Automation in Construction xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data

Xianyuan Zhan [a], Satish V. Ukkusuri [a,b,*], Chao Yang [b]

[a] Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907, USA
[b] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, School of Transportation Engineering, Tongji University, 4800 Cao'an Road, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 31 August 2015
Received in revised form 11 November 2015
Accepted 5 December 2015
Available online xxxx

Keywords:
Short-term average link travel time estimation
Trip based data with partial information
Bayesian mixture model
Path inference
EM algorithm

ABSTRACT

Accurate estimation and prediction of urban link travel times are important for urban traffic operations and management. This paper develops a Bayesian mixture model to estimate short-term average urban link travel times using large-scale trip-based data with partial information. Unlike typical GPS trajectory data, trip-based data from taxies or other sources provide limited trip level information, which only contains the trip origin and destination locations, trip travel times and distances, etc. The focus of this study is to develop a robust probabilistic short-term average link travel time estimation model and demonstrate the feasibility of estimating network conditions using large-scale trip level information. In the model, the path taken by each trip is considered as latent and modeled using a multinomial logit distribution. The observed trip data given the possible path set and the mean and variance of the average link travel times can thus be characterized using a finite mixture distribution. A transition model is also introduced to serve as an informative prior that captures the temporal and spatial dependencies of link travel times. A solution approach based on the expectation–maximization (EM) algorithm is proposed to solve the problem. The model is tested on estimating the mean and variance of the average link travel times for 30 min time intervals using a large-scale taxi trip dataset from New York City. More robust estimation results are obtained owing to the adoption of the Bayesian framework.

1. Introduction

Accurate estimation and prediction of urban link travel times are essential for various applications in urban traffic operations and management. Traditional approaches for urban link travel time estimation and prediction have largely relied on link-based data from various sources, such as loop detectors [3,18,22,27], automated vehicle identification (AVI) sensors [15,19,21], and remote traffic microwave sensors (RTMS) [24]. All of these approaches require installing corresponding fixed sensors to collect link level traffic information. Cost associated with installing and maintaining the physical facilities of sensors on each road segment limits the use of the previous approach only to major road segments or small transportation networks.

With the rapid development of pervasive computing technologies, we are at an important point in history where our ability to collect data far exceeds our ability to make useful inferences of the data. The recent availability of the massive vehicle re-identification data from urban sparse sensor networks as well as the large-scale spatio-temporal data from mobile sensors have provided a new alternative to use the trip-based data for urban link travel time estimation and predictions. Different from the traditional link-based data, the emerging trip-based data are more general and widely available in different data sources, however, only contain partial trip level information (starting and ending time and location).

For instance, the trip-based data can be easily obtained from the sparse sensor networks that have the capability of re-identifying vehicles at multiple locations, such as the automated license plate recognition (ALPR) data from multiple ALPR camera equipped intersections [26], the urban vehicle tracking system using radio frequency identification (RFID) technology [16,30]; and the vehicle re-identification data from wireless magnetic sensors [14]. These types of data all share similar characteristics that the vehicle license plate (ALPR data) or "virtual license plate" (RFID tags in RFID vehicle tracking system and the vehicle magnetic signals from wireless magnetic sensors) information and the passing timestamps of each vehicle are recorded at each monitored location. Thus when data from several sparsely located sites are available, the vehicle with the same license plate or "virtual license plate" number can be matched together, and the trip level information of vehicles can be established. Recent years, there have been rapid deployments of large-scale ALPR systems and RFID vehicle tracking systems in many countries. For example, Beijing has a ALPR system of 374 high definition (HD) ALPR cameras in 2010, and this number will increase to 3000 by 2015 [1]. On the other hand, a large-scale RFID vehicle

* Corresponding author at: Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA. Tel.: +1 765 494 2296.
E-mail addresses: zhanxianyuan@purdue.edu (X. Zhan), sukkusur@purdue.edu (S.V. Ukkusuri), tongjiyc@tongji.edu.cn (C. Yang).

http://dx.doi.org/10.1016/j.autcon.2015.12.007
0926-5805/

tracking system has been planned and will be deployed throughout Shanghai in the next few years [30]. Massive amount of streaming trip-based data can be obtained from such large-scale sparse sensor networks.

Another more promising source of urban trip-based data is from mobile sensors, such as the global positioning system (GPS) devise equipped vehicles or users' mobile phones, which can be viable sources of data for monitoring traffic conditions in large cities [8]. Commercial companies such as Inrix have already shown value by collecting and utilizing "large-scale" historical traffic data from GPS-enabled vehicles or mobile phones. As an important component of public transportation system in urban areas, taxicabs equipped with GPS devices have been increasingly considered as an ideal ubiquitous sensors in monitoring the traffic states in urban transportation networks. Currently, installing GPS devises in taxicabs has become a common practice in many cities, which is mainly used to locate taxis and track lost packages, etc. There are two nice features about the GPS data generated by taxis. Firstly, data related to taxi movements are abundant. In New York City, there are around 13,000 yellow medallion taxis serving 240 million passengers per year and transporting 71% of all Manhattan residents' trips [17]. In Hong Kong, there are 15,000 taxis by the end of 2013 and transporting more than a million passengers everyday [7]. The huge number of taxi trips generate vast amount of data, providing valuable information about the real-time traffic conditions in the urban transportation network. Secondly, the unique mobility feature of the taxi data enables the continuous monitoring of urban transportation networks with large coverage areas without the need of installing any fixed sensors. Ideally the detailed GPS trajectories information can be obtained from these mobile sources, however in practice, it is often not the case. For example, GPS devises with low updating frequency often result in extremely sparse vehicle trajectories, which is not possible to cover link level traffic condition but can be treated as the trip-based data (trips between consecutive sparse trajectory points). Also, due to technological limitations and privacy concerns from taxi operation agencies in many cities, detailed trajectory data from taxis are not always available, while the trip-based data from taxis are shared by local agencies, e.g., New York City Taxi and Limousine Commission (NYCTLC). Unlike typical GPS trajectory data, such large-scale trip-based data from mobile sensors provide limited information, which only contain the origin and destination coordinates, travel time and distance of a trip.

All the abovementioned trip-based data share several common characteristics. First, the data only contain very sparse trip level information (starting and ending time and location), hence exact path taken by a vehicle needs to be inferred. Secondly, such data have extensive amount of records, which compensates for the incompleteness of the information and makes the link travel time estimation possible. Currently, there is limited research on estimating short-term urban link travel times using large-scale trip-based data. For the trip-based data from sparse sensor networks, researchers mainly focus on estimating travel time on long stretch of arterials [14] when the actual path taken by a vehicle does not need to be inferred. For trip-based data from mobile sensors, most of the recent researches focus on estimating travel time using detailed trajectory data from GPS-equipped vehicles or mobile phones [8–10,29]. These models require high resolution trip trajectories to be known, which is inapplicable when partial information is available. Zhan et al. [25] proposed a short-term urban link travel time estimation model using the large-scale taxi trip data. The model is essentially solving a least square regression problem that minimizes the error between the observed and average link travel times. This makes the model only capable of providing point estimates for the hourly average of link travel times, which cannot incorporate the variability of link travel times. Moreover, the model relies on strong assumptions on independency of link travel times, and is unable to incorporate historical data, which limits its ability to provide robust short-term average link travel times for real world applications. The fact is that the urban link travel times are intrinsically uncertain due to the within day variability of traffic conditions, the stochastic characteristics of delays at signalized intersections, impacts from weather and other environmental attributes. This means that the variability of the travel time should be estimated in addition to the mean travel time of a roadway [20,28]. This has important implications in urban link travel time reliability and the understanding of the path choice behavior of travelers. Hence it is essential to consider the link travel time as a distribution and estimate both its mean and variance.

This study develops a robust probabilistic short-term average link travel time estimation model that only uses the limited information from large-scale trip-based data, and has the ability to capture the variability in short-term link travel times. Furthermore, a taxi trip dataset from New York City is used to demonstrate the potential and practical value of estimating network wide link states using large-scale trip-based data in urban transportation operation and management applications. This work contributes to the literature in the following aspects:

1. One of the first studies that models and solves short-term urban link travel time estimation problem using general large-scale trip-based data without detailed path information.
2. Probabilistic modeling framework that estimates both mean and variance of short-term link travel times.
3. Robust solution approach is developed, which allows for efficient real-time implementation.
4. A large-scale taxi trip dataset from New York City is used to demonstrate the proposed model, which shows the potential of utilizing various sources of trip-based data for urban link travel time estimation.

The paper is organized as follows: the next section presents the methodology of the proposed model, which includes the description of model assumption and model development. The third section proposes an EM algorithm to solve the given problem. The fourth section presents the numerical results and validation of the model. The last section concludes the paper and provides several future extensions of the model.

2. Methodology

This section presents the proposed Bayesian link travel time estimation framework. We will first introduce several modeling assumptions, followed by the definition of notations and the detailed description of the Bayesian mixture model for short-term average link travel time estimation.

2.1. Model assumptions

To reduce the modeling complexity, we first pose the following assumptions in the link travel time estimation problem:

1. Short-term average link travel times are approximated and modeled using normal distributions. Although link travel times from individual vehicles generally do not necessarily follow normal distribution, the distribution of average link travel times within a short time period behaves more similar to the normal distribution as a result of the well-known central limit theorem in probability theory. Another advantage of using normal distribution is its additivity property, which can significantly reduce model complexity and allow for tractable parameter estimation.
2. The delay at intersections caused by traffic signal is included in the link travel time. Since only limited path level information is available, it is insufficient to separate intersection delay from total travel time. Therefore, we focus on the estimation of the mean and variance of the average link travel time over a short time period.
3. The spatial (upstream and downstream neighboring links) and temporal (between consecutive time intervals) dependencies on

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

average link travel times are captured by a transition model and used as an informative prior for the model. However, in order to reduce computational complexity, the average link travel times within the same time interval are assumed to be independent among links.

4. The travel time when a taxi driver traversing part of the link is proportional to the distance he/she traveled on the link. This assumption is used to obtain the partial travel times that the driver spent on starting and ending links of the trip.
5. Driver's route choice is based on utility maximization. The assumption is that each driver minimizes both trip time and distance. Further, the path cost perceived by a driver during the route choice decision-making is based on mean link travel times and distances.

2.2. Notations

x_l^t	Average link travel time of link l in time interval t
μ_l^t	Mean of the average link travel time for link l in time interval t
σ_l^t	Standard deviation of the average link travel time for link l in time interval t
d_k	Total distance of path k
y_i^t	Actual trip travel time of trip observation i in time interval t
ρ_i^t	Predicted trip travel time of trip observations i in time interval t
R_i	Reasonable path set of trip observation i
d^i	Actual trip distance of trip observation i
z_k^i	Latent variable denoting the use of path k in observation i, $z_k^i \in \{0,1\}$
ψ	Hyperparameter measuring non-spatial correlation of average link travel times between consecutive time intervals
κ	Hyperparameter measuring spatial correlation of average link travel times between consecutive time intervals
M	Matrix representing spatial correlation among different links in the network
$\boldsymbol{\alpha}$	Vector of distance proportion parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$
$\boldsymbol{\beta}$	Vector of parameters associated with the path cost computation, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$
$\boldsymbol{x}^t$	Vector of average link travel times in the network, $\lvert\boldsymbol{x}^t\rvert = m$
$\boldsymbol{y}^t$	Vector of trip travel times, $\lvert\boldsymbol{y}^t\rvert = n^t$
$\boldsymbol{z}$	Vector of latent variable $\boldsymbol{z} = (z_k^i)$, $\forall k \in R^i, i = 1, \ldots, n^t$
$\boldsymbol{\mu}^t$	Vector for the mean of average link travel times in time interval t, $\boldsymbol{\mu}^t = (\mu_1^t, \ldots, \mu_m^t)^T$
$\boldsymbol{\Sigma}^t$	Covariance matrix for the average link travel times in time interval t, $\boldsymbol{\Sigma}^t = diag((\sigma_1^t)^2, \ldots, (\sigma_m^t)^2)$
$\boldsymbol{D}^t$	Set of all trip distances in time interval t

2.3. Observation model

In this paper, we propose a new Bayesian mixture model to estimate short-term average link travel times. The proposed model adopts an online Bayesian approach, which includes: 1) an observation model $P(\boldsymbol{y}^t \mid \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ parameterized by time-varying parameter $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$ for the behavior of observed trip travel times $\boldsymbol{y}^t$, and 2) a transition model (served as prior distribution) $P(\boldsymbol{\mu}^t \mid \boldsymbol{\mu}^{t-1})$ to describe the time-varying parameter $\boldsymbol{\mu}^t$. We begin our discussion by first introducing the observation model.

In the observation model, the short-term average link travel time x_l^t of link l in time interval t is assumed to follow a normal distribution $N(\mu_l^t, (\sigma_l^t)^2)$, with mean μ_l^t and variance $(\sigma_l^t)^2$. For simplicity, we write the mean of all average link times to be $\boldsymbol{\mu}^t = (\mu_1^t, \ldots, \mu_m^t)^T$ and $\boldsymbol{\Sigma}^t = diag((\sigma_1^t)^2, \ldots, (\sigma_m^t)^2)$. The path travel time is hence modeled as the summation of a set of average link travel times, and the probability of the actual trip travel time y_i^t of observation i in time interval t using path k is:

$$\alpha_1 x_0^t + \alpha_2 x_D^t + \sum_{l \in k \setminus \{O,D\}} x_l^t \sim N\left(\alpha_1 \mu_0^t + \alpha_2 \mu_D^t + \sum_{l \in k \setminus \{O,D\}} \mu_l^t, (\alpha_1 \sigma_O^t)^2 + (\alpha_2 \sigma_D^t)^2 + \sum_{l \in k \setminus \{O,D\}} (\sigma_l^t)^2\right) \quad (1)$$

where $(\mu_O^t, (\sigma_O^t)^2)$, $(\mu_D^t, (\sigma_D^t)^2)$ represent the mean and the variance of starting and ending links of the trip. As a driver only experiences part of the total link travel times when traversing on the starting/ending links of the trip (e.g., picking-up/dropping-off passengers in the middle of the street), follow Assumption 4 in Section 2.1, we introduce

α_1, α_2 to be the distance proportions that the vehicle traverses on the starting and ending links. The travel time of the vehicle on the starting/ending links can hence be modeled as $\alpha_1 x_0^t$ and $\alpha_2 x_D^t$. Fig. 1 provides a simple illustration of the modeling of path travel times. The notation L_O and L_D denotes the length of starting and ending links of the trip. For simplicity, denote:

$$\begin{aligned} g_k^i(\boldsymbol{\mu}^t) &= \alpha_1 \mu_0^t + \alpha_2 \mu_D^t + \textstyle\sum_{l \in k \setminus \{O,D\}} \mu_l^t, \\ h_k^i(\boldsymbol{\Sigma}^t) &= (\alpha_1 \sigma_O^t)^2 + (\alpha_2 \sigma_D^t)^2 + \textstyle\sum_{l \in k \setminus \{O,D\}} (\sigma_l^t)^2 \end{aligned} \quad (2)$$

Thus $g_k^i(\boldsymbol{x}^t) \sim N(g_k^i(\boldsymbol{\mu}^t), h_k^i(\boldsymbol{\Sigma}^t))$, and the probability of path travel time y_i^t using path k is:

$$(y_i^t \mid k, \boldsymbol{x}^t) = P\left(y_i^t \mid k, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right) = N\left(y_i^t \mid g_k^i(\boldsymbol{\mu}^t), h_k^i\left(\boldsymbol{\Sigma}^t\right)\right) \quad (3)$$

Since the detailed trajectory information is unknown in the trip based data, the actual path taken by a driver needs to be inferred. Given the origin and destination of a trip, the size of the possible path set for a trip is typically huge in a large network. To reduce the size of the problem and make the short-term travel time estimation problem tractable, we first obtain a set of candidate paths for each trip record by constructing an initial path set using a k-shortest path algorithm [23]. The path length is computed based on link distances. Only the paths with distances that do not significantly deviate from the observed trip distances will then be included in a reasonable path set R_i for model

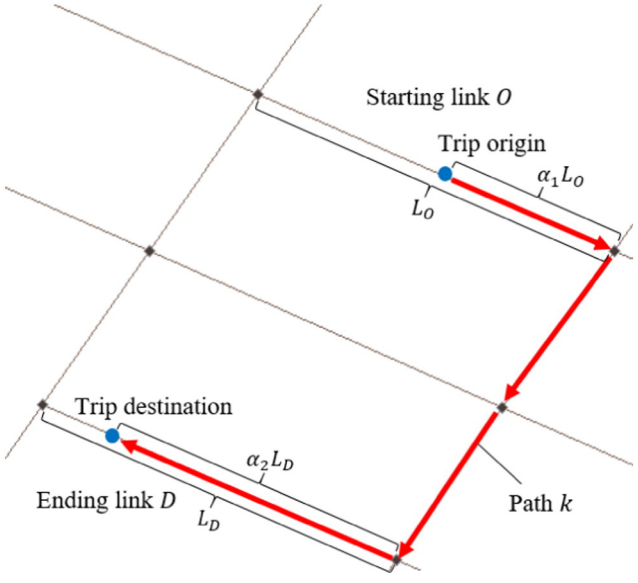

Fig. 1. Illustration of the modeling of path travel times.

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

estimation. Furthermore, if all of the path distances in the generated initial path set deviate significantly from the observed trip distance, this trip record will be dropped from the estimation. The probability of taking a particular path k is evaluated using a route choice model based on utility maximization (Assumption 5 in Section 2.1), which is formulated as a multinomial logit distribution that is widely used in route choice analysis:

$$\pi_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, \boldsymbol{D}^t) = \frac{\exp\left[-C_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, d_k)\right]}{\sum_{s\in R_i}\exp\left[-C_s^i(\boldsymbol{\mu}^t, \beta, d_s)\right]} \tag{4}$$

where d_k is the trip distance for trip i that takes path k in time interval $t, d_k \in \boldsymbol{D}^t$; $C_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, d_k)$ is the path cost function. Here we assume the perceived path cost is based on the trip distance and the mean link travel times $\boldsymbol{\mu}^t$, rather than the actual trip travel time. This will significantly reduce the model complexity, and is also more reasonable, as drivers are not possible to know the actual trip travel times before they make route choice decision. However, they may perceive the average link travel times of the network based on their experience. For the case of trip-based data from taxis, following Zhan et al. [25], we model the path cost function $C_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, d_k)$ as a combination of trip travel time and trip distance, which is given as:

$$C_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, d_k) = \beta_1 g_k^i(\boldsymbol{\mu}^t) + \beta_2 d_k \tag{5}$$

where β_1, β_2 are cost coefficients for trip travel time and distance. According to Zhan et al. [25], the estimated values for β_1, β_2 from the same large-scale taxi trip dataset are given as 0.275/min and 1.563/mile.

After developing the path travel time distribution of a particular path k and the corresponding route choice probability $P(y_i^t|k, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, the path travel time of trip observation i in time interval t can hence be modeled as the following finite mixture distribution:

$$P\left(y_i^t|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{D}^t\right) = \sum_{k\in R_i}\pi_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, \boldsymbol{D}^t)P\left(y_i^t|k, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right). \tag{6}$$

Finally, the observation model of all observations $\boldsymbol{y}^t$ given the set of link travel time parameters $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$ takes the form of:

$$H\left(\boldsymbol{y}^t|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{D}^t\right) = \prod_{i=1}^n\sum_{k\in R_i}\pi_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, \boldsymbol{D}^t)P\left(y_i^t|k, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right). \tag{7}$$

2.4. Transition model

In the urban environment, the short-term link travel times between consecutive time intervals can be highly correlated, and spatial dependencies among adjacent links also exist. It is important to incorporate such temporal and spatial dependencies into consideration. Given the limited information provided in the data, we introduce a transition model that serves as an informative prior to the observation model. Using this transition model, the historical estimation results can be incorporated to provide extra information to current estimation, and thus lead to more robust estimation results.

A typical strategy in online Bayesian learning to make use of the knowledge from historical estimation results is to introduce a transition model that is based on the square norm of differences between the current parameter θ^t and the previous parameter θ^{t-1} [4, 13], which is

$$P\left(\theta^t|\theta^{t-1}, \gamma\right) = N\left(\theta^t|\theta^{t-1}, \gamma^{-1}\right) \propto \exp\left(-\frac{\gamma}{2}\left\|\theta^t - \theta^{t-1}\right\|^2\right). \tag{8}$$

In this work, we apply the similar strategy and introduce a transition model for the mean of the short-term average link travel times $\boldsymbol{\mu}^t$, with the consideration of both temporal dependency between consecutive time periods and spatial dependencies among adjacent links. The proposed transition model is

$$P(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1}, \psi, \kappa) = N(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1}, \psi\boldsymbol{I} + \kappa\boldsymbol{M}) \tag{9}$$

where ψ, κ are predetermined hyperparameters. Specifically, ψ is the scale parameter for non-spatial error, and κ is the scale parameter that captures spatial correlation among adjacent links (both upstream and downstream neighboring links). $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{M}$ is the square matrix that represents spatial dependencies among links, which is defined as follows:

$$M_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and link } i \text{ is adjacent to link } j \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Here we only introduce the prior distribution for the mean of short-term average link travel times $\boldsymbol{\mu}^t$ rather than $\boldsymbol{\Sigma}^t$. The rationale behind this treatment is based on the observation that the short-term average link travel times $\boldsymbol{\mu}^t$ usually does not vary significantly from $\boldsymbol{\mu}^{t-1}$, while the transition between $\boldsymbol{\Sigma}^t$ and $\boldsymbol{\Sigma}^{t-1}$ usually lacks regularity.

2.5. Overall model

Following the Bayes' theorem,

posterior∝likelihood × prior

the overall Bayesian mixture model is given as

$$\begin{aligned}\Pi\left(\boldsymbol{y}^t|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\mu}^{t-1}, \boldsymbol{D}^t, \psi, \kappa\right) &= \frac{H\left(\boldsymbol{y}^t|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{D}^t\right)\cdot P(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1}, \psi, \kappa)}{\int \boldsymbol{H}\left(\mathbf{y}^t|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \mathbf{D}^t\right)\cdot P(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1}, \psi, \kappa)d\boldsymbol{\mu}^t} \\ &= \frac{1}{Z}\prod_{i=1}^n\sum_{k\in R_i}\pi_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, \boldsymbol{D}^t)P\left(y_i^t|k, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right)\cdot N(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1}, \psi\boldsymbol{I} + \kappa\boldsymbol{M})\end{aligned} \tag{11}$$

where Z is the normalizing constant. The parameter to be estimated are $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$, whereas $\boldsymbol{y}^t, \boldsymbol{D}^t$ are observed from the data and $\boldsymbol{\mu}^{t-1}$ is from the

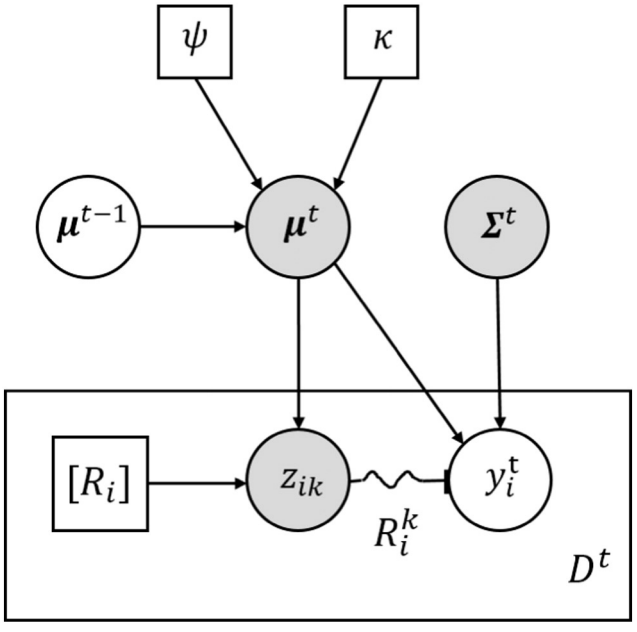

Fig. 2. Plate notation for the proposed mixture model.

estimation result of the previous time interval. The plate notation of the above model can be represented in Fig. 2. In the plate notation, the large rectangle plate represents the repetition for the data observations D^t. The squares indicate fixed parameters and the circles indicate random variables. Filled-in shapes indicate variables with unknown values which are to be inferred. The indication $\{R_i\}$ indicates the reasonable path set of observation i, and R_i^k represents a specific path in the reasonable path set. z_i^k is a latent variable that corresponds to the choice of path k for observation i, which will be further described in the following section.

3. Solution approach

The proposed problem is a big data problem involving a large number of taxi trips, with each containing a certain number of reasonable paths. An expectation–maximization (EM) algorithm [2,5] is developed to efficiently estimate the model parameters. EM algorithm is a powerful tool for finding maximum likelihood solution for models involving latent variables. The EM algorithm contains two repeated updating steps: the expectation (E) step and the maximization (M) step. The E-step takes expectation over the latent variable using current parameter values to remove latent variables from the formulation. The M-step re-estimates the model parameters by maximizing the expected value of either the complete-data log likelihood, which provides maximum likelihood estimation (MLE) solution, or the complete-data posterior distribution, which provides maximum posterior (MAP) solution [2]. In this work, we present the results for both the MLE and the MAP solution for the proposed model. The MAP (Bayesian) solution is generally more accurate compared with the MLE solution, since it incorporates additional prior information and effectively reduces the overfitting that typically occurs in MLE solutions.

The EM algorithm is generally preferred over directly maximizing the likelihood function when dealing with model involving hidden variables such as the incomplete-data log likelihood function in Eq. (7). This is because that the EM algorithm can produce more robust estimates and avoid singularities of the likelihood function in which a mixture component collapses onto a particular data point [2]. Moreover, the EM algorithm is proved to converge to a local maximum of the observed data likelihood function. For detailed discussion about the EM algorithm and its convergence to local maxima of the likelihood function, please refer to Dempster et al. [5] and Bishop [2]. The development and description of the proposed EM algorithm are presented as follows:

Step 1: Initialize model parameters: $\boldsymbol{\mu}_{old}^t$ and $\boldsymbol{\Sigma}_{old}^t$.

Step 2: E-step:

The route choice for each trip can be perceived as a latent variable in the model. Let z_k^i be a binary variable that takes values of 0 and 1, with 1 suggesting the path being utilized, then

$$P\left(z_k^i=1\right)=\pi_k^i\left(\boldsymbol{\mu}^t|\boldsymbol{\beta}|d_k\right). \tag{12}$$

Thus

$$P\left(y_i^t|z_k^i=1\right)=P\left(y_i^t|k,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right) \tag{13}$$

$$P(y_i^t|\boldsymbol{z})=\prod_{k\in R_i}P\left(y_i^t|k,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)^{z_k^i}. \tag{14}$$

Using Bayes Theorem

$$P\left(\boldsymbol{y}^t,\boldsymbol{z}|\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\propto\prod_{i=1}^n\prod_{k\in R_i}\left[\pi_k^i\left(\boldsymbol{\mu}^t,\boldsymbol{\beta},d_k\right)P\left(y_i^t|k,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\right]^{z_k^i}. \tag{15}$$

From this posterior distribution, the expected value over z_k^i can be computed as

$$\begin{aligned}\mathbb{E}\left(z_k^i\right)\equiv P\left(z_k^i=1|\boldsymbol{y}\right)&=\frac{\sum_{z_k^i}z_k^i\left[\pi_k^i\left(\boldsymbol{\mu}^t,\boldsymbol{\beta},d_k\right)P\left(y_i^t|k,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\right]^{z_k^i}}{\sum_{z_k^i}\sum_{s\in R_i}\left[\pi_s^i\left(\boldsymbol{\mu}^t,\boldsymbol{\beta},d_s\right)P\left(y_i^t|s,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\right]^{z_s^i}}\\&=\frac{\pi_k^i\left(\boldsymbol{\mu}^t,\boldsymbol{\beta},d_k\right)P\left(y_i^t|k,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)}{\sum_{s\in R_i}\left[\pi_s^i\left(\boldsymbol{\mu}^t,\boldsymbol{\beta},d_s\right)P\left(y_i^t|s,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\right]}=\gamma\left(z_k^i\right)\end{aligned}. \tag{16}$$

In the above equation, $\gamma(z_k^i)$ is evaluated using the current parameter values $\boldsymbol{\mu}_{old}^t$ and $\boldsymbol{\Sigma}_{old}^t$.

Step 3: M-step:

The expected value of the complete-data log likelihood function is given as

$$\begin{aligned}\boldsymbol{\Sigma}^t)&=\mathbb{E}_{\boldsymbol{z}}\left[\ln P(\boldsymbol{y}^t,\boldsymbol{z})|\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right]=\sum_{\boldsymbol{z}}P\left(\boldsymbol{z}|\boldsymbol{y}^t,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)1\mathrm{n}\,P\left(\boldsymbol{y}^t,\boldsymbol{z}|\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\\&=\sum_{i=1}^n\sum_{k\in R_i}\gamma\left(z_k^i\right)\left[\ln\pi_k^i\left(\boldsymbol{\mu}^t,\boldsymbol{\beta},d_k\right)+\ln P\left(y_i^t|k,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\right]\\&=\sum_{i=1}^n\sum_{k\in R_i}\gamma\left(z_k^i\right)\left\{-\beta_1 g_k^i\left(\boldsymbol{\mu}^t\right)-\beta_2 d_k-\ln\sum_{s\in R_i}\exp\left[-\beta_1 g_s^i\left(\boldsymbol{\mu}^t\right)-\beta_2 d_s\right]\right.\\&\left.-\frac{1}{2}\ln(2\pi)-\frac{1}{2}h_k^i\left(\boldsymbol{\Sigma}^t\right)-\frac{\left[y_i^t-g_k^i\left(\boldsymbol{\mu}^t\right)^2\right]}{2h_k^i\left(\boldsymbol{\Sigma}^t\right)}\right\}.\end{aligned} \tag{17}$$

Similarly, the complete-data log posterior probability function is given as

$$\begin{aligned}F\left(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)&=\mathbb{E}_{\boldsymbol{z}}\left[\ln\Pi\left(\boldsymbol{y}^t,\boldsymbol{z}\,|\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t,\boldsymbol{\mu}^{t-1},\boldsymbol{D}^t,\psi,\kappa\right)\right]\\&=\sum_{\boldsymbol{z}}P\left(\boldsymbol{z}|\boldsymbol{y}^t,\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\ln\left[\frac{1}{Z}P\left(\boldsymbol{y}^t,\boldsymbol{z}|\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)N\left(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1},\psi\boldsymbol{I}+\kappa\boldsymbol{M}\right)\right]\\&=Q\left(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)+\ln N\left(\boldsymbol{\mu}^t|\boldsymbol{\mu}^{t-1},\psi\boldsymbol{I}+\kappa\boldsymbol{M}\right)-\ln Z\\&=Q\left(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)-\frac{m}{2}\ln(2\pi)-\frac{1}{2}\left|\psi\boldsymbol{I}+\kappa\boldsymbol{M}\right|^{-1}\\&-\frac{1}{2}\left(\boldsymbol{\mu}^t-\boldsymbol{\mu}^{t-1}\right)^T\left(\psi\boldsymbol{I}+\kappa\boldsymbol{M}\right)^{-1}\left(\boldsymbol{\mu}^t-\boldsymbol{\mu}^{t-1}\right)-\ln Z\end{aligned}. \tag{18}$$

The updated parameter is obtained by either maximizing $Q(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$, which results in MLE solution; or maximizing $F(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$, which leads to the MAP (Bayesian) solution of the proposed Bayesian mixture model. This can be seen as solving one of the following two constrained optimization problems:

$$\begin{aligned}&\max_{\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t}Q\left(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\\&s.t.\quad\mu_l^t\geq t_{min}>0\end{aligned} \tag{19}$$

or

$$\begin{aligned}&\max_{\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t}F\left(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right)\\&s.t.\quad\mu_l^t\geq t_l^{min}>0\end{aligned} \tag{20}$$

where $t_l^{\min}$ is the minimum possible travel time for link l. Note that the $\ln Z$ in $F(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ is a constant, hence can be omitted from the optimization problem. As both $Q(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ and $F(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ are twice continuous differentiable, a wide range of the constrained optimization algorithm can be used to efficiently solve the short-term link travel time estimation problem.

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

Step 4: Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\left(\boldsymbol{\mu}_{new}^{t}, \boldsymbol{\Sigma}_{new}^{t}\right) = \underset{\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t}{arg\ max}\ F\left(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right) \text{ or } \underset{\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t}{arg\ max}\ Q\left(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right) \quad (21)$$
$$\boldsymbol{\mu}_{old}^{t} \leftarrow \boldsymbol{\mu}_{new}^{t}, \quad \boldsymbol{\Sigma}_{old}^{t} \leftarrow \boldsymbol{\Sigma}_{new}^{t}$$

and repeat steps 2 and 3.

4. Numerical experiments

4.1. Test data and network

The trip-based data used in this research is from a large-scale taxi trip dataset collected by NYCTLC. The data contains the information of trip origin and destination coordinates, trip distance, trip duration and other related information. Around 30,000 to 50,000 daily trips are recorded in the entire year of 2013. In this study, we extract a week's data (from 2013/10/07 to 2013/10/13) to test the proposed model. A 1175 m × 1780 m rectangle area in Midtown Manhattan is selected to serve the study region. The corresponding transportation network inside the study region is illustrated in Fig. 3, which contains 136 nodes and 254 directed links. This network includes several highly congested road segments in Midtown Manhattan, such as 5th Avenue, 7 Avenue, and Broadway. Severe congestions are expected to be observed in the estimation results. All taxi trip data with both origins and destinations fall within the study region are extracted to test the proposed model.

Due to the limited amount of data available in the study region, we split the extracted data into 30 min time intervals to perform the average link travel time estimation. The choice of using 30 min time interval is to guarantee that enough data are available to perform travel time estimation, while keeping the length of time interval as short as possible. The amount of data observed from each of the 30 min time interval over the study week is illustrated in Fig. 4. From the data, we observe as many as 1400 trips for weekdays and 800 trips for weekends within a time interval. For real world application of the proposed model, smaller time intervals can be used if the amount of data is sufficient. In the numerical experiments, we run the model from 8:00–22:00 for each of the 7 days in the tested week. For validation purposes, all trip observations in each time interval are randomly split into a training set (80% of all observations) and a test set (20% of observations). Only the data in the training set are used for estimation, which further reduces the number of observations used in the travel time estimation. In the real-world implementation of the proposed model, a larger study region and all the observations can be used for model estimation. This will result in significantly higher number of observations for training and lead to more accurate link travel time estimates.

4.2. Numerical results

The proposed model is implemented in MATLAB and parts of the codes are compiled into C to improve the computation efficiency. Before running the numerical scenarios, the k-shortest paths ($k = 20$) for each nodal pair in the network are computed. This step is necessary, since it can avoid the expensive k-shortest path computation during estimation process. The reasonable path set of the data can be then efficiently obtained by directly utilizing the already computed k-shortest path set of the network. A threshold ratio $r = 20\%$ is also introduced to filter out unqualified data which path distance is deviated more than $1 \pm r$ compared with the observed path distance. For the MAP solution, link travel times assuming vehicles traveling at constant speed of 10 mile/h are used as the initial values for the mean of the average link travel times $\boldsymbol{\mu}^0$ at first time interval. For later time intervals, the mean of the average link travel times $\boldsymbol{\mu}^{t-1}$ from previous time intervals is used in the transition model $P(\boldsymbol{\mu}^t | \boldsymbol{\mu}^{t-1}, \psi, \kappa)$. Furthermore the scale-

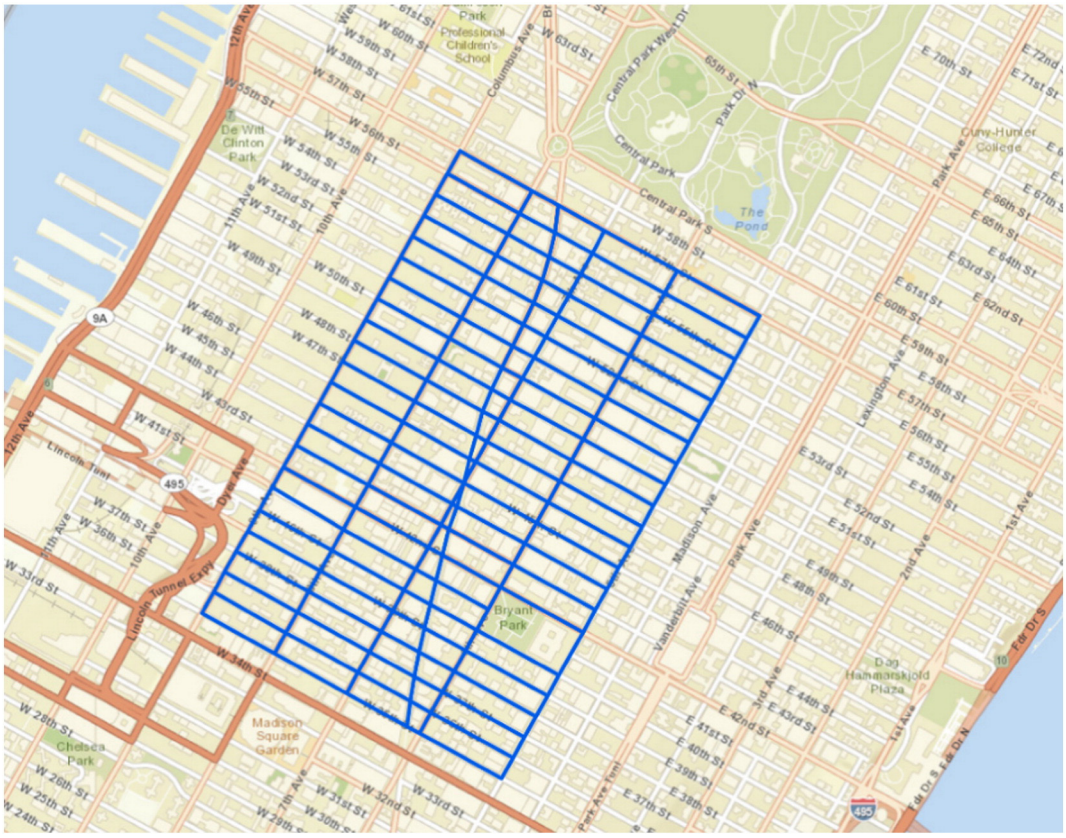

Fig. 3. Test network and study region.

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

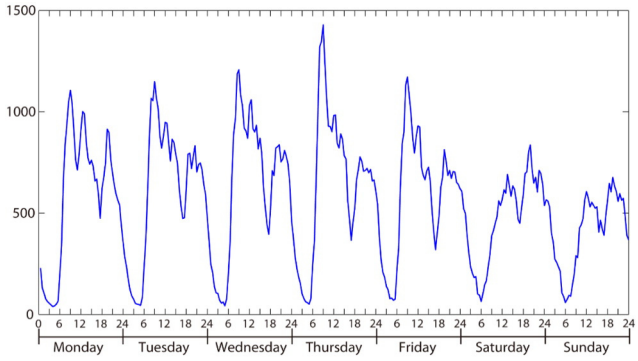

Fig. 4. Number of observed trips for each 30 min time interval in the study region.

parameter ψ and κ are set to be $\kappa = 0.1\psi$ in the numerical experiments, which suggests that the temporal correlation dominates the degree of spatial correlation from upstream and downstream neighboring links.

The proposed EM algorithm is found to be very robust and showed good convergence property in the numerical tests. As shown in Bishop [2], the EM algorithm increases the log likelihood function at each iteration, thus guarantees the convergence regardless of the choice of the initial parameter value used. However, a proper initial parameter value will substantially speed up the algorithm convergence. In this study, the initial value of both the mean and standard deviation of the average link travel times are computed as the link distances divided by 10 mile/h. All the numerical experiments successfully find the convergent solutions, which confirm the robustness of the solution approach. The convergence plot (Fig. 5) for the expected value of the complete-data log posterior probability function likelihood $F(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ plus the normalization term ln$Z$, the complete-data log likelihood function $Q(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, and the incomplete-data log likelihood LL ($\ln H(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D})$) are illustrated using the Monday 10:00–10:30 data when solving for the MAP solution. Rapid convergence is achieved during the first few iterations, and it is observed that 30 iterations are sufficient to obtain convergent solutions for all the test scenarios. The entire estimation process can be finished within 15 min using a 2.4 GHz CPU laptop. The computation time can be further reduced by implementing parallel computing techniques or using a more powerful computer.

The proposed model is tested using 30 min intervals' data from 8:00–22:00 for each day of the selected week. For brevity, we only present the estimation result of four time intervals (9:00–9:30, 13:00–13:30, 19:00–19:30 and 21:00–21:30) for a representative weekday (Wednesday) and a weekend (Saturday) in Figs. 6 and 7. The time interval from 9:00 to 9:30 contains the maximum number of trip observations in weekdays, which represent the morning peaks. The 13:00–13:30 and 19:00–19:30 time intervals correspond to another two smaller peaks in Fig. 4, and the time interval of 21:00–21:30 is selected to test for off-peak situations. To reduce repetition, only validation results for Monday, Wednesday, Friday and Saturday's numerical experiments are presented in Fig. 9, as result of the rest days in the tested week exhibit similar accuracies. The validation of the proposed model will be further discussed in the following section.

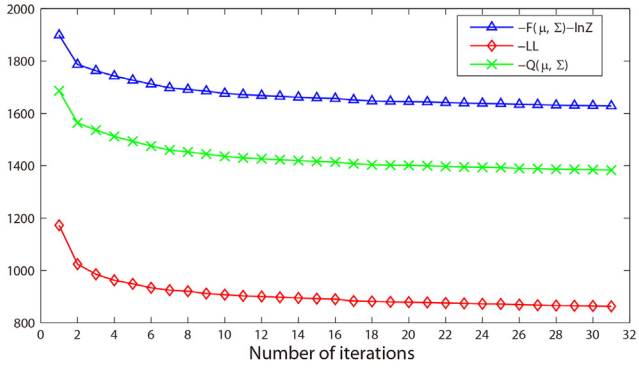

Fig. 5. Convergence plot of the proposed EM algorithm for scenario: Monday 10:00–10:30.

The estimated mean of average link speeds (referred as "mean speed" for brevity in following contents) instead of estimated mean of the average link travel times are used to give a more intuitive representation of the results (Figs. 6(a) and 7(a)). Note that since the signal delay is included into the link travel time estimation, thus the value of presented mean link speed will be lower than the usual traversing speed people experience while driving. The model estimation results are consistent with the highly congested situation of the network in the study region. For weekdays, it can be observed that the entire test network is severely congested during 9:00–9:30 and 13:00–13:30 time intervals. Most of the links in the network have low mean speeds. The results show that almost 70% of the links in the test network have mean speeds ranging from 3 to 10 mile/h in both of the two time intervals. The congestion has seen a trend of alleviation in 19:00–19:30 time interval, as more links are observed to have higher mean speeds. During 21:00–21:30 off-peak time interval, the traffic condition is observed to be greatly improved, that almost 50% of links have mean speed ranging from 5 to 18 mile/h. The situation on weekend is quite different from that on weekdays, where it is less congested during 9:00–9:30 time interval, and then becomes congested in 13:00–13:30 and 19:00–19:30 time intervals. The traffic conditions in these two time periods are still better than congested hours on weekdays, since there are about 60–65% of links have mean speeds ranging from 3 to 15 mile/h.

We also present the normalized standard deviation (estimated average link travel time standard deviation divided by link length) as a measure of the uncertainty and the variability of the estimated link travel times, presented in Figs. 6(b) and 7(b). The reason for normalizing the estimated standard deviation of link travel times is to ensure it is comparable across different links. From the perspective of statistical estimation, the normalized standard deviation is found larger in cases with few observations (e.g., 9:00–9:30 and 21:00–21:30 time intervals on weekends), as too little information is available to infer the model parameters. For time periods with higher number of observations, there is no obvious pattern for the normalized standard deviation of average link travel times. This reflects the fact that the variability of link travel times is largely dependent on corresponding traffic condition during the specific periods. Therefore, it is not very useful to introduce a specific transition model $P(\boldsymbol{\Sigma}^t|\boldsymbol{\Sigma}^{t-1})$ for the variance of average link travel times, which will also greatly increase computation complexity during model parameter estimation. By estimating the variance of the link travel times, we are able to capture the variability of short-term link travel times, and have more robust interpretation of urban traffic network conditions.

4.3. Validation

Because the ground truth data is not available in this research, we validate the result by evaluating the model predicted path travel times against observed average path travel times. Let $(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ be the estimated mean and variance for the link travel times for a given time interval. The predicted path travel time for observation i is thus estimated as the path travel time using the most likely path. Let $K = \max_k \{\pi_k^i(\boldsymbol{\mu}^t, \boldsymbol{\beta}, d_k), k = 1, 2, \ldots, |R^i|\}$, then the predicted average path travel time is given as:

$$\rho_i^t = g_K^i(\boldsymbol{x}^t) \sim N\left(g_K^i(\boldsymbol{\mu}^t), h_K^i\left(\boldsymbol{\Sigma}^t\right)\right) \quad (22)$$

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

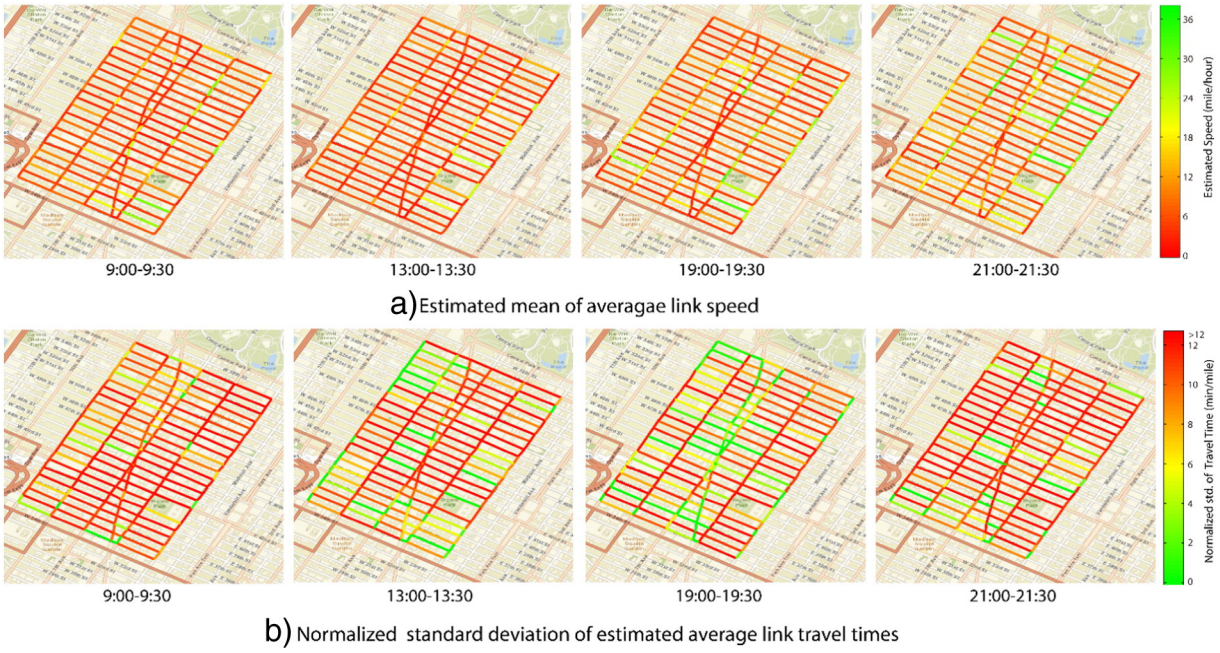

Fig. 6. Estimation results for Wednesday (2013/10/9).

with the mean and variance estimated as

$$E(\rho_i^t) = g_K^i(\boldsymbol{\mu}^t), \quad \mathrm{Var}(\rho_i^t) = h_K^i\left(\boldsymbol{\Sigma}^t\right). \tag{23}$$

The prediction accuracy is thus evaluated using the mean absolute percentage error (MAPE) of the observed path travel times against the predicted mean values:

$$\mathrm{MAPE} = \frac{1}{n^t}\sum_{i=1}^{N_{test}^t}\left|\frac{\rho_i^t - y_i^t}{y_i^t}\right| \times 100\% \tag{24}$$

where N_{test}^t is the size of the testing set in time interval t. Note that the predicted path travel time is computed using the average link travel times within a 30 min time interval, while the individual path travel time can be highly uncertain due to the stochastic intersection delay at a traffic signal. Thus certain level of discrepancy is expected between the observed path travel times and the predicted path travel times.

The validation results of numerical experiments for Monday, Wednesday, Friday and Saturday data are presented in Fig. 9. The results for the rest days of the tested week have shown similar accuracy levels thus are not presented to avoid repetition. From Fig. 9, it is observed that almost in all time intervals, the MAP (Bayesian) solution achieves lower MAPE value than the MLE solution. Furthermore, Fig. 8 presents the log likelihood value and MAPE for both MAP and MLE solutions of the estimated average link travel times. It can be seen that in almost all time intervals, the MLE solution achieves greater log likelihood (more likely according to observation model), however, the accuracy

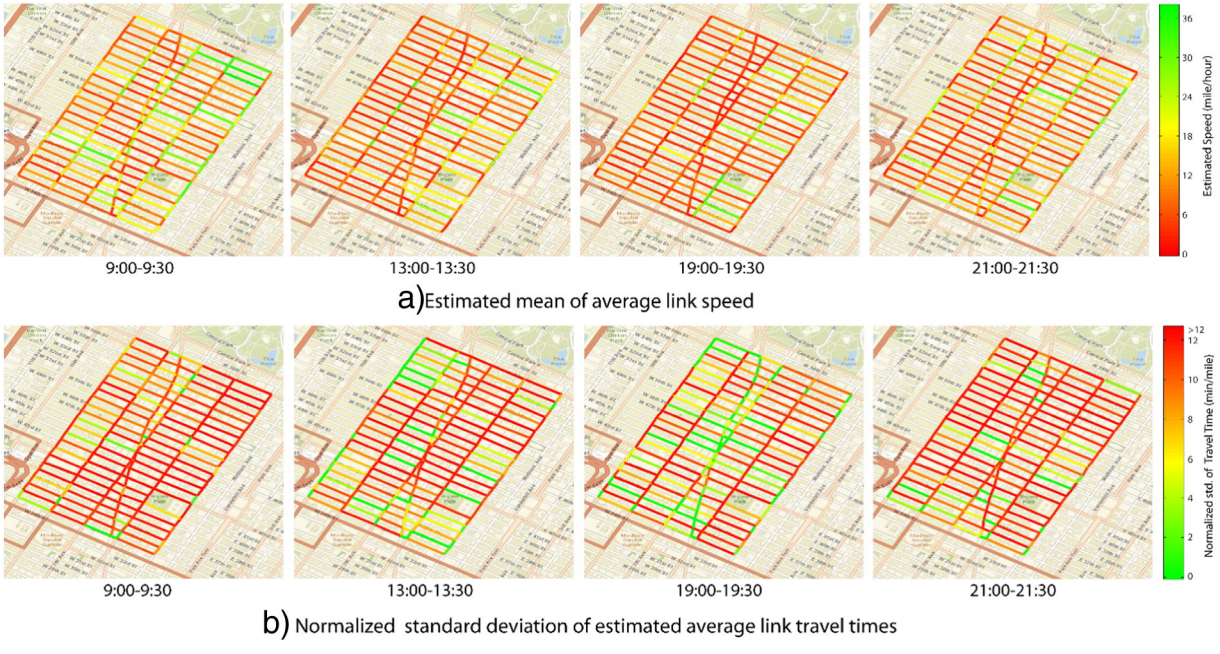

Fig. 7. Estimation results for Saturday (2013/10/12).

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

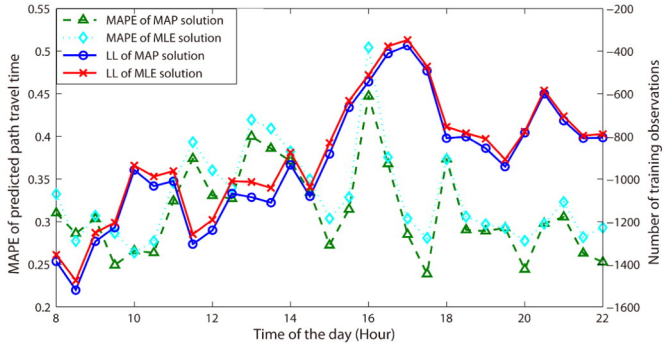

Fig. 8. Existence of overfitting in MLE solutions (results from Wednesday, 2013/10/9).

measured by MAPE is consistently worse than the MAP solution. This reveals the existence of overfitting to training data in MLE solution, which can cause biased and less accurate results. All these results show that incorporating historical information can produce more accurate and robust link travel time estimates. Consequently, in the following content, we will only discuss the results for MAP solutions.

Reasonable validation results are obtained from the MAP solutions of test scenarios. From Fig. 9, it is shown that the MAPE for most of the time intervals are ranging from 25% to 35%. The amount of observations used in parameter estimation is found to have great impact on the model accuracy. For example, the MAPE has peaks during 16:00–17:00 in all the presented four days, which at the same time, has relatively lower number of training observations. This result is intuitive, as the network conditions cannot be fully covered by the limited amount of observed trip level information. The level of congestion is also found to impact the model accuracy. For example, larger MAPE values (30–35%) are observed from more congested hours (e.g. 12:00–14:00), which is mainly due to the rapid change in traffic state during congestion within the 30 min time interval. On the other hand, for some less congested time periods (e.g., 20:00–22:00), the computed MAPEs are relatively low (25–30%). Result for Monday is an exception, where its MAPE is found to have a peak in 21:00–21:30 interval (37%). This might result from the relatively lower number of training observations (<500) available, whereas the same time interval on Wednesday and Friday have more than 600 training observations.

In the numerical results, only a small study region and 80% of all observed data within each time interval are used for training. In actual implementation of the proposed model, a larger study region with all the observed data can be used, which can significantly increase the number of observations available for model estimation and potentially lead to more accurate short-term average link travel time estimates.

5. Conclusions

This study develops a Bayesian mixture model to estimate short-term average link travel times using large-scale trip-based data. The model only needs partial information provided in the data, in this case, the origin and destination location, trip travel time and distance. The path taken by the taxi is considered as latent and inferred using a multinomial logit distribution. The observation model of the trip data given the reasonable path set and the mean and variance of the average link travel times can be then characterized using a finite mixture distribution. Also, a transition model that serves as the prior distribution is introduced to incorporate temporal correlation from historical estimation results and the spatial dependencies among neighboring links. Finally, a solution approach based on the EM algorithm is proposed to efficiently solve the problem. More robust estimation results are obtained owing to the adoption of the online Bayesian framework.

Currently, the model is validated through comparing the observed trip travel times and the predicted trip travel times. Future research can be done to verify the proposed model by comparing the estimates

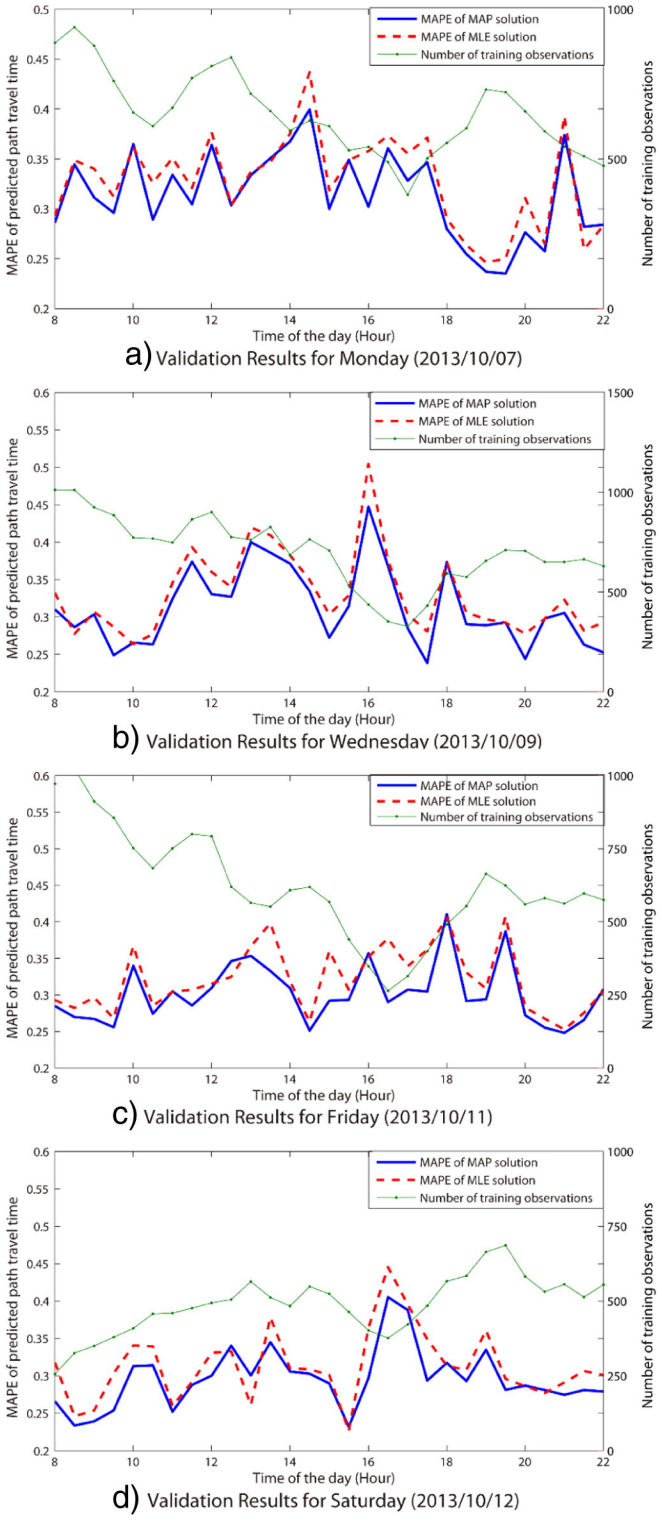

Fig. 9. Validation results for four representative days in the tested week.

either against link speed data from loop detectors, or detailed trajectory information of taxi trips, which is collected by NYCTLC, but currently not available to researchers. It should be noted that the proposed model is also applicable to trajectory data, since the intermediate trajectory points can be treated as the origin and destination pairs in the model. Using such more detailed data, the accuracy of the link travel time estimates can be greatly improved.

The proposed model provides a very flexible probabilistic framework and several extensions can be made to further improve the

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007

accuracy of the estimation. Using a more realistic link travel time distribution will greatly relax the restrictive assumption of the normal distributed average link travel times. Recent literature has shown that the link travel times are more likely to follow a bimodal or lognormal distribution due to the involvement of signal delay. Incorporating such classes of realistic distribution can be helpful to better account for impact of signal delays. Also, in this work we only use a simple prior distribution for the mean of average link travel times, more comprehensive prior distributions can be used to provide more accurate and additional prior information to the model. For example, introducing prior for both $\boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^t$, or considering heterogeneous spatial and temporal dependencies for different links. Moreover, the model proposed in this paper mainly focuses on taxi trip data, hence the path cost function in the route choice model mainly reflect the special route choice behavior for taxi drivers. When generalizing to other trip-based data (e.g., LPR data from a sparse sensor network), the path cost function can be modified accordingly based on the route choice behaviors reflected in the data. These further extensions of the proposed model would lead to more accurate and robust link travel time estimation for urban traffic operation and management and fully utilize the abundant large-scale trip-based data available in big cities.

Acknowledgments

The authors would like to thank the constructive comments of the two reviewers. This work is partly funded by the US National Science Foundation grants 1017933 and 1520338 for which the first two authors are grateful.

References

[1] Beijing Municipal Committee, The twelfth five-year plan for the transportation development of BeijingRetrieved from http://zhengwu.beijing.gov.cn/ghxx/sewgh/t1237237.htm July 2012.
[2] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[3] B. Coifman, Vehicle reidentification and travel time measurement on congested freeways, Transp. Res. A Policy Pract. 36 (10) (2002) 899–917.
[4] J.F. De Freitas, M. Niranjan, A. Gee, A. Doucet, Sequential Monte Carlo methods to train neural network models, Neural Comput. 12 (4) (2000) 955–993.
[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Methodol. (1977) 1–38.
[7] Government of Hong Kong, Transport — Hong Kong: the Facts, 2013.
[8] J.C. Herrera, D. Work, X. Ban, R. Herring, Q. Jacobson, A.M. Bayen, Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment, Transp. Res. C Emerg. Technol. 18 (2010) 568–583.
[9] R. Herring, A. Hofleitner, P. Abbeel, Estimating Arterial Traffic Conditions Using Sparse Probe Data, Proceedings of the ITS (September, 2010) 19–22.
[10] T. Hunter, R. Herring, P. Abbeel, Path and Travel Time Inference from GPS Probe Vehicle Data, Neural Information Processing Systems Foundation (NIPS) (December 2009) Vancouver, Canada.
[13] T. Kurihara, Y. Nakada, K. Yosui, T. Matsumoto, Bayesian On-line Learning: A Sequential Monte Carlo with Importance Resampling, Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop 2001, pp. 163–172.
[14] K. Kwong, R. Kavaler, R. Rajagopal, P. Varaiya, Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors, Transp. Res. Part C Emerg. Technol. 17 (2009) 586–606.
[15] R. Li, G. Rose, Incorporating uncertainty into short-term travel time predictions, Transp. Res. Part C Emerg. Technol. 19 (6) (2011) 1006–1018.
[16] M. Lu, W. Chen, X. Shen, H.C. Lam, J. Liu, Positioning and tracking construction vehicles in highly dense urban areas and building construction sites, Autom. Constr. 16 (2007) 647–656.
[17] NYCTLC, New York City Taxi and Limousine Commission 2012 Annual Report, 2012.
[18] Oh, J.S., Jayakrishnan, R., Recker, W., 2003. Section Travel Time Estimation from Point Detection Data. 82nd Annual Meeting of Transportation Research Board, Washington, DC, USA.
[19] D. Park, L.R. Rilett, Forecasting multiple-period freeway link travel times using modular neural networks, J. Transp. Res. Board (98) (1998) 163–170.
[20] H. Rakha, I. El-Shawarby, M. Arafeh, F. Dion, Estimating Path Travel-time Reliability, Proceedings of the IEEE Intelligent Transportation Systems Conference 2006, Canada, Toronto, 2006.
[21] H.D. Sherali, J. Desai, H. Rakha, A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times, Transp. Res. B Methodol. 40 (10) (2006) 857–871.
[22] C.-H. Wu, J.-M. Ho, D.T. Lee, Travel-time prediction with support vector regression, IEEE Trans. Intell. Transp. Syst. 5 (4) (2004) 276–281.
[23] J.Y. Yen, Finding the K shortest loopless paths in a network, Manag. Sci. 17 (1971) 712–716.
[24] J. Yeon, L. Elefteriadou, S. Lawphongpanich, Travel time estimation on a freeway using discrete time Markov chains, Transp. Res. B Methodol. 42 (4) (2008) 325–338.
[25] X. Zhan, S. Hasan, S.V. Ukkusuri, C. Kamga, Urban link travel time estimation using large-scale taxi data with partial information, Transp. Res. Part C Emerg. Technol. 33 (2013) 37–49.
[26] X. Zhan, R. Li, S.V. Ukkusuri, Lane-based real time queue length estimation using license plate recognition data, Transp. Res. Part C Emerg. Technol. 57 (2015) 85–102.
[27] X. Zhang, J. Rice, Short-term travel time prediction, Transp. Res. Part C Emerg. Technol. 11 (3–4) (2003) 187–210.
[28] F. Zheng, H. Van Zuylen, Uncertainty and predictability of urban link travel time, Transp. Res. Rec. J. Transp. Res. Board 2192 (2010) 136–146.
[29] F. Zheng, H. Van Zuylen, Urban link travel time estimation based on sparse probe vehicle data, Transp. Res. Part C Emerg. Technol. 31 (2013) 145–157.
[30] H. Zhu, M. Li, Y. Zhu, L.M. Ni, HERO: online real-time vehicle tracking, IEEE Trans. Parallel and Distrib. Syst. 20 (2009) 740–752.

Please cite this article as: X. Zhan, et al., A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction (2015), http://dx.doi.org/10.1016/j.autcon.2015.12.007