



VeMo: Enable Transparent Vehicular Mobility Modeling at Individual Levels With Full Penetration

Yu Yang , Xiaoyang Xie, Zhihan Fang, Fan Zhang, *Member, IEEE*,
Yang Wang , *Member, IEEE*, and Desheng Zhang, *Member, IEEE*

Abstract—Understanding and predicting real-time vehicle mobility patterns on highways are essential to address traffic congestion and respond to the emergency. However, almost all existing works (e.g., based on cellphones, onboard devices, or traffic cameras) suffer from high costs, low penetration rates, or only aggregate results. To address these drawbacks, we utilize Electric Toll Collection systems (ETC) as a large-scale sensor network and design a system called VeMo to transparently model and predict vehicle mobility at the individual level with a full penetration rate. Our novelty is how we address uncertainty issues (i.e., unknown routes and speeds) due to sparse implicit ETC data based on a key data-driven insight, i.e., individual driving behaviors are strongly correlated with crowds of drivers under certain spatiotemporal contexts and can be predicted by combining both personal habits and context information. We evaluate VeMo with (i) a large-scale ETC system with tracking devices at 773 highway entrances and exits capturing more than 2 million vehicles every day; (ii) a fleet consisting of 114 thousand vehicles with GPS data as ground truth. Compared with state-of-the-art benchmark mobility models, the experimental results show that VeMo outperforms them by 10 percent on average.

Index Terms—Vehicular mobility modeling, real-time locations, stationary sensors, toll systems, destination, route, speed

1 INTRODUCTION

UNDERSTANDING and modeling individual vehicular mobility on highways have various applications, e.g., congestion prediction [1], route planning [2] and ramp metering [3]. However, modeling and predicting individual vehicle locations in fine spatial-temporal granularity are extremely challenging due to a large number of vehicles and limited infrastructures on highways compared to cities [4], [5].

The existing approaches for vehicle location prediction can be basically categorized into two groups: (i) mobile infrastructure based solutions such as cellphones (e.g., Online Map Services [6]) and onboard devices (e.g., OBD devices [7]), and (ii) static infrastructure based solutions: traffic cameras [8], loop sensors [9], and RFID [10]. For mobile infrastructure based solutions, they typically have privacy issues since they require real-time GPS locations of vehicles [11]; for static infrastructure based solutions, they typically introduce low spatial coverage or high costs for a complete highway system coverage [12]. Further, both of them may suffer low penetration rates, e.g., some commuters do not use

navigation apps when traveling some familiar routes [13]; traffic cameras are not pervasive on highways in some countries [14].

In this paper, to address these drawbacks, we utilize a highway Electric Toll Collection (ETC) system as a sensor network for vehicular mobility modeling and prediction. Compared to the existing approaches, our ETC based solution has the following advantages: (i) it requires no additional infrastructure since it relies on data already gathered in real time over highway networks for toll collections; (ii) it poses no additional privacy threats because it does not collect vehicle-specific GPS data; (iii) it does not suffer from low penetration rates since all vehicles have to be charged by an ETC system when using highway systems. Even some highways are installed with induction loops, they cannot achieve individual-level modeling compared to the ETC system, since they cannot distinguish individual vehicles.

However, since an ETC system is deployed for toll collections instead of mobility modeling, it brings new challenges. (i) An ETC system only logs when and where a vehicle enters and leaves a highway system for billing purposes and it leads to extremely sparse location records for each vehicle, i.e., only two data points per trip, which makes it challenging to predict destinations without intermediate locations. (ii) In a complicated highway network, given an entrance and exit, there are many potential routes as shown by our later analyses, and ETC data do not log any information regarding which route was taken during a particular origin and destination pair. Based on our data, we found that the shortest routes are not the first choices for many vehicles due to congestion. Without any historical routes or speeds, it is difficult to train a model as conventional machine learning tasks.

- Yu Yang, Xiaoyang Xie, Zhihan Fang, and Desheng Zhang are with the Department of Computer Science, Rutgers University, Piscataway NJ 19403 USA. E-mail: yu.yang@rutgers.edu, [xx88](mailto:xx88@rutgers.edu), [zhihan.fang](mailto:zhihan.fang@rutgers.edu), [desheng.zhang](mailto:desheng.zhang@cs.rutgers.edu)@cs.rutgers.edu.
- Fan Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China. E-mail: zhangfan@siat.ac.cn.
- Yang Wang is with the University of Science and Technology of China, Hefei 230000, China. E-mail: Angyan@ustc.edu.cn.

Manuscript received 10 Jan. 2020; revised 1 Sept. 2020; accepted 25 Nov. 2020.
Date of publication 21 Dec. 2020; date of current version 3 June 2022.
(Corresponding authors: Yu Yang.)
Digital Object Identifier no. 10.1109/TMC.2020.3044244

(iii) Further, driving speeds vary by personal habits and different spatiotemporal contexts, and ETC data do not directly log real-time driving speeds. Straightforward solutions (e.g., assuming real-time speeds vary near the speed limit) usually do not perform well because of various driving behaviors under different contexts.

To address these challenges, we perform a systemic investigation of a large-scale ETC system. We design a model called VeMo to model and predict individual vehicular mobility based on sparse observations on real-time origins as well as historical origins and destinations. To achieve this, we divide the ultimate objectives into three subproblems including destination prediction, route prediction, and speed prediction. In each of the subproblems, we investigate three key kinds of features that are strongly correlated to vehicular mobility including personal features, crowd features, and context features. The key insight we found is: even with complicated highway networks and real-time context, individual travel behaviors are strongly correlated with crowds under certain spatiotemporal contexts and can be predicted by combining both personal habits and context information. Combining the results of the three subproblems, we model the vehicular mobility patterns on highways. We summarize the key contributions as follows.

- To our knowledge, we conduct the first systematic investigation of vehicular mobility modeling and prediction based on large-scale ETC and GPS data. Our investigation is based on ETC data from 7.8 million vehicles and GPS data from 114 thousand vehicles. A number of useful features are explored that reveal the general vehicular mobility patterns.
- We analyze both ETC and GPS data providing in-depth discussions on vehicular mobility patterns. Based on our analyses, we design a mobility prediction system called VeMo with three key components to predict destinations, routes, and speeds for individual vehicles based on both historical and real-time ETC data. Technically, we extract unobserved routes and speeds through joint optimization. By studying mobility features at both the individual and crowd level, we fuse them based on a Mondrian Forests model to address the uncertain mobility issue.
- We implement and evaluate the VeMo in the Guangdong Province, China with (i) an ETC system covering 1,439 highway entrances and exits, and it captures around 2 million vehicles per day; (ii) a vehicle fleet and its GPS data including 114 thousand vehicles for evaluation only, where 20 percent of vehicles have the trajectories on highways. Compared with state-of-the-art solutions, VeMo provides a 10 percent accuracy gain on average.
- Based on our system, we design a real-world application to detect ongoing anomaly events on highways. It achieves automatically anomaly event detection and event location inference at the same time without extra human resources. The result shows we detect 85.7 percent of the anomaly events and locate them within 300 meters of the event locations.

Authorized licensed use limited to: Florida State University. Downloaded on July 29, 2023 at 17:32:55 UTC from IEEE Xplore. Restrictions apply.

2 MOTIVATION

2.1 Use Cases

VeMo aims to predict the real-time locations of individual vehicles, which enables various applications that cannot be achieved by previous solutions. As a collaboration with the highway administrators, we give two exemplary applications that matter a lot to highway management.

- *Highway anomaly event detection*: One important work of highway administrators is to detect the anomaly event for emergency responses, such as traffic accidents and road maintenance. However, it is quite expensive to arrange regular road check manually or cannot detect anomalies in time. By predicting the real-time location of a vehicle, we can know when the vehicle is expected to leave the highway in a regular situation. If most vehicles do not leave the highways as expected, there are potential ongoing anomaly events on highways. Based on which vehicles are affected, we could further infer the most likely location of the anomaly source, which could guide the highway administrators for fast emergency responses. Compared to other approaches, our approach utilizes existing infrastructures without introducing much extra cost. We will describe the detailed implementation in Section 6.
- *Hazmat cargo tank tracking*: Improving driving safety on highways is always an important topic for highway administration companies. There are more than 6 million accidents on highways in the United States during 2015 [15]. A special case is when the vehicles are hazmat highway cargo tanks, which may result in even worse accidents. By predicting the locations of these special vehicles, highway administration companies can better understand the potential risk, which is not achievable by group-based prediction.

Uniqueness of ETC Based Systems. To implement those applications, previous work based on smartphones or traffic cameras either requires extra installed infrastructures or suffers from low penetration rates of vehicles. On highways, the key strength of the ETC system is it detects all the vehicles (i.e., a high penetration rate). Compared to GPS-based approaches, our approach utilized regular billing data for mobility modeling, which minimizes the extra expose of users' privacy.

2.2 Challenges

It is not trivial to predict the real-time locations of vehicles because of the uncertainties caused by various traffic conditions and driving behaviors. To show these challenges, we study one-month data (both ETC transactions and trajectories of sample vehicles) in the Guangdong province of China and identify several challenges regarding three key factors including destinations, routes, and speeds. The detailed data description is presented in Sections 3 and 5.

2.2.1 Destination Uncertainty

To predict the real-time locations of vehicles, it is important to understand the destinations and routes. However, it is not trivial to predict routes and destinations. To characterize

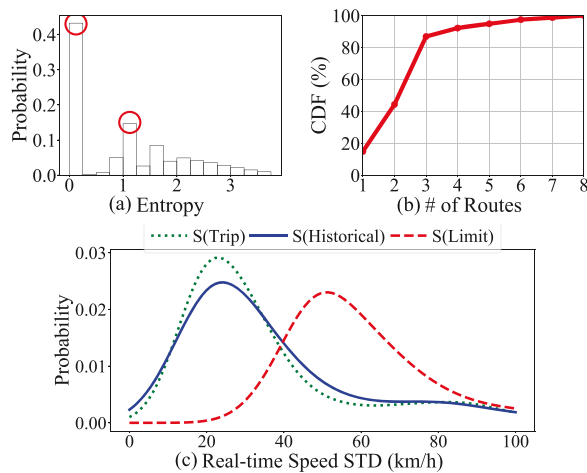


Fig. 1. Challenges: (a) destination entropy; (b) number of routes; (c) speed STD.

the inherent predictability across vehicles, we present the destination entropy of each vehicle in Fig. 1(a). The figure reveals two peaks as the entropy equals 0 and 1, which indicates the next location of a vehicle could be found on average in any $2^0 = 1$ and $2^1 = 2$ locations, respectively. Especially, we find most vehicles travel on highways only once in one month with the entropy=0; vehicles often commute between locations with the entropy=1. Some works [16] were conducted to predict the destinations of vehicles whose entropy is greater than or equal to one since those vehicles generally have regular commute patterns or extensive historical data. However, it is not clear how to predict the destinations of vehicles with only a few historical transactions.

2.2.2 Unobserved Routes and Speeds

Previous studies have been done to model the route choices and driving speeds [2], [17]. Through studying the historical routes and speeds in the trip recorded by GPS-based devices, some sophisticated models are proposed to predict vehicular mobility in the near future. However, in our setting, one of the key characteristics of the ETC system is that it can only obtain very sparse information (i.e., the time and location when entering and exiting highways). This leads to the problem that we cannot obtain detailed routes and speeds to learn the route and speed model, which is not solved in the previous work.

Moreover, routes and speeds also vary depending on user behaviors and contexts. For a given origin and destination, people can choose different routes if the road network is not trivial (i.e., only one route from the origin to the destination). Fig. 1(b) illustrates the number of routes between the origin-destination pairs. We found that only 17 percent of station pairs have only one route based on GPS trajectories obtained from 114 thousand vehicles. It is impractical to assume only the shortest routes are used by vehicles. (Note that these trajectories are only used in the motivation and evaluation rather than the model design.) As for speeds, people empirically expect that the driving speeds of vehicles are around certain speeds (e.g., speed limit or average speed) with less variance. However, in our study, we found the real-time speed is more complicated than the empirical intuition. To illustrate the characteristics of

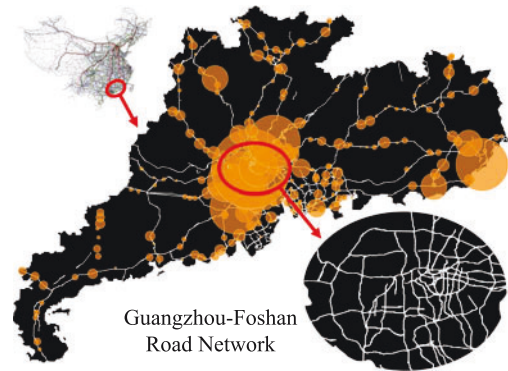


Fig. 2. ETC systems in Guangdong Province.

real-time speed, we study the real-time speed standard deviation (STD) across vehicles by replacing the mean value in the standard formula of standard deviation with the speed limit ($S(Limit)$), the historical average speed ($S(Historical)$), the current trip average speed ($S(Trip)$), respectively. Fig. 3 demonstrates both $S(Historical)$ and $S(Trip)$ have a Gaussian-like distribution with the mean STD near 20 km/h. It leads to a 330-meter offset in one-minute driving if only the average speed is utilized to obtain the real-time location. It also reveals the fact that it is difficult for people to drive at the speed limit (e.g., can only drive at 60 km/h compared to the speed limit of 120 km/h) because of the heavy traffic.

3 ETC SYSTEM AND DATA DESCRIPTION

3.1 Notations

Given ETC data on the vehicle’s trip levels,

- An *edge* e is a highway segment between two adjacent toll stations, i.e., the finest spatial unit for ETC-based modeling.
- A *route* r is a set of adjacent edges, which connect the origin toll station and the destination toll station of a particular trip.
- A *K-edge trip* is a trip of a particular vehicle with K edges in its route between the origin and the destination. Specifically, a *single-edge trip* has only one edge.

3.2 Infrastructure Overview

Fig. 2 shows the road structure and the locations of toll stations in the Guangdong province, which has 69 highways and 773 ETC toll stations with 1,439 highway entrances and exits covering an area of 179,800km². The circles represent toll stations and the larger the icon, the heavier the daily traffic volume. It shows the traffic mainly concentrates on the central area and the road structure in that area is also complex as shown in the Guangzhou-Foshan Road Network. Each toll station detects all vehicles when they enter the highway, and then logs the records as transactions after they leave the highway. The toll station identifies a vehicle by ETC RFID devices (for regular charging) or cameras (for the purpose of detecting escaping charges).

As shown in Table 1, each generated transaction contains information including entering and exit station, entering and exiting time, vehicle id, vehicle type (i.e., car, bus, truck), and

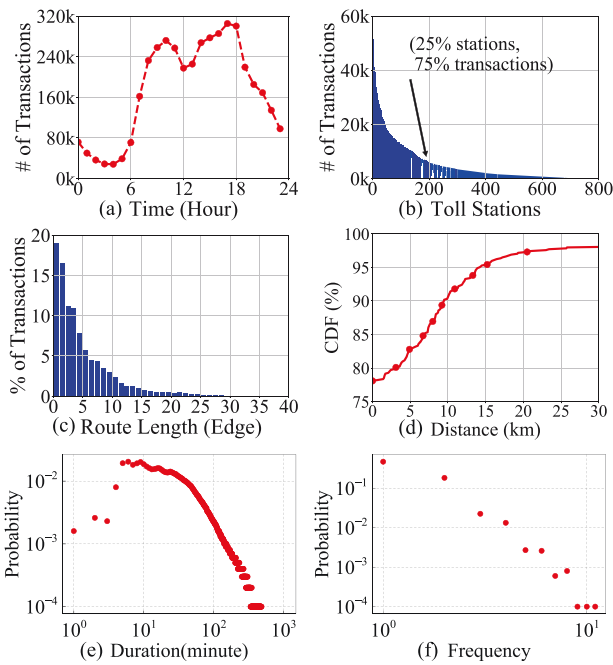


Fig. 3. Statistic descriptions: (a) volume over time; (b) volume over stations; (c) route length (#edges); (d) nearest station; (e) duration; (f) nearest station.

TABLE 1
ETC Transaction Description

Field	Value
Entering/Exit Toll Station	Humen Station
Entering/Exit Time	2016-07-01 13:00:01
Vehicle Id	F37SS1D4GU
Vehicle Type	Car/Bus/Truck
Weight	1500 kg
Number of Daily Transactions: 4 millions	
Number of Daily Vehicles: 2 millions	

weight. Such a transaction was generated when a vehicle enters and exits the highway network with both ETC cards or cash. On average, there are more than 4 million transactions generated every day from 2 million vehicles.

3.3 Statistic Description

Fig. 3(a) plots the average traffic volume in a day. It shows there are two peak hours (i.e., 10 am and 6 pm), which potentially make prediction challenging due to uncertainty (e.g., route choice, traffic jam, etc) introduced by high traffic volume. Fig. 3(b) depicts the daily transaction volume of all the toll stations, where 25 percent of the stations contribute 75 percent of the transactions. It suggests the major number of vehicles enter the highway from a limited number of stations, indicating prediction related to unpopular stations may suffer from a lack of historical and real-time data.

Fig. 3(c) describes the route length of the trip in terms of the number of edges. Since we do not know the actual route of trips, the statistic result is based on the shortest route assumption. We find only 18 percent of the trips are between the adjacent stations with only one edge. Fig. 3(d) shows the distance

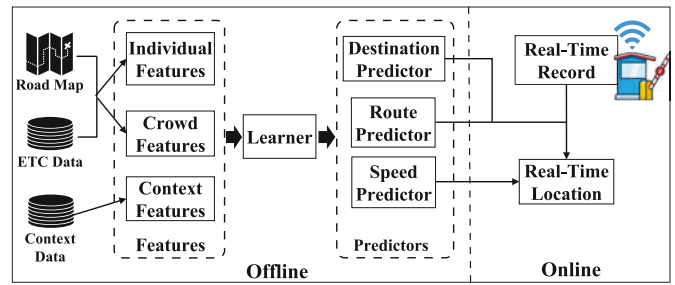


Fig. 4. Framework.

between a station to its nearest station. It shows 78 percent of the nearest stations are within 1 km of each other, which reveals the very dense spatial distribution of stations.

We further investigate how a vehicle moves on highways. Fig. 3(e) plots the duration of vehicles on highways. We found most vehicles only spend 7 to 10 minutes on a single trip. This is caused by the heavy volumes of inner-city mobility. Fig. 3(f) shows the frequency of vehicles using highways in one month. It shows most vehicles only use highways once a month, which implies the functionality of the local roads still dominates regular mobility. It is dramatically different from the mobility patterns in other countries such as the U.S. that people heavily rely on highways on daily commuting.

4 VEMO DESIGN

4.1 Framework

Fig. 4 shows the framework of our system, which consists of two parts: offline learning and online prediction.

In offline learning, all the data come from three data sources including the road map, historical ETC transactions, and context data. In the feature extraction, we divide all the features into three categories, which are individual features, crowd features, and context features. By fitting these features into the learner, we train three predictors for destinations, routes, and speeds. By combining these predictors together, we predict the real-time locations of vehicles. In the next three subsections, we introduce three predictors for destinations, routes, and speeds from a feature perspective respectively, and then unify them together with a prediction model based on Mondrian Forest.

4.2 Destination Predictor

Destination prediction has been intensively studied in the past few years [18], [19]. The existing approaches for the vehicle destination prediction mainly rely on transition probabilities between different locations through learning historical trajectories using various Markov chain based models [20], [21]. One of the key prerequisites is that there should be enough historical data of individuals to learn the transition probabilities. However, in our context, most vehicles only have limited historical data (as we discussed in Section 2), which makes it hard to directly apply the Markov chain based models. To address this issue, we explore more individual features, crowd features, and context features.

Individual Features. We list several individual features:

- *Historical Destinations:* As shown in Fig. 1(a), the mobility patterns of most individuals in terms of

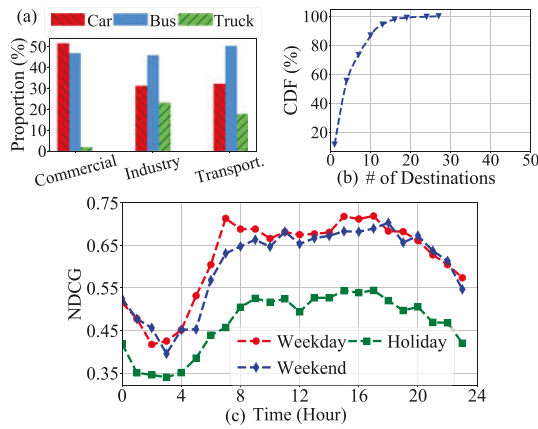


Fig. 5. Destination predictor features: (a) destination variance; (b) crowd destinations; (c) context impacts.

destinations are relatively stable. Therefore, historical destinations may largely represent their future destinations.

- *Time Factor:* Considering the commuting pattern in Fig. 1(a) with the entropy equals to 1, by introducing the entering time, the uncertainty of destinations is reduced. We use half-hour time windows to split one day into 48-time slots.
- *Vehicle Type:* It has three values: cars, buses, and trucks. Intuitively, the trucks most probably go to areas with high cargo demand (e.g., industry parks) and buses often go to areas with a dense population (e.g., commercial districts or transportation hubs). Fig. 5(a) shows the proportion of different vehicle types in different types of areas. We select three exemplary areas and calculate the proportion of different types of vehicles whose destinations are in the area. We found only a few trucks go to the commercial areas; cars and buses contribute major volume in the commercial and transportation hub areas, respectively.

Crowd Features. The individual vehicle’s historical data can be very sparse (as we suggested in Section 2). we try to use the crowd destinations to provide complementary information. Fig. 5(b) shows the possible destinations from the same origins by half of all the vehicles. We found almost 50 percent of vehicles go to at most 10 destinations. It indicates lots of vehicles from the same origins share similar destinations, which can be used to infer the destination of a vehicle without any historical destination data.

Context Features. We further consider other context features, i.e., the day of the week, weekday/weekend, holidays, that may have impacts on the destination choices. We choose the 10 most popular destinations for each origin and compare the rank of these destinations on a regular day with that in other days using the measurement of Normalized Discounted Cumulative Gain *NDCG* [22]. The lower the *NDCG*, the lower similarity of the destination choices. Fig. 5(c) shows that the measurement between weekdays, weekends, and holidays. The holiday has very different destination choices compared to other days. In the early morning and the late afternoon of weekends, the *NDCG* is also lower than that of weekdays. It suggests these factors have an impact on the choice of destinations.

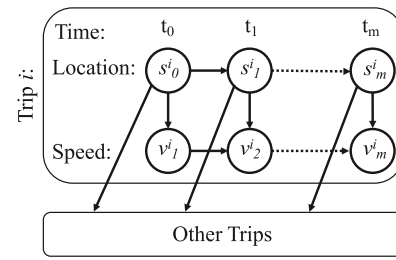


Fig. 6. Route and Speed correlation.

4.3 Historical Route and Speed Learning

As we discuss in Section 2, the reason that previous works are not feasible in our setting is that the historical routes and speeds of individual vehicles cannot be observed by the ETC system. In order to learn the mobility of individual vehicles, we propose a joint learning approach to obtain the historical routes and speeds of vehicles simultaneously, which are utilized as training data to model the route choices and real-time speeds in Sections 4.4 and 4.5.

Several studies [17], [23] have been done to investigate the relationship between the travel routes and real-time speeds, which found the route of vehicles can be inferred with only speeds information. This finding indicates the strong correlation between the routes and speeds, which inspires us to learn the routes and speeds simultaneously.

To achieve this, we first present a few preliminaries.

- *Time:* we divide a day of 24 hours into K time slots (t) (i.e., each time slot is equal to 10 minutes).
- *Location:* we split the highway road networks into M equal length road segments (s) (i.e., 1 km).
- *Speed:* we discretize the speed into H discrete integer speed (v) by the smallest unit of 1 km/h (e.g., if the speed limit is 120 km/h, then we have 121 different values between 0 and 120 km/h).

In this way, the states of vehicles in each trip on highways can be presented as a sequence of states (t, s, v) between the origin and the destination. As an example of the trip i in Fig. 6, the vehicle enters the highway from the road segment s_0 at the time t_0 and exits the highway from the road segment s_m at the time t_m . It is worth mentioning that, in other trips, vehicles can be at the same location as the same time as the trip i .

Then our objective is to infer the most likely state sequence of each trip. The solution is motivated by the key observation that at the same time multiple vehicles are traveling on the same road segments and their real-time speeds can be considered as samples of the speed distribution. The following insights reveal the characteristics of the distribution.

- *Speeds distribution on the road segment:* By analyzing the sample GPS trajectories, we observe that speeds of vehicles on the same road segment follow a normal distribution, which is also validated in other contexts [24].
- *Speed STD distribution:* Moreover, as shown in Fig. 3, we also observe strong normality of the speed.

Since both insights show the normality, to quantify them, we utilize the Kolmogorov-Smirnov test to test the normality. Specifically, the states of different trips within the same time and location are grouped as samples to test the normality of speed on the road segments. For the speed STD

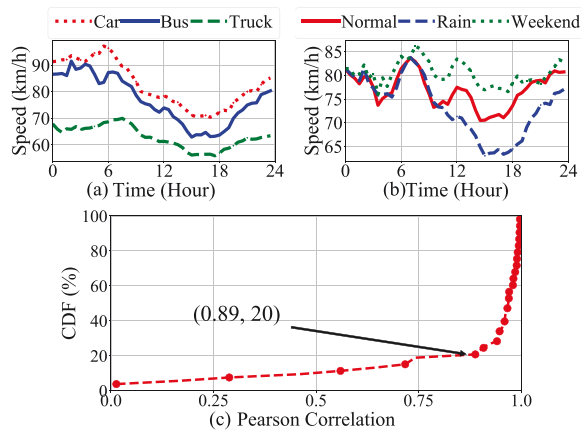


Fig. 7. Speed predictor features: (a) speed variance; (b) context impacts; (c) speed correlation.

distribution insight, it is measured as suggested in Section 2. Then all the STDs are considered as samples to test the normality.

Given the normality test of both the speed distribution in each road segment and speed STD distribution of all the vehicles, our problem can be transformed into an optimization problem to find the best state sequence combination for the maximization of the number of the acceptance of normality tests. Suppose we have N trips with J vehicles, we formulate the problem as following:

$$\underset{sc}{\text{maximize}} \sum_i^N \mathbb{1}_A(Rnorm(sc)) + \sum_j^J \mathbb{1}_A(Snorm(sc)), \quad (1)$$

where sc is the combination of the state sequences of different trips, $Rnorm$ is a test function to check the normality of the speed distributions, $Snorm$ is a test function to check the normality of the speed STD distribution. $\mathbb{1}_A$ is an indicator function of the test acceptance.

A straightforward approach to solve the optimization problem is to search for all the possible state sequence combinations. For each trip, the possible state sequence is $K \times M \times H$. Then the total search space is $O(N^{K \times M \times H})$, which is time-consuming to search. To reduce the search space, we introduce several simple but effective heuristics to guide the search.

- **State sequences constrained by routes:** Shown in Fig. 2, there is a limited number of routes between origins and destinations, which naturally reduces the search space of possible location sequences.
- **Spatial smoothness:** Constrained by the structure of the road network and the speed limit, the next location of the vehicle can be the reachable road segments under the speed limit. (e.g., suppose the speed limit is 120 km/h , the next location in 5 minutes can only be the road segments within a range of $5 \text{ minutes} \times 120 \text{ km/h} = 10 \text{ km}$.)

Given these heuristics, we perform a standard search algorithm (e.g., DFS) to find the best combination of the state sequence. Then the historical routes can be obtained by

concatenating the locations in each trip and speeds can be directly obtained from the state sequence.

4.4 Route Predictor

Similar to the destination prediction, we study the features from three perspectives: individual features, crowd features, and context features.

Individual Features. We utilize the following features for the route prediction at the individual level.

- **Historical Routes:** Based on a previous study, people are more reluctant to change their regular routes if they have more experience with these routes [2], which indicates historical routes are most likely to be their future routes given the same origin and destination.
- **Driving Experience:** Empirically, experienced people are good at finding the best routes [2]. We quantify the experience by two factors: (i) the frequency of driving on highways, which can be obtained from historical ETC transactions; (ii) the saved travel time compared to the average travel time, which can also be computed from historical data.
- **Time Factor:** Empirically, people generally have their own estimations about the route traffic at a different time, e.g., taking a detour during the rush hour to avoid the traffic. It affects their future route choices.

Crowd Features. For those people who have no or only limited historical data, we incorporate the route choices of crowds to infer their route choice. Specifically, we use the probability of historical crowds' routes between particular origin/destination at a certain time.

Context Features. People's route choices are affected by the real-time context [25], i.e., the day of the week and real-time traffic speed, which can be estimated with ETC transactions in the recent past.

4.5 Speed Predictor

Our key idea of speed prediction is to learn the relation between individual driving speed and other features (e.g., crowd speed) in order to predict the real-time speed given all these features. We introduce our features on the individual, crowd, and context level.

Individual Features. Since the driving speed is essentially based on people's behaviors, we define a set of individual vehicle's features.

- **Historical Driving Speed:** As shown in Fig. 1(c), the driving speed is relatively stable for a particular person. We use their average speeds of historical trips to reflect their general driving speed.
- **Vehicle Type:** This feature reflects the vehicle's type (i.e., cars, buses, trucks). Intuitively, the driving speed of cars should be higher than trucks and buses. Fig. 7(a) also validates this intuition.
- **Time Factor:** Fig. 7(a) shows that the driving speed varies at different times of a day, which is mainly due to the different traffic conditions.

Crowd Features. People may behave differently under different traffic conditions. Instead of studying the detailed behavior patterns of individuals, which may have many factors to discuss, we directly investigate the correlation

between the individual speed and the crowd speed. Fig. 7(c) shows the Pearson correlation of the individual speed and crowd traffic speed. More than 80 percent of vehicles have at least a 0.89 correlation coefficient with the crowd traffic speed. Motivated by the strong correlation, the crowd traffic speed is an important feature to estimate the individual driving speeds on specific edges. Therefore, we extract the features of the vehicle speed samples, which are incorporated to estimate the crowd traffic speed. Instead of using the average crowd traffic speed (which may cause an estimation bias), we consider the statistic values of the crowd traffic speed distribution, including minimum, lower fourth, median, upper fourth, and maximum of the samples. We rely on the crowd features to learn how the driver would react under different situations, in order to predict the real-time speed.

Context Features. Besides the vehicle-related features, we also consider other factors that may have impacts on the driving speed, including weather and weekday/weekend. As shown in Fig. 7(b), the speed is decreased by 10 percent at most on a rainy day and increased by 5 percent on weekend. This is reasonable because people tend to drive slower when raining and fewer people use highways to work on the weekend, which makes the highways less congested.

4.6 Learning With Mondrian Forest

Mondrian forests [26] is an online random forest model using Mondrian processes to construct ensembles of decision trees. Compared to offline or online random forest [26], it provides the ability to process online data and online updates faster and more accurately. Compared with other algorithms, the Mondrian forests model has the following advantages:

- It is more robust to heterogeneous features. In our data input, we have both numerical variables (i.e., speed) and categorical values (i.e., vehicle type, weather, weekday/weekend). These variables can be input into the model directly without conversion or normalization.
- It provides self-check on the importance of the features during the training stage. For example, such as weather conditions and holidays, these variables would only have high importance under certain conditions with a low frequency.
- Compared to other neural-based models (e.g., deep neural network), the results are more explainable because of the internally used decision tree [27].

For different tasks, we fit all the extracted features into Mondrian forests and learn three predictors to work collaboratively on the real-time location prediction, which is illustrated in Section 4.7. More importantly, our system is flexible with other machine learning methods. The more important contribution is the analysis process and to find effective features.

4.7 Put Them All Together

In the previous sections, we have conducted an analysis of the three key tasks: destination prediction, route inference, and speed estimation. Based on multiple extracted features, we learn three predictors *d-predictor*, *r-predictor*,

s-predictor for each of the tasks, perceptively. The procedure of real-time location prediction is described in Algorithm 1. As a vehicle enters the system when it is detected in the entrance of the highway, we first predict its destination with *d-predictor*. Based on the entrance and the predicted destination, we infer its potential route with *r-predictor*. Finally, we continuously predict its speed given the real-time context inputs and the pre-trained *s-predictor* before it leaves the highway. The speed is then transformed as the distance in every equal time interval and mapped into the real-time location.

Algorithm 1. Real-Time Location Prediction

Input: *d-predictor*: the destination predictor,
r-predictor: the route predictor,
s-predictor: the speed predictor,
entrance: the entering toll station,
interval: the updating time interval
t₀: the entering time.

Output: real-time locations

```

1: destination ← d-predictor given entrance
2: route ← r-predictor given destination
3: distance = 0
4: while distance < route.length do
5:   speed ← s-predictor at ti
6:   distance += speed × interval
7:   location ← match distance to route
8: end

```

4.8 Feedback Updating

We further improve our system with online updating mechanisms. Three types of real-time feedback can be directly observed in our system, including *Wrong destination*, *Early Arrival*, and *Late Arrival*.

- *Wrong Destination*: A vehicle leaves the highways from a toll station other than the predicted one, i.e., *d-predictor* is wrong.
- *Early Arrival*: A vehicle leaves the highways earlier than predicted. It could be a wrong prediction in either the route, the speed, or both.
- *Late Arrival*: A vehicle arrives at the exit station later than expected. Similarly, it could be a wrong prediction in either the route, the speed, or both.

Instead of remaining the system unchangeable, we make use of these three types of feedback to provide online updating ability for the system's self-awareness. In particular, from wrong destinations, we directly utilize the online training mechanism of the Mondrian forests by inputting the data to update the *d-predictor*. For early arrival and late arrival, it is more complex since it involves both *r-predictor* and *s-predictor*. In general, there are three possible cases: only *r-predictor*, only *s-predictor*, or both *r-predictor* and *s-predictor*. We take a brute-force strategy by taking the new data into three cases. In each case, we re-conduct the location prediction process and calculate the arrival time error compared to the actual arrival time. The case with minimal error is selected as our updating mechanism.

5 EVALUATION

5.1 Evaluation Methodology

Ground Truth. We introduce another dataset with detailed GPS trajectories to obtain the ground truth of vehicle locations in Guangdong. It provides real-time locations of 114 thousand vehicles including 75 percent cars, 13 percent buses, and 12 percent trucks. These vehicles upload their real-time locations every 10 to 30 seconds. The detailed data format is presented in Table 2.

For each vehicle, we first apply a map matching algorithm [28] to map trajectories onto the road network. Then only the trajectories on highways are remained to obtain entering toll stations, exit toll stations, routes, and real-time locations, which occupy 20 percent of the vehicles in our dataset.

Evaluation Metrics. For each component, we define the evaluation metrics as follows:

- Destination and Route prediction

$$accuracy = \frac{\#prediction_{correct}}{\#prediction_{all}} \times 100\%, \quad (2)$$

where $\#prediction_{correct}$ is the number of corrected prediction and $\#prediction_{all}$ is the total number.

- Speed Prediction

$$accuracy = 1 - \frac{|speed_{predict} - speed_{actual}|}{speed_{actual}}, \quad (3)$$

where $speed_{predict}$ is the predicted speed and $speed_{actual}$ is the ground truth.

- Real-Time Location Prediction: we quantify the location accuracy by measuring the percentage of predicted locations within the accuracy threshold (i.e., 100 meters) of the ground truth considering the GPS errors every 15 seconds (i.e., the average uploading time interval of data in ground truth) [29]. The accuracy formula is defined as

$$accuracy = \frac{\#prediction_{correct}}{\#prediction_{all}} \times 100\%. \quad (4)$$

Baselines for Intermediate Results. For the individual prediction components, i.e., predictions for destinations, routes, and speeds, since we utilize a unified algorithm for all of them, we evaluate them from the perspective of a learning model by comparing it with the other learning models. The selected learning models are presented as follows, and each of them is representative of a group of methods with similar bases:

- *Empirical Estimation (Emp):* The baseline represents the prediction based on naive empirical knowledge. For the destination and route prediction, we consider the most frequently visited destinations and routes. For speed prediction, we utilize their historical average speed.
- *Bayesian Network (Bayes)* [27]: Bayesian network is a typical graph-based algorithm, which is representative of the probability-based models.

Authorized licensed use limited to: Florida State University. Downloaded on July 29, 2023 at 17:32:55 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Ground Truth Format

Field	Value	Field	Value
Id	POSF51B4GU	Type	Car/Bus/Truck
Longitude	113.402904	Latitude	23.167894
Time	2016-06-01 00:00:34		

- *Neural Network (Neural)* [27]: Neural network represents the models that focus on learning the linear or non-linear combination between features and targets.

Baselines for End-to-End Results. For the overall performance of the real-time locations, we choose the baselines based on two principles: (i) static infrastructure based methods; (ii) mobile sensor based methods.

- *STrack* is an approach to detect vehicular locations by cameras on the roads. Wherever a vehicle is captured by a camera, we assume its location is known to the system. Considering cameras may not accurately detect all the vehicles because of shadows, rain, or detection faults, our baseline is considered as an upper bound. Because these cameras are generally set to detect motoring offenses without open location information, we assume a given percentage of edges have been installed with cameras. For the location estimation of vehicles between cameras, we assume they are uniformly distributed on the roads between cameras.
- *CTrack* [30]: This baseline aims to track vehicles based on cellular networks by periodical communications between onboard cellphones and cell towers. Based on the locations of cell towers, we infer the locations of the cellphones (thus vehicles). The cell tower locations we use are located in Shenzhen City, where the ETC system is spread with 79 toll stations. We implement CTrack by assuming each vehicle has an onboard cellphone to interact with cell towers and follow the trajectory mapping algorithm [30].

Impacts of Factors. We evaluate several factors to show the impacts on the performance of VeMo,

- *Weather:* Weather condition is a factor that affects the driving behavior such as driving speed. We evaluate the accuracy in both regular day and extreme weather day (e.g., heavy rain).
- *Accuracy threshold:* Given different accuracy threshold to declare the accuracy, the performance may be varied. We choose several threshold values to show the accuracy.
- *Time factors:* The performance may vary at different times. We evaluate it on weekdays, weekends, and holidays.
- *Spatial factors:* Different areas have different densities of toll stations and different volumes of traffic. We evaluate VeMo in different areas in Guangdong, i.e., both the downtown areas and suburban areas.
- *Vehicle Types:* Different types of vehicles may have different challenges of prediction. We evaluate it by applying VeMo on different types of vehicles, i.e., car, bus, truck.

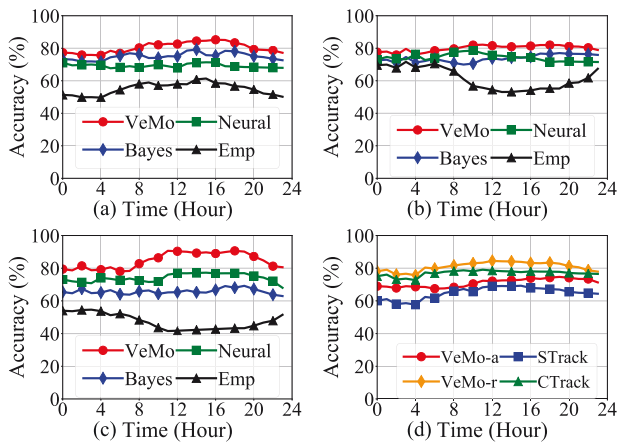


Fig. 8. Comparison to baselines: (a) destination prediction; (b) Route prediction; (c) Speed prediction; (d) location prediction.

5.2 Evaluation Results

5.2.1 Real-Time Edge-Cloud Design

Since most of the applications built on our system require real-time responses, it is necessary to have real-time cloud components. Even it is feasible to conduct prediction in a powerful server, however, it is challenging to update the model in the cloud in real time. Our solution is to combine both the cloud (i.e., center servers) and the edges (i.e., computer systems in the toll stations).

Cloud. All the data is stored in the cloud system for security issues. As the new data collected in the edges, the data is transmitted to the cloud through Ethernet. All the trained models are also stored in the cloud to distribute to the edges.

Edge. Given the truth that a vehicle only appears in a few toll stations, we could pre-distribute the trained individual models to top frequent edges according to historical records. Considering the online updating feature of our model, we update the model directly in the edge devices and transmit it back to the cloud. Generally, a vehicle leaving the toll station would not get back to the highways immediately. There is enough time to transmit the model to the cloud.

5.2.2 Comparison to Baselines

We evaluate both individual predictors and overall location predictor. For each of the individual components, we evaluate it by comparing it to the three baselines, respectively. Then three predictors work collaboratively to predict the locations of vehicles.

(i) *Destination Prediction.* Fig. 8(a) plots the destination prediction result, where VeMo has better performance than other three models with an average gain of 11 percent. *Bayes* performs better than *Neural*, which means the probability relationship is better to model the destination prediction problems. Moreover, *Emp* achieves 60 percent accuracy during the day time, which suggests the destination choices are relatively stable.

(ii) *Route Prediction.* Fig. 8(b) presents the result of route prediction. Compared to the other baselines, VeMo achieves an average performance gain of 6 percent. It suggests the performance does not vary much in terms of

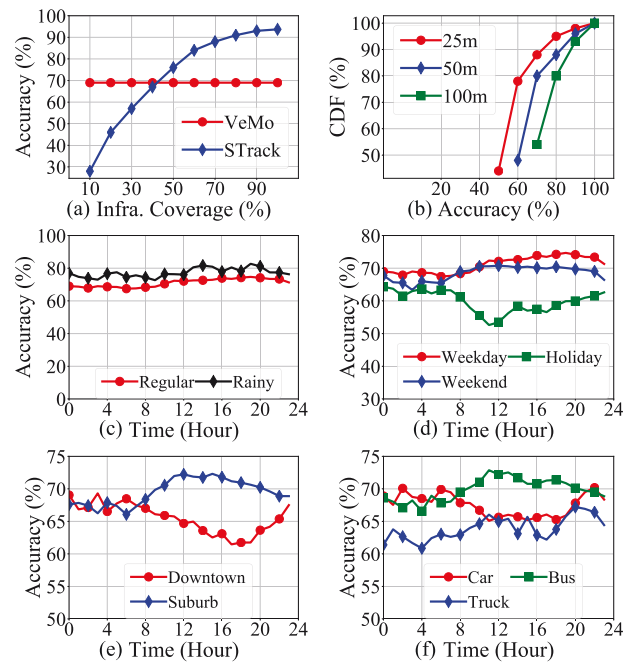


Fig. 9. Impact factors: (a) % of infrastructure; (b) accuracy threshold; (c) weather impact; (d) time impact; (e) spatial impact; (f) type impact.

different learning models. *Emp* has a similar performance in the morning but poor performance during the daytime, which means the route choices are flexible when there is heavy traffic.

(iii) *Speed Prediction.* Fig. 8(c) shows the result of speed prediction. VeMo has an average performance gain of 17 percent. During the day time, the accuracy is higher, because the heavy traffic constrains the speed variation. *Emp* shows poorer performance during the daytime because the empirical knowledge cannot obtain the real-time traffic information. Moreover, *Neural* is better than *Bayes*, which suggests the advantage of linear combination based method on the speed prediction tasks.

(iv) *Location Prediction.* We combine individual predictors together to evaluate the real-time locations of vehicles. Since the route has dominating impacts on the locations of vehicles, to show more sophisticated evaluations, we test the accuracy of both (i) the vehicles (*VeMo-a*) and (ii) those vehicles with correctly predicted routes (*VeMo-r*). Then we compare them with *STrack* and *CTrack*. Fig. 8(d) plots the evaluation results. Considering the vehicles with correctly predicted routes, VeMo (shown as *VeMo-r*) has the average accuracy about 82 percent. The reason that VeMo has similar accuracy as *CTrack* is that the baseline experiment is conducted inner city, which has a dense cell tower distribution. Even including all the vehicles (shown as *VeMo-a*), VeMo achieves an average accuracy of 70 percent, which is still at the same level of *STrack*, which means VeMo can be an alternative solution of *STrack* without introducing extra infrastructures.

We also evaluate the impacts of coverage percentage of *STrack*, and show the result in Fig. 9(a). After the coverage percentage increases to 50 percent, *STrack* achieves better performance. Since it is expensive to provide such high infrastructure coverage, VeMo outperforms *STrack* in terms of feasibility.

TABLE 3
Individual Component Efficiency

Relative Improvement	Emp	Bayes	Neural
Destination	-56%	+9%	+23%
Route	-39%	+8%	+19%
Speed	-53%	+14%	+11%

5.2.3 Impacts of Factors

Five factors are evaluated including accuracy threshold, weather, time factors, spatial factors and vehicle types. The metrics are the same as the Equation (4).

(i) *Impacts of Threshold.* We choose accuracy threshold including 25, 50, 100 meters to show how the accuracy changes in Fig. 9(b). The lower the line, the better the accuracy. We found higher thresholds lead to higher accuracy. The 100-meter threshold has higher accuracy while 25-meter and 100-meter thresholds have obvious lower accuracy.

(ii) *Impacts of Weather.* We select one day with heavy rain and compare it to a regular day in Fig. 9(c). We surprisingly found the rain even increase the prediction accuracy. Since people tend to drive slowly in the heavy rain, the individual speed is reduced and there is a smaller range of speed variance on the way, which benefits the prediction accuracy.

(iii) *Impacts of Time Factors.* Fig. 9(d) shows the performance of VeMo in weekday, weekend and holiday. The accuracy in weekday and weekend is similar. Moreover, the performance of the holiday is different than other days, especially during the morning. This is because the destination choices are less predictable on the holidays when people generally do not follow regular mobility patterns.

(iv) *Impacts of Spatial Factors.* We investigate the performance of VeMo in both downtown areas and suburb areas and show the result in Fig. 9(e). In the early morning, two areas have similar accuracy. During the daytime starting at 8 am, the performance in the downtown areas decreases. This is because the road structure is more complex in those areas, which makes the route prediction less accurate.

(v) *Impacts of Vehicle Types.* Fig. 9(f) shows the performance of different types of vehicles. Trucks have the lowest accuracy because they have longer travel distances and irregular mobility patterns (e.g., one truck may travel between different areas for cargo services as long as there are demands of cargo transportation). Buses have higher accuracy because they have the most regular mobility patterns compared to trucks and cars. Cars' accuracy decreases during the daytime because they generally travel inner cities, which is impacted by both traffic conditions and road structures.

5.3 Overhead

5.3.1 Overall Efficiency

We implement VeMo on a server with Intel Xeon E5-1660 3.00GHz CPU and 32 GB RAM in 16 threads. After loading all the data, the training process takes 450 seconds. The speed prediction is 500 times per thread every second on average, which can satisfy the real-time need of 4 million daily transactions.

5.3.2 Sub-Components Efficiency

We compare the efficiency of the individual components with three baselines. The result is summarized in Table 3, where each cell is the relative improvement (i.e., prediction speed) compared with each baseline. For the destination prediction, *Emp* has the best efficiency. The reason is that *Emp* is based on the most frequent history, which can be implemented with nearly constant time by a HashTable data structure. Compared with *Bayes* and *Neural*, our method performs with 9 and 23 percent faster speed, respectively. A similar result is found in the route prediction. As for the speed prediction, *Emp* is also the fastest because *Emp* is implemented by the historical average. Compared with *Bayes* and *Neural*, our method outperforms them by 14 and 11 percent, respectively. Overall, our method has a lower prediction speed than *Emp* but provides much better accuracy. Compared with *Bayes* and *Neural*, our method outperforms them both in accuracy and efficiency.

6 APPLICATION

6.1 Overview

Based on VeMo's output, we design a novel application to detect highway anomaly events. Traditionally, there are two approaches to detect highway anomalies for emergency responses: (i) administrators dispatch vehicles to manually check each road segment [31]; (ii) nearby drivers report anomalies by chance (e.g., call 911). However, it is always desirable to have a system automatically report the event without relying on human participation that is not predictable. To this end, we aim to detect the highway anomaly events automatically based on our infrastructure (i.e., the ETC system) and mobility model (i.e., VeMo). Specifically, we answer two questions: (i) whether there is an anomaly event on highways; (ii) if so, where the anomaly event happens. Our key idea is that (i) the anomaly event would increase the travel time that leads to travel delay of many vehicles; (ii) different vehicles would be impacted given different locations of the anomaly event. We obtained the highway anomaly events from the online traffic alarm reports of Shenzhen Transportation Administration. In total, we collected 21 highway anomaly events in January 2016, including 20 accidents and 1 road maintenance event.

6.2 Anomaly Event Detection

One of the key impacts of the anomaly event on highways is the travel delay. For example, a vehicle passing through anomaly areas would have a longer travel time to its exit station compared to regular traffic conditions. If many vehicles encounter the same issue, we may conclude that there is a potential anomaly event on highways. This gives us an opportunity that by analyzing vehicular travel time on highways (i.e., travel time between entering and exiting stations), we could know if there is an ongoing anomaly event happening. Further, each vehicle exiting highways brings the latest updates about the anomaly event whether it is still lasting.

To show the travel time difference under regular conditions and anomaly conditions, we select a traffic accident as an exemplary scenario. Fig. 10(a) shows the location of the accidents at 6:50 am caused by the crash between a car and

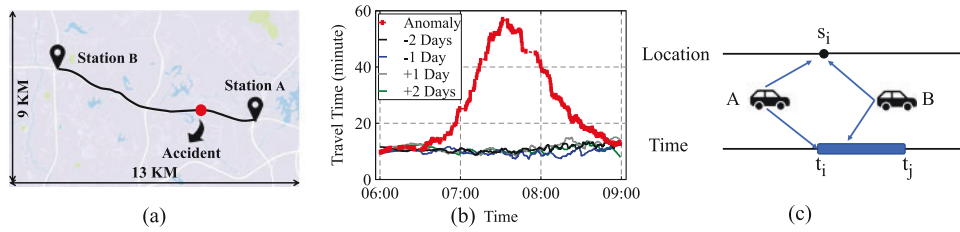


Fig. 10. Application scenario: (a) event location; (b) travel time; (c) anomaly location inference.

a bus. We analyze the travel time between station A and station B to demonstrate the impact of the accident in Fig. 10 (b). We found the travel time dramatically increases from 7 am. At the worst time near 7:30 am, the travel time increases to 1 hour. We also plot the travel time of the same time period in the previous two days and the next two days. We found the travel time is stable near 10 minutes. In summary, Fig. 10(b) shows a significantly different travel time observation when accidents happen.

Based on the observation, we design a simple and effective rule to detect the anomaly event. Empirically, we implement it with a sliding window of θ (i.e., 15 minutes). When each vehicle exit an ETC station, we compare the average travel time with the historical average travel time of the same time window. If the travel time is α (i.e., 2) times than the historical average, we claim there is an anomaly event happening on highways. Note that θ and α are tunable in different contexts. Considering the alarm reports may not cover all the anomaly events, we utilize *recall* as our performance metric. In our experiment, we successfully detect 18 anomaly events, which counts for recall of 85.7 percent.

6.3 Anomaly Location Inference

After detecting an anomaly event, we next infer where the anomaly event actually happens. For each vehicle on highways, our system VeMo outputs its real-time locations, which can be presented as a spatiotemporal sequence (i.e., $\langle t_0, s_0 \rangle, \dots, \langle t_n, s_n \rangle$). We demonstrate the basic idea in Fig. 10(c). Suppose there is an anomaly event happening on the location s_i between time t_i and t_j . If vehicle A did not pass s_i before t_i , then its travel time would be impacted. In contrast, vehicle B was not impacted. Given the observations from ETC stations, we could obtain which vehicles' travel time is impacted. Then our goal is to find the location s_i that maximizes the probability of our observations. Mathematically, we aim to optimize the objective function

$$\arg \min_{s \in S, t \in T} \left| \sum_i^N I(v_i | s, t) - Obser \right|, \tag{5}$$

where S is the set of locations, T is the time slots, N is the number of observed vehicles, v_i is the i_{th} vehicle, $Obser$ is the number of observed impacted vehicles, $I(v_i | s, t)$ is the indicator function representing whether v_i is impacted defined in Eq. (6).

$$I(v_i | s, t) = \begin{cases} 1 & \text{if } v_i \text{ passes } s \text{ before } VeMo(v_i, t), \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where $VeMo(v_i, t)$ is the output of VeMo that represents the location of v_i at time t .

The solution is straightforward that we apply a breadth-first search algorithm to iterate all the possible s and t to select the ones with minimized Eq. (5). To improve the computing efficiency, we prune the search space by two strategies: (i) considering impacted the highways between all the possible stations, we consider the common road segment of them, which prunes the space of S ; (ii) given the anomaly event first detected time, we only consider the time period of 15 minutes before, which prunes the space of T . In our experiment, considering the accurate time missing in the alarm reports, we only evaluate the location inference accuracy as the absolute distance between the inference and the actual location. The final result shows an average error of 286.5 meters, which demonstrates the effectiveness of our inference result. Combining it with the real-time anomaly event detection, the traffic administration could dispatch staff to the anomaly location from the nearest entering stations as early as possible.

6.4 Comparison With Group-Level Approaches

There are also other works detecting anomaly events by two approaches: (i) GPS-based approaches [32], [33]; (ii) group-level travel time based approaches [34], [35]. For GPS-based approaches, if a number of vehicles get stuck on the same road, there may be an anomaly event. However, they require a large number of participants with real-time GPS collection to provide enough spatial coverage, which does not apply to our scenario. For travel time based approaches, they can detect the happening of anomaly events if a number of vehicles' travel time are longer than regular time. However, they cannot provide fine-grained anomaly location inference while only coarse-grained road-level locations. In contrast, our method does not require GPS collection, which makes it easy to deploy. By modeling the mobility of each individual vehicle, we can infer where anomaly events actually happen, which can help the authority quickly reach these locations from the nearest entrances.

7 DISCUSSIONS

Lessons Learned. We list several lessons we learned in our study as follows.

- vehicle's mobility pattern in terms of destinations can be identified as three major groups, single-time travel vehicles, commuting vehicles and multi-destination vehicles;

- the overall distributions of both speed STDs cross vehicles and speeds on the road segment follow strong normality, which can be considered as constraints to infer the routes and speeds simultaneously;
- individual highway speeds vary based on driving behaviors but are highly correlated with generic traffic speeds. The overall derivation of individual speeds follows a Gaussian-like distribution;
- anomaly events are detectable by the travel time of vehicles given the stable historical travel time in regular situations. The event locations can be inferred by the real-time locations of vehicles considering which vehicles are impacted.

Why ETC Data Only? In this work, we focus on ETC data only because ETC systems provide a full penetration rate transparently based on data already collected. Moreover, the ETC based toll system is universal and exists almost everywhere even in developing countries where satellite images or mobile infrastructure is not well penetrated. If combined with other datasets even with small scale, e.g., GPS data from highway service vehicles or traffic camera data, we may be able to further improve our accuracy.

Data Collection and Privacy Protection. The ETC data used in our work have been anonymized, which is under the consent agreement of users. In the agreement, the users are notified that their data will be collected and used for analyses to improve highway management. The GPS data we utilized have been anonymized and collected by an insurance company as a part of usage-based insurances for discounts that their data would be used for research purposes under users' agreement. Further, the GPS data only cover the period on highways, which do not expose users' information such as home or work locations. All the data are collected by an opt-out policy that users can ask to opt out from the research-purpose data collection.

Generalization. Our approach is based on the highway system context. According to the survey [36], [37], 48 major countries and regions have toll roads which are managed by similar ETC systems or manual toll collection, where we envision our approach can be adapted in all these similar systems. For the discovered insights, it reflects how humans drive in general, which depends on the inherent driving behaviors instead of a specific dataset. Further, all the features we use in our approach can be found in other similar contexts. For other contexts such as local vehicles, our system can be potentially adapted to other infrastructures. For example, if a local region is deployed with surveillance cameras, vehicles could be detected by cameras, which can be an analogy to our toll stations.

Limitations and Open Problems.

- Our system work in a controlled environment, i.e., a highway system with both entering and existing records. Therefore the same technique may not be applied to local streets without toll booths to track every vehicle enter or leave a street. In this case, additional data, e.g., partial GPS, can be combined with our solution for prediction. However, we believe our solution can be generalized to stationary sensors that can capture the vehicle's passing such as cameras.

TABLE 4
Vehicular Mobility Survey

	Aggregate	Individual	
Mobile	[44], [45], [46], [47]	[48], [49], [50], [51], [52], [53]	
Static	[35], [38], [54], [55]	Partial Penetration [30], [39], [56]	Full Penetration Our work

- Even ETC systems only capture the vehicle twice, it still has the privacy issue of exposing locations. However, compared with GPS based solutions, it is a better/low-cost privacy-reserved approach since ETC data have already been collected as a mandatory process for billing; whereas other approaches need new devices or dedicate data collection process with potential continuous location collection.
- Our solution can help detect the anomaly events when there are a certain number of vehicles have abnormal travel time. But in this work, we focused on the fundamental location prediction and did not try to explicitly handle the anomalies, which would be a good direction for our future work.
- Our solution relies on the historical data of vehicles to learn their driving behaviors. In the ETC system, we found there is only 9 percent of the new vehicles without any historical data after 10-day data accumulation, which is a very small number of vehicles. For those without historical data, we can only infer their behaviors according to majority behaviors (i.e., crowd features). It is still an interesting open problem.

8 RELATED WORK

We divide most related works into two major parts (shown in the Table 4): mobile sensing based approaches and stationary sensing based approaches.

Static Infrastructure. Static infrastructures, e.g., traffic cameras [38], cell towers [30], WiFi access point [39], are widely used for vehicle mobility modeling. Some communication related works are also studies based on the static infrastructure [40], [41], [42], [43]. Compared with existing work, our approach makes use of the existing infrastructures to predict vehicle mobility without extra cost. All the vehicles entering the highways are detected, which does not require the installation of interaction apps. The requirement of only single real-time observations, e.g., entrance to a highway, largely increases the feasibility of our approach in the real world. Some approaches such as cell phone networks may have the potential to infer traffic conditions at low cost. But normally the cellphone data are not available for highway administrators. They can only use the data collected by themselves. Further, the cell phone network cannot be narrowed down to vehicular mobility since the driver and passage cannot be distinguished from the cell phone data only, which may introduce extra bias.

Mobile Infrastructure. Mobile infrastructures, i.e., smartphones and onboard devices, are extensively studied to understand vehicular mobility. [39], [49], [57] use smartphones to track vehicles in real time. The inference on mobility is studied in detail by smartphone data [50]. [51]

estimates the urban traffic using vehicular fleets with onboard devices. [58] implements regular vehicle tracking through commercial vehicles with onboard devices. Other works [59], [60], [61] such as crowdsourcing information collection and energy issues can also benefit from the mobile infrastructures. However, these approaches are either limited by low penetration rates of apps [13] or focus on the aggregated level [51], and typically raise privacy issues of exposing vehicle GPS data [49].

Vehicular mobility on the highways is also studied in the transportation community (i.e., the destination and speed prediction [62], [63], [64], [65]). However, previous works mainly focused on the aggregated traffic characteristics, such as origin-destination matrix or traffic speed on the road segments. Different from these works, our system aims at the mobility model of individual vehicles, which requires a microscope analysis of the vehicle mobility pattern.

9 CONCLUSION

In this work, we focus on vehicular mobility modeling on large-scale highway systems. In particular, we motivate and design a novel system called VeMo with three components for the destination, route, and speed inference. Combining them together, we infer the real-time locations of vehicles on highways. More importantly, we implement and evaluate VeMo based on the large-scale data in the Guangdong highway network in China, utilizing a large-scale ETC system with 773 stations and a large-scale vehicle fleet with GPS data as ground truth. We advance state-of-the-art vehicle mobility modeling approaches by some key lessons we learned. Based on the modeling result, we implement a real-world application to detect ongoing anomaly events on highways. In the future, we look forward to investigating personal driving behaviors and design coexistence strategies in the scenario of heterogeneous vehicles including regular and autonomous vehicles.

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable comments. This work was supported in part by NSF 1849238, 1932223, 1952096, and 2003874.

REFERENCES

- [1] D. Huang, S. Shere, and S. Ahn, "Dynamic highway congestion detection and prediction based on shock waves," in *Proc. 7th ACM Int. Workshop Veh. Internetw.*, 2010, pp. 11–20.
- [2] E. Ben-Elia and Y. Shiftan, "Which road do I take? A learning-based model of route-choice behavior with real-time information," *Transp. Res. A, Policy Pract.*, vol. 44, no. 4, pp. 249–264, 2010.
- [3] T. Schmidt-Dumont and J. H. van Vuuren, "Decentralised reinforcement learning for ramp metering and variable speed limits on highways," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 8, 2015, Art. no. 1.
- [4] 511NJ, "511NJ: Get connected and go!" 2017. [Online]. Available: <http://www.511nj.org/cameras.aspx>
- [5] NYC Department of Transportation, "New york city camera," 2017. [Online]. Available: <http://dotsignals.org>
- [6] Google, "Google map," 2017. [Online]. Available: <https://www.google.com/maps>
- [7] N. Capurso, E. Elskan, D. Payne, and L. Ma, "Poster: A robust vehicular accident detection system using inexpensive portable devices," in *Proc. 12th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2014, Art. no. 367.
- [8] Y. Wan, Y. Huang, and B. Buckles, "Camera calibration and vehicle tracking: Highway traffic video analytics," *Transp. Res. Part C: Emerg. Technol.*, vol. 44, pp. 202–213, 2014.
- [9] S. Taghvaeeyan and R. Rajamani, "Portable roadside sensors for vehicle counting, classification, and speed measurement," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 73–83, Feb. 2014.
- [10] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu, "Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 237–248.
- [11] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw.*, 2011, pp. 145–156.
- [12] US DOT, 2017. [Online]. Available: [https://www.itscosts.its.dot.gov/ITS/benecost.nsf/DisplayRUCByUnitCostElementUnadjusted?ReadForm&UnitCostElement=CCTV+Video+Camera&Subsystem=Roadside+Detection+\(RS-D\)](https://www.itscosts.its.dot.gov/ITS/benecost.nsf/DisplayRUCByUnitCostElementUnadjusted?ReadForm&UnitCostElement=CCTV+Video+Camera&Subsystem=Roadside+Detection+(RS-D))
- [13] Statista, "Number of Internet users who used the Internet for route planning, to access maps or road maps (e.g., Google maps) in Germany from 2013 to 2016, by frequency (in millions)," 2017. [Online]. Available: <https://www.statista.com/statistics/432169/online-route-planning-and-map-usage-eg-google-maps-germany>
- [14] GHSA, "Speed cameras on highways," 2018. [Online]. Available: <https://www.ghsa.org/state-laws/issues/speed-and-red-light-cameras>
- [15] Insurance Information Institute, "Highway safety," 2017. [Online]. Available: <https://www.iii.org/fact-statistic/facts-statistics-highway-safety>
- [16] R. Dewri, P. Annadata, W. Eltarjaman, and R. Thurimella, "Inferring trip destinations from driving habits data," in *Proc. 12th ACM Workshop on Privacy Electron. Soc.*, 2013, pp. 267–272.
- [17] J. Yu *et al.*, "SenSpeed: Sensing driving conditions to estimate vehicle speed in urban environments," *IEEE Trans. Mobile Comput.*, vol. 15, no. 1, pp. 202–216, Jan. 2016.
- [18] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, 2008, Art. no. 779.
- [19] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 201–212.
- [20] M. Li, A. Ahmed, and A. J. Smola, "Inferring movement trajectories from GPS snippets," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 325–334.
- [21] T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 163–172.
- [22] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.
- [23] X. Gao, B. Firner, S. Sugrim, V. Kaiser-Pendergrast, Y. Yang, and J. Lindqvist, "Elastic pathing: Your speed is enough to track you," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 975–986.
- [24] B. Hellenga, P. Izadpanah, H. Takada, and L. Fu, "Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments," *Transp. Res. C, Emerg. Technol.*, vol. 16, no. 6, pp. 768–782, 2008.
- [25] E. Ben-Elia, R. Di Pace, G. N. Bifulco, and Y. Shiftan, "The impact of travel information's accuracy on route-choice," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 146–159, 2013.
- [26] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh, "Mondrian forests: Efficient online random forests," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3140–3148.
- [27] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. New York, NY, USA: Springer, 2001.
- [28] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2009, pp. 336–343.
- [29] M. S. Grewal, L. R. Weill, and A. P. Andrews, *Global Positioning Systems, Inertial Navigation, and Integration*. Hoboken, NJ, USA: Wiley, 2007.
- [30] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod, "Accurate, low-energy trajectory mapping for mobile devices," in *Proc. 8th USENIX Conf. Netw. Syst. Des. Implementation*, 2011, pp. 267–280.
- [31] California Highway Patrol, "Chp traffic incident information page," 2020. [Online]. Available: <https://www.chp.ca.gov/traffic>

- [32] F. Seraj, B. J. van der Zwaag, A. Dilo, T. Luarasi, and P. Havinga, "RoADS: A road pavement monitoring system for anomaly detection using smart phones," in *Big Data Analytics in the Social and Ubiquitous Context*. Berlin, Germany: Springer, 2015, pp. 128–146.
- [33] R. Bauza, J. Gozalvez, and J. Sanchez-Soriano, "Road traffic congestion detection through cooperative vehicle-to-vehicle communications," in *Proc. IEEE Local Comput. Netw. Conf.*, 2010, pp. 606–612.
- [34] Z. Fang *et al.*, "MAC: Measuring the impacts of anomalies on travel time of multiple transportation systems," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–24, 2019.
- [35] Y. Yang, F. Zhang, and D. Zhang, "SharedEdge: GPS-free fine-grained travel time estimation in state-level highway systems," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, 2018, Art. no. 48.
- [36] Wikipedia, "List of electronic toll collection systems," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_electronic_toll_collection_systems
- [37] Wikipedia, "List of toll roads," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_toll_roads
- [38] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura, "FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3687–3696.
- [39] A. Thiagarajan *et al.*, "VTrack: Accurate, energy-aware road traffic delay estimation using mobile phones," in *Proc. 7th ACM Conf. Embedded Netw. Sensor Syst.*, 2009, pp. 85–98.
- [40] A. Balasubramanian, R. Mahajan, A. Venkataramani, B. N. Levine, and J. Zahorjan, "Interactive WiFi connectivity for moving vehicles," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 427–438, 2008.
- [41] Y. Li, C. Peng, Z. Yuan, J. Li, H. Deng, and T. Wang, "Mobileinsight: Extracting and analyzing cellular network information on smartphones," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 202–215.
- [42] F. Bai, K. R. Moghadam, and B. Krishnamachari, "A tale of two cities-characterizing social community structures of fleet vehicles for modeling V2V information dissemination," in *Proc. 12th Annu. IEEE Int. Conf. Sens. Commun. Netw.*, 2015, pp. 506–514.
- [43] Z. Fang, F. Zhang, L. Yin, and D. Zhang, "MultiCell: Urban population modeling based on multiple cellphone networks," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, 2018, Art. no. 106.
- [44] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.
- [45] A. V. Khezerlou, X. Zhou, L. Li, Z. Shafiq, A. X. Liu, and F. Zhang, "A traffic flow approach to early detection of gathering events: Comprehensive results," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 6, 2017, Art. no. 74.
- [46] M. X. Hoang, Y. Zheng, and A. K. Singh, "FCCF: Forecasting city-wide crowd flows based on big data," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2016, Art. no. 6.
- [47] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Taxi-passenger-demand modeling based on Big Data from a roving sensor network," *IEEE Trans. Big Data*, vol. 3, no. 3, pp. 362–374, Sep. 2017.
- [48] D. Zhang, J. Zhao, F. Zhang, and T. He, "UrbanCPS: A cyber-physical system based on multi-source big infrastructure data for heterogeneous model integration," in *Proc. ACM/IEEE 6th Int. Conf. Cyber-Physical Syst.*, 2015, pp. 238–247.
- [49] Y. Zhao *et al.*, "GreenDrive: A smartphone-based intelligent speed adaptation system with real-time traffic signal prediction," in *Proc. ACM/IEEE 8th Int. Conf. Cyber-Physical Syst.*, 2017, pp. 229–238.
- [50] A. Thiagarajan, J. Biagioni, T. Gerlich, and J. Eriksson, "Cooperative transit tracking using smart-phones," in *Proc. 8th ACM Conf. Embedded Netw. Sensor Syst.*, 2010, pp. 85–98.
- [51] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale traffic estimation from a roving sensor network," in *Proc. 10th ACM Conf. Embedded Netw. Sensor Syst.*, 2012, pp. 141–154.
- [52] X. Xie, F. Zhang, and D. Zhang, "PrivateHunt: Multi-source data-driven dispatching in for-hire vehicle systems," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, 2018, Art. no. 45.
- [53] L. Liu *et al.*, "BigRoad: Scaling road data acquisition for dependable self-driving," in *Proc. 15th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2017, pp. 371–384.
- [54] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," in *Proc. 25th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2017, Art. no. 1.
- [55] Z. Qin, Z. Fang, Y. Liu, C. Tan, W. Chang, and D. Zhang, "EXIMIUS: A measurement framework for explicit and implicit urban traffic sensing," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, 2018, pp. 1–14.
- [56] G. Chandrasekaran *et al.*, "Tracking vehicular speed variations by warping mobile phone signal strengths," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2011, pp. 213–221.
- [57] B.-J. Ho, P. Martin, P. Swaminathan, and M. Srivastava, "From pressure to path: Barometer-based vehicle tracking," in *Proc. 2nd ACM Int. Conf. Embedded Syst. Energy-Efficient Built Environ.*, 2015, pp. 65–74.
- [58] X. Xie *et al.*, "coSense: Collaborative urban-scale vehicle sensing based on heterogeneous fleets," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 4, 2018, Art. no. 196.
- [59] X. Chen, X. Wu, X.-Y. Li, Y. He, and Y. Liu, "Privacy-preserving high-quality map generation with participatory sensing," in *Proc. IEEE Conf. Comput. Commun.*, 2014, pp. 2310–2318.
- [60] T. Zhang, N. Leng, and S. Banerjee, "A vehicle-based measurement framework for enhancing whitespace spectrum databases," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 17–28.
- [61] S. Mathur *et al.*, "ParkNet: Drive-by sensing of road-side parking statistics," in *Proc. 8th Int. Conf. Mobile Syst. Appl. Services*, 2010, pp. 123–136.
- [62] G.-L. Chang and J. Wu, "Recursive estimation of time-varying origin-destination flows from traffic counts in freeway corridors," *Transp. Res. Part B: Methodol.*, vol. 28, no. 2, pp. 141–160, 1994.
- [63] J. Weng, R. Yuan, R. Wang, and C. Wang, "Freeway travel speed calculation model based on ETC transaction data," *Comput. Intell. Neurosci.*, vol. 2014, 2014, Art. no. 48.
- [64] X. Wang, K. An, L. Tang, and X. Chen, "Short term prediction of freeway exiting volume based on SVM and KNN," *Int. J. Transp. Sci. Technol.*, vol. 4, no. 3, pp. 337–352, 2015.
- [65] J. Zhao, Y. Gao, Y. Guo, and Z. Bai, "Travel time prediction of expressway based on multi-dimensional data and the particle swarm optimization-autoregressive moving average with exogenous input model," *Advances Mech. Eng.*, vol. 10, no. 2, 2018, Art. no. 1687814018760932.



Yu Yang is currently working toward the PhD degree from the Department of Computer Science, Rutgers University, Piscataway, New Jersey, working with professor Desheng Zhang. His research interests include mobile, sensing and networked systems in the scope of smart cities, with a major focus on vehicular mobility modeling and vehicle-based sensing. He is also interested in data mining and applied machine learning in spatiotemporal data.



Xiaoyang Xie is currently working toward the PhD degree in the Department of Computer Science, Rutgers University, Piscataway, New Jersey. He is interested in applying data analytics, data mining, machine learning on large scale urban data for urban sensing, and services design. His research interests include the technical integration on the large scale heterogeneous urban data, including taxis, buses, personal vehicles, logistical trucks, smart cards, etc.

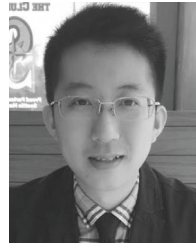


Zhihan Fang received the BS degree from the School of Software Engineering, Tongji University, China, in 2015. He is currently working toward the PhD degree in computer science from Rutgers University, Piscataway, New Jersey, since 2015. Currently, he is a research assistant in ARSENAL Lab with professor Desheng Zhang. His research interests include urban cyber physical systems, urban data mining and analytics, data fusion, and integration from heterogeneous sensor networks.



Fan Zhang (Member, IEEE) received the PhD degree in communication and information system from the Huazhong University of Science and Technology, China, in 2007. He was a postdoctoral fellow with the University of New Mexico, Albuquerque, New Mexico, and University of Nebraska-Lincoln, Lincoln, Nebraska from 2009 to 2011. He is currently a professor at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He is also the director of the Shenzhen Institute of Beidou Applied

Technology(SIBAT), China. His research interests include intelligent transportation systems, urban computing, and Big Data and AI technology.



Desheng Zhang (Member, IEEE) is currently an assistant professor with the Department of Computer Science, Rutgers University, Piscataway, New Jersey. He is broadly concentrated on bridging cyber-physical systems also known as Internet of Things under some contexts and big urban data by technical integration of communication, computation, and control in data-intensive urban systems. He is also focused on the life cycle of big-data-driven urban systems, from multi-source data collection to streaming-data processing, heterogeneous-data management, model abstraction, visualization, privacy, service design, and deployment in complex urban setting.



Yang Wang (Member, IEEE) received the PhD degree from the University of Science and Technology of China (USTC), Hefei, China, in 2007, under supervision of Prof. L. Huang. He is currently an associate professor at the University of Science and Technology of China, China. He also worked as a postdoc at the University of Science and Technology of China, China with Prof. L. Huang. His research interest mainly includes wireless (sensor) networks, distributed systems, data mining, and machine learning.

▷ **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.**