Citywide Traffic Volume Estimation Using Trajectory Data

Xianyuan Zhan, Yu Zheng, *Senior Member, IEEE*, Xiuwen Yi, and Satish V. Ukkusuri

Abstract—Traffic volume estimation at the city scale is an important problem useful to many transportation operations and urban applications. This paper proposes a hybrid framework that integrates both state-of-art machine learning techniques and well-established traffic flow theory to estimate citywide traffic volume. In addition to typical urban context features extracted from multiple sources, we extract a special set of features from GPS trajectories based on the implications of traffic flow theory, which provide extra information on the speed-flow relationship. Using the network-wide speed information estimated from a travel speed estimation model, a volume related high level feature is first learned using an unsupervised graphical model. A volume re-interpretation model is then introduced to map the volume related high level feature to the predicted volume using a small amount of ground truth data for training. The framework is evaluated using a GPS trajectory dataset from 33,000 Beijing taxis and volume ground truth data obtained from 4,980 video clips. The results demonstrate effectiveness and potential of the proposed framework in citywide traffic volume estimation.

Index Terms—Urban computing, traffic volume estimation, trajectories, traffic flow theory

1 Introduction

Traffic volume is a central traffic state measure that has a wide range of applications. For example, the citywide traffic volume pattern is often used as the basis for transportation and urban planning. Local transportation agencies also need real-time volume information to perform interventions on traffic, e.g., alter traffic signal timing or close certain road, in order to react to severe congestion or emergency events. Moreover, traffic volume serves as the input data for computing vehicle emission, which is required in many pollution monitoring systems [1].

Traditional approaches for traffic volume estimation and prediction heavily rely on data from various road-based sensors, i.e., loop detectors [2], [3], [4], [5] or surveillance cameras [6], thus mainly applicable to major road sections or limited-scale road networks. In many of these studies, traffic volume measures are directly monitored by sensors, and the volume estimation or prediction is achieved using filtering-based algorithms. Other studies utilize indirect traffic state measures, e.g., traffic density or speed, to estimate traffic volume using fundamental diagrams (FD) of traffic flow [2], [7], [8]. This approach exploits the intrinsic relationship between traffic volume, density and speed to perform estimation. The major drawback of FD-based approach, lies in the need for calibration using sufficient amount of traffic data for each individual road. Also, FD is typically designed for highways and major arterials and tend to perform poorly on small streets. Both of the previous two conventional approaches require installing a large number of sensors to achieve network-wide traffic volume estimation, which is neither cost-effective nor practical. Fortunately, the recent emergence of large-scale data generated by diverse sources in urban spaces has provided a new alternative for solving many urban and transportation problems [9], [32], [33]. Using the rich information contained in these big and heterogeneous data sources, tackling the citywide traffic volume estimation becomes possible [1], [10], [11], [12].

Estimating citywide traffic volume is a difficult task involving many challenges. First, due to the high cost of installing and maintaining road-based sensors, we typically do not have direct information about traffic volume at a city scale. Although it is possible to observe the real time traces from some sample vehicles (e.g., GPS equipped taxis [1], [11]) or mobile phone users (e.g., social media check-ins [10] or cellular record data [12]), it is generally insufficient to infer the detailed traffic volume on each road segment. As sample vehicles such as taxis only account for a small faction of the total traffic and lack representativeness of the overall traffic. Moreover, the real-time GPS trajectories from sample vehicles cover only limited proportion of the network, which lead to serious data sparsity issue that cannot be simply solved by interpolation [13] or using historical patterns [14]. Second, although the network-wide traffic speed estimation are relatively easy using the large-scale crowd-based sensing data [1], [15], [16], [17], how to utilize these speed estimates in volume inference still remains to be an issue. Clearly, there exists certain relationship between traffic speed and volume, e.g., fundamental diagrams of traffic flow [7], [8]. However, given the lack of ground truth volume data for each road and the scalability requirement of the problem prohibit the direct use of FD-based approaches. Third, many transportation and urban applications require real-time and citywide traffic volume estimation, which cannot be addressed by

- *X. Zhan and S.V. Ukkusuri are with Purdue University, West Lafayette, IN 47907. E-mail: {zhanxianyuan, sukkusur}@purdue.edu.*
- *Y. Zheng and X. Yi are with Microsoft Research, Beijing 100080, China. E-mail: {yuzheng, v-xiuyi}@microsoft.com.*

Manuscript received 20 Mar. 2016; revised 23 Sept. 2016; accepted 22 Oct. 2016. Date of publication 25 Oct. 2016; date of current version 9 Jan. 2017. Recommended for acceptance by J. Xu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2016.2621104

traditional approaches. This poses very high requirement on the scalability and the efficiency of the citywide volume estimation model.

To address the aforementioned challenges, we proposes a scalable traffic volume estimation procedure, referred as TVE. TVE achieves citywide traffic volume estimation using data from GPS trajectories, road networks, point of interest (POI) information as well as weather conditions, rather than relying on traffic data from road-based sensors. This paper is an extension of our previous paper [1] in which a volume estimation model is used as an important sub-component to infer citywide gas consumption and pollution emissions. In this paper, we further offering following contributions:

- We propose a hybrid framework that incorporates both the well-established traffic flow theories and highly scalable machine learning techniques to estimate the citywide traffic volume. This is completely new compared with [1].
- We propose new methods to extract traffic flow related features from GPS trajectory data. This incorporates prior knowledge from traffic flow theories and helps to improve volume estimation accuracy.
- We construct new probabilistic graphical models for learning the volume related high level feature. Compared with the model used in [1], we introduce separate graph structures for higher level roads (highways and major roads) and small roads, and consider more impacting factors.
- We develop a volume re-interpretation model to establish the mapping between the learned volume related high level feature and predicted traffic volume using a small amount of ground truth data. The new model removes the unreasonable assumption in [1] that uses normal distribution to infer traffic volume. We also show that the volume re-interpretation model greatly improves the estimation accuracy as compared with the method used in [1].
- We have conducted a large-scale ground truth data collection for more comprehensive model testing and evaluation. We recorded 4,980 video clips from 262 roads in Beijing road network, whereas [1] only uses data from 358 video clips. Our results shows that the proposed framework outperforms the approach in [1] and other baseline methods in terms of effectiveness. We will release the dataset and codes along with this paper.

The paper is organized as follows: the next section presents the preliminaries of the framework; Sections 3 and 4 describe the detailed methodologies for the traffic speed and volume inference model; Section 5 evaluates the performance of our method. After reviewing related work in Section 6, we conclude this paper in the final section.

2 Overview

2.1 Preliminaries

Definition 1. *A trajectory of a vehicle is a sequence of time-ordered spatial points $T_r : p_1 \rightarrow \cdots \rightarrow p_n$, where each point has a coordinate and a timestamp t, $p = (l, t)$.*

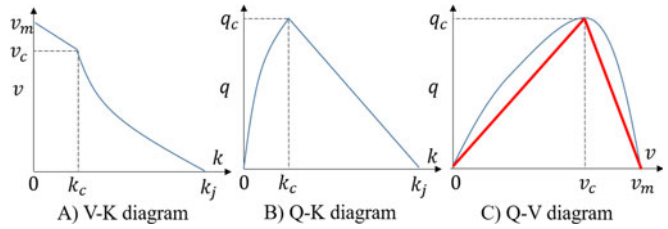


Fig. 1. A typical fundamental diagram of traffic.

Definition 2. *A point of interest (POI) is a venue, e.g., school and shopping mall in urban environment, which have a name, address, coordinates, category and other attributes.*

Definition 3. *A road network is comprised of a set of road segment $\{r\}$ connected among each other in a graph format. Each road segment r is considered as a directed edge, which contains property attributes such as length $r.l$, class $r.c$ (e.g., a highway or a street), direction $r.dir$ (e.g., one-way or bidirectional), speed limit $r.lim$, the number of lanes $r.nl$ and the number of connected road segments $r.nc$.*

Definition 4. *The traffic volume is the number of vehicles passing a reference point on a road per unit of time.*

Definition 5. *The traffic density is defined as the number of vehicles per unit length of the roadway.*

Definition 6. *The fundamental diagram of traffic flow is a diagram that characterizes the relationship between traffic volume (q), density (k) and speed (v) of a road.*

Fundamental diagram describes the empirical relationship between three key measures of traffic flow: volume, density and speed, which is the core element in traffic flow theory. FD is first observed and derived in 1930's [7], and has been verified in a large number of empirical studies. FD makes three basic observations of traffic flow: 1) the traffic speed decreases as the increase of number of vehicles (density) on a road; 2) when traffic is light (free flow regime), the traffic volume increases with the increase of density; while under congested traffic (congested flow regime), both traffic volume and speed decreases with the increase of density; 3) there exist a phase transition (critical) point separates the free flow and congested flow regime.

Fig. 1 presents a typical FD for a highway segment, which comprises three diagrams that capturing speed-density (V-K diagram), flow-density (Q-K diagram) and flow-speed (Q-V diagram) relationships. Among all the diagrams in FD, the Q-V diagram is of particular interest in this study. While the acquisition of traffic density data heavily relies on in-road traffic sensors, the network-wide road speed can be effectively estimated using many existing methods [1], [15], where Q-V diagram can provide important prior knowledge in estimating traffic volume. There are three key parameters to specify a Q-V diagram, including the capacity volume q_c, the critical speed v_c and the free flow speed v_m. The capacity volume q_c represents the maximum volume on a road under optimal traffic conditions and the critical speed v_c is the speed that corresponds to q_c on Q-V diagram. Despite the nice functional form of the FD, it should be noted that the FD of a road is impacted by a lot of factors and can be drastically different across roads. Thus

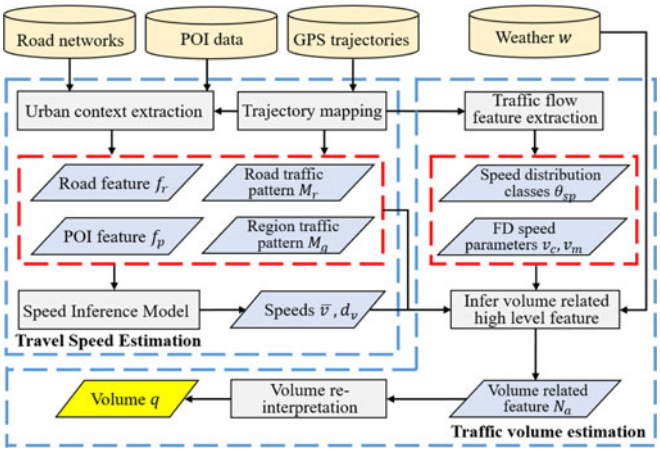


Fig. 2. Overall framework.

FD needs to be properly calibrated for each individual road segment. Furthermore, FD is typically derived for highways and major arterials, where the traffic are less impacted by nearby urban environment. The FD on small streets can be very noisy or may not exist.

2.2 Framework

Fig. 2 presents the overall framework of this paper. The framework consists of two major components: travel speed estimation (TSE) and traffic volume estimation (TVE). The TSE component contains three parts. First, the GPS trajectories from sample vehicles are mapped to the road network. The urban context features related with road network, Point of Interest as well as course-grained traffic patterns are then extracted from multiple data sources. Based on the extracted urban context features, a speed inference model is used to estimate the mean and standard deviation of road speeds for the entire network, which serves as the direct input of road traffic states. The TVE component is the main focus of this paper, which operates in three steps. First, a set of traffic flow related features are extracted from GPS trajectories, which help to establish the speed-flow relationship. Second, the volume related high level feature is learned from a partially observed Bayesian network using all the extracted features. Finally, we introduce a volume re-interpretation model to map the volume related high level feature into the predicted volume using a small amount of ground truth volume data.

3 Travel Speed Estimation

3.1 Trajectory Mapping

The GPS trajectories from sample vehicles need to be projected onto the road network before providing any useful information. We use a map-matching algorithm proposed by Yuan et al.[18] which considers both the position context of GPS points and the road network topology. Once the trajectory points p_i's are mapped to the road network, the travel speed of each point v_i, as well as the mean $\bar{v}_i$ and standard deviation d_v of travel speed on a road can be computed as

$$v_i = \frac{Dist(p_i.l, p_2.l)}{|p_{i+1}.t - p_1.t|} \quad (1)$$

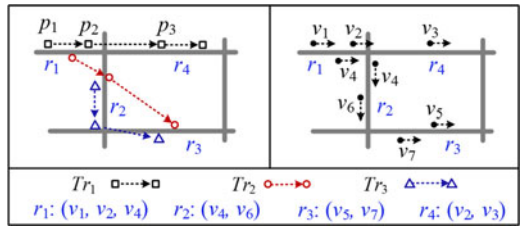


Fig. 3. Map-matching and speed computation.

$$\bar{v} = \frac{\sum_i^n v_i}{n}, \; d_v = \sqrt{\sum_i^n \frac{(v_i - \bar{v})^2}{n}}, \quad (2)$$

where $Dist$ calculates the road network distance between the two points p_i and p_{i+1}, and l, t is the location and timestamp of a data point p. The map-matching and speed computation are illustrated in Fig. 3. The value of $\bar{v}_i$ and d_v for roads covered with reasonable amount of GPS trajectories can be directly obtained. For roads covered with no or insufficient GPS trajectories, we use a travel speed estimation model to infer the speed information.

3.2 Urban Context Extraction

This subcomponent extracts two classes of urban context features from multiple data sources: 1) physical features of the road and network; and 2) historical traffic patterns to facilitate travel speed estimation.

Physical Features. The physical features of a road segment r consists of three parts: 1) road features f_r, including attribute information of length, class, direction, speed limit, number of lanes, and number of connections, etc. All the roads are further classified into three groups based on speed limit: highway (70-120 km/h), major road (50-60 km/h) and small roads (30-40 km/h). 2) POI features f_p within 50 meter radius from r's end points (see Fig. 4A). We only considered the top 10 categories that are located near road segments most frequently, namely: *Schools, Companies & Offices, Banks & ATMs, Malls & Shopping, Restaurants, Gas stations & Vehicle services, Parking, Hotels, Residences, Transportation, and Entertainment & Living Services*. For each road segment, f_p is constructed as a ten dimensional vector with each element corresponds to the number of occurrence for venues of a particular POI category. 3) Global position feature f_g, that denotes in which part of the 4×4 grids of the city a road segment belongs, which illustrated in Fig. 4B.

Traffic Patterns. Apart from the static physical features of road segments, we also construct two matrices M_r and M_g calculated from historical trajectories, which represent the fine and coarse-grained traffic patterns respectively. Specifically, each column in M_r denotes the average traffic

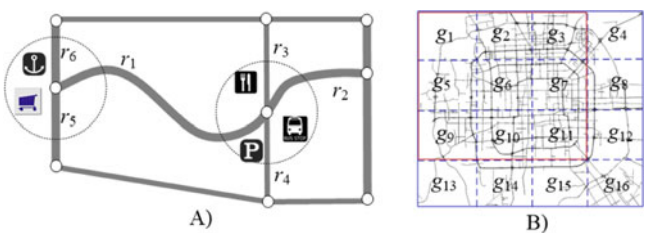


Fig. 4. Physical feature extraction.

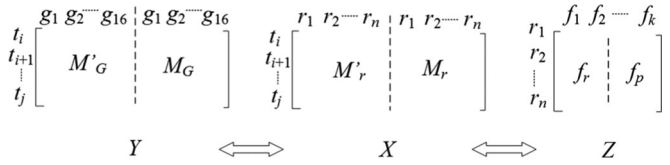


Fig. 5. Context-aware matrix factorization.

condition $(\bar{v}, d_v)$ on a particular road segment and each row represents a particular time slot (e.g., 4:00-4:10 pm) calculated using historical data over a long period (e.g. 2 months). M_g is a coarse-grained but denser representation of traffic patterns, which partition the entire road network into disjoint cells in a 4×4 grids (see Fig. 4B) instead of a road segment. Each column of M_G is the average number of vehicle traversing a specific grid cell computed based on data over long period of time, and each row denotes the corresponding time slot. M_g captures the aggregate level traffic flow pattern in the city. The similarity between two rows in the matrices models the correlation between two time slots. Due to the sparsity of sampled trajectories that covering the entire network, M_r, M_g provide important information of traffic evolution pattern and improve estimation accuracy of the travel speeds.

3.3 Speed Inference Model

The traffic conditions $(\bar{v}, d_v)$ of road segments that are not covered by GPS trajectories are inferred using a collaborative matrix factorization approach proposed in Shang et al. [1]. Matrix factorization is a widely used multivariate analysis technique in recommendation system, computer vision and other fields, which factorize a matrix into two or more matrices to learn partially observed object or latent patterns in the data [19], [20]. As depicted in Fig. 5, we formulate three matrices X, Y, and Z, where $X = [M'_r||M_r]$ represents the fine-grained traffic conditions, $Y = [M'_g||M_g]$ represents the coarse-grained traffic conditions and Z contains the physical features of road segments. Specifically, M'_r and M'_G are matrices built from the real-time trajectory data received from t_i to t_j (e.g., 1-3 pm), where t_j is the current time slot. Due to the existence of missing speeds from road segments not covered by GPS trajectories, the row t_j of M'_r in matrix X is only partially filled. M_r and M_G are constructed using historical data in the same time slots, which helps deal with the data sparsity. The rows in Z denote road segments and the columns represent different kinds of features. Our goal is to fill the missing values in M'_r with the information provided in M_r, Y and Z. This can be achieved by jointly factorize the matrices X, Y and Z as follows:

$$Y \approx T \times (G;G)^T; X \approx T \times (R \times R)^T; Z \approx R \times F^T, \quad (3)$$

where T, G, R, and F are the latent factors. X shares latent factor T with Y, which captures the temporal correlation between the two matrices. X also shares latent factor R with Z, as they correspond to the same set of roads. After collaborative factorization, we can recover X with the production of T and $(R;R)^T$, and thus obtaining the missing values in row t_j of M'_r, i.e., the unknown speeds in current time slot. There are two major advantages of the proposed approach. First, it fuses the information from multiple aspects, i.e., spatial and temporal information, fine and coarse-grained traffic pattern, as well as historical and real-time data. Second, it exploits the latent shared factor structure across different data matrices, thereby allows incorporating more information as well as control during the factorization process and results in more accurate speed estimates. The proposed collaborative matrix factorization problem is solved by minimizing following objective function

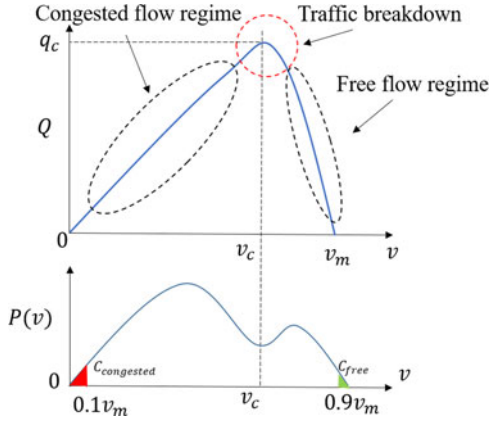


Fig. 6. Q-V diagram and road speed distribution.

$$L(T,R,G,F) = \frac{1}{2}||Y - T(G;G)^T||^2 + \frac{\lambda_1}{2}||X - T(R;R)^T||^2 + \frac{\lambda_2}{2}||Z - RF^T||^2 + \frac{\lambda_3}{2}(||T||^2 + ||R||^2 + ||G||^2 + ||F||^2), \quad (4)$$

where $||\cdot||$ denotes the Frobenius norm. The objective function can be minimized iteratively using gradient descent algorithm. Detailed discussion on the solution approach can be found in Shang et al. [1].

4 Traffic Volume Estimation

4.1 Traffic Flow Feature Extraction

GPS trajectories contain rich information about the traffic flow related features that encode the flow-speed relationship. We extract two types of traffic flow features from GPS trajectories based on the prior knowledge of traffic flow theory, namely 1) the speed distribution class θ_{sp} and speed parameters that characterize the Q-V diagram.

4.1.1 Identify Road Speed Distribution Classes

The speed distribution of a road segment reflects important information about the shape of Q-V diagram. For instance, traffic flow studies show that there exists an unstable phase transition state referred as *traffic breakdown* in real world traffic, which separates the free flow and congested flow regime [8]. When reflected in the speed distribution of a road, the traffic breakdown will lead to a low probability density region as illustrated in Fig. 6. This indicates that speed distributions of roads reveal important information which is useful to characterize the flow-speed relationship of a road segment, such as when traffic breakdown occurs and the possible ranges of free flow/congested flow regime. Moreover, the classes obtained by clustering road speed distributions can be perceived as a means to summarize

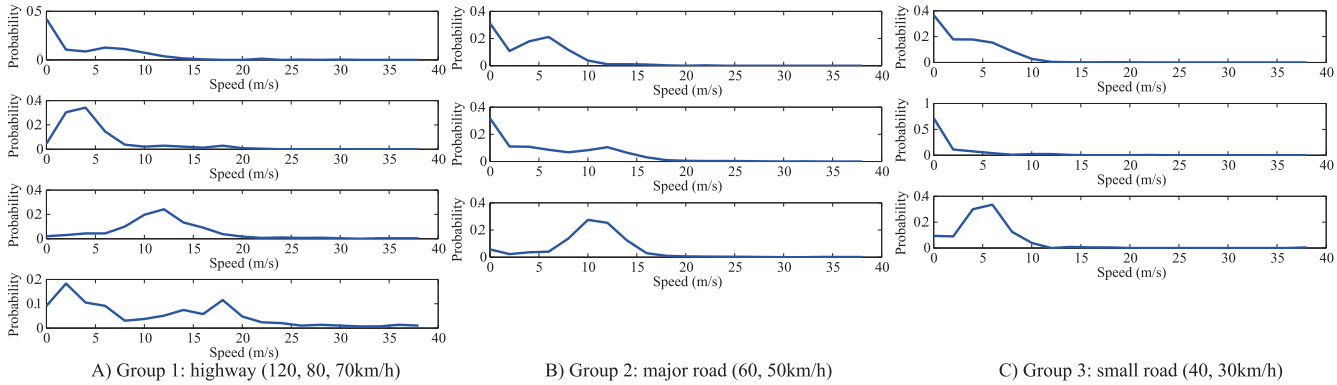


Fig. 7. Exemplars found by affinity propagation clustering.

additional information from historical vehicle trajectories and reflect the possible average congestion level of a particular road segment.

To construct the speed distribution, we only extract a single speed value from each vehicle trajectory that corresponds to the middle section of the road. This is to remove the potential impact from the upstream and downstream traffic signals. As reported in [21], the FDs obtained using data collected from sensors located in the middle section of a road are often cleaner and well-behaved, whereas using data from the sensors near downstream intersection often produce noisy FDs or do not exhibits clear pattern. The extracted sample speeds are then compiled into a sample speed set S_r for each road r and used to construct the empirical speed distributions $P_r(v)$ for roads covered by at least 20 trajectories.

We use the affinity propagation clustering [22] to cluster speed distributions of each road group into a set of road speed distribution classes θ_{sp}. Choosing affinity propagation clustering is mainly due to its capability of working under flexible distance measures and does not need to predetermine the number of clusters. Jensen-Shannon divergence is used as the distance metric for any pair of road speed distributions, which is commonly used in clustering probability distributions

$$dist_{ij} = \frac{1}{2} D\left(P_i || \frac{P_i + P_j}{2}\right) + \frac{1}{2} D\left(P_j || \frac{P_i + P_j}{2}\right), \qquad (5)$$

where $D(P||Q)$ is the Kullback-Leibler divergence. Fig. 7 shows the speed distributions of the exemplars (clustering centers in affinity propagation clustering) obtained after clustering. The results show that road Group 1 (highway) can be clustered into four classes, and both Group 2 (major road) and 3 (small road) can be clustered into three classes.

4.1.2 Extract FD Speed Parameters

As discussed previously, the speed distribution of a road encodes important information about the shape of Q-V diagram, which makes it possible to infer FD related parameters, specifically, the critical speed v_c and the free flow speed v_m. While the free flow speed of a road r can be easily estimated as $v_m = max\{v \in S_r, r.lim\}$, ($r.lim$ is the speed limit), the estimation of critical speed v_c is less straightforward.

From the empirical observations of the traffic breakdown and its associated low probability density region, the critical speed value v_c can be estimated as the partition point that separates the free flow and congested flow regime. Several empirical studies have shown the validity of using conventional clustering approach (e.g., K-means) on traffic speed and volume data to classify the free flow and congested flow regimes of the traffic [23], [24], [25]. Inspired by the special properties of Q-V diagram as well as the insights from previous studies, we use K-means to clustering the speed data in S_r for each road r with at least 20 sampled speed into classes of free flow regime C_{free} and congested flow regime $C_{congest}$. In addition, we pre-label speed data with value smaller than $0.1v_m$ as class $C_{congest}$ and speed data with value greater than $0.9v_m$ as class C_{free} (illustrated in Fig. 6). The pre-labeled speed data do not change class label during K-means update and thus ensure consistent clustering results for all roads. Finally, the critical speed v_c is estimated as $0.5 \times (max\{v \in C_{congest}\} + min\{v \in C_{free}\})$. Fig. 8 shows the K-means clustering results for two representative Group 1 (N. 4th Ring Rd. W.) and Group 2 (Zhong Guan Cun East Rd.) roads of Beijing. The solid lines in the two figures represent the empirical speed probability distribution functions (speed pdf) of the two roads. The red and blue dots are sample speeds that correspond to congested and free flow regime. It clearly shows the existence of the low probability density region due to traffic breakdown, and the effectiveness of the clustering scheme in finding the partition point (v_c) of free flow and congestion flow regime.

4.1.3 Infer Missing Traffic Flow Features

The speed distribution classes θ_{sp} and FD speed parameters v_c, v_m are obtained only for roads covered by at least 20 vehicle trajectories. This only accounts for a fraction of roads in the network (61 percent given the GPS trajectory dataset used in this study). The missing traffic flow features

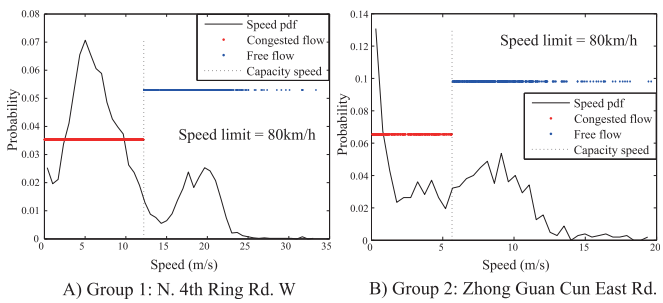


Fig. 8. K-means clustering results for two representative roads.

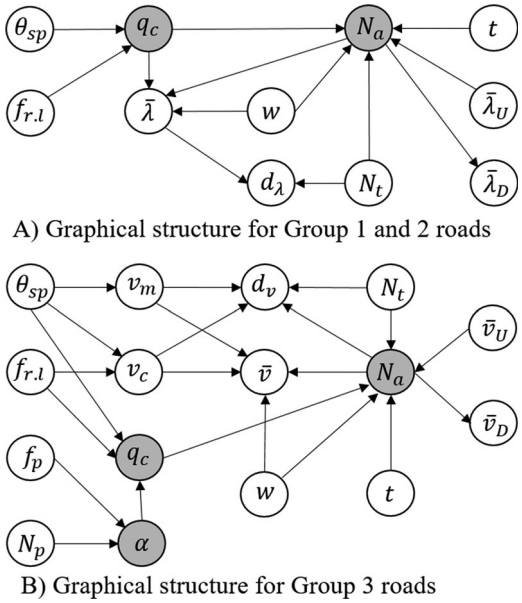


Fig. 9. The proposed Bayesian network structures.

for the rest of the roads need to be inferred. We use a simple user-based K-nearest neighbor (KNN) algorithm to infer the missing traffic flow features. The physical features associated with each road are used as the features to run the user-based KNN algorithm, including the road features f_r and POI features f_p. Since these features involves both numerical and unordered categorical variables, the Heterogeneous euclidean-Overlap Metric (HEOM) [26] is used to correctly characterize the distance between two data records. The HEOM between two data records R_i and R_j ($R_i = (f_r^T, f_p^T, f_g^T)^T$) is computed as follows:

$$\begin{aligned} d_a(x,y) &= \begin{cases} overlap_a(x,y) & \text{if feature } a \text{ is categorical} \\ diff_a(x,y) & \text{if feature } a \text{ is numerical} \end{cases} \\ overlap_a(x,y) &= \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases} \\ diff_a(x,y) &= \frac{|x-y|}{range_a} \\ HEOM(R_i,R_j) &= \sqrt{\sum_{a=1}^{m} d_a(R_i[a],R_j[a])^2}. \end{aligned} \tag{6}$$

After obtain the K-nearest neighbors for roads with missing traffic flow features, v_c, v_m are inferred as the mean values from the nearest neighbors; and the speed distribution class θ_{sp} is obtained as the maximum vote from the nearest neighbors.

4.2 Learning Volume Related High Level Feature

Given the intrinsic relationship among traffic speed, volume and density that captured in FD, a natural choice for traffic volume estimation is to characterize and utilize the functional form of FD. However, this approach is very problematic. First, the fundamental diagrams need to be calibrated for each individual road in order to produce accurate estimates. This requires sufficient amount of traffic volume data from each road segment for training, which is costly and impractical. Second, FD is observed and derived for traffic under no or low level of environmental disturbance, e.g., traffic on highways and major arterials, while FDs on small streets are either noisy or fail to capture any clear pattern. Hence a more scalable unsupervised approach needs to be developed to learn the underlying relationship between traffic speed and volume, as well as other potential impacting factors. We propose two unsupervised graphical model for different groups of roads based on partially observed Bayesian networks to learn the relationship between the volume related high level feature N_a and other relevant features. The traffic volume can then be predicted by re-interpreting N_a using a small amount of training data with ground truth traffic volume.

To fully incorporate the explicit shape of the Q-V diagram, we introduce following transformation to obtain the mean and standard deviation of a derived feature, referred as *road utilization* (λ) for each road segment. The idea of introducing λ is based on the fact that Q-V diagrams for higher level of roads (highways and major roads) can often be well approximated as a triangular shape using following linear function (illustrated as the red line in Fig. 1 C)

$$q \approx q_c \cdot \lambda(v) = \begin{cases} q_c \cdot \frac{v}{v_c} & 0 \le v \le v_c \\ q_c \cdot \max\{\frac{v_m - v}{v_m - v_c}, 0\} & v_c \le v \le v_m. \end{cases} \tag{7}$$

The road utilization λ approximates on what level the road is utilizing its total volume capacity, and explicitly encodes the shape of the Q-V diagram thereby providing extra information for volume estimation. Given the estimated mean and standard deviation $(\bar{v}, d_v)$ of speeds from the TSE component, the mean ($\bar{\lambda}$) and standard deviation (d_λ) of λ can be easily derived as

$$\begin{aligned} \bar{\lambda} &= \begin{cases} \bar{v}/v_c & 0 \le v \le v_c \\ \max\{\frac{v_m - \bar{v}}{v_m - v_c}, 0\} & v_c \le v \le v_m \end{cases} \\ d_\lambda &= \begin{cases} d_v/v_c & 0 \le v \le v_c \\ \frac{d_v}{v_m - v_c} & v_c \le v \le v_m. \end{cases} \end{aligned} \tag{8}$$

Fig. 9 presents the structures of the Bayesian network for Group 1&2 (Fig. 9 A) and Group 3 roads (Fig. 9 B). The shaded nodes represent hidden variables and the blank nodes denote observed variables. For each road group, we train a separate Bayesian network model using data of all road segments in this group. The proposed network structures are obtained after testing a series of structures constructed using the prior knowledge of traffic flow. The final network structure for each road group is selected as the one that yields the highest estimation accuracy. The results show that the dependency structure among the volume related feature N_a and other contributing features are quite different for Group 1&2 roads and Group 3 roads.

For higher level roads (Group 1 and 2), the Q-V diagrams are often well-behaved due to less disturbance from the nearby urban environment. As a result, it is observed that when incorporating features such as $\bar{\lambda}$, d_λ that encode the explicit shape of the Q-V diagram, better results are obtained. As expected, the POI related features are found not having significant impact on traffic volume and absent in the model due to higher level of access control and traffic separation for higher level roads. More specifically, the volume related high

level feature N_a of a road segment is influenced by five major factors, including the capacity volume q_c (hidden, determined by the road speed distribution class θ_{sp} and road length $f_{r.l}$), weather condition w, corresponding time slot t, the number of observed sample vehicles in the time slot N_t, and the mean road utilization level of its most congested upstream road $\bar{\lambda}_U$ (obtained as the $\bar{\lambda}$ value from the upstream road that has the same or higher road level and the lowest average speed). On the other hand, N_a influences the mean road utilization level of its most congested downstream road segment $\bar{\lambda}_D$, and co-determines the mean of the road utilization level $\bar{\lambda}$ with the capacity volume q_c.

For lower level roads (Group 3), as the FD may be more noisy or no longer exist, using the derived road utilization features $(\bar{\lambda}, d_\lambda)$ that encode explicit shape of Q-V diagram lead to inferior results. However, directly using the FD speed parameters v_c, v_m still improves the estimation accuracy. Moreover, more factors are observed to involve in the dependency structure in determining N_a, notably the surrounding POIs α which is influenced by the POI feature f_p as well as the total number of POIs N_p. For Group 3 roads, the capacity volume q_c is influenced by not only θ_{sp} and $f_{r.l}$, but also the surrounding POIs α. N_a is influenced by almost the same set of features (the mean road utilization level of most congested upstream road $\bar{\lambda}_U$ is now its mean speed $\bar{v}_U$), but influences the mean and standard deviation of the road speed $\bar{v}, d_v$, as well as the mean speed of the most congested downstream road $\bar{v}_D$.

All the variables used in the Bayesian networks are discretized to reduce computation complexity while maintaining reasonable resolution of the data. The variables are discretized either based on common knowledge or into categories representing intervals that contain equal proportion of data. For example, $\bar{\lambda}$ is discretized into five categories, i.e., $[0, 0.2]$, $[0.2, 0.4]$, $[0.4, 0.6]$, $[0.6, 0.8]$, $[0.8, 1]$, while the speed variables $\bar{v}$, d_v, v_c and v_m are discretized into K ordered categories with each contains $1/K$ proportion of data. Moreover, N_a is set to have five categories and its value for each instance will be inferred from the Bayesian network.

Due to the existence of hidden nodes, the conditional probabilities of hidden nodes $H = \{q_c, N_a, \alpha\}$ cannot be drawn by counting the occurrence of each condition. The Expectation-Maximization (EM) algorithm is hence used to learn the proposed Bayesian networks. Denote O as the set of observed nodes, $Pa(v)$ is the predecessors of node v. The algorithm for learning the Bayesian network parameters (conditional probabilities $P(v|Pa(v))$) is presented below:

The algorithm starts by randomly initializing the parameters. The first part of the main loop (Line 3-10) is the E-step, which uses simple Bayesian rule to compute the joint probability $P(h, e)$ for each possible set of values h for hidden nodes using the evidence e from the observed nodes. The hidden values h can then be inferred as the set of values h produces the maximum conditional probability $P(h|e)$. Hence the E-step is actually an inference process. In M-step (Line 11-14), the algorithm updates the conditional probability $P(v|Pa(v))$ for each node using the inferred results. The two steps keep iterating until the parameters converge. Once the Bayesian network are properly learned, the volume related high level feature N_a for each instance can be inferred by just using the E-step of Algorithm 1, which will be a probability distribution over five discrete categories. It will be then re-interpreted into the predicted volume using a re-interpretation model described in the following section.

Algorithm 1. Parameter Learning for the Partially Observed Bayesian Networks

Input: Bayesian network structure; observed evidence E
Output: $P(v|Pa(v))$ of each node
1: Randomly initialize $P(v|Pa(v))$ for each node v
2: **while** $P(v|Pa(v))$ *does not converge* **do**
3: **foreach** *evidence* $e \in E$ *from observed nodes* O **do**
4: **foreach** *values* h *for* $v \in H$ **do**
5: $P(h, e) \leftarrow \prod_{v \in H} P(v|Pa(v)) \prod_{Pa(u) \cap H \neq \emptyset, u \in O}$
6: $P(u|Pa(u))$
7: **end**
8: **foreach** *instance values* h *for* $v \in H$ **do**
9: $P(h|e) \leftarrow P(h, e) / \sum_h P(h, e)$
10: **end**
11: **end**
12: **foreach** *node* v **do**
13: $\rho \leftarrow$ the occurrences of $(v, Pa(v))$
14: $P(v|Pa(v)) \leftarrow \rho /$ the occurrences of $Pa(v)$
15: **end**
16: **end**
17: **return** $P(v|Pa(v))$

4.3 Traffic Volume Re-Interpretation

As N_a is learned from a completely unsupervised process, there is no concrete link between the categories of N_a and the actual volume categories, i.e., intervals of the actual traffic volume. Without certain interpretation, the physical meaning of N_a cannot be established. Shang et al. [1] proposed a volume estimation scheme to map the Bayesian network inference results to traffic volumes based on the normal distribution assumption. This scheme first fits the ground truth volume data into normal distributions $f(x; \mu, \sigma)$, then assumes each N_a category corresponds to ordered volume intervals with the equal probability of 0.2, and finally predicts the volume q by solving $P(N_a \in c) = \int_m^q f(x; \mu, \sigma) dx / 0.2$, where m is the lower bound of the volume interval that correspond to the N_a category c with largest probability. This scheme has several flaws. First, the normal distribution over-approximate the actual traffic volume distribution, which is inaccurate. Second, the assumption of ordered actual volume intervals with equal probability in Shang et al.'s scheme lacks theoretical support. As the volume related high level feature N_a are essentially learned as state labels, which does not preserve any inherent ordering nor equal marginal probability.

To addresses the drawbacks in Shang et al.'s scheme, we propose a new traffic volume re-interpretation model formulated as a bi-level optimization problem to re-interpret N_a using a small amount of ground truth volume data. The proposed model utilizes the empirical traffic volume distribution, and removes the assumption of ordered volume intervals and equal marginal probability of each N_a states.

We begin by perceiving the inferred results of N_a, i.e., the probability distribution over five N_a categories, as the membership probabilities $P^{N_a} = (p_1^{N_a}, p_2^{N_a}, \ldots, p_5^{N_a})$ correspond to five unknown volume categories defined over ordered disjoint intervals A_1:$[0, a_1]$, A_2:$[a_1, a_2], \ldots, A_5$:$[a_4, a_5]$. Let G:$c \in N_a \rightarrow I \in \{A_1, A_2, \ldots, A_5\}$ be a one-to-one mapping

function from the N_a categories to the actual volume categories. Our goal is to infer the mapping function G and the interval boundaries a_i's for each actual volume category.

To uncover the predicted volume value q, we assume $q \in A_s = [a_{s-1}, a_s]$, for $c = \{i : \max\{p_i^{N_a}, i = 1, \ldots, 5\}\}$ and $G(c) = s$, which means the predicted volume stays in the actual volume interval with the largest membership probability. Furthermore, denote $P_R = \sum_{G(i)>s, i=1,\ldots,5} p_i^{N_a}$ and $P_L = \sum_{G(i)<s, i=1,\ldots,5} p_i^{N_a}$ as the probabilities that q is greater and smaller than the values in interval $A_s = [a_{s-1}, a_s]$. If $P_R > P_L$, then it is reasonable to believe q should take larger values in A_s, as q has larger membership weight on intervals with higher volumes; similarly q should take smaller values if $P_R < P_L$. Ideally, if $P_R = P_L$, which means the predicted volume q is not impacted by both sides of the volume intervals, hence $q = E[q|q \in A_s] = \int_{a_{s-1}}^{a_s} x f_e(x) dx$, where $f_e(x)$ denotes the empirical probability distribution of traffic volume obtained using ground truth volume data from a particular road group. According to above assumptions, we develop following procedure to predict volume q.

Let P_s^m be the conditional cumulative probability that corresponds to $E[q|q \in A_s]$ for interval A_s, i.e., $P_s^m = \int_{a_{s-1}}^{E[q|q \in A_s]} f_e(x|x \in A_s) dx$. Then the conditional cumulative probability $\hat{P}$ of $q \in A_s$ can be modeled as following weighted averages

$$\hat{P} = \begin{cases} \frac{p_s^{N_a}}{p_s^{N_a} + |P_r - P_L|} \cdot P_s^m + \frac{|P_r - P_L|}{p_s^{N_a} + |P_r - P_L|} \cdot 0 & \text{if } P_L \geq P_R \\ \frac{p_s^{N_a}}{p_s^{N_a} + |P_r - P_L|} \cdot P_s^m + \frac{|P_r - P_L|}{p_s^{N_a} + |P_r - P_L|} \cdot 1 & \text{if } P_L < P_R, \end{cases} \tag{9}$$

or equivalently,

$$\hat{P} = \begin{cases} \frac{p_s^{N_a}}{p_s^{N_a} + |P_r - P_L|} \cdot P_s^m & \text{if } P_L \geq P_R \\ 1 - \frac{p_s^{N_a}}{p_s^{N_a} + |P_r - P_L|} \cdot (1 - P_s^m) & \text{if } P_L < P_R. \end{cases} \tag{10}$$

The traffic volume q can hence be predicted as the value that solves following equation

$$\hat{P} = \int_{a_{s-1}}^{q} f_e(x|x \in A_s) dx = \frac{\int_{a_{s-1}}^{q} f_e(x) dx}{\int_{a_{s-1}}^{a_s} f_e(x) dx}. \tag{11}$$

It can be observed that following this construction, the predicted volume $q = E[q|q \in A_s]$ if $P_L = P_R$; $q \to a_{s-1}$ if $P_L > P_R$ and $q \to a_s$ if $P_L < P_R$.

The mapping function G and the interval boundaries a_1, a_2, a_3, a_4 (a_5 is set to be the maximum volume from the ground truth data) can be inferred by minimizing the squared error between the ground truth volume q_n^a of instance n in the training set D_{tr} and the predicted volume $\hat{q}_n$, which is modeled as following bi-level optimization problem

$$\min_{G} \min_{a_1,\ldots,a_4} \sum_{n=1}^{|D_{tr}|} (q_n^a - \hat{q}_n)^2 \tag{12}$$
$$s.t.\ 0 < a_1 < a_2 < a_3 < a_4 < a_5.$$

Note that the computation of the predicted volume $\hat{q}_n$ using Equation (11) involves evaluating the empirical distribution of volume, hence forbids the use of gradient-based optimization algorithms. However, as the inner-level problem is a simple least squared error problem involving only four variables (a_1, a_2, a_3, a_4), it can be easily solved using existing derivative-free optimization algorithms. We use the pattern-search method to efficiently solve the inner-level problem. Pattern-search is a type of direct-search methods, which is applicable to objective functions that are not continuous or differentiable and enjoys global convergence as proved in [27]. For the outer-level problem, note that the mapping function G is essentially a vector generated by the permutation of {1, 2, 3, 4, 5}. Since there are only 120 possible arrangements for the permutation of {1, 2, 3, 4, 5} and the inner-level problem can be solved efficiently, we use the brutal force method to try all the possible G vectors in order to guarantee the result is global optimal. Finally, once the mapping function G and the interval boundaries $a_1, \ldots, a_4$ are learned, the predicted traffic volume can be efficiently computed using Equations (10) and (11).

TABLE 1
Statistics of Road with Mapped Trajectories

Road group	Group 1	Group 2	Group 3
Speed limit(km/h)	70-120	50-60	30-40
# segments	5,298	31,600	63,062
Total length (km)	1,450	3,991	6,934
% covered/interval	71.3%	63.1%	37.6%
# travels/interval	13.88	5.96	3.72

5 Experiments

5.1 Datasets

Road Network. The road network of Beijing is used in this study, which comprises of 186,266 nodes and 249,080 edges. The road network covers a 40×50 km area with a total length (of road segments) of 25,651 km. After filtering some very small road segments, we obtain 99,960 edges. All roads are divided into three road groups according to the speed limit, namely highway (70-120 km/h), major road (50-60 km/h) and small road (30-40 km/h). Road segments that have speed limit lower than 30km/h are removed from this study, since they mainly correspond to tiny or unpaved roads, which does not have a lot of value for traffic volume estimation.

POI Data. This dataset consists of 273,165 POIs of Beijing, which are classified into 195 tier two categories. We only choose use the top 10 categories that are located nearby road segments most frequently (see Section 3.2).

GPS Trajectories. We use a GPS trajectory dataset generated by 33,000 Beijing taxis over a period of 118 days. The total number of GPS points in the dataset is 716,019,905, and the total length of the trajectories is over 651,453,863 km. The average sampling rate is 49.9 seconds per point. After projecting the trajectories onto the road network, we come up with the statistics shown in Table 1. "% covered/interval" denotes the proportion of road segments traveled by at least one taxi in a given 30 min time interval (we partition the entire day into 30 min intervals to summarize the data). The last row of Table 1 presents the average number of traverses by taxis in a time interval.

Weather Conditions. We also collected the weather data to incorporate the impacts from weather conditions.The

TABLE 2
Ground Truth Data Collected for Evaluation

Time	6:00-10:00			10:00-16:00			16:00-20:00			after 20:00			total
Group	1	2	3	1	2	3	1	2	3	1	2	3	
Holi.	89	123	80	185	248	194	163	201	148	70	90	70	1,661
Work.	209	204	199	379	408	373	330	380	336	165	170	166	3,319
Total	904			1,787			1,558			731			4,980

weather data are compiled into a binary feature for each time slot, with 0 represents normal weather condition (no rain or only light rain), and 1 indicates severe weather condition (heavy rain).

Ground Truth Data. We recorded 4,980 videos clips with each has 5 min in length on all groups of roads (262 roads in total) over a period of 29 days (2015/6/2-6/30). The ground truth road speeds are also recorded as the average of five sample vehicle speeds measured by a laser speed gun during the 5 min period. We manually count the number of vehicles traversing the road segments by inspecting the videos clips. The summary statistics of the collected ground truth data are presented in Table 2, in which "Holi" and "Work" correspond to the number of video clips collected on holidays and workdays respectively.

5.2 Settings

As the speed inference model in Section 3.3 has already been tested in a previous study [1], we focus on evaluating the volume estimation component developed in this paper.

We set the time slot length as 10 minutes in our experiments. The ground truth volume data are converted into the number of vehicles per minute per lane to allow for consistent evaluation across different groups of roads. The Bayesian networks presented in Fig. 9 are obtained after testing 46 possible network structures and their variants (changes by adding or moving a node or edge). The final network structures are selected as the ones that result in highest estimation accuracy. We also tested a dynamic extension of the proposed Bayesian network model. The experiments show that the dynamic model does not provide noticeable accuracy improvement and needs much longer computation time in learning and inference. Hence we keep using the static Bayesian network for our experiment. When evaluating the traffic volume re-interpretation model, we randomly partition the ground truth volume data into 50 percent training data to train the model and the rest 50 percent of data for evaluation.

We use the mean absolute error (MAE) and mean relative error (MRE) to evaluate the performance of each component in the proposed framework, which are calculated as follows:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$
$$MRE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i}, \quad (13)$$

where y_i is the ground truth value for record i and $\hat{y}_i$ is the predicted values; n is the total number of records in the testing set.

TABLE 3
Results for Traffic Flow Feature Inference

Accuracy of θ_{sp} for each road group					
Road group	KNN	SVM	MLR	RF	GTB
1	**73.32%**	41.71%	63.32%	69.83%	69.43%
2	**65.99%**	42.34%	55.65%	55.02%	57.46%
3	**64.23%**	45.27%	58.18%	59.82%	60.24%
MRE of v_c, v_m for all road groups					
MRE	KNN	RR	BRR	RFR	GBTR
v_c	**14.24%**	16.51%	16.42%	14.74%	16.23%
v_m	2.94%	3.01%	2.99%	**2.47%**	2.87%

5.3 Evaluation on the Traffic Flow Feature Inference

Extracting the traffic flow features, i.e., θ_{sp}, v_c, v_m, requires the road to be covered by reasonable amount of vehicle trajectories (≥ 20). Only 61 percent of roads are covered by sufficient amount of vehicle trajectories using the GPS trajectory dataset, and the traffic flow features for the rest of the roads need to be inferred using the user-based KNN algorithm presented in Section 4.1.3. We test the performance of the user-based KNN algorithm against a set of supervised learning methods using the data from the 61 percent of roads covered by sufficient amount of vehicle trajectories. For the speed distribution classes θ_{sp}, we compare the user-based KNN algorithm against four widely used multi-class classification methods, including support vector machine (SVM), multi-class logistic regression (MLR), random forest (RF) and gradient tree boosting (GTB). For the FD speed parameters v_c, v_m, four regression methods are tested, namely ridge regression (RR), Bayesian ridge regression (BRR), random forest regressor (RFR) and gradient tree boosting regressor (GTBR). Table 3 presents the 10-fold cross-validation results of the inference quality for the traffic flow features. Specifically, we compute the accuracy of θ_{sp} for each road group, and MRE of v_c, v_m for all road groups.

The results show that the user-based KNN algorithm outperforms the tested supervised learning methods in almost all the cases. The only exception is the v_m, in which the user-based KNN algorithm yields an MRE that is only 0.47 percent higher compared with random forest regressor. The relatively good performance of the user-based KNN algorithm is due to the use of the Heterogeneous euclidean-Overlap Metric, which correctly captures the dissimilarity between both numerical and unordered categorical variables. It should also be noted that the user-based KNN algorithm allows inferring θ_{sp}, v_c, v_m at the same time, which is more effective compared with using more complex supervised classification algorithms for θ_{sp} and regression algorithms for v_c, v_m.

The user-based KNN algorithm can effectively infer the critical speed v_c with MRE less than 15 percent, and achieves very high accuracy in inferring the free flow speed v_m (MRE $<$ 3 percent). For speed distribution classes, road Group 1 has the highest accuracy of 73.32 percent while Group 2 and 3 roads have relatively larger error with accuracy around 65 percent. This is expected, as the speed distributions for highways usually demonstrates more regular patterns and fit better with the theoretical prediction of FD.

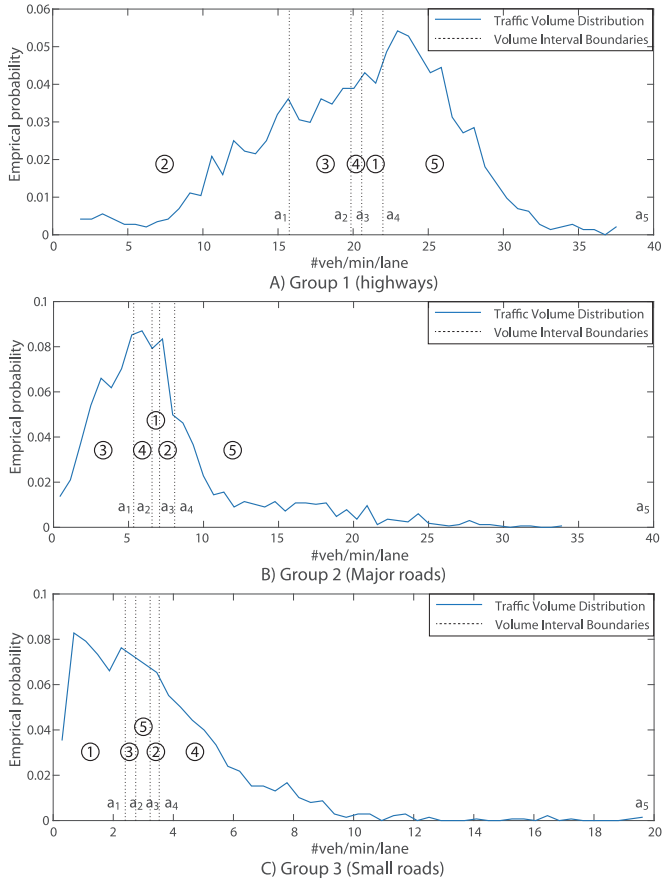


Fig. 10. Training results of the volume re-interpretation model.

While on lower level of roads, the traffic patterns are more chaotic and lead to higher level of error in the inference results of θ_{sp}. Another important factor that contributes to the error is the limited road features available. The inference accuracy can be further improved if a more comprehensive urban context dataset is used. Due to the relatively lower accuracy for Group 2 and 3 roads, we have tested in TVE by using θ_{sp} as a hidden node. However, the results show that using θ_{sp} as an observed node can greatly improve the estimation accuracy thanks to the extra road speed distribution information.

5.4 Evaluation on TVE

5.4.1 Results for Traffic Volume Re-Interpretation

Fig. 10 presents the results of the traffic volume re-interpretation model. The actual volume interval boundaries $a_1, \ldots, a_5$ and the mapping function G are obtained by solving the bi-level optimization problem defined in Equation (12). The volume interval boundaries $a_1, a_2, \ldots, a_5$ are indicated as the dotted line in Fig. 10, and the mapping function G is presented as the category numbers of the volume related high level feature N_a in the circles.

The results clearly show that the categories of N_a do not correspond to ordered actual volume intervals. This is intuitive, as N_a is inferred from a completely unsupervised process, there is no prior information to establish the mapping between N_a categories and the actual volume intervals. It is also observed that some N_a categories (e.g., Category 1 and 4 for Group 1 roads; Category 1, 2 and 4 for Group 2 roads; Category 2, 3 and 5 for Group 3 roads) correspond to very small and similar volume intervals, which means these categories of N_a all represent similar or possibly overlapping traffic volume conditions. As we enforce non-overlapping volume intervals in our volume re-interpretation model to reduce modeling complexity, such categories that reflect similar traffic volume conditions will result in consecutive small intervals as shown in Fig. 10.

Using the same set of training and testing data, we compare the volume re-interpretation model with four baseline methods: linear regression, Bayesian ridge regression, regression using random forest and the volume re-interpretation scheme used in Shang et al. [1] (described in Section 4.3). For all the four baseline methods, we use the volume related high level feature N_a (probability distribution over five categories) learned from the Bayesian network as input to predict the actual traffic volume. Furthermore, we conducted two additional tests, which apply the volume re-interpretation model as well as the volume estimation scheme in Shang et al. [1] on the Bayesian network structure used in Shang et al. These two tests are designed to evaluate the performance of the proposed Bayesian network structure compared with the one used in Shang et al. [1].

The testing results of the volume re-interpretation model and the baseline methods are presented in Table 4. The results show that the volume re-interpretation model with the proposed Bayesian network structure achieves the best performance in almost all road groups. The only exception is for Group 3 small roads, where regression using random forest achieves slightly lower MRE. This is mainly due to the fact that traffic flow patterns on small roads usually lack regularity, hence ensemble based approach that involves multiple classifiers, e.g., regression based on random forest, tends to perform better. Using the same Bayesian network structure, the volume re-interpretation model works much

TABLE 4
Performance of the Volume Re-Interpretation Model

Methods	Bayesian network structure	Group 1		Group 2		Group 3	
		MAE	MRE	MAE	MRE	MAE	MRE
Re-interpretation	Proposed	**4.75**	**23.6%**	**3.40**	**45.5%**	**1.72**	50.0%
Re-interpretation	Shang et al.	4.91	24.3%	3.73	48.1%	1.96	54.9%
Shang et al.	Proposed	5.75	28.5%	4.42	58.2%	2.25	68.0%
Shang et al.	Shang et al.	6.24	30.8%	4.57	60.5%	2.28	68.8%
Linear regression	-	4.99	25.1%	3.79	52.0%	1.90	54.4%
Bayesian ridge	-	5.02	25.3%	3.82	52.4%	1.91	54.7%
Random forest	-	4.88	24.6%	3.62	49.8%	**1.72**	**49.8%**

TABLE 5
Overall Performance of TVE

Methods	MAE	MRE	Time (μs/r)
TVE	**3.32**	**32.2%**	160.5
TVE $w/o\ d_v$	3.65	35.3%	157.2
TVE $w/o\ w$	3.67	35.4%	129.3
LR	3.57	34.4%	0.15
FD-Ind	2.46	21.4%	0.13
FD-Ind-half	2.95	25.7%	0.13
FD-All	4.10	37.7%	0.13
*FD-G*12	21.3	290.4%	0.13

better than the scheme used in Shang et al. [1]. The method used in Shang et al. yields larger error mainly due to the over approximation of the volume distribution using normal distribution, and the assumption of ordered actual volume intervals with equal probability. The results also show that the proposed Bayesian network structure achieves lower MAE and MRE in all the road groups compared with the one used in Shang et al. [1]. This is mainly due to the use of the traffic flow related features as well as more accurate characterization of the relationship among the impacting factors.

It is observed in Table 4 that the proposed model achieves the highest accuracy for Group 1 roads. The MRE of which is only 23.6 percent. This is primarily due to the well-behaved FD and higher level of regularity in traffic volume-speed relationship for these roads. The model yields relatively lower accuracy for Group 3 roads (MRE = 50.0 percent). This is as expected, as the FD for small roads are often more noisy compared with higher level roads. Moreover, as reported in Table 1, only 37.6 percent of road segments are covered by at least one taxi in 30 min time intervals, which is much lower compared with the other two groups. This leads to even less information available from trajectories. Furthermore, the inaccuracies in the traffic flow feature inference for road segments with insufficient coverage of vehicle trajectories (Table 3) may also produce erroneous inference for some road segments. All these factors lead to extra inaccuracies in the extracted traffic flow features (θ_{sp}, v_c, v_m) and eventually causing higher level of error in the volume estimates. Using a larger trajectory dataset covering a longer time period (currently we used a trajectory dataset of 118 days) will remedy the inaccuracy caused by the data sparsity problem and can potentially improve the traffic estimation quality for all road groups in the model.

5.4.2 Overall Performance

Table 5 presents the overall performance of TVE on the testing set of all road groups. We compare TVE with seven baselines: we first tested removing the standard deviation of speed d_v as well as the weather condition w from the Bayesian networks in TVE, denoted as *TVE* $w/o\ d_v$ and *TVE* $w/o\ w$ respectively. We also compared the TVE with other existing methods, including using the linear regression (LR) and fundamental diagram.

The results show that TVE outperforms the ones that removing the standard deviation of the speed d_v or weather condition w. This is intuitive, as the standard deviation of speed provides extra information about the traffic conditions. It is also known that severe weather conditions, e.g., heavy rain or snow, will cause slower traffic thereby lead to smaller traffic volume. Hence including d_v and w in the Bayesian network help to improve the inference accuracy.

We also compare TVE with two conventional approaches, which are linear regression and infer volume through calibrated fundamental diagrams, specifically, the Q-V diagram of FD. For LR, we use the same set of training and testing data for evaluation. The results show that the TVE outperforms the simple linear regression with lower MAE and MRE, which shows the benefit of incorporating the prior knowledge of traffic flow theories. For the tests using FD, we use the classic flow-speed function proposed by Greenshields [7] to estimate traffic volume. As FD is impacted by a lot of factors and the FD parameters can be drastically different across road segments, thus FD needs to be trained separately for each road segment. We consider four different settings. In the first setting, we select five road segments for each road group that have largest amount of ground truth volume and speed data (number of ground truth data records ranging from 28-36 for each road). We separately train the FD for each individual road segment using half of its ground truth data and test using the remaining data. The results of the overall MAE and MRE from the 15 selected road segments with the individually trained FDs are presented as *FD-Ind* in Table 5. We further test the setting referred as *FD-Ind-half*, which reduces the training set of the test *FD-Ind* to half and use the same testing set. This setting is to evaluate the performance of FD approach under limited ground truth volume data, as well as its sensitivity to the size of training data. We also test the setting that uses all the volume and speed data in the training set to train the FD for each road group, and the evaluation results is presented as *FD-All*. It can be easily observed that the FD approach yields very accurate volume estimation results if trained and used for each individual road segment. However, when estimating the FD using data from all road segments, the performance quickly decreases even the road segments are from the same road group. Moreover, if we use the FD parameters learned from the Group 1 roads to infer the volumes for Group 2 roads (*FD-G*12), high level of error is obtained. It is also found that reasonable amount of training data are required in order to accurately calibrate the FD for each road, as when reducing the training set to half (*FD-Ind-half*), both MAE and MRE are increased by 0.49 and 4.3 percent respectively. These results show that FD-based approach is not scalable and cannot be applied for network-wide traffic volume estimation. On the other hand, although TVE still needs a small amount of ground truth data to train the volume re-interpretation model, it is performed on the global level. Once properly trained, TVE can be applied to all roads in each road groups.

Table 5 also presents the average inference time for a single road segment. The experiments are conducted on a Quad-Core 3.6 GHz CPU and 16 GB RAM server. From the computation side, TVE is more costly compared with the inference approach using LR and FD. The largest amount of inference time is spent on inferring the Bayesian network, while the volume re-interpretation process can be finished efficiently in $0.11\,\mu$s. The computation

TABLE 6
Performance of TVE on Different Days

	Group 1	Group 2	Group 3
MAE-Weekday	5.00	3.84	1.99
MRE-Weekday	24.74%	46.9%	52.24%
MAE-Weekend	4.59	3.11	1.52
MRE-Weekend	22.9%	47.7%	52.76%

time can be further improved by implementing a more efficient inference algorithm for the Bayesian network. Furthermore, since the volume inference for each road is independent, the estimation process can run in parallel on a multi-core machine or a cluster, which can further reduce the overall computation time.

5.4.3 Temporal Performance

We also explored the temporal performance of TVE in different types of days and different time periods of the day. Table 6 presents the MAE and MRE of TVE on weekdays and weekends for each road group. It shows that TVE has better performance for Group 1 roads on weekends. This might due to less congestion on highways on weekends, which leads to simpler traffic conditions and more likely to be captured in the dependency structure learned in the Bayesian network. Also, the MAE values are consistently lower for all road groups on weekends, which is mainly due to lower traffic volumes on weekends. As for MRE, it is observed that TVE has similar overall performance for Group 2 and 3 roads regardless of weekday or weekends.

Fig. 11 provides a more detailed illustration of the relative temporal performance of TVE. Four time periods are tested, including 6:00-10:00 (contains morning peak), 10:00-16:00, 16:00-20:00 (contains evening peak) and after 20:00. The overall performance of TVE is similar in different time periods of the day, however, slight differences exist. It is observed that during the daytime, Group 1 roads have lower MRE in non-peak hours, e.g., 10:00-16:00 for weekdays and 6:00-10:00 for weekends. This is again due to less congestion on highways and more regular traffic patterns. On the other hand, higher MRE values are observed for Group 2 and 3 roads in 6:00-10:00 for weekends. This is caused by the limited taxi trajectories covering these lower level of roads during this time period, causing less or inaccurate information available for model training and inference. In addition, higher MRE are found for all road groups and different types of days at night (after 20:00). This might be associated with both lower GPS trajectory coverage as well as the change of drivers' driving behavior at night. As drivers tend to drive slower and more cautious at night, which could lead to potentially different volume-speed relationship and thus result in higher error in the volume inference.

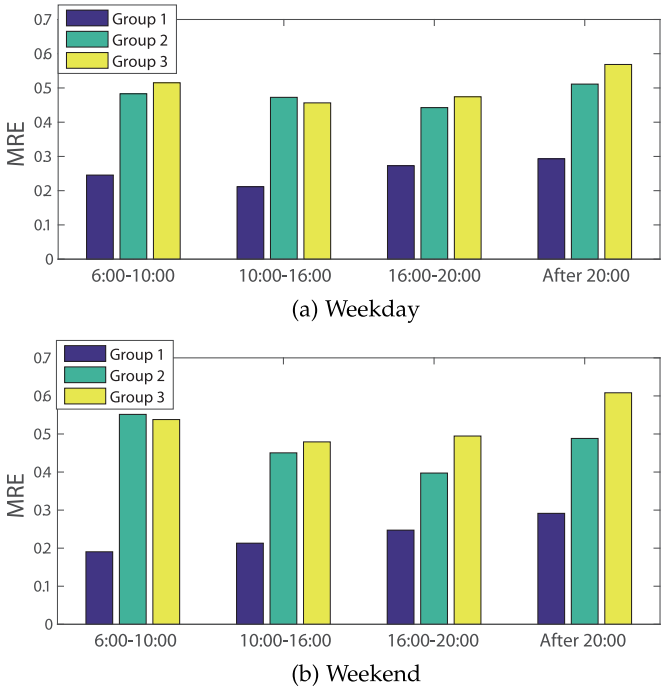


Fig. 11. Temporal performance of TVE.

6 Related Work

6.1 Traditional Traffic Volume Estimation Approaches

Traffic volume estimation and prediction are important engineering problems which are studied by many researchers. Traditional approaches rely heavily on data from various fixed road-based sensors, i.e., loop detectors [2], [3], [4], [5] or surveillance cameras [6]. Since actual traffic volumes can be directly measured from these sensors, most studies using this type of data mainly focus on predicting future traffic volume for dedicated roads using filtering-based algorithms, such as Kalman Filter [3], [5] and autoregressive integrated moving average (ARIMA) [28]. The cost associated with installing and maintaining road-based sensors restricts the applicability of these approaches only to a small fraction of major road segments in the network. To fully utilize the road-based sensor data, some researchers use hidden Markov model [29] to infer traffic volume on unmonitored road segments by exploiting the network topology and the correlation structure among roads. However, applying such methods typically lead to large model with huge number of parameters, and the model performance is still greatly impacted by the network coverage of the road-based sensors. Despite the studies that use direct traffic volume measures, another stream of research utilizes indirect traffic state measures, e.g., traffic density or speed, to estimate traffic volume using fundamental diagrams of traffic flow [2], [7], [8], [14]. This approach exploits the underlying relationship between traffic volume, density and speed to perform estimation. The major drawback of FD based approach is the need for sufficient amount of ground truth data for each road segment during the calibration phase, which is not scalable for citywide applications.

6.2 Traffic Volume Estimation Using Mobile Sensor Data

In recent years, many researchers have turned to using emerging spatiotemporal data generated from various mobile sensors to estimate citywide traffic volume. The data

sources include: cellular record data, social media data and GPS trajectory data from probe vehicles (e.g., GPS equipped taxis and floating cars). The dynamic nature and network-wide coverage offer mobile sensor data unmatchable benefit in traffic modeling over traditional fixed sensor data. Several researchers have explored the possibility of using mobile phone [12] and social media data [10] in traffic volume estimation. However, as these data are indirect approximations to real world traffic, only coarse-grained estimates of traffic volume can be obtained, such as region level volume [12] or congestion condition [10].

In contrast, probe vehicle data attract more attention of researchers, as probe vehicles can serve as dynamic probes in actual traffic flow. However, as probe vehicles only account for a small fraction of the actual traffic, serious data sparsity issue emerges. Early studies address the data sparsity issue by using interpolation-based methods, such as Kriging [13] for unmonitored road segments. Some researchers also explored combining both probe vehicle data and loop detector data in traffic volume estimation [11]. The traffic volume is estimated by utilizing the correlation structure captured in a regression model constructed using both data sources. However, as loop detector data play an important role in the model training, hence this approach is not fully scalable. Other studies [1], [14] take an alternative approach, which first infer network-wide speeds, and then estimate traffic volume by exploiting the underlying relationship between traffic volume and speed. Estimating network-wide traffic speed using probe vehicle data has been shown to be a relatively easier problem compared with volume estimation and has been explored in a large number of studies [15], [16], [17], [30], [31]. This study follows the similar alternative approach. Moreover, we proposed new methods to incorporate well-established traffic flow theories in feature extraction and model construction in addition to mining large-scale GPS trajectory data, which provide extra prior knowledge of traffic flow condition and help to improve estimation quality.

7 Conclusion

We develop a new framework that integrates both highly scalable machine learning techniques and well-established traffic flow theories to estimate the citywide traffic volume using data from GPS trajectories and several other sources. We extract a set of traffic flow features from GPS trajectories based on the traffic flow theory, which lead to improved estimation quality. The relevant features that involved in the dependency structure in determining traffic volume are also investigated using partially observed Bayesian networks. It is found that higher level roads (Group 1&2) possesses very different Bayesian network structure compared with small roads and impacted by fewer factors. The framework is evaluated using a GPS trajectory dataset from 33,000 Beijing taxis over a period of 118 days and ground truth volume data from 4,980 video clips. The results show that the framework achieves overall MRE of 32.2 percent on all groups of roads. Moreover, when applied to highways, the MRE can be as low as 23.6 percent. The evaluation results show our framework outperforms the baseline approaches in terms of effectiveness and scalability. Currently, our framework is less accurate in estimating traffic volume for lower level roads (MRE=50 percent for Group 3 roads). This is mainly because the traffic on small roads lacks regularity and the fundamental diagrams for these roads are usually not well-behaved. Hence the extracted traffic flow related features are not sufficient to fully capture the speed-flow relationship. Future work can be done to explore better methods for lower level roads. It will also be meaningful to extend the framework by incorporate traffic data from existing traffic sensors in the city, which will further improve the estimation accuracy.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant No. 61672399 and No. U1401258) and the China National Basic Research Program (973 Program, No. 2015CB352400). Yu Zheng is the corresponding author. This study is an extension of [1], which appeared in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2014).

References

[1] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1027–1036.

[2] L. Muñoz, X. Sun, R. Horowitz, and L. Alvarez, "Traffic density estimation with the cell transmission model," in *Proc. Amer. Control Conf.*, 2003, vol. 5, pp. 3750–3755.

[3] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. Part B*, vol. 18, no. 1, pp. 1–11, 1984.

[4] J. Kwon, P. Varaiya, and A. Skabardonis, "Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1856, pp. 106–117, 2003.

[5] D. Wilkie, J. Sewall, and M. Lin, "Flow reconstruction for data-driven traffic animation," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 89:1–89:10, Jul. 2013.

[6] X. Zhan, R. Li, and S. V. Ukkusuri, "Lane-based real-time queue length estimation using license plate recognition data," *Transp. Res. Part C: Emerg. Technol.*, vol. 57, pp. 85–102, 2015.

[7] B. Greenshields, W. Channing, H. Miller, and J. Bibbins, "A study of traffic capacity," in *Proc. 14th Annu. Meeting Highway Res. Board*, 1935, pp. 448–477.

[8] B. S. Kerner, "Three-phase traffic theory and highway capacity," *Physica A: Statist. Mech. Appl.*, vol. 333, no. 1–4, pp. 379–440, 2004.

[9] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[10] S. Wang, L. He, L. Stenneth, P. S. Yu, and Z. Li, "Citywide traffic congestion estimation with social media," in *Proc. 23rd SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2015, pp. 1–10.

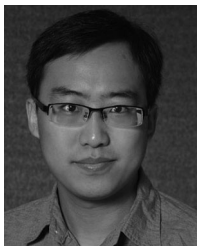
[11] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale traffic estimation from a roving sensor network," in *Proc. 10th ACM Conf. Embedded Netw. Sensor Syst.*, 2012, pp. 141–154.

[12] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. del Castillo, "Traffic flow estimation models using cellular phone data," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1430–1441, Sep. 2012.

[13] H.-X. Zou, Y. Yue, Q.-Q. Li, and A. G.-O. Yeh, "Traffic data interpolation method of non-detection road link based on Kriging interpolation," *Jiaotong Yunshu Gongcheng Xuebao*, vol. 11, no. 3, pp. 118–126, 2011.

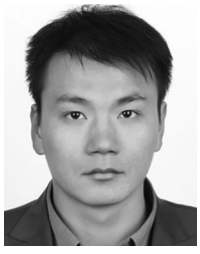
[14] A. Gühnemann, R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner, *Monitoring Traffic and Emissions by Floating Car Data*. Leeds, U.K.: Inst. Transport Studies, 2004.

[15] Y. Wang, Y. Zheng, and Y. Xue, "Travel time estimation of a path using sparse trajectories," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 25–34.
[16] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transp. Res. Part C: Emerg. Technol.*, vol. 33, pp. 37–49, 2013.
[17] X. Zhan, S. V. Ukkusuri, and C. Yang, "A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data," *Autom. Construction*, 2015. [Online]. Available: doi:10.1016/j.autcon.2015.12.007
[18] J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G. Z. Sun, "An interactive-voting based map matching algorithm," in *Proc. IEEE Int. Conf. Mobile Data Manage.*, 2010, pp. 43–52.
[19] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 42–49, 2009.
[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[21] X. Wu, H. X. Liu, and N. Geroliminis, "An empirical analysis on the arterial fundamental diagram," *Transp. Res. Part B: Methodological*, vol. 45, no. 1, pp. 255–266, 2011.
[22] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
[23] J. Kianfar and P. Edara, "A data mining approach to creating fundamental traffic flow diagram," *Procedia-Social Behavioral Sci.*, vol. 104, pp. 430–439, 2013.
[24] L. Sun and J. Zhou, "Development of multiregime speed-density relationships by cluster analysis," *Transp. Res. Rec.*, vol. 1934, no. 1, pp. 64–71, 2005.
[25] J. Xia and M. Chen, "A nested clustering technique for freeway operating condition classification," *Comput.-Aided Civil Infrastructure Eng.*, vol. 22, no. 6, pp. 430–437, 2007.
[26] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *J. Artif. Intell. Res.*, vol. 6, pp. 1–34, 1997.
[27] V. J. Torczon, "On the convergence of pattern search algorithms," *SIAM J. Optimization*, vol. 7, no. 1, pp. 1–25, 1997.
[28] S. Lee and D. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1678, pp. 179–188, 1999.
[29] J. Kwon and K. Murphy, "Modeling freeway traffic with coupled HMMs," Tech. Rep., Univ. California, Berkeley, CA, USA, May 2000.
[30] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton, "An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices," in *Proc. 47th IEEE Conf. Decision Control*, 2008, pp. 5062–5068.
[31] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, "A compressive sensing approach to urban traffic estimation with probe vehicles," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2289–2302, Nov. 2013.
[32] Y. Zheng, "Trajectory data mining: an overview," *ACM Trans. Intell.Syst. Technol. (TIST)*, vol. 6, no. 3, p. 29, 2015.
[33] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, 2015.



Xianyuan Zhan is currently working toward the PhD degree in the Lyles School of Civil Engineering, Purdue University. His research interests include big data analytics in transportation, urban computing, complex networks, and transportation network modeling.

Yu Zheng is a research manager at Microsoft Research, passionate about using big data to tackle urban challenges. He currently serves as the editor in chief of the *ACM Transactions on Intelligent Systems and Technology*. He is also the founding secretary of the SIGKDD China Chapter and has served as chair on over 10 prestigious international conferences, e.g. as the program co-chair of ICDE 2014 (Industrial Track). He received five best paper awards from ICDE13 and ACM SIGSPATIAL10, etc. His book, titled "Computing with Spatial Trajectories", has been used as a text book in universities world-widely and awarded the Top 10 Most Popular Computer Science Book authored by Chinese at Springer. In 2013, he was named one of the Top Innovators under 35 by MIT Technology Review (TR35) and featured by *Time Magazine* for his research on urban computing. In 2014, he was named one of the Top 40 Business Elites under 40 in China by *Fortune Magazine*, because of the business impact of urban computing he has been advocating since 2008. He is also a visiting Chair Professor at Shanghai Jiao Tong University and an adjunct professor at Hong Kong University of Science and Technology.



Xiuwen Yi is currently working toward the PhD degree in the School of Information Science and Technology, Southwest Jiaotong University. He has been working with the Urban Computing Group, Microsoft Research Asia, as a full time research intern since Feb. 2014, mentored by Dr. Yu Zheng. His research interests include urban computing and big data analytics.

Satish V. Ukkusuri is a professor in the Lyles School of Civil Engineering, Purdue University. He is the director of the Interdisciplinary Transportation Modeling and Analytics Lab, Purdue University. His current areas of interests include: dynamic network modeling, large-scale data analytics, disaster management issues, and freight transportation and logistics.