2019 IEEE Intelligent Transportation Systems Conference (ITSC)
Auckland, NZ, October 27-30, 2019

Estimating Link Travel Time Distribution Using Network Tomography Technique

Peibo Duan[1], Guoqiang Mao[2], Baoqi Huang[3] and Jun Kang[4]

Abstract— **Recently, link travel time distribution (LTTD) estimation has gained a lot of interest since the probabilistic model not only captures the dynamic features of link travel time but also provides abundant knowledge like the mean and variance which can be used as indicators to analyze link travel time reliability. However, existing methods still suffer from a number of problems: 1) most studies employ parametric models, e.g., Gaussian, which is only suitable in the limited traffic conditions like free flow or congestion. 2) many techniques heavily rely on the measurements detected on the roads. They cannot be applied to the whole road network since there is absence of observations in some roads due to the limited number of traffic detectors installed in the road network. In lieu of the aforementioned challenges, in the paper, we employ kernel density estimator (KDE) to model LTTD which is validated to be effective in any state of traffic condition. Further, motivated by the network tomography techniques, we propose an expectation maximization (EM) based algorithm to estimate model parameters only with end-to-end (E2E) measurements detected by traffic detectors at or near some road intersections. With 3.0e+07 GPS trajectories collected by the taxicabs in Xi'an, China, the experimental results show that the LTTD estimated by our proposed method are in excellent agreement with the empirical distributions, and better than its counterparts adopting Gaussian and log-normal models.**

I. Introduction

Link travel time is a kind of intuitive and easily understood traffic parameter which can be used as an indicator to measure traffic state of each link in the road network. On one hand, accurate estimation of link travel time helps travelers to make decisions for departure time and travel route in order to minimize overall trip travel time. On the other hand, it benefits the transportation agencies by identifying key bottlenecks in a road network, thereby allowing proactive traffic control like dynamic traffic signal operation.

In the last several years, it has attracted great research interest in estimating link travel time distribution (LTTD) using probabilistic models [5], [9], [16] since link travel time is stochastic and varies frequently due to the heterogeneous and dynamic nature of traffic in the complex road conditions [8], [9], [18]. Particularly, advanced data collection technology makes great progress on the techniques of LTTD estimation. These data sources [7] includes global position system (GPS), Bluetooth device, loop detectors and traffic cameras [17], [8], [9]. However, existing work still has some major challenges that have to be addressed. More precisely, these challenges exist in the following aspects:

Model selection: The widely used probabilistic models are mainly parametric, having a fixed structure (the number of parameters is finite). They have nice properties and make the parameters (the mean and covariance) estimation and reliability analysis simpler. However, the shortcomings are also obvious, that is, a parametric model is usually feasible under a particular traffic condition, e.g., free flow or congestion. In this case, given a road network including thousands of links, the selection of probabilistic models become more intricate since the traffic conditions of the links may be different from each other and vary with the time of the day.

Practicability: Many techniques of LTTD estimation heavily rely on direct active or passive measurements detected on the links. Direct active measurements refer to that link travel time is calculated by the time difference that the vehicles pass through (near) two endpoints of the link. Passive measurements mean that link travel time is calculated using the link distance divided by the space mean speed. Unfortunately, all these methods are only available for the links with sufficient observed data. They cannot be applied to the whole road network since it is impractical and cost prohibitive to cover all the links with traffic detectors.

Computational resource: To estimate travel time distribution of links without traffic detectors, a common method is to infer the data of these links by means of geospatial, temporal and historical contexts learned from the data in spatial-temporally correlated links [14]. However, a lot of computing resource and time will be consumed for missing data estimation. Also, the spatial-temporal correlation between links varies with the time of the day and road topology, and are difficult to capture accurately.

In lieu of the aforementioned problems, in this paper, we propose a framework to estimate a non-parametric model of link travel time with limited traffic detectors. The main contributions are briefly summarized as follows:

- We employ kernel density estimator (KDE) to model LTTD, which is a non-parametric model that can capture the dynamics of LTTD in any state of traffic condition. Further, an expectation maximization (EM) based algorithm is proposed to estimate model parameters.
- To improve the practicability of the estimator and save computational resource, we employ network tomography techniques to estimate LTTD with the aid of end-

[1] Peibo Duan is with School of Electrical and Data Engineering, University of Technology, Sydney, Australia Peibo.Duan@student.uts.edu.au

[2] Guoqiang Mao is with School of Electrical and Data Engineering, University of Technology, Sydney, Australia Guoqiang.Mao@uts.edu.au

[3] Baoqi Huang is College of Computer Science, Inner Mongolia University, Hohhot, China cshbq@imu.edu.cn

[4] Jun Kang is with School of Information Engineering Chang'an University, Xi'an, China. He is the corresponding author junkang@chd.edu.cn

978-1-5386-7024-8/19/$31.00 ©2019 IEEE

to-end (E2E) measurements collected at or near the road intersections. As there is no need to estimate missing data for the links without traffic detectors, the computational resource is much reduced.

- We validate the proposed method based on a data set including over 3.0e+07 GPS trajectories collected by the taxicabs in Xi'an, China. The experimental results show that the LTTD estimated using our proposed model are in excellent agreement with empirical distribution, and have better performance than its counterparts adopting Gaussian and log-normal distributions.

The organization of the paper is as follows. In Section II, the related work is briefly introduced. In Section III, we describe our approaches including the principle of network tomography and model building. The EM based algorithm is illustrated in Section IV. After that, we explore the performance of the proposed methods in Section V. Finally, we draw the conclusion in Section VI.

II. Related Work

In recent years, growing interest is motivating a shift toward estimating LTTD with different kinds of probabilistic models. In these models, Gaussian and log-normal are the most extensively researched. Particularly, Li et al. [6] suggested Gaussian was appropriate to model travel time in the presence of free flow, small time interval (e.g., 5 minutes), whereas log-normal was appropriate to model travel time in the presence of congestion, large time interval (half an hour). Other probabilistic models were also used to model link travel times such as Weibull distribution and Burr distribution [3]. Based on the analysis of real data, Hamdar et al. [4] found LTTD had a shorter right tail under free-flow conditions. Therefore, a hazard-based modeling approach was proposed with consideration of lane-changing behavior. Similarly, Moylan and Rashidi [8] constructed different hazard-based models of LTTD in the congested and non-congested state where different states were modeled by a latent-class-style approach.

The estimation of model parameters is based on the traffic data collected by different kinds of data detectors mentioned in Section I. A common weakness of these data detectors is that the collected traffic data has limited coverage, which leads to the fact that there are few or no observations in some links. To cope with this problem, techniques of missing data estimation have been used, but these techniques are mainly applied for mean travel time estimation. For instance, with GPS data, Wang et al. [14] and Tang et al. [11] applied a tensor to model link travel times as multi-linear manner geometric vectors. The tensor without GPS data was estimated based on the geospatial, temporal and historical contexts learned from the neighboring tensors. Unfortunately, these methods consumed a lot of computational resources. Another method is proposed by Zhang et al.'s [18] who estimated mean travel time with the data collected from limited number of traffic cameras. Note that this work also applied network tomography techniques. However, it could not be used for LTTD estimation since travel time was viewed as a deterministic variable, as opposed to a random variable. The limited coverage of traffic data is also considered in some research of travel time distribution estimation. A distinguished work was done by Prokhorchuk et al. [9] which developed a Gaussian copula graphical model. However, the travel time distribution estimated in these research was at the path level.

Network tomography uses the information derived from end-to-end (E2E) measurements to explore the internal characteristics of an Internet, e.g., the packet transmission delay. Specifically, a series of literature concentrates on the research of estimating link delay distribution, which is similar as LTTD estimation in the road network. However, these techniques cannot be directly used in our work since most of them are based on parametric models such as Gaussian or exponential distributions with the shortcomings discussed in Section I. Although bin size model, as a kind of non-parametric model [12], was used in the network tomography, it can vary wildly with the different configuration of bins, especially with a relatively small number of data. Therefore, in this paper, we use kernel density estimator (KDE) which provides similar distribution even with varying bandwidth and/or kernel type.

III. KDE Based Model

The probability density function (PDF) of the kernel density estimator (KDE) is defined as:

$$p(\boldsymbol{x}) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x_i - \boldsymbol{x}}{h}), \tag{1}$$

where $\boldsymbol{x}$ is the random variable, n is the number of samples, $h > 0$ is called the smoothing bandwidth that controls the amount of smoothing, x_i is the i-th sample, $K(\boldsymbol{x})$ is named the kernel (function) that is generally a smooth and symmetric function. In this paper, we use the Gaussian kernel, which has been widely used in the literature. In particular, $K(\frac{x_i - \boldsymbol{x}}{h})$ follows the standard normal distribution of $\mathcal{N}(0, 1)$.

We model a road network as a digraph. Specifically, we partition the road network into a set of *links* where each link is an one-way road segment bounded by two road intersections and there is no intersection within a link. Drawing from the graph theory, the digraph model is represented as $G = (V, E)$ where V is the set of vertices and E is the set of directed edges. Each vertex $V_i \in V$ represents an intersection. There exists an edge $e_{ij} \in E, e_{ij} = (V_i, V_j)$ if there is a link with traveling direction from V_i to V_j. We name a vertex V_i as a measurement point if there are observations detected by the traffic detectors like traffic cameras at V_i. Then $V = \{V_{meas}, V_{unmeas}\}$ where V_{meas} is the set of measurement points and $V_{unmeas} = V \backslash V_{meas}$. Obviously, the travel time distribution on e_{ij} can be estimated if both end points of e_{ij}, $V_i, V_j \in V_{meas}$. However, in real life, it is impractical to cover $\forall V_i \in V$ with traffic detectors. As a result, there is always a sequence of links between two measurement points. With the principle of graph theory, we define a travel route between two intersections as a path, denoted by:

$$r = \{e_1, e_2, ..., e_{d_r}\}, \tag{2}$$

where d_r is the number of links and the edges in r are all distinct from each other. Given a path r between $V_i, V_j \in V_{meas}$, we can obtain the travel time on r with the observations at V_i and V_j. For instance, a vehicle traveling from V_i to V_j through r is captured by the traffic cameras at V_i and V_j at time t_1 and t_2, then the travel time $t = t_2 - t_1$. In this paper, we also name t as an E2E measurement.

Consider the situation that the positions of the traffic data collected by some traffic detectors are not exactly located at the road intersections, but somewhere nearby, e.g., GPS data collected by the probe vehicles. We use a distance and time proportion method to estimate E2E measurements. More details will be illustrated in Section V. We use $\boldsymbol{t}$ to represent a random variable of travel time. The objective of our work is to estimate the distribution of $\boldsymbol{t}_{e_k}$ for $\forall e_k \in E$ with the E2E measurements detected by the limited traffic detectors.

Travel time in the neighboring links is spatially and temporally correlated to a greater or lesser extent. For simplicity, in this paper, we assume the travel time of a vehicle on different links is spatially independent. Furthermore, we assume that the travel time of different vehicles on the same link is also independent. Ignoring dependencies have a lot of benefits on the assumptions. For instance, in [14], to simply the objective function of path travel time estimation, Wang et al. assumed the travel time on different links are independent. Based on the above analysis, we are now ready to derive the estimates of LTTD.

We model the distribution of $\boldsymbol{t}_{e_k}$ with KDE as follows:

$$p(\boldsymbol{t}_{e_k}|\Theta_{e_k}) = \frac{1}{n_{e_k} h_{e_k}} \sum_{i=1}^{n_{e_k}} \mathcal{N}(\boldsymbol{t}_{e_k}|\mu_{e_k,i}, h_{e_k}^2), \tag{3}$$

where $\Theta_{e_k} = \{n_{e_k}, h_{e_k}, \mu_{e_k}\}$ is the set of parameters characterizing the KDE. More precisely, n_{e_k} is the number of vehicles traveling through e_k during a given time interval (e.g., 8:00am-8:30am), h_{e_k} is the bandwidth and $\mu_{e_k} = \{u_{e_k,i}|i = 1, 2, ..., n_{e_k}\}$ where $u_{e_k,i} \in \mu_{e_k}$ is the travel time when the i-th vehicle traverses e_k. As $\boldsymbol{t}_r = \sum_{k=1}^{d_r} \boldsymbol{t}_{e_k}$, the distribution of $\boldsymbol{t}_r$ conditioned on Θ_{e_k} can be parameterized as follows:

$$p(\boldsymbol{t}_r|\Theta_r) = p(\boldsymbol{t}_{e_1}|\Theta_{e_1}) * ... * p(\boldsymbol{t}_{e_{d_r}}|\Theta_{e_{d_r}}), \tag{4}$$

where $*$ represents the convolution operation and $\Theta_r = \{\Theta_{e_k}|k \in d_r\}$.

In the network tomography, the transmission route of a packet is always known. However, in our work, the path r where an E2E measurement is collected is usually unknown because of the following two reasons: i) the limited number of traffic detectors makes the travel route unobservant, and ii) there may be multiple paths between two measurement points. We use $R = \{r_1, r_2, ..., r_{|R|}\}$ to denote the alternative paths between two measurement points where $|\cdot|$ is the carnality of a set. Given an E2E measurement t, we introduce a binary variable $p_{t|r_j}, r_j \in R$ where $p_{t|r_j} = 1$ if t is collected from $r_j \subseteq R$ and $p_{t|r_j} = 0$ otherwise. Obviously, $\sum_{r_j \in R} p_{t|r_j} = 1$ since an E2E measurement is collected only from a unique route. We use $P_{t|R} = \{p_{t|r_1}, p_{t|r_2}, ..., p_{t|r_{|R|}}\}$ to represent the set of binary variables for t based on routes R, so that the probability of t conditioned on $P_{t|R}$ and Θ_R is modeled by:

$$p(t|P_{t|R}, \Theta_R) = \prod_{r_j \in R} p(t_{r_j}|\Theta_{r_j})^{p_{t|r_j}}, \tag{5}$$

where $\Theta_R = \{\cup\Theta_{r_j}|r_j \subseteq R\}$. Given the set of E2E measurements between two measurement points in a time interval, T, we define $P_{T|R} = \cup_{t \in T} P_{t|R}$, then the log-likelihood of T is formulated as:

$$\mathcal{L}(T|P_{T|R}, \Theta_R) = \sum_{t \in T} \ln p(t|P_{t|R}, \Theta_R). \tag{6}$$

In a road network, supposing that we have M pairs of measurement points, then we use $\mathrm{T} = \{t \in T_m|m \in M\}$ to denote the set of all E2E measurements over the whole study site. The log-likelihood of T is formulated as:

$$\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T}|\mathrm{R}}, \mathbf{\Theta}_{\mathrm{R}}) = \textstyle\sum_{m \in M} \mathcal{L}(T_m|P_{T_m|R_m}, \Theta_{R_m}), \tag{7}$$

where $\mathrm{R} = \{R_m|m \in M\}$ is the set of paths with measured data in the road network, $\mathbf{P}_{\mathrm{T}|\mathrm{R}} = \cup_{m \in M} P_{T_m|R_m}$ and $\mathbf{\Theta}_{\mathrm{R}} = \cup_{m \in M} \Theta_{R_m}$. By substituting (4) and (6) into (7), we obtain $\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T}|\mathrm{R}}, \mathbf{\Theta}_{\mathrm{R}})$ as follows:

$$\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T}|\mathrm{R}}, \mathbf{\Theta}_{\mathrm{R}}) = \sum_{m \in M} \sum_{t \in T_m} \sum_{r_j \in R} p_{t|r_j} \ln p(t_{r_j}|\Theta_{r_j}). \tag{8}$$

To simplify (8), we introduce $\mathbb{R} = \cup_{m \in M} R_m$. Moreover, we define $\mathbf{P}_{\mathrm{T}|\mathbb{R}} = \{p_{t|r_j}|t \in \mathrm{T}, r_j \subseteq \mathbb{R}\}$ where $p_{t|r_j} = 1$ if and only if t is collected on route r_j and otherwise, $p_{t|r_j} = 0$. Obviously, the significance of $\mathbb{R}$ and $\mathbf{P}_{\mathrm{T}|\mathbb{R}}$ are similar as R and $P_{\mathrm{T}|\mathrm{R}}$. Meanwhile, we introduce $\mathbf{\Theta}_{\mathbb{R}} = \{\cup\Theta_{r_j}|r_j \subseteq \mathbb{R}\}$. As both R and $\mathbb{R}$ should cover. all the edges in G, we have $\mathbf{\Theta}_{\mathbb{R}} = \mathbf{\Theta}_{\mathrm{R}} = \{\cup\Theta_{e_k}|e_k \in E\}$. In this case, (8) can be rewritten as

$$\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T}|\mathbb{R}}, \mathbf{\Theta}_{\mathbb{R}}) = \sum_{t \in \mathbf{T}} \sum_{r_j \in \mathbb{R}} p_{t|r_j} \ln p(t_{r_j}|\Theta_{r_j}). \tag{9}$$

The parameters of KDE, $\mathbf{\Theta}_{\mathbb{R}}$, heavily rely on $\mathbb{R}$ and $\mathbf{P}_{\mathrm{T}|\mathbb{R}}$, which has a close relationship with the placement of traffic detectors and the road topology. In this paper, we estimate $\mathbb{R}$ and $\mathbf{P}_{\mathrm{T}|\mathbb{R}}$ based on Google Map and the method proposed by Zhang et al. [18]. To simply the problem, we only consider one path with the shortest length between two intersections. It can be easily extended to the case of multiple possible paths between two intersections. The difference only relies on the increment of the computational complexity. The estimation procedure is described as follows:

Step 1: Obtain the candidate paths between any pair of intersections using Google Map Javascript API. After that, we get the routing matrix W.

Step 2: Calculate the bases of the routing matrix W, denoted by $\mathbf{B}_W$, where each basis $B_W \in \mathbf{B}_W$ is defined as a maximal subset of linearly independent paths.

Step 3: Define $\mathbb{R}_{B_W}$ as the set of paths in B_W. Meanwhile, assign a weight to each road intersection with the significance of being the cost for deploying the traffic detectors. By deploying the traffic detectors at both ends of each path in $\mathbb{R}_{B_W}$, the costs of traffic detectors deployment for $\forall B_W \in \mathbf{B}_W$ can be calculated. The basis with the minimal cost, denoted by $B_{opt} \in \mathbf{B}_W$, is selected and then $\mathbb{R} = \mathbb{R}_{B_{opt}}$.

Step 4: As there is only one path between two intersections, $\mathbf{P}_{\mathrm{T}|\mathbb{R}}$ is known with the estimated $\mathbb{R}$.

With above algorithm, the paths in $\mathbb{R}$ cover all the links E in the road network G. Thus, we can estimate travel time distribution of all the links from the E2E measurements detected by the traffic detectors deployed at the ends of each path in $\mathbb{R}$. However, this deployment strategy of traffic detectors are not the optimal one since we can further reduce the number of traffic detectors with a little sacrifice of estimation accuracy. Therefore, the trade-off between the estimation accuracy and the number of deployed traffic detectors is the problem that has its own merit and warrants a separate study. It will be researched in our future work. In next section, we will introduce the approach to estimating $\mathbf{\Theta}_{\mathbb{R}}$ with $\mathbb{R}$ and $\mathbf{P}_{\mathrm{T}|\mathbb{R}}$.

IV. EM Based Algorithm

Recall (3), $\forall \Theta_{e_k} \subseteq \mathbf{\Theta}_{\mathbb{R}}$ has the parameters $\{n_{e_k}, h_{e_k}, \mu_{e_k}\}$. n_{e_k} is related to the number of data collected on the paths that cover e_k. We define a $|\mathbb{R}|$ dimensional vector $P_{t|\mathbb{R}} = (p_{t|r_j} | r_j \subseteq \mathbb{R})$. Then, n_{e_k} can be estimated by

$$n_{e_k} = \sum_{t \in \mathrm{T}} P_{t|\mathbb{R}} \cdot W^k, \tag{10}$$

where W^k is the k-th column of W.

In order to estimate h_{e_k} and μ_{e_k}, we first simplify the representation of (3) based on: 1) the associative property of convolution, that is, $f_1(x)*(f_2(x)+f_3(x)) = f_1(x)*f_2(x)+f_1(x)*f_3(x)$, and 2) the property that the convolution of two Gaussian distributions, i.e. $\mathcal{N}(\mu_1, \sigma_1^2) * \mathcal{N}(\mu_2, \sigma_2^2)$, is also a Gaussian distribution: $\mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. With these two properties, (4) can be rewritten as

$$p(\boldsymbol{t}_r|\Theta_r) = (\prod_{k=1}^{d_r} \frac{1}{n_{e_k} h_{e_k}}) \sum_{z=1}^{\mathcal{Z}_r} \mathcal{N}(\boldsymbol{t}_r | \mu_{r,z}, h_r^2), \tag{11}$$

where $\mathcal{Z}_r = \prod_{k=1}^{d_r} n_{e_k}$, $\mu_{r,z} = \sum_{k=1}^{d_r} u_{e_k,i}, \forall i \in n_{e_k}$ and $h_r^2 = \sum_{k=1}^{d_r} h_{e_k}^2$.

Given the natural logarithm of $p(\boldsymbol{t}_r|\Theta_r)$:

$$\ln p(\boldsymbol{t}_r|\Theta_r) = \sum_{k=1}^{d_r} \ln \frac{1}{n_{e_k} h_{e_k}} + \ln \sum_{z=1}^{\mathcal{Z}_r} \mathcal{N}(\boldsymbol{t}_r | \mu_{r,z}, h_r^2), \tag{12}$$

we obtain $\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T|R}}, \mathbf{\Theta}_{\mathrm{R}})$ as follows:

$$\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T|R}}, \mathbf{\Theta}_{\mathrm{R}}) = \sum_{t \in \mathbb{T}} \sum_{k=1}^{d_{r_j}} \ln \frac{1}{n_{e_k}} + \mathcal{L}(\mathbb{T}|\mathbf{\Theta}_{\mathbb{R}}), \tag{13}$$

where $\mathbb{T} = \{t | t \in \mathrm{T}, p_{t|r_j} = 1\}$.

Algorithm 1 EM algorithm

```
Initialization: $\mathbf{\Theta}_{\mathbb{R}}^{(0)}$
for $q \in 1, 2, ...$
E-step:
    $\gamma_{t_{r_j}}^{(q)}(y_z)$: Being updated using (16) with $\mathbf{\Theta}_{\mathbb{R}}^{(q-1)}$
M-step:
    for each $\mu_{e_k,i}^{(q)}$ in $\Theta_{e_k}^{(q)} \subseteq \mathbf{\Theta}_{\mathbb{R}}^{(q)}$ and $h_{e_k}^{(q)}, e_k \in E$
        $\mu_{e_k,i}^{(q)} \leftarrow \frac{\sum_{r_j \subseteq \mathbb{R}_{e_k}} \sum_{t_{r_j} \in \mathbb{T}_{r_j}} \sum_{z \in \mathcal{Z}_{r_j}(\mu_{e_k,i})} \gamma_{t_{r_j}}^{(q)}(y_z) t_{r_j}}{N_{r_j}}$
        $(h_{e_k}^{(q)})^2 \leftarrow \frac{\sum_{r_j \subseteq \mathbb{R}_{e_k}} \sum_{t_{r_j} \in \mathbb{T}_{r_j}} \sum_{z \in \mathcal{Z}_{r_j}} \gamma_{t_{r_j}}^{(q)}(y_z)(t_{r_j} - u_{e_k,i})^2}{N_{r_j}}$
    endfor
Terminal:
    if $\mathbf{\Theta}_{\mathbb{R}}^{(q)}$ converges to a local optimum
        return $\mathbf{\Theta}_{\mathbb{R}}^{(q)}$
    endif
endfor
```

From (13), we can observe that the parameters $\mathbf{\Theta}_{\mathbb{R}}$ are only included in $\mathcal{L}(\mathbb{T}|\mathbf{\Theta}_{\mathbb{R}})$. Thus, setting the derivative of $\mathcal{L}(\mathbf{T}|\mathbf{P}_{\mathbf{T}|\mathbb{R}}, \mathbf{\Theta}_{\mathbb{R}})$ with respect to $\mathbf{\Theta}_{\mathbb{R}}$ to zero, we have

$$\frac{\mathrm{d}\mathcal{L}(\mathrm{T}|\mathbf{P}_{\mathrm{T|R}}, \mathbf{\Theta}_{\mathrm{R}})}{\mathrm{d}\mathbf{\Theta}_{\mathbb{R}}} = \frac{\mathrm{d}\mathcal{L}(\mathbb{T}|\mathbf{\Theta}_{\mathbb{R}})}{\mathrm{d}\mathbf{\Theta}_{\mathbb{R}}} = 0. \tag{14}$$

Unfortunately, there is no closed form solution for (14) due to the logarithm of cumulative Gaussian distribution. As a result, the Maximum Likelihood (ML) method does not work here. To address this problem, we employ the EM algorithm to estimate $\mathbf{\Theta}_{\mathbb{R}}$ (Algorithm 1) based on the following assumption:

Assumption 1: The $h_{e_k}s$ of the KDE models for the travel time in $\forall e_k \in E$ are same.

To begin with, we introduce the latent variables. For $\forall r_j \subseteq \mathbb{R}$, we define $\mathcal{Z}_{r_j}$-dimensional latent variables as $\boldsymbol{y}_{r_j}$ in which $\forall y_z \in \boldsymbol{y}_{r_j}$ satisfies $y_z \in \{0, 1\}$ and $\sum_{y_z \in \boldsymbol{y}_{r_j}} y_z = 1$. Given the definition that the marginal distribution over $\boldsymbol{y}_{r_j}$ is $p(y_z = 1) = \mathcal{Z}_{r_j}^{-1}$, we formulate the distribution of $\boldsymbol{y}_{r_j}$ as $p(\boldsymbol{y}_{r_j}) = \prod_{y_z \in \boldsymbol{y}_{r_j}} \mathcal{Z}_{r_j}^{-y_z}$. We also define the conditional distribution of an E2E measurement t_{r_j} as a Gaussian distribution with $p(t_{r_j}|y_z = 1) = \mathcal{N}(t_{r_j}|\mu_{r_j,z}, h_{r_j}^2)$. The joint distribution of t_{r_j} is given by:

$$\begin{aligned} p(t_{r_j}) &= \sum_{\boldsymbol{y}_{r_j}} p(\boldsymbol{y}_{r_j}) p(t_{r_j}|\boldsymbol{y}_{r_j}) \\ &= \mathcal{Z}_{r_j}^{-1} \sum_{z \in \mathcal{Z}_{r_j}} \mathcal{N}(t_{r_j}|\mu_{r_j,z}, h_{r_j}^2) \end{aligned} \tag{15}$$

We define $\gamma_{r_j}(y_z) \equiv p(y_z = 1|t_{r_j})$, which can be calculated based on Bayes theorem:

$$\begin{aligned}\gamma_{r_j}(y_z) &= \frac{p(y_z=1)p(t_{r_j}|y_z=1)}{p(t_{r_j})} \\ &= \frac{\mathcal{N}(t_{r_j}|\mu_{r_j,z},h_{r_j}^2)}{\sum_{z\in\mathcal{Z}_{r_j}}\mathcal{N}(t_{r_j}|\mu_{r_j,z},h_{r_j}^2)} \qquad (16)\end{aligned}$$

In Algorithm 1, $\mathbf{\Theta}_{\mathbb{R}}^{(0)}$ are the initial values of $\mathbf{\Theta}_{\mathbb{R}}$. In line 4, $\mathbb{R}_{e_k} = \{r_j|e_k \in r_j, r_j \subseteq \mathbb{R}\}$, $\mathcal{Z}_{r_j}(u_{e_k,i}) \triangleq \{z|z \in \mathcal{Z}_{r_j}, \mu_{e_k,i} \in u_{r_j,z}\}$ and N_{r_j} is

$$N_{r_j} = \sum_{r_j\subseteq\mathbb{R}_{e_k}} \sum_{t_{r_j}\in\mathbb{T}_{r_j}} \sum_{z\in\mathcal{Z}_{r_j}(\mu_{e_k,i})} \gamma_{r_j}(y_z(t_{r_j})). \qquad (17)$$

As the performance of the EM algorithm heavily relies on $\mathbf{\Theta}_{\mathbb{R}}^{(0)}$, we use the initialization strategy given in [1]. Convergence is achieved when $\mathbf{\Theta}_{\mathbb{R}}^{(q)} \approx \mathbf{\Theta}_{\mathbb{R}}^{(q-1)}$.

V. Experimental Results

To validate our proposed method, we use the GPS trajectories anonymously reported by over 11,000 taxicabs on Sep. 5th, 2016 (Mon.) in Xi'an, China. With an average sampling frequency of 30 seconds, we yield over 3.0e+07 raw data records. Noises exist in the collected GPS data, mainly due to erroneous measurements. Thus, we carry out data preprocessing as follows: 1) map matching with a weight-based topological algorithm proposed by Velaga et al. [13]. 2) Outliers filtering such as the data where the detected locations are outside the scope of Xi'an city. Techniques of GPS data preprocessing has been further researched over past years [14], [15]. To save space, we do not include the details of these techniques. The travel time when a taxi traverses a link is calculated in two different ways, which depend on the number and the positions of GPS data reported by this taxi: 1) Only one GPS data record reported by a vehicle in a link. The travel time is the link length divided by the space mean speed inferred from the instantaneous speed recorded by the GPS using the method in [2]. 2) Multiple GPS data records reported by a vehicle on a link. These GPS data might not exactly reside at the endpoints of the link. To counter this effect, we apply the method, namely distance and time proportion proposed by Sanaullah et al.'s [10]. We use the method to evaluate the E2E measurements.

TABLE I

The number of links and travel states in each time interval

	Time intervals	Travel state	No. of links	No. of intersections
τ_{17}	8:00am-8:30am	Congestion	4934	3545
τ_{19}	9:00am-9:30am	Congestion	5271	4031
τ_{23}	11:00am-11:30am	Free flow	5178	3804
τ_{31}	15:00pm-15:30pm	Free flow	5023	3752
τ_{35}	17:00pm-17:30pm	Congestion	5201	3957
τ_{41}	20:00pm-20:30pm	Free flow	4885	3524

We divide a day into 48 equal time intervals, denoted by $\{\tau_i|i = 1, 2, ..., 48\}$ where $\forall\tau_i$ represents half an hour, e.g., the time interval between 8:00am-8:30am. After that, we set up the model in each τ_i and implement the LTTD estimation in Java and Matlab. We use the following two ground truths to validate the accuracy of our estimation: *1) Opt-KDE*: the travel time distributions estimated by the KDE with the optimal bandwidth. As discussed in Section III, Opt-KDE can fit the distribution of empirical data better than any other models. Thus, we compare the results of our proposed method with Opt-KDE to validate estimation accuracy; *2) Empirical CDF*: the CDF of empirical data. With the KS test defined in Section III, we can observe whether the estimated probability distribution is accepted or not.

TABLE II

The percentage of intersections that should deploy traffic detectors

Time intervals	τ_{17}	τ_{19}	τ_{23}	τ_{31}	τ_{35}	τ_{41}
Percentage	59%	61%	63%	53%	57%	60%

TABLE III

Average KL divergence for different models

Time	KDE-E2E	Gaussian	log-normal	GMM	Hazard
τ_{17}	0.92	1.51	1.24	0.93	1.13
τ_{19}	**1.03**	1.46	1.37	**0.89**	1.07
τ_{23}	0.96	1.73	1.19	1.01	1.15
τ_{31}	**0.90**	1.93	0.91	**0.87**	0.97
τ_{35}	0.94	1.49	1.25	0.98	1.14
τ_{41}	0.86	1.28	0.94	0.91	1.02

To assess the deviation between an estimated LTTD and the ground truth (Opt-KDE) in a link, we use the metric named Kullback Leibler (KL) divergence, which is defined as follows:

$$D_{KL}(P_{opt}||P_{es}) = \sum_{t\in T_{e_k}} p_{em}(t)\ln\frac{p_{em}(t)}{p_{es}(t)}, e_k \in E. \qquad (18)$$

In (18), p_{opt} represents the LTTD with Opt-KDE, and p_{es} is LTTD with our proposed model and the counterparts including Gaussian, log-normal, GMM (Gaussian mixture model) and hazard based model. In particular, the GMM has three components of Gaussian, formulated by $\sum_{i=1}^{3}\phi_i\mathcal{N}(\mu_i,\Sigma_i)$. The hazard based model was proposed by Emily and Taha in [8]. It provides a good performance in estimating LTTD by considering such factors as travel speed, weather, traffic condition, etc. As we do not have the data like the weather, we cannot re-build the model as the one in [8], which considered numerous factors that potentially affect the variation of travel times. In our paper, we only use the traffic speed and traffic condition to set up the hazard based model. To evaluate the performance of the estimated results over the whole study site, we calculate the average KL divergence by $\sum_{e_k\in E} D_{KL}(P_{opt}||P_{es})/|E|$.

Due to page limit, we only present the estimated results in six representative time intervals listed in Table I, including the traffic states of free flow and congestion. The study site (number of links and number of intersections) in each time interval are presented in the third and fourth column where each link has enough observations. After estimating $\mathbb{R}$, Table

TABLE IV

THE KS TEST BASED ON DIFFERENT PROBABILISTIC MODELS WITH SIGNIFICANCE LEVEL $\alpha = 0.01$.

$D_{em,es}$	τ_{17} $c(\alpha) = 0.120$	τ_{19} $c(\alpha) = 0.085$	τ_{23} $c(\alpha) = 0.093$	τ_{31} $c(\alpha) = 0.115$	τ_{35} $c(\alpha) = 0.960$	τ_{41} $c(\alpha) = 0.120$
Opt-KDE	0.027	0.033	0.021	0.029	0.045	0.030
KDE-E2E	0.042	**0.059**	0.037	0.045	0.064	0.038
Gaussian	0.231 (reject)	0.189 (reject)	0.204 (reject)	0.176 (reject)	0.271 (reject)	0.266 (reject)
log-normal	0.117	0.134 (reject)	0.802	0.095	0.131 (reject)	0.128 (reject)
GMM	0.051	**0.046**	0.043	0.055	0.088	0.039
Hazard	0.109	0.087	0.091	0.103	0.882	0.130 (reject)

II shows the percentage of intersections that should deploy traffic detectors. In the best scenario, approximately 53% of intersections require deploying traffic detectors, and in the worst scenario, approximately 63% of intersections require deploying traffic detectors.

From Table III, we can observe that the performance of our proposed method, namely KDE-E2E, is always better than Gaussian, log-normal and hazard based model in each time interval, but a little worse than GMM in τ_{17} and τ_{31}. This can be explained by the fact that GMM has a similar structure with Opt-KDE. However, GMM, together with other counterparts are feasible based on the assumption that there should be observed data in each link. It is difficult to estimate their parameters once there are no observations.

Table IV presents the results of the KS test implemented on a randomly selected link. Clearly, our proposed model is accepted at each time interval. Compared to KDE-E2E and GMM, we can find that $D_{em,es}$ of GMM in τ_{19} is smaller than $D_{em,es}$ obtained from our proposed model. This result is consistent with the result in Table III.

VI. CONCLUSION

Motivated by the network tomography, in this paper, we proposed a KDE based method to estimate LTTD over the whole road network. As the method only needs the E2E measurements detected by the traffic data collected at or near both ends of the path, only a limited number of traffic detectors are necessary. With the real data, the proposed method has better performance than the widely used parametric models, e.g., Gaussian and log-normal in both traffic states of free flow and congestion. From Algorithm 1, we can observe that the complexity of the EM based algorithm depends on the number of parameters in the KDE, which has a close relationship with the number of E2E measurements. In our future work, we aim to come up with a sampling algorithm to control the number of available E2E measurements in order to improve the efficiency of the EM based algorithm.

REFERENCES

[1] C. Biernacki, G. Celeux, and G. Govaert, “Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models,” *Computational Statistics & Data Analysis*, vol. 41, no. 3-4, pp. 561–575, 2003.

[2] P. Duan, G. Mao, C. Zhang, and S. Wang, “Starima-based traffic prediction with time-varying lags,” pp. 1610–1615, 2016.

[3] Y. Guessous, M. Aron, N. Bhouri, and S. Cohen, “Estimating travel time distribution under different traffic conditions,” *Transportation Research Procedia*, vol. 3, pp. 339–348, 2014.

[4] S. H. Hamdar, A. Talebpour, and J. Dong, “Travel time reliability versus safety: A stochastic hazard-based modeling approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 264–273, 2015.

[5] M. Li, X. Zhou, and N. M. Rouphail, “Quantifying travel time variability at a single bottleneck based on stochastic capacity and demand distributions,” *Journal of Intelligent Transportation Systems*, vol. 21, no. 2, pp. 79–93, 2017.

[6] R. Li, G. Rose, and M. Sarvi, “Using automatic vehicle identification data to gain insight into travel time variability and its causes,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1945, pp. 24–32, 2006.

[7] U. Mori, A. Mendiburu, M. Álvarez, and J. A. Lozano, “A review of travel time estimation and forecasting for advanced traveller information systems,” *Transportmetrica A: Transport Science*, vol. 11, no. 2, pp. 119–157, 2015.

[8] E. K. Moylan and T. H. Rashidi, “Latent-segmentation, hazard-based models of travel time,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2174–2180, 2017.

[9] A. Prokhorchuk, V. P. Payyada, J. Dauwels, and P. Jaillet, “Estimating travel time distributions using copula graphical lasso,” pp. 1–6, 2017.

[10] I. Sanaullah, M. Quddus, and M. Enoch, “Developing travel time estimation methods using sparse gps data,” *Journal of Intelligent Transportation Systems*, vol. 20, no. 6, pp. 532–544, 2016.

[11] K. Tang, S. Chen, and Z. Liu, “Citywide spatial-temporal travel time estimation using big and sparse trajectories,” *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[12] Y. Tsang, M. Coates, and R. D. Nowak, “Network delay tomography,” *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2125–2136, 2003.

[13] N. R. Velaga, M. A. Quddus, and A. L. Bristow, “Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems,” *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 672–683, 2009.

[14] Y. Wang, Y. Zheng, and Y. Xue, “Travel time estimation of a path using sparse trajectories,” pp. 25–34, 2014.

[15] D. Woodard, G. Nogin, P. Koch, D. Racz, M. Goldszmidt, and E. Horvitz, “Predicting travel time reliability using mobile phone gps data,” *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 30–44, 2017.

[16] Q. Yang, G. Wu, K. Boriboonsomsin, and M. Barth, “A novel arterial travel time distribution estimation model and its application to energy/emissions estimation,” *Journal of Intelligent Transportation Systems*, pp. 1–13, 2017.

[17] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, “Urban link travel time estimation using large-scale taxi data with partial information,” *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37–49, 2013.

[18] R. Zhang, S. Newman, M. Ortolani, and S. Silvestri, “A network tomography approach for traffic monitoring in smart cities,” *IEEE Transactions on Intelligent Transportation Systems*, 2018.