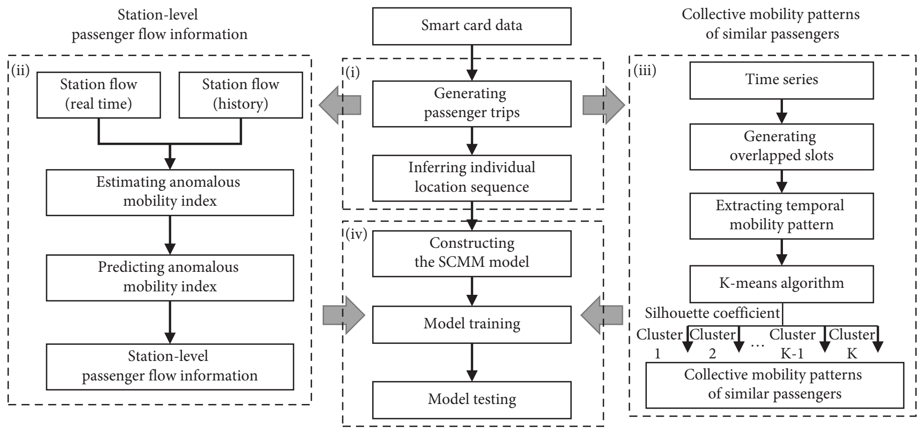


Figure 3: The framework of the proposed SCMM model.

$$D_t = \frac{1}{D'} \sum_{d=1}^{D'} \mathrm{MAX}\left(H_t^d, H_{t+1}^d, H_{t+2}^d\right),$$
$$F_t = \frac{1}{D'} \sum_{d=1}^{D'} \left(H_t^d + H_{t+1}^d + H_{t+2}^d\right), \quad (3)$$

where the operator MAX returns the maximum value. The overlapped slots $S$ are sorted in a descending order of $F_t$. The sorted overlapped slots $S' = \{S_1', S_2', \ldots, S_j', \ldots, S_{22}'\}$ are used to denote the temporal mobility pattern of the passenger. Specifically, we iteratively select $S_t'$ from $S_1'$ to $S_{22}'$ to generate the nonoverlap slots $S'' = \{S_1'', S_2'', S_3'', S_4''\}$, which have no overlapping in time.

The detailed method to generate $S''$ is as follows: firstly, $S_1'$ is the first nonoverlap slot in $S''$ (i.e., $S_1''$). Secondly, $S_2'$ is compared with $S_1''$. If $S_1''$ and $S_2'$ are not overlapped in time, $S_2'$ is set to the second nonoverlap slot in $S''$; otherwise, $S_2'$ is not added to $S''$. Next, we check if $S_3'$ has overlapped in time with the existing nonoverlap slots in $S''$ ($\{S_1', S_2'\}$ or $\{S_1'\}$) to determine whether $S_3'$ will be added to $S''$. This process continues until we find the four nonoverlap slots $\{S_1'', S_2'', S_3'', S_4''\}$. The proportions of active days with trips $\{D_{\mathrm{top}_1}, D_{\mathrm{top}_2}, D_{\mathrm{top}_3}, D_{\mathrm{top}_4}\}$ of the identified four nonoverlap slots $\{S_1'', S_2'', S_3'', S_4''\}$ of each passenger are used as the features for clustering passengers [50].

The K-means algorithm is used for clustering the passengers. The selected features $\{D_{\mathrm{top}_1}, D_{\mathrm{top}_2}, D_{\mathrm{top}_3}, D_{\mathrm{top}_4}\}$ generate the feature space. Each passenger with the feature vector is a data sample. We use the silhouette coefficient [51] to determine the suitable number of passenger groups. For each passenger $p$, the silhouette coefficient is calculated as follows:

$$s(p) = \frac{b(p) - a(p)}{\max\{a(p), b(p)\}}, \quad (4)$$

where $a(p)$ is the average Euclidean distance between passenger $p$ and the other passengers in the same group, $b(p)$ is the minimum average Euclidean distance between passenger $p$ and passengers in any other groups. The average value of $s(p)$ of all passengers is defined as the silhouette coefficient of the groups and used to determine the optimal value of $K$. The number of passenger groups is tested from 2 to 17 (i.e., $K = 2, 3, \ldots, 17$), and the value of $K$ that achieves the highest silhouette coefficient is used. When predicting the location of a passenger, the synthetic mobility probability is calibrated using the collective mobility probability of passengers in the same group.

*3.5. Predicting Passenger Locations.* The Markov chain (MC) model is a commonly used mobility prediction model which can achieve high prediction accuracy [52, 53]. In this study, we develop a Markov chain model-based individual mobility prediction model which also combines station-level passenger flow information with collective mobility patterns of similar passengers. In the SCMM model, $P_p(l_t | l_{t-1})$ represents the synthetic mobility probability that a passenger $p$ is at location $l_t$ in time slot $t$ under the condition that the passenger is at location $l_{t-1}$ in time slot $t-1$:

$$P_p\left(l_t | l_{t-1}\right) = \alpha P_i\left(l_t | l_{t-1}\right) + (1-\alpha) P_c\left(l_t | l_{t-1}\right), \quad (5)$$

where $P_i(l_t | l_{t-1})$ is the individual mobility probability that passenger $p$ is at location $l_t$ in time slot $t$ under the condition that the passenger is at location $l_{t-1}$ in time slot $t-1$, $P_c(l_t | l_{t-1})$ is the collective mobility probability that passengers in the same group are at location $l_t$ in time slot $t$ under the condition that they are at location $l_{t-1}$ in time slot $t-1$, and $\alpha$ is a weight index balancing the weights of $P_i(l_t | l_{t-1})$ and $P_c(l_t | l_{t-1})$. Given the location $l_{t-1}$ in current time slot, the location $l_t$ with the greatest probability $P_p(l_t | l_{t-1})$ is selected as the predicted location.

The individual mobility probability $P_i(l_t | l_{t-1})$ combines individual mobility patterns with station-level passenger flow information:

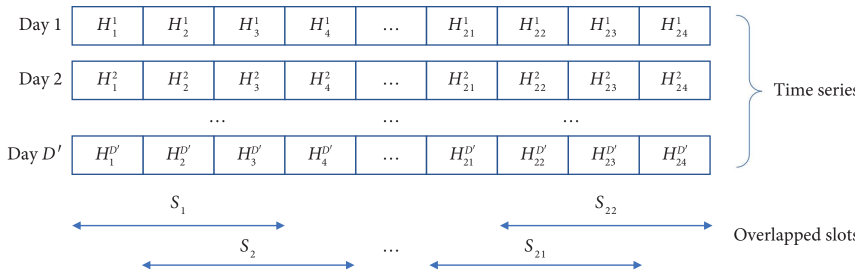


Figure 4: Illustration of the temporal pattern extracting procedure.

$$P_i\left(l_t|l_{t-1}\right) = \frac{c\left(l_{t-1},l_t\right)}{\sum_{l_t\in L} c\left(l_{t-1},l'_t\right)} \cdot \delta_{t,l_t}, \quad (6)$$

where $c\,(l_{t-1,}l_t)$ is the number of times that the passenger is at location $l_{t-1}$ in time slot $t-1$ and at location $l_t$ in time slot $t$, $L$ is the set of locations that a passenger has visited, and $\delta_{t,l_t}$ is the anomalous mobility index of location $l_t$ in time slot $t$.

The collective mobility probability $P_c\,(l_t\,|l_{t-1})$ is calculated as follows:

$$P_c\left(l_t\,|l_{t-1}\right) = \frac{C\left(l_{t-1},l_t\right)}{\sum_{l_t\in L} C\left(l_{t-1},l'_t\right)}, \quad (7)$$

where $C\,(l_{t-1,}l_t)$ is the number of times that passengers in the same group are at $l_{t-1}$ in time slot $t-1$ and at $l_t$ in time slot $t$.

In order to determine the optimal value of weight index $\alpha$, training data are further divided into two parts, namely, the training part and the validation part. The value of the weight index $\alpha$ is tested from 0 to 1 with a tolerance of 0.1 (i.e., $\alpha = 0,\ 0.1,\ 0.2, \ldots, 1$). We select the $\alpha$ that achieves the highest median of prediction accuracy rates for all passengers, where the prediction accuracy rate of an individual passenger is defined as follows:

$$\text{Accuracy} = \frac{N_{\text{true}}}{N_{\text{all}}}, \quad (8)$$

where $N_{\text{all}}$ is the total number of time slots in the validation part and $N_{\text{true}}$ is the number of time slots in which the individual location prediction is correct.

*3.6. Comparative Models.* In this study, three benchmark models are introduced to validate the proposed SCMM model, i.e., a first-order Markov chain model (MC model), a Markov chain model incorporating the proposed indices (SMM model), and a random forest (RF) model.

In the RF model, the input features are composed by the information of time slot $t$ and the location of the passenger in time slot $t-1$, $l_{t-1}$. Here, the information of time slot is treated as categorical variables and is converted to 24 binary variables.

In the MC model, the mobility probability is calculated as follows:

$$P'_p\left(l_t\,|l_{t-1}\right) = \frac{c\left(l_{t-1},l_t\right)}{\sum_{l_t\in L} c\left(l_{t-1},l'_t\right)}, \quad (9)$$

where $c\,(l_{t-1,}l_t)$ is the number of times that the passenger is at location $l_{t-1}$ in time slot $t-1$ and at location $l_t$ in time slot $t$.

A Markov chain model established by (6) is denoted as Markov chain model with station-level information (SMM), which is another benchmark model used in this study.

## 4. Results

The individual mobility prediction models are trained and validated using the individual location sequences inferred by the smart card data collected from November 1 to December 23, 2014.

Individuals tend to travel in a way like their "similar strangers" [27]. In order to take advantage of such collective mobility patterns, passengers are clustered based on their temporal mobility patterns using the K-means algorithm. The silhouette coefficient is used to determine the optimal number of clusters. As shown in Figure 5(a), the silhouette coefficient reaches its peak value when the number of clusters is set to 2, so passengers are clustered into two groups. We analyze the temporal mobility patterns (denoted by $\{D_{\text{top}_1}, D_{\text{top}_2}, D_{\text{top}_3}, D_{\text{top}_4}\}$) of the two groups of passengers. Figure 5(b) shows that passengers in Group 1 travel during multiple nonoverlap slots, while passengers in Group 2 have two dominant active nonoverlap slots, implying that passengers in Group 2 might be commuters. The obtained temporal mobility patterns are similar to the findings in the previous work [50].

In the proposed SCMM model, individual and collective mobility information are balanced using the weight index $\alpha$. To obtain the optimal value of $\alpha$, the training data are further divided into two parts. Data collected from November 1 to November 30 are used as the training part and data collected from December 9 to December 23 are used as the validation part. As shown in Figure 6, the median of prediction accuracy rates reaches its peak at $\alpha = 0.9$. Therefore, the weight index $\alpha$ is set to 0.9, implying that while a passenger's mobility is mainly affected by his/her historical mobility patterns, the collective mobility information also plays a significant role. Interestingly, the prediction accuracy will decrease if the prediction model only relies on individual mobility information (i.e., $\alpha = 1$) or collective mobility information (i.e., $\alpha = 0$). This finding highlights the necessity of combining individual's own historical mobility data with collective mobility information, which is the key for improving prediction accuracy.

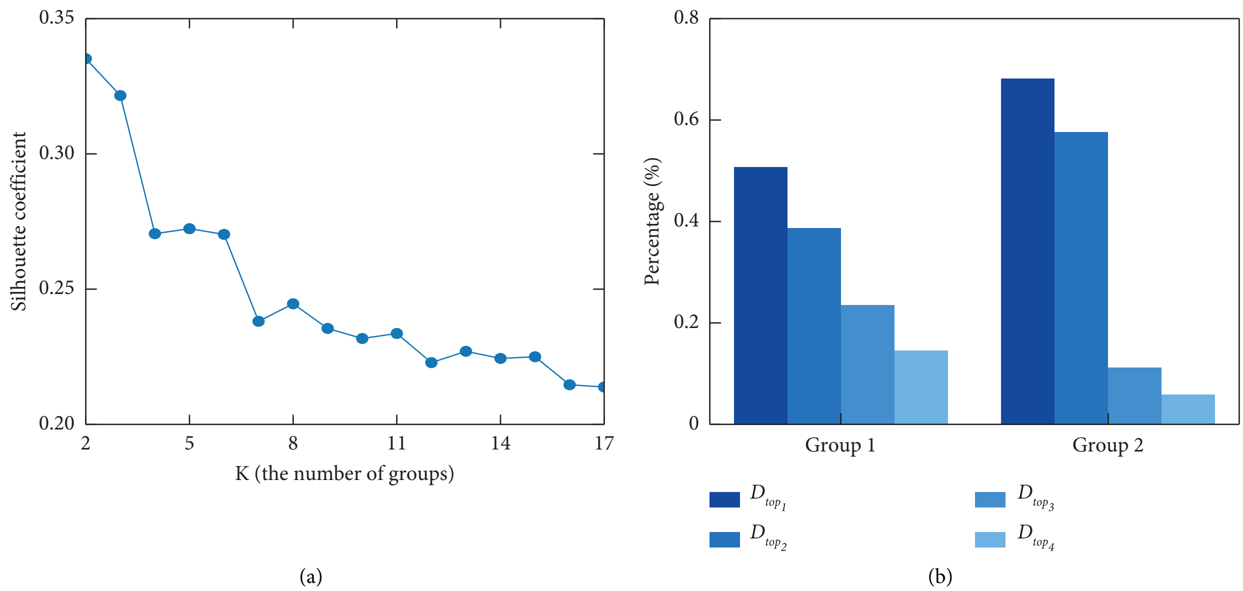


Figure 5: (a) The silhouette coefficient under different number of passenger groups. (b) For each group of passengers, the median proportion of active days with trips $D_{\text{top}_i}$ during the top four nonoverlap slots.

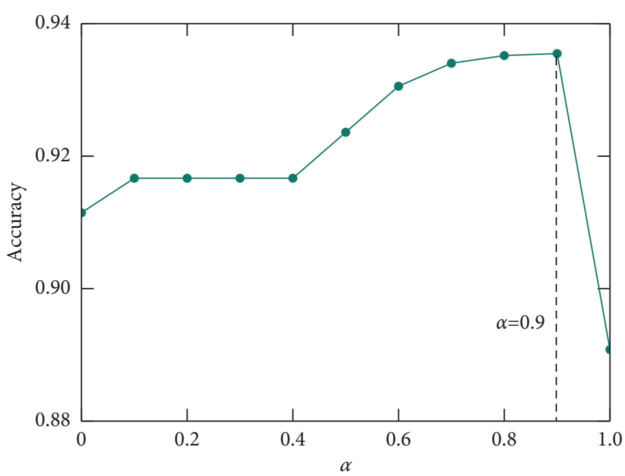


Figure 6: The median of prediction accuracy rates under different settings of weight index $\alpha$.

The effectiveness of the SCMM model is validated from the following three aspects: (1) prediction of individual mobility, (2) prediction of out-passenger flow $f_{\text{out}}$ at the crowding station, and (3) prediction of the spatial distribution of passenger sources during the large crowding events. Here, the random forest (RF) model, the Markov chain (MC) model, and the Markov chain model with station-level information (SMM) are used for comparison. Mobility patterns predicted using the RF model are pretty active among the benchmark models, and passengers are predicted to make trips in possible time slots. Therefore, the median of prediction accuracy rates of RF model is the lowest on ordinary days (Table 2). The median of prediction accuracy rates of the MC model is 90.1%, implying that MC model can predict individual mobility pretty well. The median of prediction accuracy rates of the SMM model in ordinary days is the same as that of the MC model (i.e., 90.6%), while the median of prediction accuracy rates in event days increases from 90.1% to 91.5%. The results imply that the proposed anomalous mobility index can well distinguish crowding events from ordinary conditions and capture the anomalous mobility patterns on event days. Moreover, by taking the collective mobility patterns of similar passengers into consideration, the SCMM model can further improve the median of prediction accuracy rate to 93.2% (Table 2). A possible explanation is that a passenger's willingness to explore a new place can to some extent be captured by the collective mobility patterns of his/her similar passengers.

Table 2: The median of prediction accuracy rates of different models.

| Model | All days (%) | Ordinary days (Dec. 25–Dec. 30) (%) | Event days (Dec. 24 and Dec. 31) (%) |
|---|---|---|---|
| RF | 88.9 | 89.2 | 91.7 |
| MC | 90.1 | 90.6 | 90.1 |
| SMM | 90.1 | 90.6 | 91.5 |
| SCMM | 93.2 | 93.2 | 93.4 |

The performance of the proposed SCMM model is also tested on different groups of passengers. Group 2 passengers are featured with the highest prediction accuracy (the median of prediction accuracy rates is 94.3%). This can be explained by the fact that a majority of passengers in Group 2 are commuters who make routine commuting trips on weekdays [50]. However, given that passengers in Group 1 are the most active passengers, their mobility is more difficult to predict. The median of prediction accuracy rates of passengers in Group 1 decreases to 91.4%.

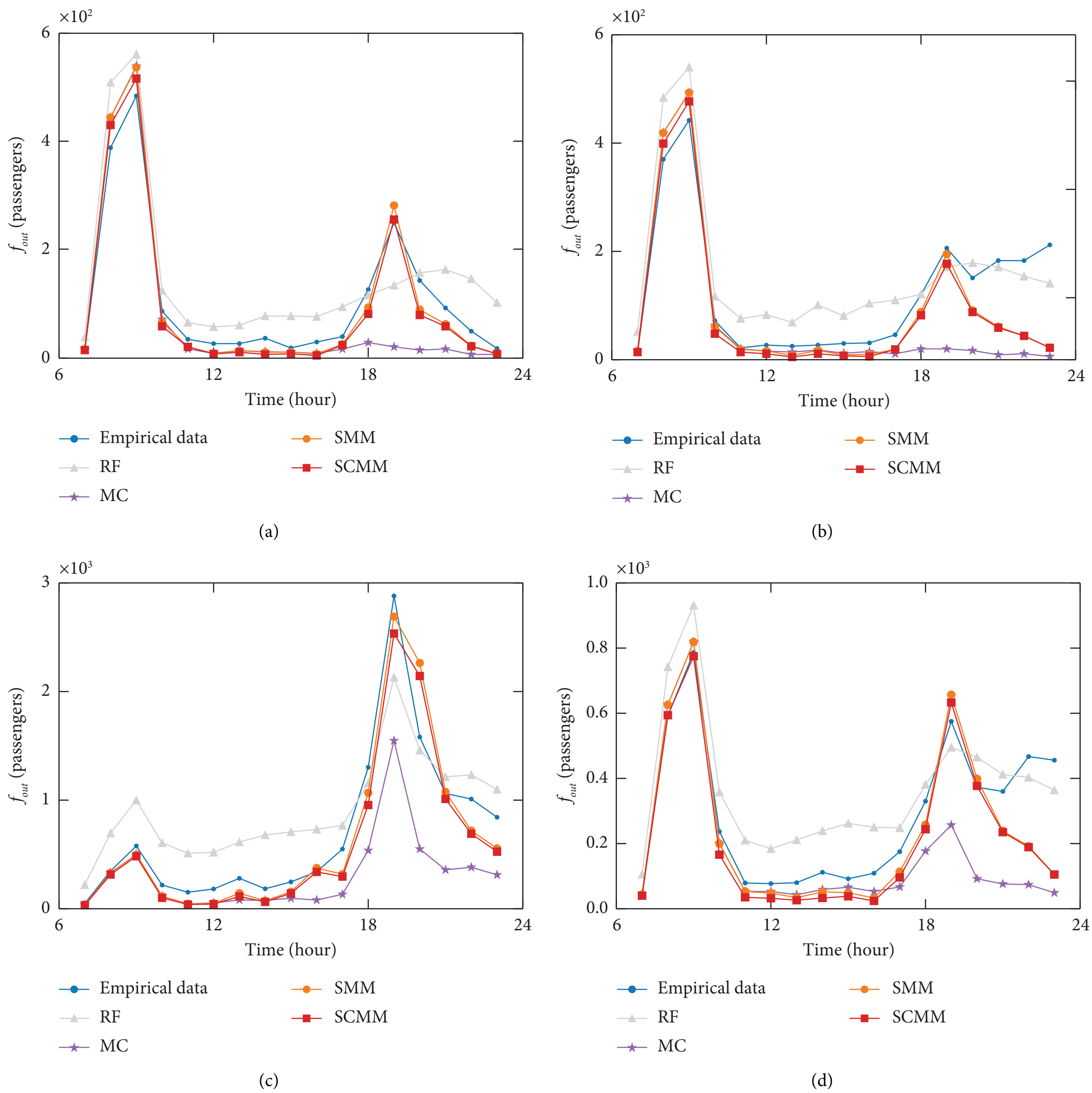


FIGURE 7: The out-passenger flow $f_{\text{out}}$ at the crowding stations. (a) The out-passenger flow of Sea World station on December 24. (b) The out-passenger flow of Sea World station on December 31. (c) The out-passenger flow of Window of world station on December 31. (d) The out-passenger flow of Qiaocheng East station on December 31.

The out-passenger flow $f_{\text{out}}$ at the crowding station is an important index for crowd management and safety control. Using the SCMM model, we can predict the anomalous mobility index and the out-passenger flow at the crowding stations. As Figure 7 shows, the out-passenger flow at crowding stations $f_{\text{out}}$ started to increase at 4:00 p.m. and reached its peak value at 7:00 p.m. in the studied crowding events. We find that the mobility patterns predicted using the RF model are too active to predict ordinary passenger flows, and the MC model is unable to predict the anomalous passenger flow in large crowding events. However, by introducing the anomalous mobility index, the SMM model and the SCMM model can well reproduce the anomalous growth of passenger flow. Then, we can close the crowding station or adjust train operation schemes [54] to prevent crowds from entering the overly crowded area and protect the safety of event participants.

The performances of the four models are further evaluated quantitatively using the mean absolute percentage

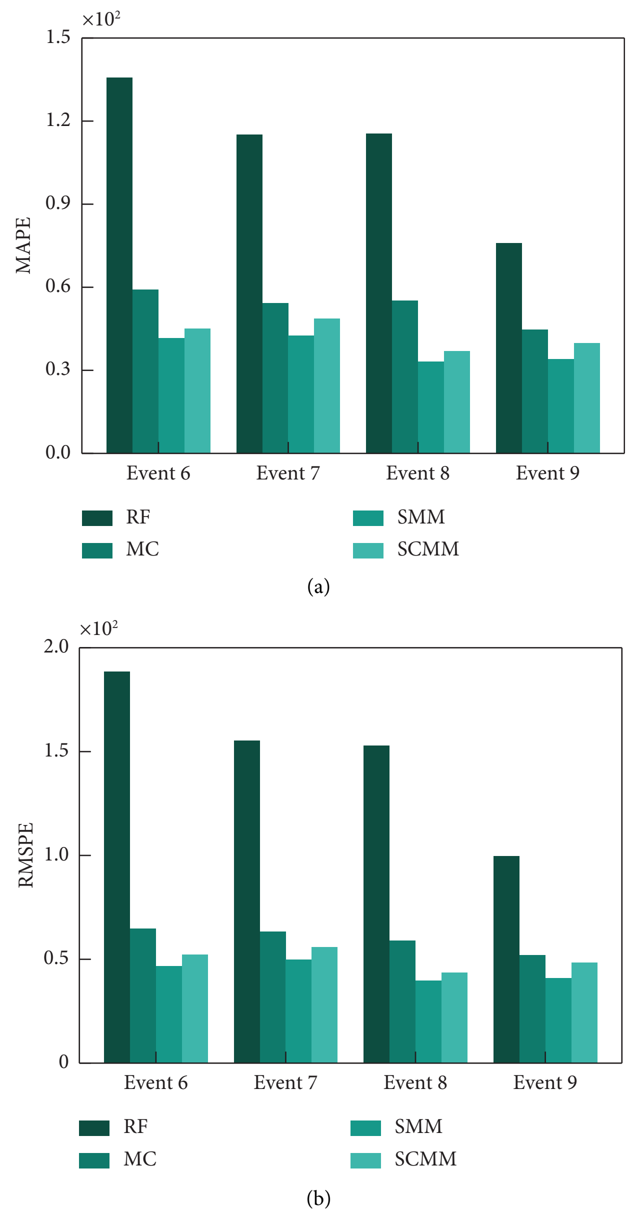


Figure 8: Performance on predicting the out-passenger flow at the crowding stations. (a) MAPE. (b) RMSPE.

error (MAPE) and the root mean square percentage error (RMSPE):

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$
$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \times 100\%, \tag{10}$$

where $n$ is the number of time slots, $y_i$ is the actual out-passenger flow $f_{\text{out}}$ in the $i^{\text{th}}$ time slot, and $\hat{y}_i$ is the predicted out-passenger flow $\hat{f}_{\text{out}}$ in the $i^{\text{th}}$ time slot. Figure 8 shows that the MAPE and the RMSPE of the RF model are much larger than the MAPEs and the RMSPEs of other models, indicating that the RF model cannot well predict the anomalous passenger flows at the crowding stations. The MAPEs and the RMSPEs of the SMM model and the SCMM model are smaller than the MAPEs and the RMSPEs of the RF model and the MC model, indicating that the SMM model and the SCMM model can well capture the anomalous passenger flows during the crowding events (Figure 8).

Next, we apply the four models to predict the spatial distribution of the passenger sources of the crowding station(where the passengers started their trips to the crowding station). Taking crowding event 7 as an example, according to our analysis of the empirical data, the sources of passengers were widely distributed in the city at 6:00 p.m. on the event day, covering most of the stations in the subway network (Figures 9(a) and 9(b)). Meanwhile, passengers

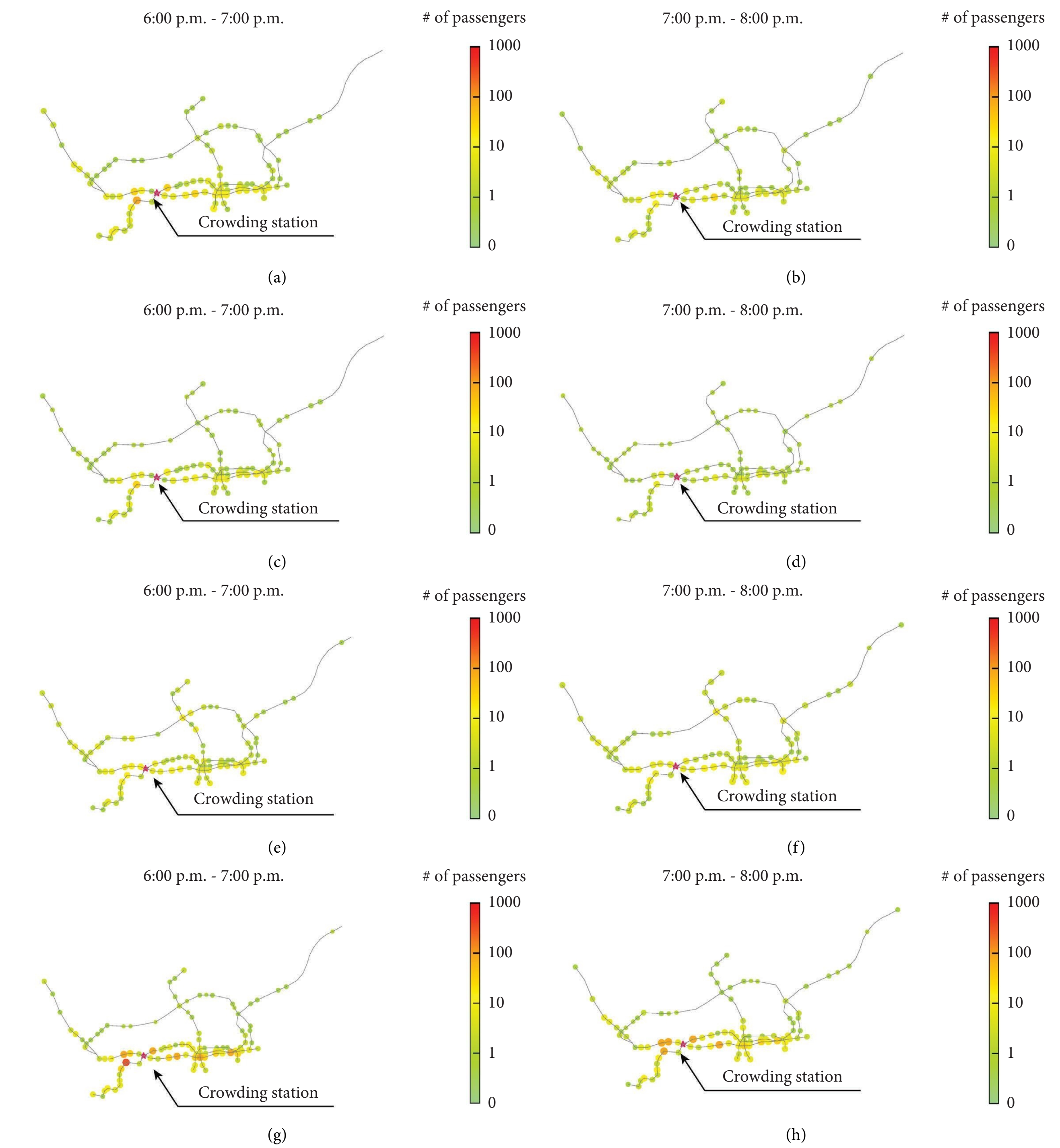


Figure 9: The spatial distribution of passenger sources at the crowding station. (a, b) Results obtained from the empirical data; (c, d) prediction of the MC model; (e, f) prediction of the RF model; (g, h) prediction of the SCMM model.

mainly came from the subway stations near the crowding station. At 7:00 p.m., the number of passenger sources considerably decreased. Figures 9(g) and 9(h) show the distribution of passenger sources and the number of passengers from each source predicted using the SCMM model, both of which are highly consistent with the empirical results shown in Figures 9(a) and 9(b). However, we find that the MC model (Figures 9(c) and 9(d)) and the RF model (Figures 9(e) and 9(f)) fail to capture the distribution of passenger sources on the event day.

## 5. Conclusions

Considering the difficulty of the MC model in predicting individual mobility under anomalous mobility situations, we combine the MC model with station-level passenger flow information and the collective mobility patterns of similar passengers to develop the SCMM model. The proposed SCMM model has two advantages. First, the anomalous mobility index captures the attractiveness of a station to passengers, which helps reproduce the crowd gathering mobility patterns during large crowding events. Second, the collective mobility patterns of similar passengers are employed to predict an individual's location in the next time slot, which further improves the prediction accuracy. Our results highlight the importance of combining an individual's own historical mobility information with station-level anomalous passenger flow information, which could be the key ingredient for predicting individual mobility in anomalous passenger flow conditions. Moreover, our methods suggest the appropriate weights of individual mobility information and collective mobility information used in the SCMM model, which could provide useful insights for future individual mobility modeling. Finally, the out-passenger flow at the crowding station and the passenger source distribution can be well predicted using the proposed SCMM model, which further validates the effectiveness of the model in predicting individual mobility under ordinary and anomalous passenger flow situations.

Given that the majority of event participants usually come to the crowding events by taking the subway [30], only subway passengers are considered in this study. Our future work will focus on incorporating multiple types of mobility data into individual mobility analysis and prediction. For instance, taxi GPS data and bus smart card data are potential data sources which can be further incorporated into the present modeling framework. In addition, the proposed SCMM model provides a general framework for individual mobility prediction and could be extended to other non-ordinary situations, such as extreme bad weather and interruptions of public transit systems.

## Data Availability

The subway smart card data and network data used to support the findings of this study have not been made available because of the confidentiality agreement.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

P. W. is supported by the Hunan Provincial Natural Science Fund for Distinguished Young Scholars (grant no. 2022JJ10077), the National Natural Science Foundation of China (grant no. 71871224), and the Science and Technology Progress and Innovation Plan of Department of Transportation of Hunan Province (grant no. 202102).

## References

[1] P. Wang, "Bridging human mobility and urban growth," *Nature Computational Science*, vol. 1, no. 12, pp. 778-779, 2021.

[2] K. Jin, W. Wang, X. Li, X. Hua, S. Chen, and S. Qin, "Identifying the critical road combination in urban roads network under multiple disruption scenarios," *Physica A: Statistical Mechanics and Its Applications*, vol. 607, Article ID 128192, 2022.

[3] P. Wang, J. Lai, Z. Huang, Q. Tan, and T. Lin, "Estimating traffic flow in large road networks based on multi-source traffic data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5672–5683, 2021.

[4] P. Wang, Z. Huang, J. Lai, Z. Zheng, Y. Liu, and T. Lin, "Traffic speed estimation based on multi-source GPS data and mixture model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 1–13, 2021.

[5] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[6] M. C. González, C. A. Hidalgo, and A. L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[7] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located Twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science*, vol. 41, no. 3, pp. 260–271, 2014.

[8] J. Tang, C. Zhao, F. Liu, W. Hao, and F. Gao, "Analyzing travel destinations distribution using large-scaled GPS trajectories: a spatio-temporal Log-Gaussian Cox process," *Physica A: Statistical Mechanics and Its Applications*, vol. 599, Article ID 127305, 2022.

[9] 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.

[10] R. Wang, N. Li, and Y. Wang, "Does the returners and explorers dichotomy in urban human mobility depend on the observation duration? An empirical study in Guangzhou, China," *Sustainable Cities and Society*, vol. 69, Article ID 102862, 2021.

[11] C. Song, T. Koren, P. Wang, and A. L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[12] X. Y. Yan, W. X. Wang, Z. Y. Gao, and Y. C. Lai, "Universal model of individual and population mobility on diverse spatial scales," *Nature Communications*, vol. 8, no. 1, pp. 1639–9, 2017.

[13] X. Y. Yan, C. Zhao, Y. Fan, Z. Di, and W. X. Wang, "Universal predictability of mobility patterns in cities," *Journal of The Royal Society Interface*, vol. 11, pp. 20140834–20141100, 2014.

[14] F. Xu, Y. Li, D. Jin, J. Lu, and C. Song, "Emergence of urban growth patterns from human mobility behavior," *Nature Computational Science*, vol. 1, no. 12, pp. 791–800, 2021.

[15] M. Nogal, A. O'Connor, B. Caulfield, and B. Martinez-Pastor, "Resilience of traffic networks: from perturbation to recovery via a dynamic restricted equilibrium model," *Reliability Engineering and System Safety*, vol. 156, pp. 84–96, 2016.

[16] B. Guo, M. Li, M. Zhou, F. Zhang, and P. Wang, "A new anomalous travel demand prediction method combining Markov model and complex network model," *Physica A:*

*Statistical Mechanics and Its Applications*, vol. 619, Article ID 128697, 2023.

[17] S. Maity and S. Sundar, "A coupled model for macroscopic behavior of crowd in flood induced evacuation," *Physica A: Statistical Mechanics and Its Applications*, vol. 607, Article ID 128161, 2022.

[18] B. M. Pastor, M. Nogal, M. Nogal, N. Connor, and R. Teixeira, "Transport network resilience: a mapping and sensitivity analysis strategy to improve the decision-making process during extreme weather events," *International Journal of Critical Infrastructures*, vol. 17, no. 4, pp. 330–352, 2021.

[19] J. H. Cho, D. K. Kim, and E. J. Kim, "Multi-scale causality analysis between COVID-19 cases and mobility level using ensemble empirical mode decomposition and causal decomposition," *Physica A: Statistical Mechanics and Its Applications*, vol. 600, Article ID 127488, 2022.

[20] P. Liu and Y. Zheng, "Temporal and spatial evolution of the distribution related to the number of COVID-19 pandemic," *Physica A: Statistical Mechanics and Its Applications*, vol. 603, Article ID 127837, 2022.

[21] H. Zhou, Z. Zheng, X. Cen, Z. Huang, and P. Wang, "A data-driven urban metro management approach for crowd density control," *Journal of Advanced Transportation*, vol. 2021, Article ID 6675605, 14 pages, 2021.

[22] L. Feng and E. Miller-Hooks, "A network optimization-based approach for crowd management in large public gatherings," *Transportation Research Part C: Emerging Technologies*, vol. 42, pp. 182–199, 2014.

[23] Z. Huang, P. Wang, F. Zhang, J. Gao, and M. Schich, "A mobility network approach to identify and anticipate large crowd gatherings," *Transportation Research Part B: Methodological*, vol. 114, pp. 147–170, 2018.

[24] S. Gambs, M. O. Killijian, and M. N. Del Prado Cortez, "Next place prediction using mobility Markov chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, pp. 1–6, Bern, Switzerland, April 2012.

[25] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 911–918, Pittsburgh, PA, USA, September 2012.

[26] B. Hawelka, I. Sitko, P. Kazakopoulos, and E. Beinat, "Collective prediction of individual mobility traces for users with short data history," *PLoS One*, vol. 12, no. 1, pp. e0170907–e0170914, 2017.

[27] F. Alhasoun, M. Alhazzani, and M. C. González, "City scale next place prediction from sparse data through similar strangers," *Knowledge Discovery and Data Mining*, vol. 8, no. 8, 2017.

[28] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 19–34, 2018.

[29] B. Guo, H. Yang, F. Zhang, and P. Wang, "A hierarchical passenger mobility prediction model applicable to large crowding events," *Journal of Advanced Transportation*, vol. 2022, Article ID 7096153, 12 pages, 2022.

[30] Z. Huang, X. Ling, P. Wang et al., "Modeling real-time human mobility based on mobile phone and transportation data fusion," *Transportation Research Part C: Emerging Technologies*, vol. 96, pp. 251–269, 2018.

[31] N. Eagle and A. S. Pentland, "Eigenbehaviors: identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 11, pp. 1689–1066, 2009.

[32] J. Zheng and L. M. Ni, "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 153–162, Pittsburgh, PL, USA, September 2012.

[33] F. Li, Z. Gui, Z. Zhang et al., "A hierarchical temporal attention-based LSTM encoder-decoder model for individual mobility prediction," *Neurocomputing*, vol. 403, pp. 153–166, 2020.

[34] A. Al-Molegi, M. Jabreel, B. Ghaleb, and Stf-Rnn, "Space time features-based recurrent neural network for predicting people next location," in *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Athens, Greece, December 2016.

[35] H. P. Hsieh, C. T. Li, X. Gao, and T-gram, "T-gram: a time-aware language model to predict human mobility," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 1, pp. 614–617, 2021.

[36] B. Mo, Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction in mass transit systems using smart card data: an interpretable activity-based hidden Markov approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 1–13, 2021.

[37] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, pp. 312–317, Funchal, Portugal, September 2010.

[38] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 25–33, Chicago, Il, USA, November 2011.

[39] C. Yang, F. Yan, and S. V. Ukkusuri, "Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system," *Transportmetrica: Transportation Science*, vol. 14, no. 7, pp. 576–597, 2018.

[40] X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 Haiti earthquake," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 29, pp. 11576–11581, 2012.

[41] F. Zhang, Z. Li, N. Li, and D. Fang, "Assessment of urban human mobility perturbation under extreme weather events: a case study in Nanjing, China," *Sustainable Cities and Society*, vol. 50, Article ID 101671, 2019.

[42] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Using data from the web to predict public transport arrivals under special events scenarios," *Journal of Intelligent Transportation Systems*, vol. 19, no. 3, pp. 273–288, 2015.

[43] F. Rodrigues, S. S. Borysov, B. Ribeiro, and F. C. Pereira, "A bayesian additive model for understanding public transport usage in special events," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2113–2126, 2017.

[44] X. Ni, Z. Sun, Y. Gu, H. Cui, and H. Xia, "Assembly of a novel biosynthetic pathway for gentamicin B production in Micromonospora echinospora," *Microbial Cell Factories*, vol. 15, no. 6, pp. 1–10, 2016.

[45] Y. Li, X. Wang, S. Sun, X. Ma, and G. Lu, "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 306–328, 2017.

[46] Z. Zheng, X. Ling, P. Wang, J. Xiao, and F. Zhang, "Hybrid model for predicting anomalous large passenger flow in urban metros," *IET Intelligent Transport Systems*, vol. 14, no. 14, pp. 1987–1996, 2020.

[47] Z. Cheng, M. Trépanier, and L. Sun, "Incorporating travel behavior regularity into passenger flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 128, Article ID 103200, 2021.

[48] G. Goulet Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 1–16, 2016.

[49] B. Guo, H. Yang, H. Zhou et al., "Understanding individual and collective human mobility patterns in twelve crowding events occurred in Shenzhen," *Sustainable Cities and Society*, vol. 81, Article ID 103856, 2022.

[50] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3135–3146, 2017.

[51] P. J. Rousseeuw and Silhouettes, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[52] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, vol. 3, pp. 2923–2929, 2013.

[53] T. E. Huillet, "On a Markov chain model for population growth subject to rare catastrophic events," *Physica A: Statistical Mechanics and Its Applications*, vol. 390, no. 23–24, pp. 4073–4086, 2011.

[54] Y. Li, X. Yang, J. Wu, H. Sun, X. Guo, and L. Zhou, "Discrete-event simulations for metro train operation under emergencies: a multi-agent based model with parallel computing," *Physica A: Statistical Mechanics and Its Applications*, vol. 573, Article ID 125964, 2021.