Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 9401630, 12 pages
https://doi.org/10.1155/2019/9401630

Research Article

Data-Driven Prediction System of Dynamic People-Flow in Large Urban Network Using Cellular Probe Data

Xiaoxuan Chen,[1] **Xia Wan**,[2] **Fan Ding**,[3] **Qing Li**,[4] **Charlie McCarthy**,[5] **Yang Cheng**,[6] **and Bin Ran**[7]

[1]*Ford Motor Company, 22000 Michigan Ave, Dearborn, MI 48124, USA*
[2]*GlobalFoundries, 400 Stone Break Rd Extension, Malta, NY 12020, USA*
[3]*TOPS Laboratory, University of Wisconsin-Madison, 1415 Engineering Drive, Room 1217, Madison, WI 53706, USA*
[4]*BMW Technology Inc., 540 W Madison St Suite 2400, Chicago, IL 60661, USA*
[5]*TranSmart Technologies Inc., 411 S Wells St, Chicago, IL 60607, USA*
[6]*TOPS Laboratory, University of Wisconsin-Madison, 1415 Engineering Drive, Room 1249A, Madison, WI 53706, USA*
[7]*TOPS Laboratory, Department of Civil and Environmental Engineering, University of Wisconsin-Madison, USA*

Correspondence should be addressed to Xiaoxuan Chen; xchen324@wisc.edu

Received 10 October 2018; Accepted 10 December 2018; Published 13 January 2019

Academic Editor: Yair Wiseman

Cellular probe data, which is collected by cellular network operators, has emerged as a critical data source for human-trace inference in large-scale urban areas. However, because cellular probe data of individual mobile phone users is temporally and spatially sparse (unlike GPS data), few studies predicted people-flow using cellular probe data in real-time. In addition, it is hard to validate the prediction method at a large scale. This paper proposed a data-driven method for dynamic people-flow prediction, which contains four models. The first model is a cellular probe data preprocessing module, which removes the inaccurate and duplicated records of cellular data. The second module is a grid-based data transformation and data integration module, which is proposed to integrate multiple data sources, including transportation network data, point-of-interest data, and people movement inferred from real-time cellular probe data. The third module is a trip-chain based human-daily-trajectory generation module, which provides the base dataset for data-driven model validation. The fourth module is for dynamic people-flow prediction, which is developed based on an online inferring machine-learning model (random forest). The feasibility of dynamic people-flow prediction using real-time cellular probe data is investigated. The experimental result shows that the proposed people-flow prediction system could provide prediction precision of 76.8% and 70% for outbound and inbound people, respectively. This is much higher than the single-feature model, which provides prediction precision around 50%.

1. Introduction

Dynamic people-flow in this paper refers to the estimated number of people moving into or out of a zone, which reflects the real-time travel demand. Due to the trend of increasing urbanization shifts, people-flow monitoring data has become an essential source of information for decision-making in urban planning, urban disaster and emergency management, and urban roadway operations. More broadly, dynamic people-flow can provide critical decision-making insights applicable to all industries, such as targeting a specific audience for advertisements or selection of optimal store location.

Traditionally, urban people-flow is estimated using a 4-step method based on survey data, which is both labor and capital intensive, and also gets updated infrequently. Some studies processed video data collected from single or multiple closed-circuit television (CCTV) cameras, which

provide high accuracy people-flows in real-time. However, this method is impractical to apply to a large network because of the infrequency of installed surveillance cameras. Some other studies using passive data collection methods feature GPS, Bluetooth, or a social media network. However, they are suffering due to sample size limitations, and the bias of user attributes.

Cellular network operators collect cellular probe data of mobile phone users daily. In recent years, thanks to the rapid development of cellular communication technology, most people in developed and developing countries own mobile phones. For instance, as of June 2016, 77.3% of the total population in China owns at least one mobile phone [1]. The anonymous mobile phone traces record the location of mobile phone users when they text, call, connect to the Internet, and even passively when the mobile phone communications to the cellular network. This provides an opportunity to study and monitor human activity using cellular data, but there are still some issues with deriving people-flow from cellular data.

The first issue is that the update frequency of the mobile phone user's location relies on the mobile phone activity frequency, which is not uniformly distributed temporally or spatially. Temporally, people usually use their mobile phones more frequently during the day than during the night. Spatially, some people use their mobile phones more frequently near work while others use their mobile phones more frequently at home. Therefore, it is hard to use a statistical method to estimate the real-time people-flow based on the people movement detected by cellular probe data. The second issue is the efficiency of data processing and model calibration, since the cellular probe dataset is extremely large.

To address the issues above, a machine-learning based data-driven system is designed to predict the grid-based inbound/outbound people-flow. The study area is divided into square grids, which integrate multiple data sources as the input features for the machine-learning model. The inbound/outbound flow of each grid is estimated with real-time cellular data that is aggregated into 5-minute increments as the real-time people movement feature. To calculate the model, individual trajectories were inferred by a trip-chain model and integrated 5-minute people-flow for each grid. Random forest method is used in the data-driven system result from the performance in processing a large dataset. The proposed data-driven system predicts the inbound/outbound people-flow of each grid for 30-minutes into the future.

The rest of this paper is organized as follows. Section 2 reviews the current state of analysis for people-flow estimation, the existing studies on cellular probe modeling, and recent studies on data-driven methods using passively collected data. Section 3 presents the methodology for this paper, including the grid-based data integration model, the trip-chain based individual trajectory inferring model, and the machine-learning based data-driven model. Section 4 presents a case study on a real network in a large-scale urban area. Section 5 summarizes this paper and discusses the result.

2. Related Work

2.1. People-Flow Estimation. Conventional methods for people-flow estimation are usually derived from data collected by survey, roadside detectors, surveillance video, and other passive data collection methods. The conventional travel demand between each pair of traffic analysis zone is inferred from the city Original-Destination matrices, which are estimated from the citywide survey. The survey is usually expensive and updated only once every five years. A classic four-step regional survey forecasting model is able to estimate and predict the people-flow at a large scale [2, 3]. Beside the survey data, recent studies showed that there are several others methods capable of deriving OD matrices from emerging passive data collection methods, such as traffic count data, vehicle plant matching data, GSP data, and social media data [4–9]. The traffic count data and vehicle plant matching data rely on the data collection infrastructure, which is costly, requires maintenance, and usually specific to freeway networks. The GPS-based OD derivation method has lower cost and higher accuracy, but suffers from issues including limited sample size and coverage area, sampling bias, and privacy concerns, which is why it is not widely used for OD estimation [10, 11]. The social media service as a data source for human activity studies also suffers from the sample size and sample bias issues [12]. In a study of Origin-destination demand in a large-scale network, the real-time OD demand is estimated and predicted with a data-driven method using real-time demand data in Korea. Three strategies of implementing the features for the k-nearest neighbor algorithm are compared and presented [13]. Cellular data is also widely used in the field of trip distribution estimation, traffic state estimation, and traffic flow monitoring in freeway networks [14–17].

2.2. Data-Driven Approaches. In the last 20 years, the data-driven approach has been applied to the field of intelligent transportation system (ITS) and improved the efficiency and performance of ITS [18]. The data-driven approach refers to the algorithms which are compelled by data, rather than the model driven method. It solves the problem progression in an algorithm compelled by data, while the traditional methods depend on human experiences and historical data. Taking advantage of the widely deployed ITS sensors and multiple real-time enabled data sources for individuals, vehicles, and roadway networks, the real-time data-driven based ITS system would improve the accuracy and efficiency of conventional ITS systems [19]. The method has been widely used in many subjects of current ITS systems. Some of the studies work on the short-term travel time prediction on freeway networks using speed and traffic count data [20, 21]. The data-driven based dynamic simulation approaches have been studied using real-time traffic data to estimate roadway traffic volumes across various time intervals [22]. A dynamic data-driven approach is applied to the surface transportation system [23]. Benefiting from increasing data volumes and computing power, the data-driven approach has been widely applied on transportation systems.

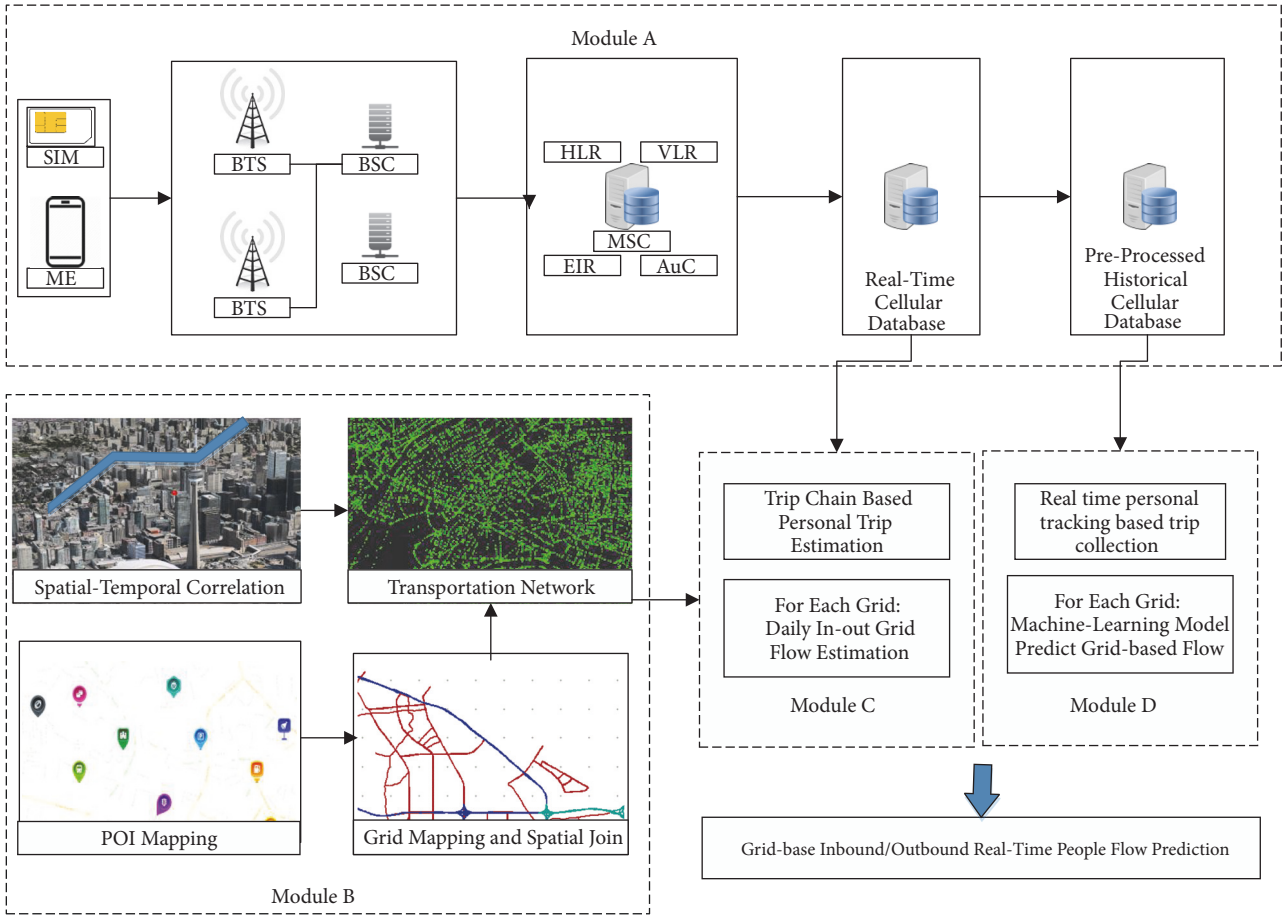


Figure 1: System architecture of people-flow prediction system.

In summary, the studies applying the data-driven approach in the field of transportation have been focused on travel time prediction, traffic statue monitoring, and travel demand estimation. Few studies have investigated the feasibility of applying the data-driven approach for people-flow prediction to a large-scale area for real-time service using cellular signaling data.

3. Dynamic People-Flow Prediction Framework

3.1. System Architecture. This paper described a data-driven based online people-flow prediction system, as shown in Figure 1. The system contains four modules.

Module a: Cellular Probe Data Preprocessing Module. This module processes the real-time cellular data and stores the preprocessed cellular data.

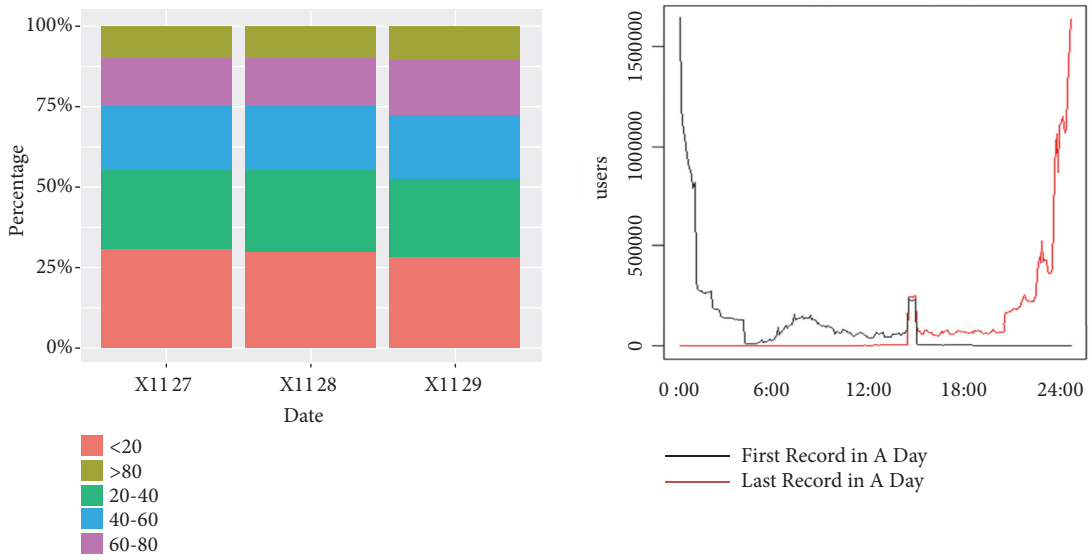
Module b: Grid-Based Data Transformation and Integration Module. This module integrates the multiple data sources as the attributes of each grid. Input features are generated to train the machine-learning model in module d.

Module c: Trip-Chain Based Human-Daily-Trajectory Inferring Module. This module provides the daily trajectories of each mobile subscriber. By integrating the trajectories, the people-flow (inbound/outbound) of the grids could be estimated as the labels for the machine-learning model in module 4.

Module d: Machine-Learning Based Online People-Flow Prediction Module. This module uses a random forest model for offline learning using the input feature from module b and input label from module c. Real-time cellular data is the input of the online prediction model.

3.2. Cellular Probe Data Preprocessing Module

3.2.1. Cellular Probe Data. Cellular network operators collect the location of cellular network subscribers for the billing and operational purposes. The location is not a highly accurate user location but a virtual location represented by the user-connected base station (BS). Each BS has a corresponding coordinate and a unique combination of cell identification code (CI) of BS and location area code (LAC) of the connected location area. The cellular data will be stored in the database by mobile switching center (MSC). For



(a) Distribution of daily cellular records

(b) First record and last record

FIGURE 2: Preliminary of cellular data temporal coverage.

each row of cellular data, it includes LAC, CI, timestamp, and event type. The GSM network signaling data of mobile subscribers is stored in a two-level hierarchy database, home location register (HLR) and visitor location register (VLR). Because of the nature of cellular phone communication, the preprocessing algorithms should be applied to generate more accurate data with fewer redundancies.

Because user location is based on the location of cellular records, the update frequency and temporal coverage of the cellular data is critical in this study. Based on sample data of Shanghai from one of the major cellular carriers of China, in Figure 2(a), more than 75% users have 20 or more records per day. Figure 2(b) shows the time for each user's first and last record. It shows that most of the users have the first record earlier than 6 AM and last record later than 10 PM.

The event of cellular probe data includes two basic types: ***Location Update*** (LU) and ***Handover*** (HO). The LA processing can be triggered in the flowing conditions: mobile phone is on, the mobile phone moves from one location area to the other location area, and a periodic location update occurs generally once per hour. Handover (HO) is triggered when the mobile phone is in communication status and travels from one BS to an adjacent BS. Both BSs are recorded in the cellular signaling database when HO is triggered. Thus, when a mobile phone user makes a phone call and has a trip through several base stations, a series of timestamped with estimated locations will be tracked. Each mobile phone user has a unique mobile station ID (MSID).

3.2.2. Cellular Probe Preprocessing. Based on the attributes of cellular phone data, the raw data will generate whether the mobile phone is moving or stationary. In the data-driven system, the quality of input cellular probe raw data is critical. There are three types of errors defined below, that will be processed in this module.

Definition 1 (duplicated data). Duplicated data is three or more pieces of continuous raw data with same MSID, Cell ID, and LAC. The processing procedure shows in Box 1.

Definition 2 (Ping-Pong switching). When the mobile phone moves to the edge of the cellular coverage area, the connection to the current BS becomes weak as the signal from the adjacent BS grows stronger. In this case, the mobile phone will terminate the connection to the current BS to connect the new BS. However, the signal attenuation and the BS-cellphone distance are not linearly changing. So, at the adjacent boundary of the two cellular coverage areas, the mobile phone may be covered by multiple BSs, with the signal intensities of each BS being similar. In this case, the cellular phone may switch the connection between two BSs even if it is stationary, shown in Box 1.

Definition 3 (drift switching processing). Occasionally, during the current process of cellular data, the mobile phone can switch to a BS which is very far from the previous BS and then switch to another BS near the first BS. The reasons that drift switching gets triggered are complex and unpredictable. The major reasons of drift switching are BS signal blocking and unstable antenna environment, shown in Box 1.

$$\Delta d_i = Arc\cos\{\sin(Lat_i) * \sin(Lat_{i+1}) * \cos(Lon_i - Lon_{i+1}) + \cos(Lat_i) * \cos(Lat_{t+1})\} * R * \frac{Pi}{180} \quad (1)$$

$$v_i = \frac{\Delta d_i}{\Delta t_i} \quad (2)$$

where Δd = distance between two points,

Algorithm1: Redundant data processing (Data ordered by MSID, Timestamp)
t: timestamp in second; i = order number of raw data;
for all MSID u
  If u.i.t = u.(i+1).t
    Delete u.i;
  end if
end for
If u.i.lat = u.(i+1).lat && u.i.lon = u.(i+1).lon && u.i.lat = u.(i-1).lat && u.i.lon = u.(I - 1).lon
  Delete u.i;
end if
Algorithm2: Ping-Pong switching processing (Data ordered by MSID, Timestamp)
for MSID u
  if —u.(i-1).t - u.(i+1).t— <T && u.(i-1).lat = u.(i+1).lat && u.(i-1).lon = u.(i+1).lon
    Delete row: u.i;
  end if
end for
Algorithm3: Drift switching processing (Data ordered by MSID, Timestamp)
for MSID u
  if —u.(i-1).t - u.(i+1).t— <T && $\Delta d/\Delta t > V$ && $\Delta d > D$
    Delete row: u.i
  end if
end for

Box 1: Cellular raw data preprocessing algorithm.

Δt =time interval between two records,

Lat, Lon = the latitude and Longitude of row i of MSID u,

R = earth radius.

Box 1. shows three data preprocessing algorithms. The preprocessed data is the input of module 3 and module 4.

3.3. Feature Integration Module. The data-driven system predicts the fine graded-based inbound/outbound people-flow for real-time service. The module input datasets are multiple grid attributes and the cellular raw data. The module output is the generated features for the data-driven model. It is because the population flow patterns for each particular area are highly related to the attributes of that area. For instance, the subway line may have larger people-flow than the vacant area. In this paper, the study area is divided into squares, which represents the "grid" in this paper. The data sources are integrated into the grids in the study area.

3.3.1. Point of Interests (POI) Features (F_p). A point of interest is a specify location that serving a particular purpose, such as restaurants or hospitals. Each POI has a coordinate, a name, a category, and address. The POI categories in this study include hotel(P_1), school(P_2), government(P_3), bank (P_4), hospital (P_5), market and mall(P_6), restaurant(P_7), stadium(P_8), transportation hub(P_9), and factory(P_{10}). For each part of the grid, the number of POIs will be calculated by POI category.

3.3.2. Transportation Network Features (F_r). The transportation network in this paper refers to the major road network (R_1), light rail network (R_2), and the subway network (R_3). Since there is a strong correlation between a transportation network and the people-flow, the links of the transportation network are mapped on each part of the grid.

3.3.3. Temporal Features (F_t). Beside the grid-based features, there are some other features that may influence the dynamic changes in people-flow. There are two binary features in this study: peak hour (T_1), Work time (T_2), and night time (T3).

3.3.4. People Movement Level Collection from Real-Time Cellular Data (F_c). The real-time people movement in this study refers to the sequence of mobile phone user locations inferred from the cellular raw data in a 5-minute time interval. There are two major events of cellular signal transition event: (1) Location Update and (2) Hand Over should both be correlated with the grids. Due to the nature of cellular data, the hand over event can locate the mobile phone more accurately than the location update.

Location Update. Based on the attributes of raw cellular data, the coverage area of each cellular tower could be calculated by Voronoi graph. Spatial join analysis is used to calculate the percentage of the Voronoi graph mapping on each grid.

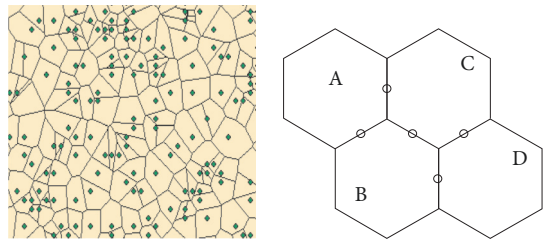
$$B_k = \left\{x \in \mathbf{X} \mid d(x, P_k) \leq d\left(x, P_j\right)\right\}; \quad j \neq k \tag{3}$$

where $\mathbf{X}$ = the study space, it is study area,

d = the distance function,

B_k = Voronoi area k,

P_j = The set associated with B_k.



(a) Voronoi map (b) Handover data to grid mapping

Figure 3: Grid mapping of cellular probe event.

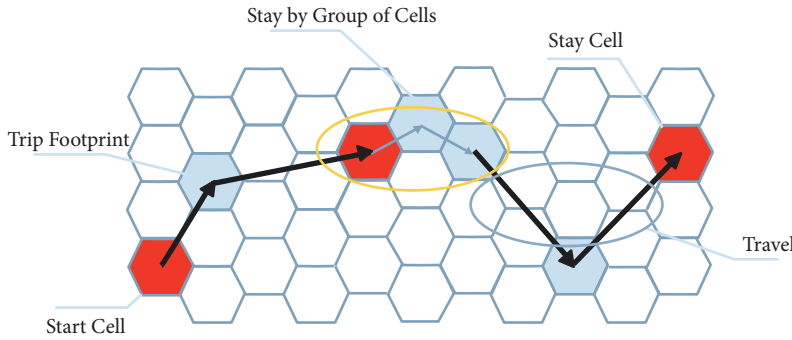


Figure 4: Illustration of individual activity inferring.

The result of the Voronoi in a study area is shown in Figure 3(a). Each dot is the location of BS. Each polygon around the dots is the cellular coverage area.

The contribution rate of each BS to the grid is shown in Figure 3(b). The hexagon represents the coverage area of BS. For grid A, the contribution rate is a%, which is calculated in

$$\text{Coverage percentage ratio} = \frac{\text{BS coverage area}}{\text{grid area}} \quad (4)$$

Hand Over. Hand over location should be the middle point between each pair of overlapping BS coverage areas. The coordinate of the middle point at the boundary of cell coverage areas is calculated as a handover point, which is shown in Figure 3(b). Each hexagon represents an estimated coverage area of a cell tower. Each dot in Figure 3(b) is the calculated HO location in the overlapping area. The calculated individual movements are aggregated in a 5-minute time interval for each cellular phone as the people movement feature.

3.4. Daily Individual Traveler Trajectory Estimation Module. This study uses a random forest (RF) model as the data-driven model. The features are the training data set, which were acquired in the previous section. In this section, the validation data set is calculated using the daily cellular data.

The proposed transportation mode shares driven model in this study is the combination of a trip-chain based microscopic mode choice model and a model transportation shares aggregating process. The mode choice decision of a mobile user within one day for every trip within is the output of the mode choice model at the individual level. The trip-chain based rules reflect the temporal-spatial and private vehicle usage constraints within one day. Then the mode choice results of the individual mobile phone users are aggregated with the characteristic to obtain the transportation mode shares at the macroscopic level. The daily individual trajectory should be inferred.

3.4.1. Inferring Individual Stays and Travels. A rule-based model is used for the home location detection and activity inferring. Figure 4 shows the stay and trips.

Home Location Detection. Mobile phone users are classified to the daytime-active users and the nighttime-active users to apply the home location detection process separately. If the user stays in a zone between 12:00 AM and 8:00 AM for sequential days, the user is classified as a daytime-active user. Otherwise, he/she is classified as a nighttime-active user. Then, the home location detection rules are set as follows: for the daytime-active user, the most frequently pinned station during 12:00 AM to 8:00 AM is set as the representative home location of the user; for the nighttime user, the most frequently pinned station between 8:00 AM and 12:00 AM is set as the representative home location of the user.

Activities Inferring. After getting the home location, the activities of the mobile phone user are extracted by inferring the Potential Stays. The location update data and phone bill data are both included in the following inferring process. A Potential Stay point is identified by a sequence of consecutive

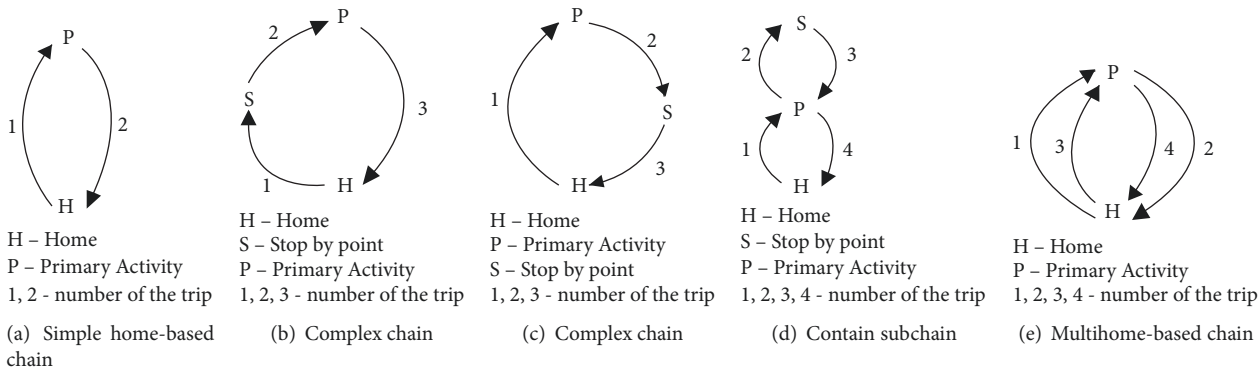


Figure 5: Typical trip-chains for urban travelers.

mobile phone records bounded by both spatial and temporal constraints as shown in Figure 4.

The spatial constraint is the roaming distance between the first and the last record in a stay location. The roaming distance should be related to the distance between base transceiver stations in that area. For example, the roaming distance in the Shanghai central area is set as 520 meters by considering that the average distance between the neighbor base transceiver stations is 260 meters in that area. The temporal constraint is the required minimum duration stay in a stay location. In this stay, only mobile pin records satisfy the spatial constraint, and duration greater than 30 minutes qualifies as a potential stay.

The activates of the mobile phone user are correspondingly extracted from the final stay detection results. The stay location and stay duration are imported as the feature of the day. If the daytime-active user stays in the same location for more than seven hours between 8:00 AM and 6:00 PM in a day, the stay location is marked as the work location, and the relative activity is marked as a work activity. The land use data could help with inferring the activity purpose.

Travel Detection. After the stay points are gradually detected, the connection between the stay points is the travel of the mobile phone user. The combination of the phone bill data and the location update data could record the movement trace between the origin and destination. There are two situations: (1) when the origin and destination location are different, the travel is the connection between the two activates; (2) when the origin and destination are the same locations, composing a trip-chain, the furthest pinned position is set as the Stop-By point for the trip. Then there are two travels for this connection. One is from the origin to the Stop-By point, and the other is from the Stop-By point to the destination. The travel distance and travel time are recorded based on the broken line connecting the sequential pinned position.

3.4.2. Extracting Trip-Chains. Trip-chain for the mobile phone users is composed in the previous section. With the activity and trip-chain theory, the typical trip-chain modes of the travelers are presented in Figure 5. The home-based trip-chains of mobile phone users within one day could be classified by: the simple home-based chain, complex chain, containing subchain, and multihome-based chain. In Figure 5(d), for example, trips 1 and 4 are the main chain and trips 2 and 3 are the subchain. If the travel of a user cannot compose a trip-chain, the travel is treated as a trip separately in the latter mode choice step.

3.4.3. Travel Model Detection and Trajectory Map Matching. In this study, the travel mode detection of the mobile phone user is at the trip-chain level. The mobile phone user's travels with a private vehicle and public vehicle are significantly different. The nonprivate vehicle user could change between nonprivate travel modes freely. Considering the rapidly growing usage of the private car in developing countries, the accuracy of the mode choice for the first home-based trip is critical for step 3. Two assumptions are made in this step.

Subway Trips. The subway mobile stations have been labeled as "subway station" or "underground lane" in the GSM network. Because each of the subway lines has a unique location update code, the mobile phone user will connect LAC of the current subway line. A mobile phone trajectory with subway mark T= (. The nearest path from could be generated. From the point to point, the subway link with the right subway line should be selected. If select multiple subway line, the nearest link should be elected as the starting point. From the subway network, the traveled trajectories could be inferred from the starting link and ending link.

Highway Trips. Map the individual trajectory on the highways. The highway solution is, if the trip-chain based travel mode selection flagged a trip as highway trips, the Dijkstra would be used to find the best highway-based routes. A set of possible routes is restricted to a corridor to estimate the area where the mobile phone subscribers would able to travel. A shape-file map of the study area which contains the roadway links and edge points

Nonhighway Trips. The trips are not on the freeway for nonmotor travel mode. In this case, the trajectory treated as a straight line. The starting point and the ending point of the straight line are the connected BS location.

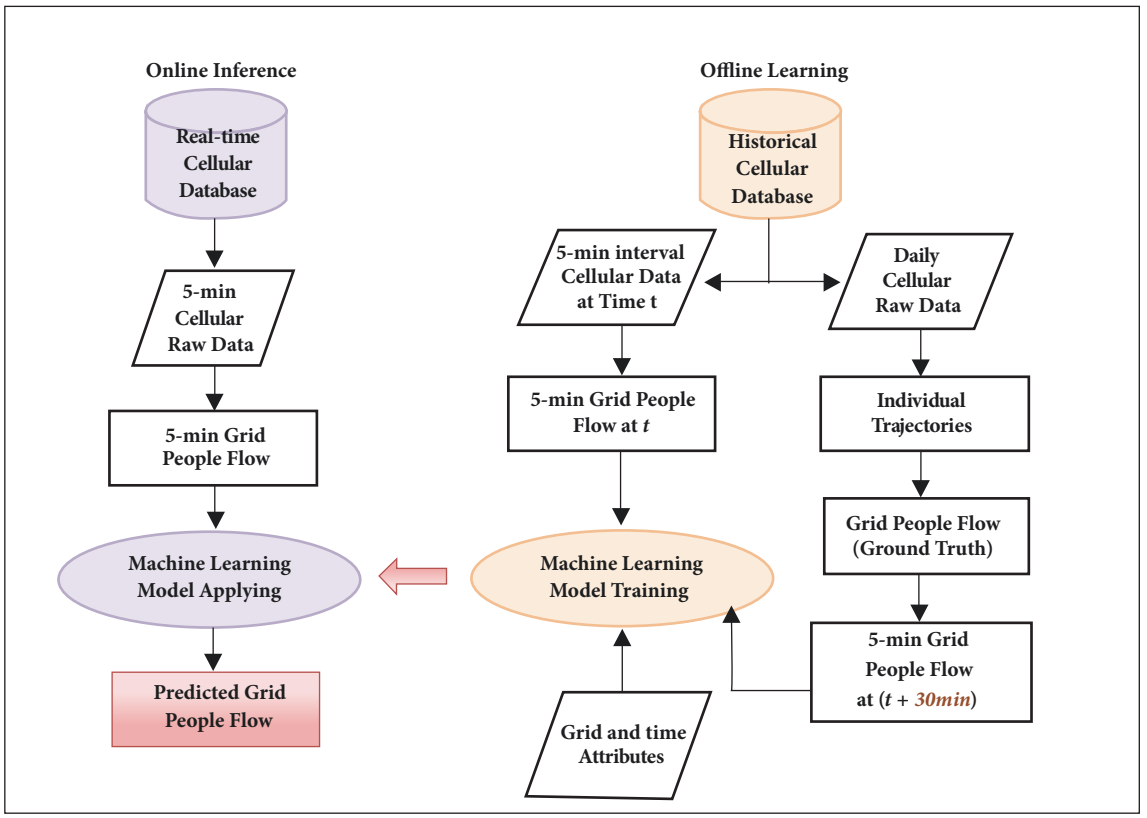


Figure 6: The work flow of data-driven process.

3.5. Data-Driven Process. There are two parts of the data-driven process shows in Figure 6. A classification random forest (RF) machine-learning model is used as the data-driven model in this paper. There are two parts of the data-driven process: the first part is the offline learning, which calibrates the RF model using the historical data. The second part is the online inferring, which calibrate the real-time cellular probe data to predict the grid-based people-flow.

3.5.1. Random Forest Model. The RF model is the major data-driven model we use in this paper. The RF algorithm firstly proposed by Breimen in 2001, which is so-called ensemble method, a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution of all trees in the forest [24].

In this study, the random forest approach is the major classification method. For each tree, the training data sets with selected features and the related features are required in each of the trees. In the tree build procedure, attributes will be split in each of the nodes in a tree from the top level to the leaf level. An entropy index is used to determine the best features in each of the nodes in

$$E = \sum_{i=1}^{n} -p_i \log_2 p_i \tag{5}$$

where E is the entropy of each feature,

n is number of values in each of the features,

p is proportion of class *I*.

Each of the individual tree classifiers results will be collected for voting. The most popular results will be the RF output result. The randomization approach is based on two parts: bagging and random selection process. In the first part of the RF method, a bootstrapping process, which is the training data will be selected randomly for each of the tree training, is used for the tree generation. The features for each of the tree trainings are also selected randomly to replace the existing features for each tree.

In the second part of RF method, the RF will be build up by undertaken the trees. The importance of each feature will be measured from the total data set. Then, the permuted data set will be used in the development and model refine. The mean decrease accuracy index (MDAI) will be calculated for each of the features. The variable is of importance of each feature (x_i) based on the calculate in

$$\partial\left(x_j\right) = \frac{1}{n\,tree} \sum_{f} \left(E_f - E_{f_j}\right) \tag{6}$$

where $\partial(x_j)$ is importance of attribute x_j,

E_f is error rate before the permutation process,

E_{f_j} is error rate before the permutation process for feature j.

Table 1: RF prediction result evaluation.

		Inbound Flow		Outbound Flow	
No.	Features	Precision	Recall	Precision	Recall
1	$\boldsymbol{F_c}$	0.51	0.39	0.49	0.36
2	$\boldsymbol{F_t}$	0.37	0.32	0.42	0.39
3	$\boldsymbol{F_c} + \boldsymbol{F_n}$	0.53	0.39	0.51	0.38
4	$\boldsymbol{F_c} + \boldsymbol{F_p}$	0.59	0.38	0.55	0.4
5	$\boldsymbol{F_c} + \boldsymbol{F_t}$	0.54	0.46	0.53	0.48
6	$\boldsymbol{F_c} + \boldsymbol{F_n} + \boldsymbol{F_p} + \boldsymbol{F_t}$	0.768	0.61	0.7	0.54

3.5.2. Offline Learning and Online Inferring. In the offline learning part, features are selected and calibrated in the RF model. Input features are selected and calibrated in this process, shows in Figure 6.

Because of the nature of each zone, the maximum number of people-flow is different. For example, the grid in transit area people-flow may reach 20,000 people per hour, but in some community area the people-flow will less than 100 people per hour. Since the RF model is calculated based on the weights of each feature, the extremely wide data distribution will affect the accuracy of the result. In this case, the people-flow estimated by 5-minutes time interval and the people-flow estimated by daily trajectory is divided into 6 levels. The number of each level of the flow is the max number of flow in each grid divided by 5. If the input estimated flow is larger than the regular maximum flow, the flow will be assigned as class 6. By multiple the predicted level with the level interval for each grid, the number of people-flow will be estimated.

Since the people-flow within an area may change because of the date type. For example, the central business district attracts fewer people-flow during the holiday than that during the work day. For better accurate, the model is calibrated based on the holiday type. In first day and last day of the holiday and last day of the holiday, the people-flow may have distributed differently.

The model is calibrated based on the date shows in Equ.(7).

$$\theta_d = \left[d_{holiday}, d_{holiday\_dates}, d_{weekday}\right] \tag{7}$$

where θ_d= the date types, which contains 5 key features of date,

$d_{holiday}$ = holiday or not,

$d_{holiday_{dates}}$ = date in holiday,

$d_{weekday}$ = weekday.

In the online part, real-time cellular data will be integrated into a 5-minute time interval. The right model is selected based on the date of cellular data. The people movement will be mapped on each of the grid. Thus, the online inferring modeling calibrates the cellular probe data and output the predicted people-flow.

4. Case Study

4.1. Data Source. The study area covered 1,000km^2, with more than 20 million people in the coverage area. The area was divided into 10,000 grids (100$*$ 100 grids). Each grid is a square with 600-meter side length. There are four data sources available in the study area.

(i) Cellular raw data: the data is collected by one of the top three major cellular carriers in Shanghai from November 27th to November 29th of 2013 (Wednesday to Friday). 3.1 million mobile phone users are extracted to test the proposed system and validate the models. There are 6.7 billion pieces of cellular data from the three days to test the proposed system.

(ii) Transportation Network data: the dataset includes network links of the subway network and major highways.

(iii) POI data: the POI dataset in the study area was collected in 2015. There are 3696 hospitals, 6395 schools, 4436 hotels, 3499 government agencies, 34495 markets, and 21928 restaurants in the study area.

(iv) Time Data: the peak hours in the study area are 6:00 AM – 9:00AM and 5:00 PM – 8:00 PM; the working hours are 8:00AM – 5:00PM; and the nighttime is 11:00PM – 5:00AM.

4.2. Prediction Results

Feature Evaluation and Selection. It is the critical process in machine-learning modeling, which selects a subset of the relevant features as the input for modeling. There are four features and six figure combinations evaluated in this section. Because the model is calibrated in real-time, the real-time people-flow ($\boldsymbol{F_c}$) and temporal feature ($\boldsymbol{F_t}$) should be primary features. Table 1 shows the combination of the primary features and two secondary features: Transportation Network Feature ($\boldsymbol{F_r}$) and POI Feature ($\boldsymbol{F_p}$). The RF result from six feature-combination scenarios are listed in Table 1

Based on the result, with more feature data set into the RF model, both precision and recall are improved. The recall improved less because the category of flow data is divided equally. The number of records for each category is not uniformly distributed.

Table 2: Predicted inbound/outbound flow level.

Inbound Flow							
	Predictions						Recall
Base Data	1	2	3	4	5	6	
1	89251	3128	201	64	18	1	0.963
2	15871	79102	3460	544	37	2	0.799
3	5124	16212	68191	2205	109	5	0.742
4	749	8987	10215	48970	1987	17	0.690
5	112	3958	8021	9090	15871	37	0.428
6	7	699	737	952	1034	162	0.045
Precision	0.803	0.706	0.751	0.792	0.833	0.723	0.768
Outbound Flow							
	Predictions						
Base Data	1	2	3	4	5	6	
1	91886	4388	278	118	2	1	0.950
2	25374	43012	3498	5252	88	4	0.557
3	6293	8183	17341	31223	872	7	0.271
4	973	2067	2187	104104	2948	26	0.927
5	99	444	946	28369	11951	62	0.285
6	9	41	58	1892	893	241	0.077
Precision	0.737	0.740	0.713	0.609	0.713	0.707	0.703

Confusion Matrix of combination #6 with four input features ($F_c + F_n + F_p + F_t$) for both inbound and outbound people-flow result is shown in Table 2. From the matrix, average precision for inbound predicted flow is 0.7268, and for outbound predicted flow is 0.703. Overall, the average precision for inbound/outbound people-flow is 0.73, which is much higher than the single-feature prediction.

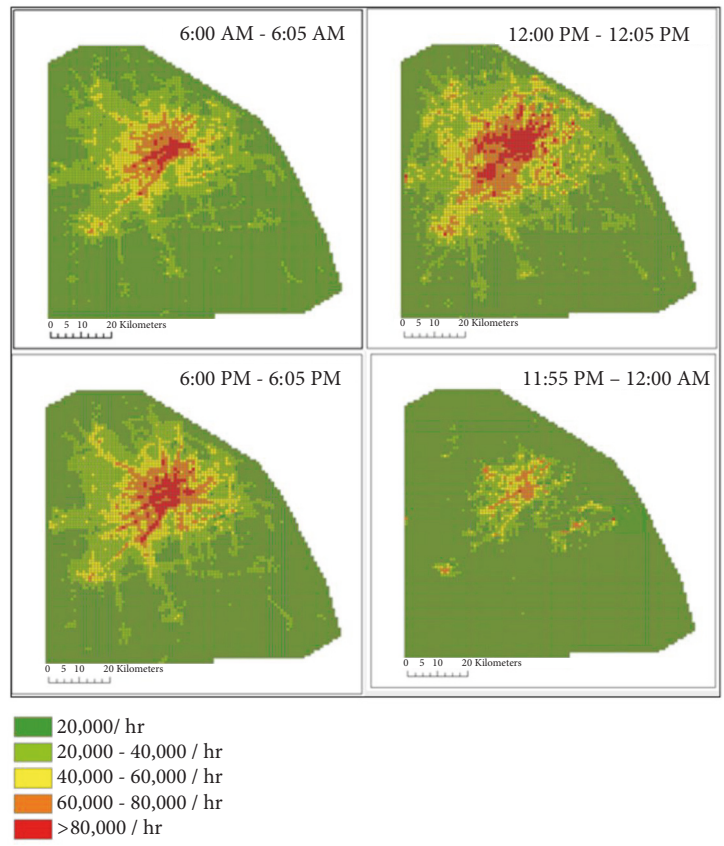
Visualization of the level of people-flow is shown in Figure 7. The combined inbound/outbound flow in Shanghai is predicted using the proposed approach. The green areas represent people-flow of less than 20,000 per hour while the red grids indicate a flow larger than 80,000 per hour. Compare the predicted result using the online inferring in Figure 7(b), with the validate people-flow inferred from the cellular probe data in Figure 7(a). From the visualized data, obviously, during peak hours (6:00 AM and 6:00 PM), the people-flow on the transportation network is very large. The middle area of the city shows the highest people-flow during the day. Additionally, the higher people-flow zones reflect the location of freeway and subway lanes.

5. Conclusion and Discussion

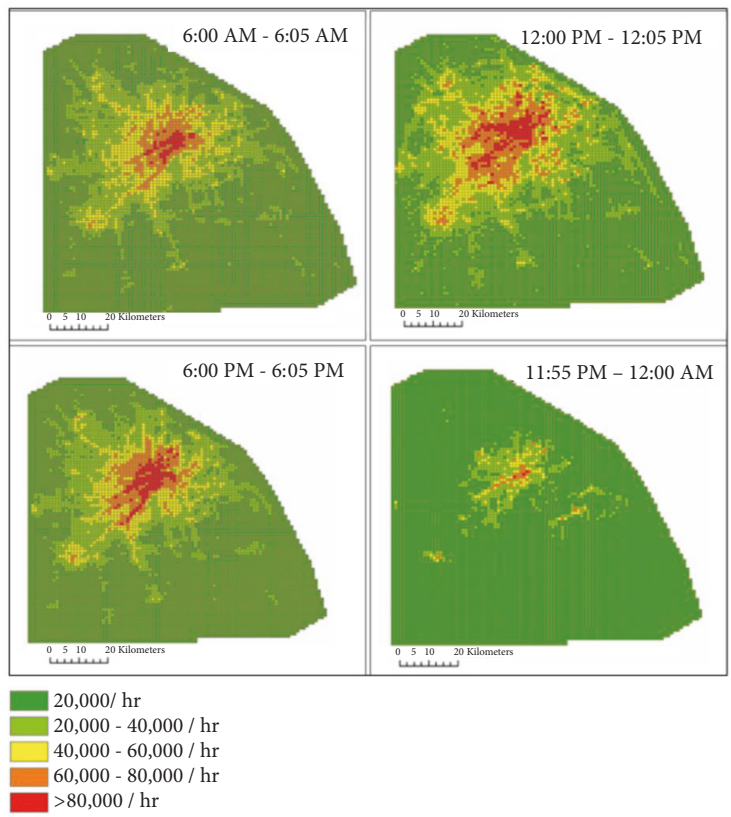
Zonal inbound and outbound people-flow is a major output of travel demand modeling. It is a critical data source for transportation planning, operations, and management and is usually estimated by travel surveys and GPS data. The travel surveys take tremendous labor and capital resources, so it is usually only taken every 3-5 years. Additionally, GPS data, including cellphone GPS or vehicle GPS, usually has a low sample size, which makes it hard to reflect the people-flow for a whole population. The cellular signaling data, which can be passively collected in real-time at low cost with a high sampling rate, has great potential to improve upon the weaknesses of GPS data and survey data. However, because cellular data is temporally and spatially sparse, few of the previous studies focused on extracting the real-time people-flow using cellphone signaling data.

This study presents a data-driven based people-flow prediction system. The benefits of the proposed prediction system are the efficient and accurate real-time people-flow prediction service. Since the cellular signaling data in a large-scale network is extremely large, the calibration efficiency for the real-time service is critical. The proposed trip-chain model provided a possibility of identifying the missing trips and calibrated the people-flow in real-time. A grid-based data integration module is used for data integration and feature extraction. Multiple data sources, including POI features, temporal features, real-time people movement level features, and the transportation network features, are integrated into a grid-level system. In this way, the model calibration process is efficient because the calibrated model could be applied on all grids with different attributes.

The online inference RF model with four types of features provides precision of 76.8% and 70% for outbound and inbound people-flow, respectively, which are much higher than the results of a single-feature prediction model. Hence, the data-driven approach in this paper using an offline training model and an online inference model is able to predict the people-flow in a real-time, efficient, and accurate way.



(a) Offline estimated people-flow inferred from daily cellular data



(b) Online predicted people-flow from real-time cellular data

Figure 7: Cooperation of based people-flow data and predicted result.

Data Availability

The cellular phone data, POI data, and transportation network data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Xiaoxuan Chen, Xia Wan, Fan Ding, and Bin Ran contributed to study conception and design; Xiaoxuan Chen and Fan Ding contributed to data collection; Xiaoxuan Chen, Qing Li, and Charlie McCarthy contributed to analysis and interpretation of results; Xiaoxuan Chen, Xia Wan, Yang Cheng, and Charlie McCarthy contributed to draft manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

References

[1] Statista, *Mobile phone penetration in China as share of the population from 2013 to 2019**, 2016, http://www.statista.com/statistics/233295/forecast-of-mobile-phone-user-penetration-in-china/.

[2] M. D. Meyer and E. J. Miller, *Urban transportation planning: a decision-oriented approach*, 1984.

[3] M.-H. Wang, S. D. Schrock, and N. Vander Broek, "The Feasibility of Using Cellular Phone Location Data in Traffic Survey on Inter-City Trips," in *Proceeding of the 92nd Annual Meeting of Transportation Research Board*, Washington, USA, 2013.

[4] F. Breu, S. Guggenbichler, and J. Wollmann, *Esitmation of Origin-Destination Matrices Using Traffic Counts - A Literature Survey*, Vasa, 2008.

[5] S. Bera and K. V. K. Rao, "Estimation of origin-destination matrix from traffic counts: The state of the art," *European Transport - Trasporti Europei*, no. 49, pp. 3–23, 2011.

[6] K. Parry and M. L. Hazelton, "Estimation of origin-destination matrices from link counts and sporadic routing data," *Transportation Research—Part B: Methodological*, vol. 46, no. 1, pp. 175–188, 2012.

[7] J. Barceló, L. Montero, L. Marqués, and C. Carmona, "Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring," *Transportation Research Record*, vol. 2175, no. 1, pp. 19–27, 2010.

[8] A. Das, A. Ghose, A. Razdan, H. Saran, and R. Shorey, "Enhancing performance of asynchronous data traffic over the Bluetooth wireless ad-hoc network," in *Proceedings IEEE INFOCOM 2001*, vol. 1, pp. 591–600, Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213), Anchorage, AK, USA, 2001.

[9] S. E. Young, *State highway administration research report bluetooth traffic detectors for use as permanently installed travel time instruments project number SP909B4D*, 2012.

[10] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw GPS data for geographic applications on the web," in *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, p. 247, New York, NY, USA, April 2008.

[11] L. Huan and I. Mygistics, *Using Mobile Phone Data to Analyze Origin-Destination Travel Flow Dynamics*, 2013.

[12] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 2014.

[13] C. Pan, J. Lu, S. Di, and B. Ran, "Cellular-based data-extracting method for trip distribution," *Transportation Research Record*, no. 1945, pp. 33–39, 2006.

[14] P. Cheng, Z. Qiu, and B. Ran, *Particle Filter Based Traffic State Estimation Using Mobile phone Network Data*, 2006.

[15] Z. Qiu, P. Cheng, and B. Ran, *A Linear Regression Model of Estimating Traffic State Using Real-Time Mobile phone Data*, 2006.

[16] Z. Qiu and B. Ran, *Applying Cellular Network Data to Detect Freeway Traffic State Using Virtual Sensor Network*, 2007.

[17] Y. Zhang, X. S. Qin, Dong., and B. Ran, *Daily OD matrix estimation using cellular probe data*, 2010.

[18] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[19] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, "Dynamic data-driven local traffic state estimation and prediction," *Transportation Research Part C: Emerging Technologies*, vol. 34, pp. 89–107, 2013.

[20] J. C. van Lint, "Reliable real-time framework for short-term freeway travel time prediction," *Journal of Transportation Engineering*, vol. 132, no. 12, pp. 921–932, 2006.

[21] J. W. C. van Lint, "Online learning solutions for freeway travel time prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 38–47, 2008.

[22] M. P. Hunter, R. M. Fujimoto, W. Suh, and H. K. Kim, "An investigation of real-time dynamic data driven transportation simulation," in *Proceedings of the 2006 Winter Simulation Conference, WSC*, pp. 1414–1421, USA, December 2006.

[23] R. Fujimoto, R. Guensler, M. Hunter et al., "Dynamic Data Driven Application Simulation of Surface Transportation Systems," in *Computational Science – ICCS 2006*, vol. 3993 of *Lecture Notes in Computer Science*, pp. 425–432, Springer, Berlin, Heidelberg, Germany, 2006.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.