Journal of Intelligent Transportation Systems

Technology, Planning, and Operations

ISSN: 1547-2450 (Print) 1547-2442 (Online) Journal homepage: https://www.tandfonline.com/loi/gits20

# A novel arterial travel time distribution estimation model and its application to energy/emissions estimation

Qichi Yang, Guoyuan Wu, Kanok Boriboonsomsin & Matthew Barth

**To cite this article:** Qichi Yang, Guoyuan Wu, Kanok Boriboonsomsin & Matthew Barth (2018) A novel arterial travel time distribution estimation model and its application to energy/emissions estimation, Journal of Intelligent Transportation Systems, 22:4, 325-337, DOI: 10.1080/15472450.2017.1365606

**To link to this article:** https://doi.org/10.1080/15472450.2017.1365606

Accepted author version posted online: 14 Aug 2017.
Published online: 16 Nov 2017.

JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS, 2018
VOL. 22, NO. 4, 325–337
https://doi.org/10.1080/15472450.2017.1365606

# A novel arterial travel time distribution estimation model and its application to energy/emissions estimation

Qichi Yang[a], Guoyuan Wu[b], Kanok Boriboonsomsin[b], and Matthew Barth[b]

[a]Google Inc., Mountain View, CA, USA; [b]Centre for Environmental Research and Technology, University of California at Riverside, Riverside, CA, USA

**ABSTRACT**

Arterial travel time information is crucial to advanced traffic management systems and advanced traveler information systems. An effective way to represent this information is the estimation of travel time distribution. In this paper, we develop a modified Gaussian mixture model in order to estimate link travel time distributions along arterial with signalized intersections. The proposed model is applicable to traffic data from either fixed-location sensors or mobile sensors. The model performance is validated using real-world traffic data (more than 1,400 vehicles) collected by the wireless magnetic sensors and digital image recognition in the field. The proposed model shows high potential (i.e., the correction rate are above 0.9) to satisfactorily estimate travel time statistics and classify vehicle stop versus non-stop movements. In addition, the resultant movement classification application can significantly improve the estimation of traffic-related energy and emissions along arterial.

**ARTICLE HISTORY**
Received 4 November 2015
Revised 6 August 2017
Accepted 6 August 2017

**KEYWORDS**
arterial travel time distribution; energy and emissions; Gaussian mixture model (GMM)

## Introduction

Estimating arterial travel time is crucial to the development and application of both advanced traffic management systems (ATMS) and advanced traveller information systems (ATIS) which rely on real-time traffic information to make better decisions, such as traffic signal control (Pandit et al., 2013) and dynamic vehicle routing (Feijer, Savla, & Frazzoli, 2012). Compared with the case of freeway segments, estimating link travel times along arterials is much more challenging because traffic conditions are more complicated and vehicles' movements can be often interrupted by control devices (e.g., traffic signals) and other disturbances, such as random mid-block crossings. More challenging issues on understanding the travel time variability may raise from the "close-loop" effects from the travellers' (re)scheduling behavior as a response (Noland & Polak, 2002). On the other hand, link travel time distributions (TTDs) can better represent the stochastic properties of traffic states along signalized corridors (Hofleitner, 2013); therefore, its modeling has been attracting significant research interest (Chen, Sun, & Qi, 2017; Sanaullah, Quddus, & Enoch, 2016).

Most of the TTD modeling algorithms can be categorized into either model-based or data-driven. The former category is built on top of physical principles, such as hydrodynamics and queuing theory (Olszewski, 1994). Derivation of these algorithms is usually accompanied with some fundamental assumptions, e.g., no lane-changing and/or known arrival distribution(s), which restrict their practicality. The data-driven algorithms, however, largely rely on fitting a significant amount of data with well-established probability distributions (Hofleitner, Herring, & Bayen, 2012a; Uno, Kurauchi, Tamura, & Iida, 2009; Zheng & van Zuylen, 2010). The major issue is the availability of enough data for model training. Although a few studies have proposed hybrid strategies to train parameters of the physical model(s) using machine learning techniques (Hofleitner, Herring, Abbeel, & Bayen, 2012b), these strategies can only be applied to one type of data sources: either fixed-location sensors (e.g., wireless magnetic sensors) or mobile sensors (e.g., probe vehicles) just like most of the data-driven algorithms.

In this paper, we modify the generic Gaussian mixture model (GMM) (McLachlan, 1988) to estimate TTDs along arterials. The proposed modified Gaussian mixture model (MGMM) can be trained very efficiently using real-world data from either fixed-location sensors or mobile sensors as well as roadway geometric features (e.g., intersection spacing) that can be extracted from most of the existing geographic information systems (GIS). The resultant good fitting of travel time probability density functions and high accuracy rate of vehicle

**CONTACT** Guoyuan Wu gywu@cert.ucr.edu Centre for Environmental Research and Technology, University of California at Riverside, Riverside, CA 92507, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/gits.

movement (stop versus non-stop) classification further validate the proposed MGMM. In addition, application of our algorithm to some simulation studies exhibits great potential to improve estimation of arterial traffic energy consumption and emissions, compared to existing methods, e.g., Yang, Boriboonsomsin, & Barth (2011).

The remainder of this paper is organized as follows: The "Background" section presents some background information for developing the MGMM-based arterial link TTD estimation algorithm. The "Modeling approach" section details the proposed model in the cases of both fixed-location sensors and mobile sensors, followed by the validation in the "Validation by field data" section using real-world measurements. The field data of fixed-location sensors are obtained from the Sensys wireless magnetic sensors network and have been pre-processed by the algorithm in Kwong, Kavaler, Rajagopal, & Varaiya (2009). The data of mobile sensors are emulated from the measurements archived in the Next-Generation Simulation (NGSIM) Program (FHWA, 2006). The "Application to energy/emissions estimation" section describes an application of the proposed model to the traffic energy consumption and emissions estimation along arterials. The last section concludes this paper with recommendations on future research.

## Background

### Link travel time measurement

Within the framework of measuring/estimating arterial travel times, a link is usually defined as the segment between two signalized intersections (including either the upstream or downstream intersection), as depicted in Figure 1. In this paper, we consider a link as the segment including the downstream intersection (i.e., between two fixed-location sensors).

As aforementioned, although advanced traffic surveillance techniques now offer a variety of ways to measure the arterial travel time, most of such data sources can be divided into two categories: fixed-location sensors and mobile sensors. Fixed-location sensors such as wireless magnetic sensors (Kwong et al., 2009), micro-loop detectors (Ndoye, Totten, Krogmeier, & Bullock, 2011), bluetooth (Araghi, Krishnan, & Lahrmann, 2016; Wasson, Sturdevant, & Bullock, 2008), and radio frequency identification (RFID), are usually installed right after intersections in order to capture any delay due to queuing effects. The link travel time, which is defined as the temporal difference between the passages of two consecutive sensors, can be measured by vehicle re-identification via the unique signature.

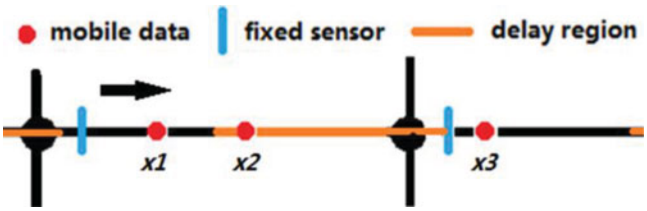

Figure 1. A typical layout of an arterial (one way) with signalized intersections and sensors is illustrated. Fixed-location sensors are usually installed right after the intersections. Red spots represent data sampling locations by mobile sensors. The black arrow points out the traffic flow direction. The orange segment represents a delay region where queues usually occur.

Mobile sensors, such as probe vehicles equipped with global positioning systems (GPS), can report second-by-second position information. By applying virtual detection lines whose locations can be chosen the same as those of fixed-location sensors (e.g., the vertical "blue bars" in Figure 1), we can estimate the link travel times with ease. Due to the rapid proliferation of smart-phone users, data streaming from GPS-enabled mobile devices has become a more and more important source of travel time measurements. However, most of the mobile sensor data for real-time traffic information collection (especially for commercial use) today are sampled at a relatively low frequency (ranging from 5 to 60 seconds in the sampling interval (BeattheTraffic, 2014)) due to the high costs in data storage and transmission. Such sparseness issue may bring about some challenge for link travel time measurement because the spatial coverage of two consecutive samples is longer than a link. In this case, a heuristic remedy is to estimate the link travel time by interpolation (based on link length and travel distance). On the other hand, if some data samples fall into the so-called intersection delay region (orange segment in Figure 1 where queues usually occur). The estimation of link travel time using such samples may be biased because the intersection delay cannot be fully captured. In this study, we filter out such data samples by applying some predefined speed threshold(s).

Besides the aforementioned data sources, the NGSIM dataset is another valuable source which is a particular application of digital image recognition and contains detailed trajectories (every tenth of seconds) of all vehicles traveling along certain roadway segments (around 1.5 miles in length) during one or more 15-min time intervals. We will use this dataset to generate the mobile sensor like data for model validation.

### Existing models to estimate link travel time distribution

With a variety of data sources, numerous studies have been conducted to estimate the arterial link TTD. Guo, Rakha, and Park (2010) proposed a GMM to estimate link travel time and the traffic state transitions between links. Using the NGSIM dataset, Ramezani and Geroliminis (2012) studied the joint TTDs on consecutive links, which can be used to improve route level travel time estimation/prediction. However, the methods mentioned above are only applicable to the data collected from fixed-location sensors. For mobile sensor based data whose sampling locations may vary, some assumptions have to be relaxed and a more generalized framework for parameter fitting needs to be developed.

Recent success on estimating TTD based on sparse mobile sensor data is archived in Hofleitner et al. (2012a, 2012b). The authors developed mixture models to capture dual-mode TTDs along signalized intersections. As described in Hofleitner et al. (2012b), for example, the delayed time was modeled as a mixture of a Dirac Delta distribution and a uniform distribution under the assumption of uniform arrival pattern of the upstream traffic. A close form of arterial link TTD was then developed, which turned out to be a mixture of a Gaussian and a quasi-uniform distribution, given the fact that the probability distribution of the sum of two independent random variables is the convolution of their individual distributions. Such simplification is of significant meaning in theory, but the real-world situation is so complicated that some strong assumptions (e.g., uniform arrival) do not hold true. In addition, when a large dataset needs to be fitted to such models, the computational load becomes overwhelming due to the presence of double integrals in the probability density function.

To overcome the aforementioned issues, we propose herein a MGMM to estimate TTD. The proposed model is not only able to handle data from either fixed-location sensors or mobile sensors, but also is more flexible (no constraints on arrival distribution) and more computationally attractive.

## Modeling approach

### Travel time decomposition

As aforementioned, there are two major types of sensors for travel time data collection. Without loss of generality, we assume that the data from both types of sensors have the same format ($tt_{x1,x2}$, $l_{x1,x2}$), where $tt_{x1,x2}$ is the travel time measurement between locations $x1$ and $x2$, while $l_{x1,x2}$ is the travel distance measurement associated with $x1$ and $x2$. For the case with fixed-location sensors, $x1$ and $x2$ represent the locations of a pair of sensors. Thus, $tt_{x1,x2}$ and $l_{x1,x2}$ (constant) are the passage time and distance, respectively, between these two sensors. For the case with mobile sensors, however, $x1$ and $x2$ are the sampling locations of two data points. Therefore, $tt_{x1,x2}$(constant) is simply the sampling interval, and $l_{x1,x2}$ is the distance between two samples.

In addition to the predetermined factors such as the speed limit, the variability in travel time along arterial links may result from: (1) the variation of individual driver's behavior and (2) the delayed time due to traffic signals or queue dissipation. All these disturbances make it difficult for arterial travel time to be modeled by a single distribution. Therefore, it is desirable to decompose arterial travel time into two or more components including the free-flow travel time and delayed time, each of which can be modeled by different single distribution and/or mixture distributions. In this paper, we formulate the travel time, $tt_{x1,x2}$ as

$$tt_{x1,x2} = tt_{ff\,x1,x2} + d_{x1,x2} \tag{1}$$

where $tt_{ff\,x1,x2}$ represents the free-flow travel time component, while $d_{x1,x2}$ is the delayed time component. As presented in the next subsection, these two (random) variables can be estimated from the field data.

### Modified Gaussian mixture model

In this subsection, we will show how to estimate $tt_{ff\,x1,x2}$ and $d_{x1,x2}$ using a MGMM. First, we define the free-flow travel time between $x1$ and $x2$ as

$$tt_{ff\,x1,x2} \triangleq l_{x1,x2} * p_{ff} \tag{2}$$

where $p_{ff}$ is the so-called free-flow pace (in sec/meter), which is the inverse of free-flow speed. It may be modeled by a variety of distributions (e.g., Gaussian or Gamma distributions (Hofleitner, Herring, & Bayen, 2011)). In this study, we model $p_{ff}$ as a Gaussian random variable, which can be justified to some degree by the histogram in Figure 2.

$$p_{ff} \sim N\left(\mu_{ff},\ \sigma_{ff}^2\right) \tag{3}$$

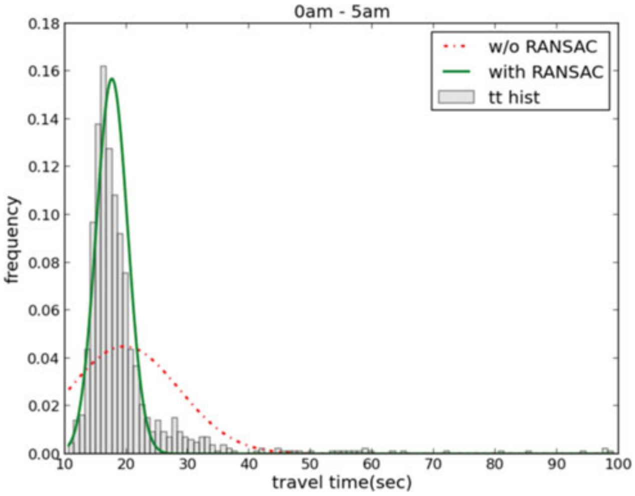

Figure 2. Link travel time histogram 0 a.m–5 a.m. over one month, Aug 2011, from Washington Ave (upstream) to Solano Ave (downstream) along San Pablo Ave, Berkeley, California.

where $\mu_{ff}$ and $\sigma_{ff}$ are the associated mean and standard deviation, respectively, and can be estimated from the data (see the "Initialization" section under the "Modeling approach" section). By substituting Eq. (3) into (2), we can get

$$tt_{ff\,x1,x2} \overset{\Delta}{=} l_{x1,x2} * p_{ff} \sim N(\mu_{ff} \cdot l_{x1,x2}, \sigma_{ff}^2 \cdot l_{x1,x2}^2) \quad (4)$$

As to the distribution of delayed time, $d_{x1,x2}$, we select a mixture of Gaussians (Bishop & Nasrabadi, 2006) to model it,

$$d_{x1,x2} \sim \sum_{k=1}^{K} \pi_k N\left(\mu_k,\ \sigma_k^2\right) \quad (5)$$

$$\mu_1 = \sigma_1 = 0$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Notice that we keep the first component to be $N(0,0)$, which corresponds to the movements without delay and with zero mean and variance during the model fitting process. $\pi_k$ is the weight of $k$th component and must sum to 1.

Since the total travel time is the sum of $tt_{ff\,x1,x2}$ and $d_{x1,x2}$, its probability is the convolution of these two. It is noted that the convolution of two Gaussians is still a Gaussian with the addition of mean and variance, i.e.,

$$\begin{aligned} P(tt_{x1,x2}) &= P(tt_{ff\,x1,x2}) * P(d) \\ &= \sum_{k=1}^{K} \pi_k N(tt_{x1,x2} \mid \mu_k + \mu_{ff} \cdot l_{x1,x2},\ \sigma_k^2 + \sigma_{ff}^2 \cdot l_{x1,x2}^2). \end{aligned} \quad (6)$$

It is well known that a GMM has multiple solutions for the same training dataset (known as identifiability problem (Bishop & Nasrabadi, 2006)), which a challenge in interpreting each component after the model is fitted. As observed in Eq. (6), however, $tt_{x1,x2}$ is a mixture of Gaussians in which the mean and variance of the first component (free-flow) are coupled with other components. This helps us identify the free-flow component with ease, because it is just the one with the smallest mean.

To solve Eq.(6), we followed the *Expectation-Maximization* (EM) Algorithm (Bishop & Nasrabadi, 2006), which is an iterative technique to obtain the maximum likelihood estimates (MLE) of parameters of a stochastic system with latent variables. The subject parameters are initialized and iteratively updated using an *expectation* (E) step and a *maximization* (M) step until the estimates between two consecutive iterations converge. More specifically, the expected value of the log-likelihood function is evaluated using the up-to-date estimates for the parameters in the current E-step, and the M-step is performed to calculate the best estimates of parameters that maximize the expected log-likelihood. These estimates are then applied to determine the distributions of the latent variables in the next E step. In the following subsections, we will introduce some special treatment in the initialization of the parameters, followed by the presentation of solution methods for both fixed and mobile sensor data.

### *Initialization*

Notice that the EM algorithm can only guarantee local optima. Therefore, the selection of initial parameter set is important to the system performance. In particular, since the free-flow component is coupled with others in the proposed GMM, the initialization of $\mu_{ff}$ and $\sigma_{ff}$ becomes even more critical. Field observations reveal that the traffic volume is often quite low in the early morning period (e.g., from 1 a.m. to 5 a.m.) and most vehicles are traveling at the free-flow speed. Therefore, we use data samples collected during this time period to identify the TTD under free-flow conditions. Figure 2 presents an example histogram of vehicle travel times collected from a pair of fixed-location sensors during the period from 0 a.m. to 5 a.m.

As shown in the figure, most of the travel time data samples fall into the region with short travel times (to the left of the distribution), while only a few of them have long delays (to the right). The dashed curve represents the results fitted with Gaussian distribution. However, the results are not satisfactory due to the delay impacts. To mitigate such effects, we apply the RANSAC (RANdom SAmple Consensus) technique (solid curve) (Fischler & Bolles, 1981) which is an iterative and robust method for model fitting or parameter estimation on a dataset with outliers, when fitting the Gaussian distributions for deriving the initial values of $\mu_{ff}$ and $\sigma_{ff}$. To apply the RANSAC technique to the mobile sensor data where $l_{x1,x2}$ may vary greatly, we assume the vehicle is traveling at free-flow speed, and apply the free-flow pace to the link length.

The initialization of $\mu_k$ and $\sigma_k$, however, can be conducted by randomly sampling from the right side of the distribution, i.e., between $\mu_{ff} \cdot l_{x1,x2} + \alpha \cdot \sigma_{ff} \cdot l_{x1,x2}$ and $\max(tt_n)$. We chose $\alpha = 3$ in this study.

Another critical parameter that needs to be initialized is the number of components, $K$, in the model. In Feng (2011), the author suggested four components, representing "free-flow," "slow free-flow," "fast delayed," and "delayed" samples. Although such segregation is not rigorous in theory, we use it as a reference in this study. Furthermore, we conduct sensitivity analysis on this parameter (including $K = 2$, 3, 4, and 5) and the results are presented in the following sections.

### Solving modified GMM

The proposed MGMM can be solved by maximizing the total log likelihood:

$$\underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k N\left(tt_n \mid \mu_{nk}, \sigma_{nk}^2\right) \quad (7)$$

$$\mu_{nk} = \mu_k + \mu_{ff} \cdot l_n$$

$$\sigma_{nk}^2 = \sigma_k^2 + \sigma_{ff}^2 \cdot l_n^2$$

$$\theta = (\mu_{ff}, \sigma_{ff},\ \mu_k, \sigma_k, \pi_k)$$

where $\theta$ is the parameter vector to be estimated; $N$ is the sample size; $l_n$ is the travel distance of $n$th data sample; and $\mu_{nk}$ and $\sigma_{nk}$ are the mean and standard deviation of $k$th Gaussian component associated with $n$th data sample. Notice that the measured travel distance may vary with data samples (e.g., from mobile sensors). In other words, each data sample may have its unique probability distribution, resulting in $n \cdot k$ pairs of different means, $\mu_{nk}$'s, and variance, $\sigma_{nk}^2$'s. Fortunately, according to Eq. (7), the calculation of $\mu_{nk}$'s and $\sigma_{nk}^2$'s (in total, $2 \cdot n \cdot k$ parameters) only depends on $(2 + 2 \cdot k + n)$ parameters, i.e., $\mu_{ff}, \sigma_{ff}^2, \mu_k, \sigma_k^2$, and $l_n$. This significantly reduces the computational load when applying to large datasets. Next, we will show the solutions to this model by differentiating the data sources (i.e., fixed-location sensors versus mobile sensors).

### Data from fixed-location sensors

As aforementioned, if data come from fixed-location sensors, $l_n$'s of all the samples are identical. They are simply the distance between the upstream and downstream sensors, or the length of the link, $l$. This implies that the mean and variance of each component in the mixture of Gaussians should be the same for each sample. Let the common mean and variance be $\mu_{lk}$ and $\sigma_{lk}^2$, respectively. Then, our model can be solved in the same way as a generic GMM using the EM algorithm, where each parameter has a closed-form solution:

$$\hat{\mu}_{lk} = \hat{\mu}_k + \hat{\mu}_{ff} \cdot l = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \cdot tt_n \quad (8)$$

$$\hat{\sigma}_{lk}^2 = \hat{\sigma}_k^2 + \hat{\sigma}_{ff}^2 \cdot l = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \cdot \left(tt_n - \hat{\mu}_{lk}\right)^2 \quad (9)$$

$$\hat{\pi}_k = N_k / N \quad (10)$$

where $\gamma_{nk}$ is the probability of sample $n$ generated by component $k$, which is also referred to as the responsibility of $k$th component to $n$th sample; and $N_k$ is the effective number corresponding to component $k$.

### Data from mobile sensors

Unlike the data from fixed-location sensors, each mobile sensor data sample has its own distance, $l_n$, which leads to $n \cdot k$ different pairs of Gaussian parameters. If we follow the same EM algorithm as the generic GMM, then the responsibility in the E-step will be written as

$$\gamma_{nk} = \frac{\pi_k N(tt_n | \mu_{nk}, \sigma_{nk}^2)}{\sum_{j=1}^{K} \pi_j N(tt_n | \mu_{nj}, \sigma_{nj}^2)} \quad (11)$$

Therefore, the objective in the M-step turns out to be maximizing the posterior total log likelihood given the responsibility $\gamma_{nk}$:

$$\log L(\theta | tt_n, l_n, \gamma_{nk}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \quad \log\pi_k + \log\left(N\left(tt_n \mid \mu_{nk}, \sigma_{nk}^2\right)\right)\} \quad (12)$$

Because of the existence of $l_n$ in $\mu_{kn}$ and $\sigma_{kn}$, the closed form solution cannot be obtained for $\mu_{ff}, \sigma_{ff},\ \mu_k, \sigma_k$. Instead, we apply the gradient descent technique (Byrd, Lu, Nocedal, & Zhu, 1995) to tackle this issue. Also, notice that the cost function can be chosen from one of the following perspectives:

(a) Use the total log likelihood as in Eq. (7);
(b) Keep the framework of the EM algorithm by finishing the E-step and deriving $\pi_k$ at the beginning of the M-step. Then set Eq. (12) as the cost function to estimate the rest of the parameters;
(c) Similar to (b), but optimize Eq. (7) at the last step instead.

It turns out that compared to (c), the selection of (a) or (b) is less reliable and the resultant problem is prone to be singular, because one or more Gaussian components may collapse into a single data sample, giving rise to zero variance during the optimization process. Therefore, we choose (c) as the cost function when handling the mobile sensor data, and apply the L-BFGS-B algorithm (Byrd et al., 1995) to the optimization process.

## Validation by field data

### Field experiment setup and data description

In this study, the fixed-location sensor data were collected from a wireless traffic sensor network (Sensys Networks, 2015) installed on the segment from Washington Ave (upstream) to Solano Ave (downstream) along San Pablo Ave, Berkeley, California. The segment length is around 0.3 miles. Five to seven magnetic sensors were installed as an array 12 feet after each intersection (recommended by Sensys Networks), recording the magnetic signature and timestamp of vehicles passing over them. The vehicles can be re-identified by comparing the peak features in the resultant signatures from two consecutive sensor arrays and taking into account the sequence of vehicles in the queue. According to Kwong et al. (2009), it is too ambitious to obtain 100% matching rate (around 70% in this study), due to the shifted relative positions between vehicles and sensors when vehicles pass the upstream and downstream sensor arrays. Although the original dataset covers entire two months from Jul. 15th to Sep. 15th, 2011, the data between 6 a.m. and 9 p.m. on weekdays were selected and divided into 15 groups (15 h, each over 5 weekdays). The statistical travel time pattern of each hour over weeks were learned separately. Notice that the signal control is actuated in this study and it is overwhelming for us to get detailed status. However, as can be observed in the following, its impacts on the TTD estimation have been implicitly (well) modeled by the proposed MGMM.

As aforementioned, an intensive set of probe data are usually not accessible. Therefore, we use the NGSIM Lankershim Blvd. dataset (from 8:30 a.m. to 8:45 a.m.) (FHWA, 2006), and emulate the data collection from mobile sensors by sampling the vehicles of interest every 10 seconds. In addition, we test the model performance under different penetration rates of probe vehicles, i.e., 10%, 50%, and 100%. Figure 3 shows the schematics of study area.

In the following subsections, we will focus on the presentation and analyses of results from fixed-location sensor data in the "Results on link travel time distribution estimation" section to "Vehicle stop versus non-stop movement classification" section given under the "Validation by field data" section. For the mobile sensor case, results will be illustrated in the "Validation through sampled data from the NGSIM dataset" section, and the sampling interval issue will be further investigated.

### Results on link travel time distribution estimation

By applying the proposed MGMM approach in the "Modeling approach" section, we estimate the TTDs over the aforementioned field data. Figure 4 shows examples of such distributions under different congestion levels over a typical week (from Sep. 5th to Sep. 9th, 2011).

It can be observed that if the traffic volume is high (bottom subplot), then link TTD will exhibit multi-mode properties. The left-most mode (i.e., the portion fitted by the green solid curve) represents the free-flow data samples, while others may result from the impacts of downstream signal timings and/or upstream arrival pattern. When the traffic condition is uncongested (top subplot), the free-flow mode of TTD stands out. The long and heavy tail (on the right) represents other data samples with various intersection delays.

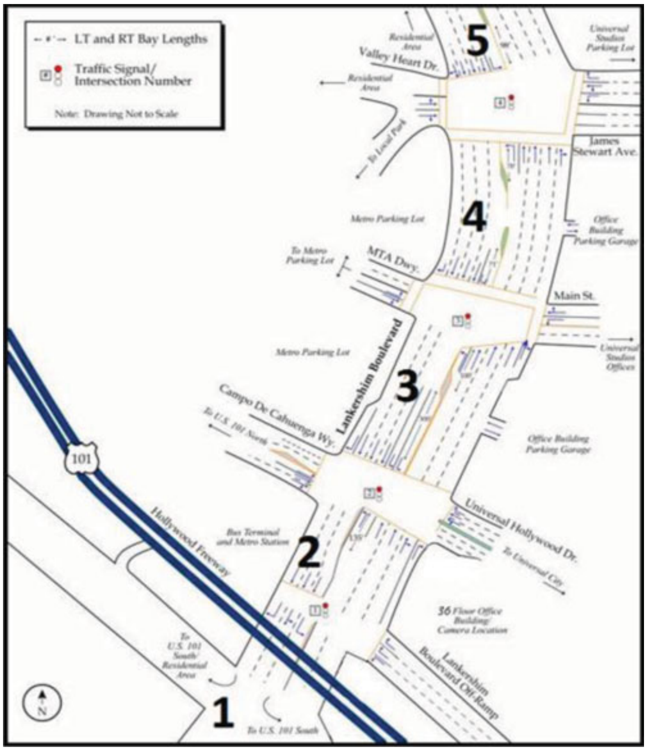

Figure 3. Map of Lankershim Blvd. archived in NGSIM source. Testing is focused on segment 3 for (both Northbound and Southbound traffic).

### Model's goodness-of-fit evaluation

In this subsection, we apply the Kolmogorov–Smirnov (K–S) test (Corder & Foreman, 2009) to evaluate the goodness-of-fit of our MGMM on the empirical TTD. The K–S test is a non-parametric test of the equality of continuous, one-dimensional probability distributions based on K–S statistic. Upon selection of significance level, $\alpha(=0.10$ in this study), the null hypothesis, $H_0$: the data follow the specified distribution, is rejected if the K–S statistic is larger than the critical value associated with that significance level, or if the corresponding $p$-value is smaller than the significance level. The smaller the $p$-value is, the less confidence in the similarity between the empirical distribution and the test distribution. In this study, we use the $p$-value as the measure of goodness-of-fit of our models.

Figure 5 shows the test results on fixed-location sensor data where the number of components varies from 2 to 5. As shown in the figure, the mixture model with two components has very small $p$-values (less than the

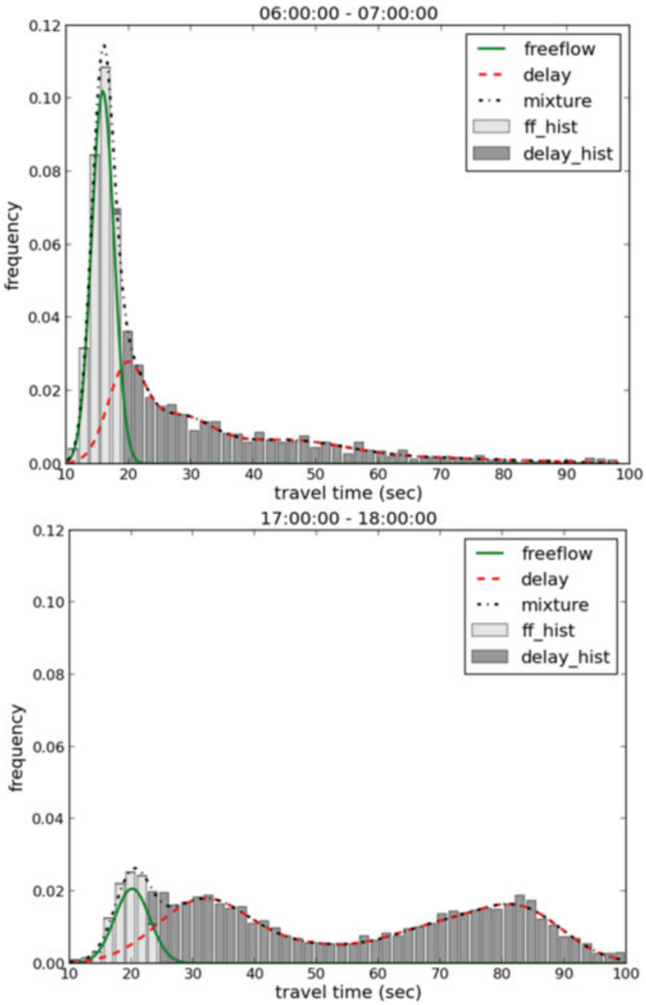

Figure 4. TTD estimation under different congestion levels. The mixture distributions with free flow and delay components result from the fixed-location sensor data during the off-peak period 6 a.m.–7 a.m. (top) and peak period 5 p.m.–6 p.m. (bottom) over the week of Sep. 5th–Sep. 9th, 2011.

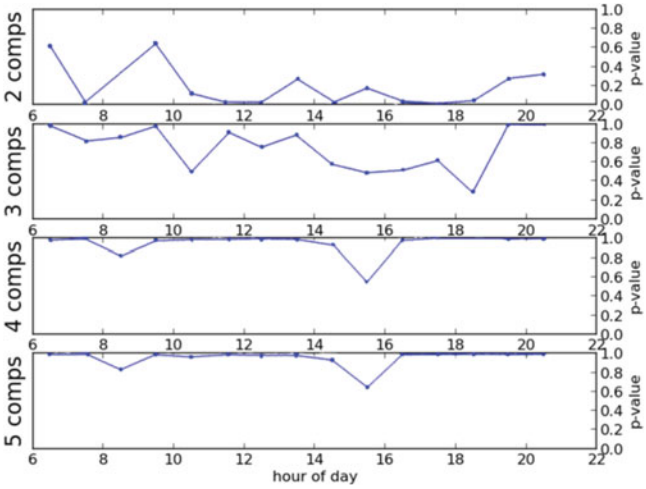

Figure 5. The $p$-values from K–S tests. The models are trained on one week data between Sep. 5th and Sep. 9th, 2011. The model with four components is preferable in this case. Note: the threshold of significance level is 0.10.

selected threshold of significance level, 0.10) in most of the time, implying poor fitting of the data. Although all the $p$-values of the mixture model with three components are above the significance level over the entire period, the mixture model with four components exhibit much better performance (especially during the periods of around 10 a.m.–2 p.m. and 5 p.m.–7 p.m.). It can be also observed that there is no obvious improvement when the number of components is increased from four to five. It should be pointed out that a GMM with a larger number of components generally provides better probability density estimation, but it may lead to more computational load (for gradient descent) and higher risk of over-fitting (i.e., lack of generalization ability) Therefore, the above results indicate that the model with four components is preferable in this study. Notice that at around 9 a.m. and 3 p.m., the fitting results from the models (even with four or five components) are not favorable. Further investigation reveals that the traffic conditions during these hours are much more complicated than other time periods: there were much more frequent queue spillbacks (due to higher traffic volume and relatively short intersection spacing) and more occurrences of pedestrian crossing the road in the mid-block. These could bring about additional challenges to the models to segregate different travel patterns (e.g., free-flow and stop-and-go). A possible solution is to differentiate such scenarios from the others and to apply different models for data fitting, which would be one of potential research topics in the future.

## *Vehicle stop versus non-stop movement classification*

Upon fitting of TTDs with the proposed MGMM, we can further classify vehicle stop versus non-stop movements. One way is to classify the data sample $(tt_{x1,x2}, l_{x1,x2})$ based on its responsibility from each component, $\gamma_{ik}$. In this study, if the free-flow component has higher responsibility than the sum of others, then the sample is classified as free-flow (or non-stop) movement. Otherwise, the sample is classified as delayed (or stop) movement. However, as shown in Figure 6, some samples in the lower left corner (with very small travel times) have higher responsibility of the sum of other components and may be misclassified as delayed samples. To remedy this issue, we further assume that the travel time of a delayed vehicle cannot be smaller than the free-flow travel time in the early morning. Therefore, the criteria for a data sample $(tt_n, l_n)$ to be classified as the free-flow movement are

$$\gamma_{n1} > \sum_{k=2}^{K} \gamma_{nk} \quad \text{or } tt_n < \mu_{ff} \cdot l_n. \tag{13}$$

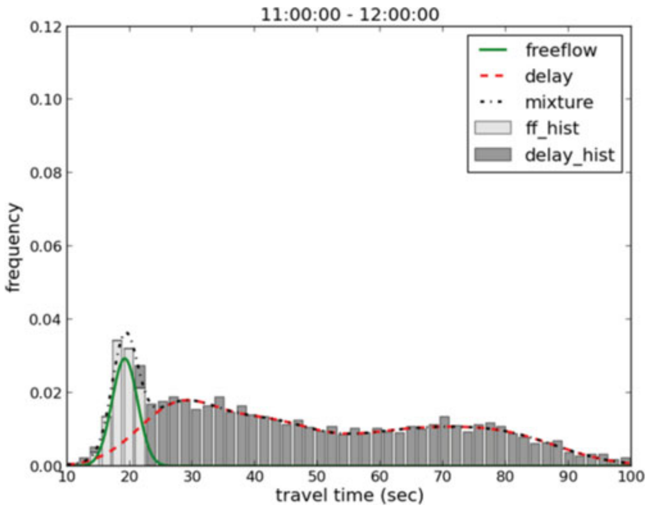

**Figure 6.** TTD and its estimation during the time period 11 a.m.–12 p.m. over one month (Aug. 2011). The solid curve represents free-flow component, while the dash curve is the delay distribution formed by all the other components. Light bars are the data classified as free flow samples, while dark bars are classified as delayed samples.

In order to evaluate the classification results, ground truth data was collected from 11 a.m. to 4 p.m. on September 8th, 2011 using a video camera (mounted on the top of a light post). Vehicle movements were then videotaped (Figure 7) when the wireless sensors were simultaneously recording the travel times. The movement types, either free-flow or delayed, of around 1,400 vehicles were visually verified.

Based on correct rates, we evaluate the performance of MGMMs with different number of components (from two to five) across different sizes of training dataset (i.e., one day, one week and one month). As shown in Figure 8, the mixture model with only two components cannot provide satisfactory results for vehicle movement classification, where the correct rate is just around 60%. The models with three and more components can achieve more than 90% correct rates even though just one-day data is used for training. This shows great potential for many traffic-related applications that require the knowledge of the percentage of stopped vehicles (i.e., stop rate). It should be pointed out that a poor-fit model does not necessarily lead to bad classification results.

**Figure 7.** Snapshot of video clip to record the vehicle movements on Sep. 8th, 2011 at intersection of San Pablo Ave and Solano St, Berkeley, CA.

### *Validation through sampled data from the NGSIM dataset*

As stated in the "Field experiment setup and data description" section, we draw on the NGSIM Lankershim Blvd. dataset (from 8:30 a.m. to 8:45 a.m.), and process it into the mobile sensor data form (with the 10-second sampling interval). Results from different penetration rates of probe vehicles (i.e., the vehicles equipped with mobile sensors), including 10%, 50%, and 100% are illustrated in Figure 9.

It can be observed that although the results are better for higher penetration rates, the overall TTD estimation performance of the proposed MGMM is satisfactory and quite robust to the penetration rate of mobile sensors. A hypothesis is that the traffic volume in this case is not so high and the free-flow component dominates, which makes it less challenging to fit the data even under low penetration rates (e.g., 10%). Further analyses show that as the sampling interval increases to 20 seconds (see Figure 10), the model performance deteriorates. When the penetration rate is set as 10%, we cannot even obtain the estimation of TTD because there are very limited data samples available. The major reason is that if the sampling interval is too long, the number of "unqualified" cases (i.e., two consecutive samples cover more than one link, resulting no data sample for that link) increases, or less "qualified" data samples can be used for model training. Even worse, those "qualified" samples may be biased because they usually come from the delayed probe vehicles.

## Application to energy/emissions estimation

### *Conceptual description*

If only the link length and travel time are available, one heuristic way to estimate vehicle energy consumption and emissions along arterials is (a) to assume the vehicle always travels at the average speed (i.e., link length divided by the travel time); and (b) to apply a microscopic emissions model, such as MOVES (motor vehicle emission simulator) (USEPA, 2017) or CMEM (comprehensive modal emissions model) (Barth et al., 2000) to the second-by-second hypothetical cruise speed profile

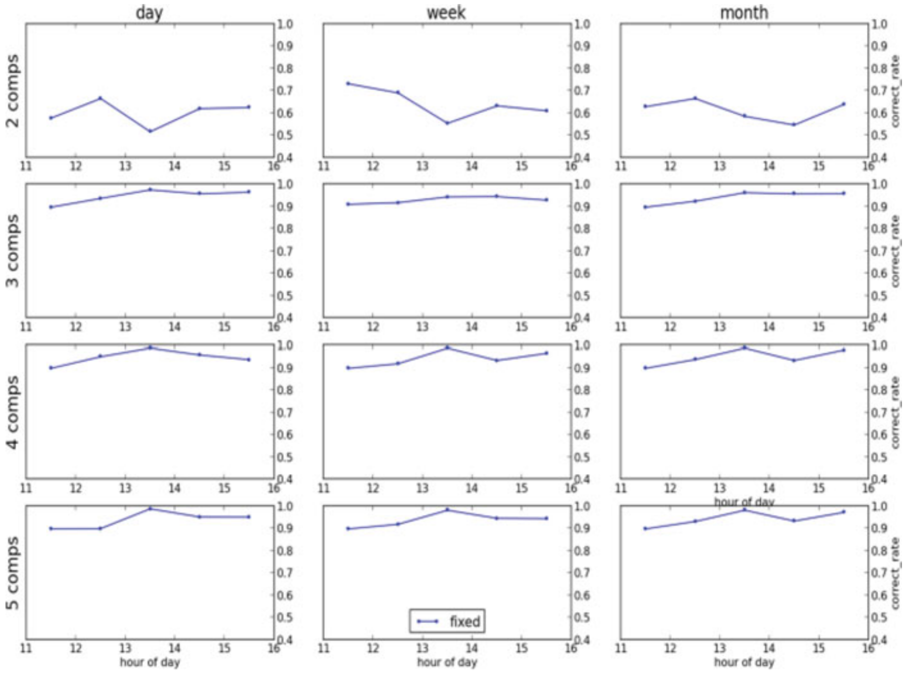

**Figure 8.** Correct rate of classification by the proposed MGMM with 2–5 components which are trained on the datasets of one day (Sept. 8th, 2011), one week (Sept. 5th –Sep 9th, 2011), and one month (Aug. 5th – Sept. 9th, 2011), respectively.

(to estimate the second-by-second fuel consumption and pollutant emissions). As shown by the dot line in Figure 11, the vehicle consumes much more fuel in the acceleration mode (between 75 and 85 seconds) than in the cruise mode (even at a high speed at the beginning of the real trajectory). Therefore, the aforementioned heuristic method (called "average speed method" in the following) may significantly underestimate the vehicle fuel consumption and emissions along arterials where stop-and-go maneuvers occur frequently.

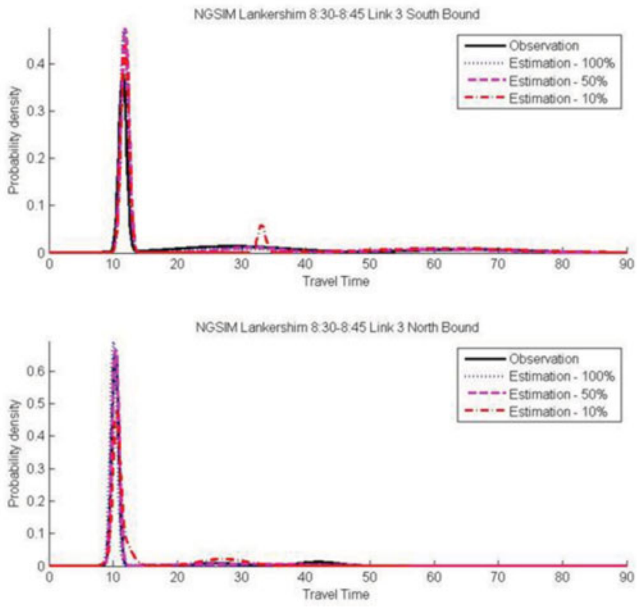

**Figure 9.** Verification of the proposed MGMM's validity on mobile sensor like datasets by estimating TTD from sampled data (every 10 seconds) in the NGSIM Lankershim Blvd. dataset (from 8:30 a.m. to 8:45 a.m.). The results are for Southbound (top) and Northbound (bottom) traffic on Link 3 (see Figure 3). Levels of penetration rate include 10%, 50%, and 100%.

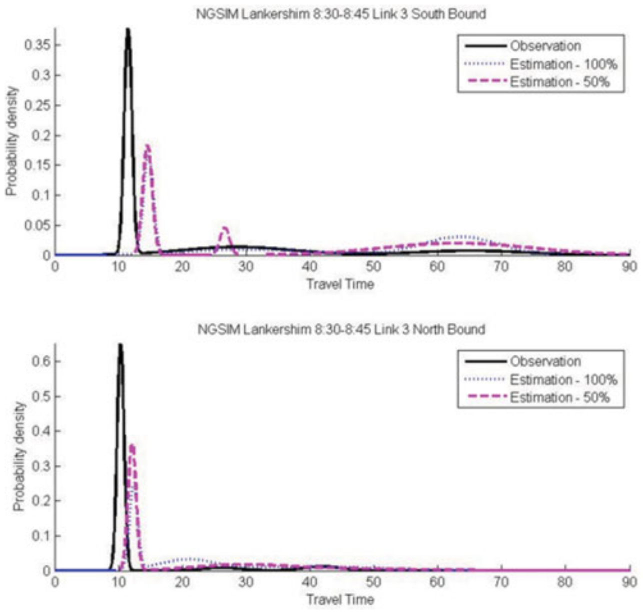

**Figure 10.** TTD estimation results from the proposed MGMM on mobile sensor like datasets (with sampling interval of 20 seconds) processed from the NGSIM Lankershim dataset (from 8:30 a.m. to 8:45 a.m.) for Southbound (top) and Northbound (bottom) traffic on Link 3. Levels of penetration rate include 50% and 100%. There are not enough samples in the case of 10% penetration rate.

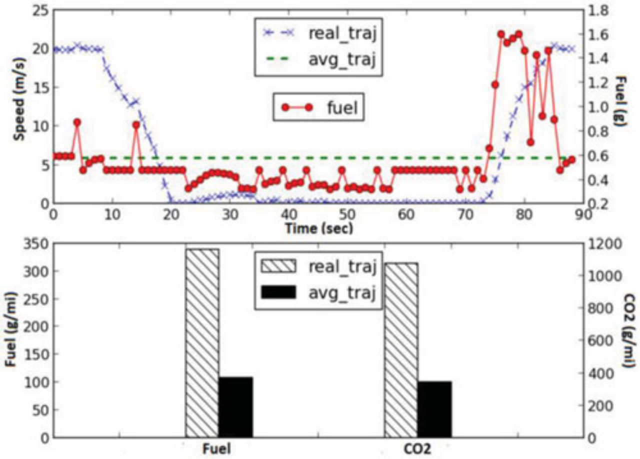

Figure 11. Comparison between the real trajectory and hypothetical cruise trajectory of a typical light-duty vehicle in terms of fuel consumption and emissions. Speed profiles and fuel consumption are shown in the top subplot. Aggregate fuel consumption and $CO_2$ emissions are shown in the bottom subplot.

Using the proposed MGMM, we can well estimate the link TTDs from the field data (collected from either fixed-location sensors or mobile sensors). Furthermore, vehicles' movements (stop or non-stop) can be satisfactorily classified. With this information, performance of the modal-based trajectory reconstruction method developed in Yang et al. (2011) can be significantly enhanced. Compared to the "average speed method," the modified modal-based method using the proposed MGMM is expected to provide much more accurate estimation of traffic-related energy consumption and emissions along arterials. In the following subsections, we will describe the simulation study for evaluating the effectiveness of energy and emissions estimation by applying the modified modal-based method based on the proposed MGMM.

## *Simulation setup*

Due to the unavailability of second-by-second vehicle trajectory data for a large-scale traffic network in the field, we resort to a microscopic traffic simulation tool, PARAMICS (Quadstone, 2017), to conduct the evaluation. In addition, the CMEM model (in the form of software plug-in to PARAMICS) is used to estimate the traffic-related fuel consumption and emissions.

A simulation network (calibrated in a previous study (Yin et al, 2007)) of a three-intersection (from top to bottom: Ventura, Los Robles, and Maybell) segment on El Camino Real in Palo Alto, CA is used in this study (see Figure 12). The intersection spacing varies from 200 to 500 m and the speed limit is 40 mph. Vehicle demands and their origin-destination (OD) patterns have been calibrated using the real-world data of a typical weekday morning (between 7:15 a.m. and 9:30 a.m.) in July 2005. In addition, there are two vehicle types defined in this network: light-duty vehicles versus buses, with a split of around 98:2. In this study, we focus on the TTD estimation along the link between the yellow bars (traffic flow direction is indicated by the red arrow). To evaluate the performance under different congestion levels, we select four volume-to-capacity (v/c) ratios for the simulation: 0.3, 0.5, 0.7, and 0.9.

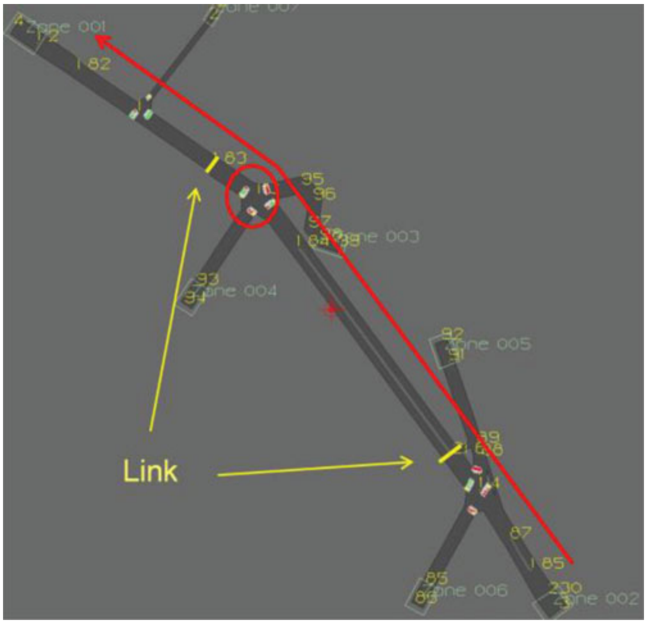

Figure 12. The three-intersection segment on El Camino Real, California coded in PARAMICS. From top to bottom: Ventura, Los Robles, and Maybell.

Data (mobile sensor like) are sampled from each vehicle (i.e., 100% penetration rate) with a time interval of 20 seconds. The sampling start time of each vehicle is randomly drawn from the range between 0 and 20 seconds after the vehicle enters the link of interest. To reduce the disturbances from randomness, we sample the same simulated trajectory repository 20 times. Then, we train 20 classifiers on each sampling dataset, and present the results of classification and energy consumption/emissions estimation in terms of the mean and standard deviation.

## *Simulation results*

We first assess the vehicle movement classification results from the four-component MGMM under different congestion levels (i.e., different v/c values). As shown in Table 1, the stop rate (i.e., the share of stopped vehicles in the overall traffic volume) increases as the congestion level increases. In addition, the correct rates (i.e., the fraction of vehicles correctly classified as stopped or non-stopped) are above 0.90 across all the congestion levels.

**Table 1.** Classification results over different congestion levels.

| | | Correct rate[b] | |
|---|---|---|---|
| v/c | Stop rate[a] | Mean | Std[c] |
| 0.3 | 0.66 | 0.96 | 0.02 |
| 0.5 | 0.73 | 0.95 | 0.00 |
| 0.7 | 0.82 | 0.96 | 0.02 |
| 0.9 | 0.99 | 0.91 | 0.03 |

[a] Fraction of stopped vehicles in the overall traffic flow.
[b] Fraction of vehicles correctly classified as stopped or non-stopped.
[c] Standard deviation.

Based on the vehicle movement classification results, we reconstruct the modal-based trajectories (see Yang et al. (2011)) and estimate their fuel consumption/$CO_2$ emissions. Figure 13 presents the average (over 20 simulation runs) relative difference (with respect to the ground truth, i.e., the cross-dash curve in Figure 11) in fuel consumption/$CO_2$ emissions across different v/c ratios for (a) the "average speed method" (dash curve in Figure 11); and (b) the modified modal-based method based on the movement classification results from the propose MGMM. As shown in the figure, both methods underestimate the energy consumption and emissions, due to the unrealistic smoothness of modeled trajectories (compared to a real trajectory). However, the modified modal-based method significantly outperforms the "average speed method" across all the congestion levels (improved by as much as 57% at v/c = 0.3 or 0.5).

Table 2 summarizes the estimation results on other criteria pollutants, such as CO, HC, and $NO_x$, by both the "average speed method" and the modified modal-based method. Unsurprisingly, the estimation by the latter has exhibited much smaller relative difference than that of the former across different congestion levels, when compared to the ground truth data. For example, when v/c = 0.5, the relative estimation errors by the "average speed method" are as high as −67%, −75%, and −97% for CO, HC, and $NO_x$, respectively, while the results from the modified modal-based method range from 0% to −13%. In addition, the estimation by the "average speed method" is not sensitive to the congestion level, while the results from the modified modal-based method follows the same trends as in the ground truth data, especially under the low v/c ratios. As the traffic gets congested (e.g., v/c = 0.9), results from the modified modal-based method also deviate nontrivially (ranging from −25% to −39%) from the ground truth. This may be caused by the occurrence of multiple stops per vehicle that is not captured in the modified modal-based method.

**Table 2.** Comparison results on average CO, HC, and $NO_x$ emissions per vehicle (gram) over different congestion levels.

| | | | Avg. speed method | | Modal-based method | |
|---|---|---|---|---|---|---|
| v/c | MOE | Ground truth | Absolute | % | Absolute | % |
| 0.3 | CO | 0.92 | 0.33 | − 64.1 | 0.87 | − 5.4 |
| | HC | 0.08 | 0.02 | − 75.0 | 0.07 | − 12.5 |
| | $NO_x$ | 0.25 | 0.01 | − 96.0 | 0.23 | − 8.0 |
| 0.5 | CO | 0.98 | 0.32 | − 67.3 | 0.94 | − 4.1 |
| | HC | 0.08 | 0.02 | − 75.0 | 0.08 | 0.0 |
| | $NO_x$ | 0.30 | 0.01 | − 96.7 | 0.26 | − 13.3 |
| 0.7 | CO | 1.10 | 0.31 | − 71.8 | 0.99 | − 10.0 |
| | HC | 0.09 | 0.02 | − 77.8 | 0.08 | − 11.1 |
| | $NO_x$ | 0.35 | 0.01 | − 97.1 | 0.29 | − 17.1 |
| 0.9 | CO | 1.45 | 0.30 | − 79.3 | 1.03 | − 29.0 |
| | HC | 0.12 | 0.02 | − 83.3 | 0.09 | − 25.0 |
| | $NO_x$ | 0.49 | 0.00 | − 100.0 | 0.3 | − 38.8 |

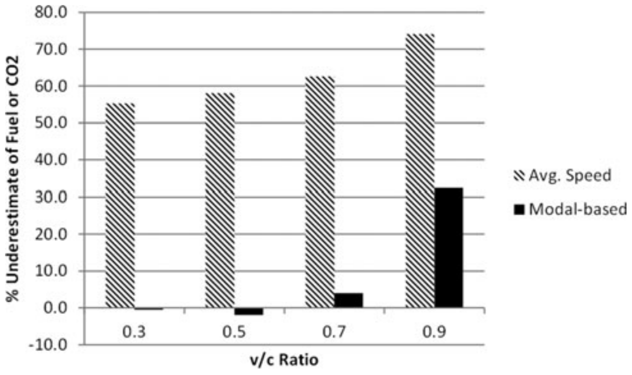

**Figure 13.** Percentage of underestimate of fuel consumption/$CO_2$ emissions using "average speed method" and modified modal-based method, compared to the ground truth.

## Conclusions and future work

In this paper, we propose a MGMM to estimate the TTDs and to classify the vehicle stop versus non-stop movements along arterials. We use field measurements from wireless magnetic sensors to assess the effectiveness of the proposed MGMM in the presence of fixed-location sensors, in terms of the goodness-of-fit measures (i.e., $p$-value in this study). The results are quite promising, especially for the model with four components. We also validate the model performance for the mobile sensor case by using the NGSIM Lankershim Blvd. dataset. In this case, it turns out that the model is quite robust to the penetration rate (ranging from 10% to 100%) of probe vehicles, but the performance is sensitive to the sampling interval. The sparseness (in terms of sampling interval) issue of the mobile sensor data still brings about challenges to application of the proposed model, due to the assumption that the data should be sampled closely enough in space (not more than one link) to qualify for training the model. One of the future steps would be to develop a robust interpolation method to address such data sparseness.

Furthermore, we use the estimated TTDs by the proposed MGMM to classify the stop versus non-stop movements and validate the results using the ground truth data of 1,400 vehicles as well as microscopic traffic simulation data (across different congestion levels). The correct rates (i.e., fraction of vehicles that are correctly classified as stopped or non-stopped) are all above 0.9 for the models with four or more components.

Based on the classification results from the proposed MGMM, we apply the modal-based method to estimate traffic-related energy consumption and emissions along an arterial corridor in the microscopic simulation. Significant improvements in terms of estimation errors are witnessed in comparison with the conventional "average speed method."

From the perspective of potential extended works, we can at least expect the following:

- The proposed MGMM should be validated using real-world mobile sensor data. Further investigation of some practical concerns (e.g., data sparseness, heterogeneous sampling frequency, and fusion with fixed-location sensor data) would provide in-depth understanding. For example, applying cluster analysis to identify periods with similar travel time patterns may help address the issue of low penetration rate of probe vehicles.
- With the estimated TTDs by the proposed model, more comprehensive algorithms could be developed to better estimate arterial traffic states, such as average idling time, and average stop rate (including the multi-stops scenario under oversaturated conditions). This may further improve the energy/emissions estimation along arterials.

## Acknowledgments

We thank Dr Peng Hao for his support in numerical validation. The contents of this paper reflect the views of the authors and do not necessarily indicate acceptance by the sponsors.

## Funding

We would like to acknowledge Sensys Networks Inc., the University of California Multi-campus Research Program Initiative (MRPI) in Sustainable Transportation, and the U.S. Department of Energy's Vehicle Technology Program (Grant number DE-EE0005448) for sponsorship of this research.

## References

Araghi, B., Krishnan, R., & Lahrmann, H. (2016). Mode-specific travel time estimation using bluetooth technology. *Journal of Intelligent Transportation Systems*, *20*(3), 219–228.

Barth, M., An, F., Younglove, T., Scora, G., Levine, C., Ross, M., & Wenzel, T. (2000). *The development of a comprehensive modal emissions model*. NCHRP Web-Only Document 122, Contractor's final report for NCHRP Project 25–11 (p. 307). National Cooperative Highway Research Program.

BeattheTraffic. (2014). http://www.beatthetraffic.com/

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Vol. 1. New York: Springer.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208.

Chen, P., Sun, J., & Qi, H. (2017). Estimation of delay variability at signalized intersections for urban arterial performance evaluation. *Journal of Intelligent Transportation Systems*, *21*(2), 94–110.

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: A step-by-step approach*. Hoboken, NJ: Wiley.

Federal Highway Administration (2006). *Next Generation Simulation (NGSIM)*. http://ngsim.fhwa.dot.gov/.

Feijer, D., Savla, K., & Frazzoli, E. (2012). *Strategic dynamic vehicle routing with spatio-temporal dependent demands*. Paper presented at the 2012 American Control Conference (ACC), Montreal, Canada, June 27–29

Feng, Y. (2011). *Arterial travel time distribution estimation and applications* (Ph.D. thesis). University of Minnesota.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*, 381–395.

Guo, F., Rakha, H., & Park, S. (2010). Multistate model for travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 2188, 46–54.

Hofleitner, A., Herring, R., & Bayen, A. (2011). *A hydrodynamic theory based statistical model of arterial traffic*. CCIT Research Report, UCB-ITS-CWP-2011–2.

Hofleitner, A., Herring, R., & Bayen, A. (2012a). *Probability distributions of travel times on arterial networks: A traffic flow and horizontal queuing theory approach*. Paper presented at the 91st Transportation Research Board Annual Meeting, Washington D.C., January.

Hofleitner, A., Herring, R., Abbeel, P., & Bayen, A. (2012b). *Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1679–1693.

Hofleitner, A. (2013). *A hybrid approach of physical laws and data-driven modeling for estimation: The example of queuing networks* (Ph.D. thesis) UC Berkeley: University of California.

Kwong, K., Kavaler, R., Rajagopal, R., & Varaiya, P. (2009). Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies*, *17*, 586–606.

McLachlan, G. J. (1988). *Mixture models: Inference and applications to clustering*. Statistics: Textbooks and Monographs, New York: Dekker.

Ndoye, M., Totten, V. F., Krogmeier, J. V., & Bullock, D. M. (2011). *Sensing and signal processing for vehicle reidentification and travel time estimation. IEEE Transactions on Intelligent Transportation Systems*, *12*, 119–131.

Noland, R. B., & Polak, J. W. (2002). Travel time variability: A review of theoretical and empirical issues. *Transport Reviews*, *22*(1), 39–54.

Olszewski, P. S. (1994). Modeling probability distribution of delay at signalized intersections. *Journal of Advanced Transportation*, *28*, 253–274.

Pandit, K., et al. (2013). Adaptive traffic signal control with vehicular ad hoc networks. *IEEE Transactions on Vehicular Technology*, *62*(4), 1459–1471.

Quadstone (2017). *PARAMICS*, http://www.paramics-online.com/

Ramezani, M., & Geroliminis, N. (2012). On the estimation of arterial route travel time distribution with Markov chains. *Transportation Research Part B*, *46*(10), 1576–1590.

Sanaullah, I., Quddus, M., & Enoch, M. (2016). Developing travel time estimation methods using sparse GPS data. *Journal of Intelligent Transportation Systems*, *20*(6), 532–544.

Sensys Networks Inc. (2015). http://www.sensysnetworks.com/

Uno, N., Kurauchi, F., Tamura, H., & Iida, Y. (2009). Using bus probe data for analysis of travel time variability. *Journal of Intelligent Transportation Systems*, *13*(1), 2–15.

U.S. Environmental Protection Agency. (2017). *MOVES (Motor Vehicle Emission Simulator)*. http://www.epa.gov/otaq/models/moves/.

Wasson, J. S., Sturdevant, J. R., & Bullock, D. M. (2008). Real-time travel time estimates using media access control address matching. *ITE Journal*, *78*, 2008.

Yang, Q., Boriboonsomsin, K., & Barth, M. (2011). *Arterial roadway energy/emissions estimation using modal-based trajectory reconstruction*. Paper presented at IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 809–814.

Yin, Y., et al. (2007). *Development of an integrated microscopic traffic simulation and signal timing optimization tool*. California PATH Research Report UCB-ITS-PRR-2007–2.

Zheng, F., & van Zuylen, H. (2010). *Reconstruction of delay distribution at signalized intersections based on traffic measurements*. Paper presented at IEEE Conference on Intelligent Transportation Systems, pp. 1819–1824.