

Prediction on Travel-Time Distribution for Freeways Using Online Expectation Maximization Algorithm

Nianfeng Wan
Department of Mechanical Engineering,
Clemson University
nwan@clemson.edu

Gabriel Gomes
University of California, Berkeley
gomes@path.berkeley.edu

Ardalan Vahidi
Department of Mechanical Engineering,
Clemson University
avahidi@clemson.edu

Roberto Horowitz
Department of Mechanical Engineering,
University of California, Berkeley
horowitz@berkeley.edu

Paper submitted to TRB Annual Meeting 2014
August 1, 2013

3300 words + 8 figure(s) + 0 table(s) \Rightarrow 5300 'words'

ABSTRACT

This paper presents a stochastic model-based approach to freeway travel-time prediction. The approach uses the Link-Node Cell Transmission Model (LN-CTM) to model traffic and provides a probability distribution for travel time. On-ramp and mainline flow profiles are collected from loop detectors, along with their uncertainties. The probability distribution is generated using Monte Carlo simulation and the Online Expectation Maximization clustering algorithm. The simulation is implemented with a reasonable stopping criterion in order to reduce sample size requirement. Results show that the approach is able to generate an accurate multimodal distribution for travel-time. Future improvements are also discussed.

Keywords: Travel-time Distribution, Model-based Prediction, Link-Node Cell Transmission Model, Monte Carlo, Online Expectation Maximization.

INTRODUCTION

Travel-time is among the most important traffic performance measures. Accurate travel-time information enables drivers to understand the traffic conditions, and hence to choose routes or manage their trip schedules to avoid congested road sections. Most of today's state-of-the-art navigation systems like Google Maps provide travel-time information. From the traffic control prospective, travel-time information also helps in monitoring and controlling traffic with traffic lights, ramp metering, etc. (1)

Challenging however is accurate prediction of travel time for a given route as it is influenced by many different traffic and road parameters: flow, density, speed, route length, geometry, to name a few. These parameters are obtained through various sources, which carry different kinds of uncertainties, making the prediction more challenging. Most existing methods for computing travel-times rely on data-mining from historical data. Those include methods based on linear regression (2), time series (3), Kalman filter (4) and (5), and artificial neural networks (6). Such data-based prediction methods require large amount of traffic data with gaps and inaccuracies due to missing or bad sensors. Instead of large investments in fixing the sensor network, traffic flow models can be built as an alternative mean, to fill the data gap and for obtaining travel-time forecasts through simulation. Model-based prediction does not depend as much on real time measurements as data-based techniques; but models can be re-calibrated on the fly when new data becomes available. To the authors' best knowledge, most current model-based prediction methods are based on microscopic simulation (7) and (8), which model the behavior of each individual vehicle. As compared with macroscopic models, microscopic models are computationally more demanding and are often difficult to calibrate.

To overcome the above-mentioned challenges, in this paper we propose to estimate travel times via a macroscopic model, which formulates the relationships among aggregate traffic quantities. Because the parameter and inputs of the model are influenced by various sources of uncertainty, it makes more sense to estimate a probability distribution for travel-time rather than a single deterministic value. A single travel-time sample along a route is usually not helpful, since it does not provide a sense of the reliability of the information. Instead a travel-time probability distribution has important uses in traveler information as well as in traffic control systems.

When estimating a probability distribution for travel-time, a challenge is the multidimensionality of the problem. Travel-time is affected by various factors, each of which may have a different kind of distribution (Gaussian, uniform, etc.), and small changes of the parameters may significantly alter the outcome. Because of the nonlinearity of the traffic model, real travel-time distributions often present multiple modes, and may be sensitive to the inputs. Finally there is the challenge of the finite availability of computation time and memory. In this paper we employ Monte Carlo simulations for generating travel-time samples under demand uncertainties. A probability distribution for travel-time is then adaptively generated, via the Online Expectation Maximization clustering method, as new samples become available from Monte Carlo simulations. A stopping criterion for sampling process is also introduced in order to reduce sample size and computational requirement.

The rest of the paper is organized as follows: The model of the macroscopic traffic simulator is introduced first. Next we describe the computation of travel-time samples via Monte Carlo simulations followed by estimation of a travel-time distribution via an Online Expectation Maximization algorithm. The effectiveness of the proposed approach is evaluated in simulations for a real-world highway segment where demand data is available. We conclude with a discussion of

the results and remaining future work.

MODEL DESCRIPTION

This paper uses BeATS (Berkeley Advanced Transportation Simulator) as the traffic simulator. BeATS is an implementation of the Link Node Cell Transmission Model (LN-CTM), described in
 5 (9) and (10).

The LN-CTM is a macroscopic model of traffic suitable both for freeways and arterials. It is an extension of CTM which simulates traffic behavior specified by volume (flow), density, and speed. In LN-CTM, the traffic network is modeled as a directed graph. Links represent road segments and nodes are road junctions. Source links introduce traffic to the network and sink links
 10 absorb traffic. The fundamental diagram, a diagram relating densities to flows, is used to specify the parameters of each link. A split-ratio matrix at each node defines how vehicles are directed from input to output links. The required data can be obtained from the Performance Measurement Systems (PeMS): an online repository, which provides a rich archive of sensor detector data for freeways in California.

15 In general, the LN-CTM requires mainline and on-ramp demand profiles, calibrated fundamental diagrams and split ratio matrices as inputs. The model can be calibrated to match actual observation results(11).

METHODOLOGY

In this section, we first discuss how to calculate travel-time in simulation. Then we introduce the
 20 Monte Carlo sampling method to obtain travel-time samples. Then we illustrate how the Online EM algorithm computes distribution parameters. Finally the methodology flow process is given.

Travel-time Calculation

In microscopic simulation, one can track individual vehicles to estimate travel-time. Macroscopic models, because they compute only aggregate quantities, cannot provide direct estimates of travel-
 25 time for individual travelers. They are better suited, however, for estimating the probabilistic characteristics of travel-time.

We next describe the technique for calculating travel-time for a driver starting a trip at time t_{start} and traveling over a route R .

The route R is composed of a sequence of links $\{r_i\}, i = 1, 2, \dots, n$. The driver starts at the
 30 beginning of link r_1 at time $t = t_{start}$. The objective is to find the time t_{end} when the driver will exit link r_n as a function of the history of macroscopic flows and densities along the route. Then a sample of travel-time for route R at time t_{start} is:

$$TT(R, t_{start}) = t_{end} - t_{start} \quad (1)$$

The process can be repeated over an ensemble of simulations to obtain the distribution of $TT(R, t_{start})$. The steps below are followed to obtain the distribution of travel-time:

35 1. Initialization: $\rho_i(t_{start})$, the initial state of each link i at time t_{start} , must be computed, using either a state estimator or by advancing the simulator from a previously known state. For the purpose of this paper, the simulation was started with an empty initial condition at midnight and advanced deterministically to the starting time. Thus, the ensemble of runs was given a deterministic initial condition at time $t = t_{start}$.

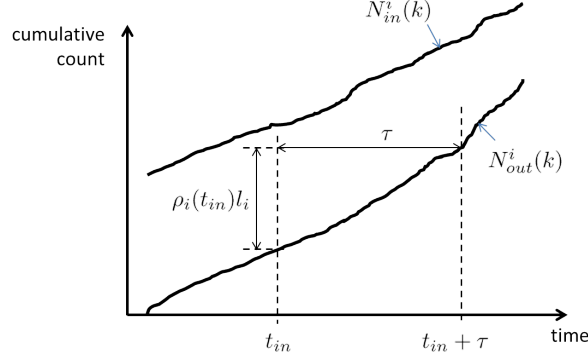


FIGURE 1 Cumulative Counts for a Link

2. Take the i th link in the route and denote its incoming and outgoing flow with $f_{in}^i(k)$ and $f_{out}^i(k)$ at time step k . Then the “cumulative counts” into the link $N_{in}^i(k)$ and those out of the link $N_{out}^i(k)$ are:

$$N_{in}^i(k) = \sum_{\alpha=0}^k f_{in}^i(\alpha) \cdot \Delta t + N_{in}^i(0) \quad (2)$$

$$N_{out}^i(k) = \sum_{\alpha=0}^k f_{out}^i(\alpha) \cdot \Delta t + N_{out}^i(0) \quad (3)$$

5 where Δt is the length of time step, $N_{in}^i(k)$ and $N_{out}^i(k)$ are in vehicle units.

Since the flows are non-negative, the cumulative counts are non-decreasing functions of time. They are shown in Figure 1 for a particular link. The travel-time in the link is the time it takes for the output flow to accumulate the total number of vehicles present in the link when the vehicle entered. Thus the travel-time τ is the solution to the following equation:

$$\rho_i(t_{in}) \cdot l_i = N_{out}^i(t_{in} + \tau) - N_{out}^i(t_{in}) \quad (4)$$

10 where t_{in} is the time when the vehicle entered the link, $\rho_i(t_{in})$ is the initial density at time t_{in} , and l_i is the length of the link i . Equation 4 is solved numerically by searching the $N_{out}^i(k)$ vector for the value $\rho_i(t_{in}) \cdot l_i + N_{out}^i(t_{in})$. The only subtlety that arises is that the initial time t_{in} and/or the final time $t_{in} + \tau$ may not fall on the time grid. In this case we also count N_{in}^i and use linear interpolation to calculate accurate t_{in} .

15 The computation of travel-time on a route is performed by computing travel times on each link of the route in sequence, and noting that the exit time for link i is the entering time for link $i + 1$.

Monte Carlo Sampling

As mentioned before, travel-time is affected by factors such as capacity and demand. Considering
 20 that they are themselves non-Gaussian random quantities, and the system is inherently nonlinear, the travel-time estimation problem becomes analytically intractable. Therefore, in this paper, the Monte Carlo method is chosen to obtain the travel-time results.

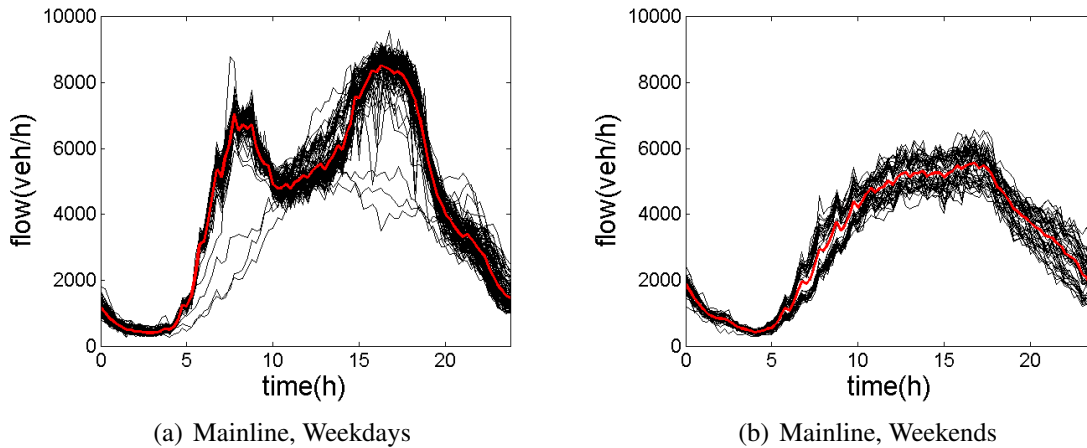


FIGURE 2 Six Months of Mainline Flow Data, and Their Average

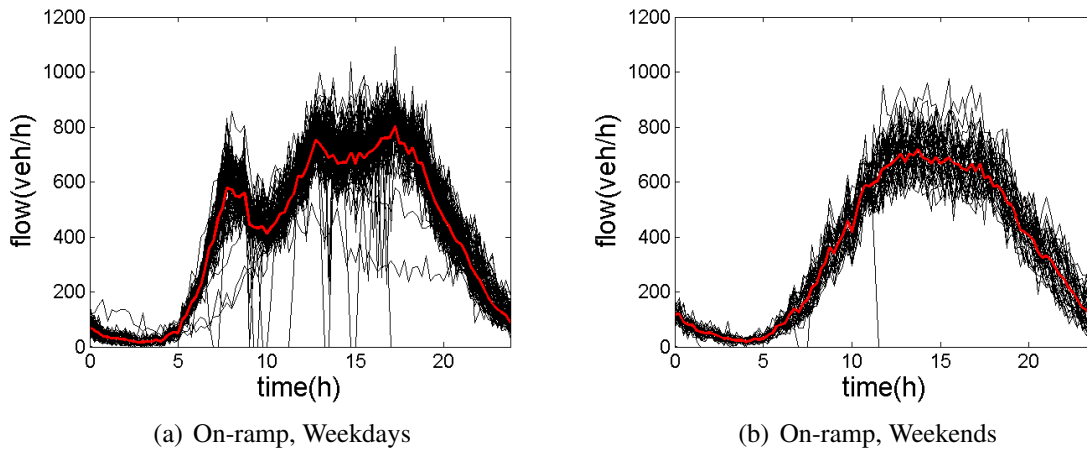


FIGURE 3 Six Months of On-ramp Flow Data, and Their Average

The Monte Carlo method samples randomly from a probability distribution. It approximates the distribution when it is infeasible to apply a deterministic method. The number of samples needed in Monte Carlo does not depend on the dimension of the problem, making it suitable for solving multidimensional problems. Another feature of Monte Carlo is that it is easy to estimate the order of magnitude of statistical error(12).

In this paper, the uncertainties are added to the mainline and on-ramp demand profiles. The on-ramp uncertainty is assumed to be Gaussian, and the deviation is generally on the order of 5% of the demand in the morning rush hour. Mainline demands are also considered as a Gaussian distribution, and the reasonable deviation is around 2.5%. Figures 2 and 3 show six months of loop detector readings gathered every five minutes from detectors on I-15 in California. These plots illustrate the typical variations in 5-minutes average flows.

With each simulation, the Monte Carlo method randomly samples from the demand distributions. Each simulation generates one sample travel-time. Because each simulation requires considerable computation and execution time, it is important to estimate how many samples will be

needed to produce statements about the travel-time distribution with a given level of confidence. Also, it will be important to parameterize the distribution in a way that captures its important features, while using a relatively small number of parameters.

Clustering via the EM Algorithm and Bayesian Inference Criterion

5 The shape of the travel-time probability distributions is not known a-priori. Based on current literature(13), the travel-time distribution can be represented as a Gaussian Mixture Model (GMM). GMM represents the data as a sum of several Gaussian distributions. The probability density distribution is then:

$$P(x|\pi, \mu, \Sigma) = \sum_{i=1}^K \pi_i \mathcal{N}(x|\mu_i, \Sigma_i) \quad (5)$$

10 where P is the probability density function, \mathcal{N} is the Gaussian distribution, K is the number of the components or clusters, μ_i is the mean, Σ_i is the covariance matrix, and π_i is the weight. The weights are such that,

$$\sum_{i=1}^K \pi_i = 1 \quad (6)$$

We use a clustering technique to find out the values of the parameters from a group of samples. The Expectation Maximization (EM) method was chosen here for clustering the data for GMM(14). The EM method starts with a random guess of the unknown parameters, and iteratively alternates between an expectation (E) step and a maximization (M) step. The E step produces the responsibilities $\{\gamma_i(x)\}$, $i = 1, 2, \dots, K$, where $\gamma_i(x)$ represents the conditional probability that the data x came from the i th cluster, given the current parameters $\{\mu_i, \Sigma_i, \pi_i\}$. That is,

$$\gamma_i(x) = \frac{\pi_i \mathcal{N}(x|\mu_i, \Sigma_i)}{\sum_{i=1}^K \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)} \quad (7)$$

The M step updates the parameters $\{\mu_i, \Sigma_i, \pi_i\}$ to maximize the expectation of the log-likelihood. The parameters are updated with,

$$\mu_i = \frac{\sum_{j=1}^N \gamma_i(x_j) \cdot x_j}{\sum_{j=1}^N \gamma_i(x_j)} \quad (8)$$

20

$$\Sigma_i = \frac{\sum_{j=1}^N \gamma_i(x_j) \cdot (x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^N \gamma_i(x_j)} \quad (9)$$

$$\pi_i = \frac{1}{N} \sum_{j=1}^N \gamma_i(x_j) \quad (10)$$

where N is the number of data points.

By iterating sufficiently between the E step and the M step, the parameters can converge. However, EM is not guaranteed to converge to a global maximum of the log-likelihood function. In this paper, we initiate several different random guesses to avoid getting stuck in local maxima(15).

25

Another important point is that the number of clusters in the distribution is unknown. This paper uses a Bayesian Inference Criterion (BIC) to estimate the optimal number of clusters(16). The BIC criterion can be represented as:

$$BIC = -\ln P(D|\mu, \Sigma) + \frac{KQ + 1}{2} \ln N \quad (11)$$

where $\ln P(D|\mu, \Sigma)$ is the log-likelihood function, D represents the samples, Q is the number degrees of freedom (here since the travel-time has one degree of freedom, $Q = 1$), K is the number of clusters, and N is the sample size. The optimal cluster number would generate the maximum BIC value. With different initial conditions, BIC may converge to different optimal numbers of clusters. In this paper, we choose the most frequent result as the optimal one.

Online EM Algorithm

Statistically, in Monte Carlo sampling, more samples provide more accurate results. However, the computation time of the simulation has a linear relation with the number of samples. The more it samples, the longer time it requires. If the prediction time is too long, the traffic condition may significantly change, and the “delayed” prediction is less reliable. Moreover, for the method to be applicable to real-time systems, it must be capable of handling streaming data. That is, given a new data packet (30 samples, for example), the clustering method should be able to use that to update a running estimate. It stops requesting new data only if the results meet a stopping criterion. The advantage of this “data stream” structure is that it minimizes the number of simulations as well as the amount of memory needed to store the samples. Both of these are essential requirements for travel advisory as well as traffic management systems.

Since the target distribution is assumed to be GMM, the Online Expectation Maximization (Online EM) method(17) is suitable for clustering the data. In this paper, the Online EM method applies EM only to newly arrived data rather than to the whole historical data. And the incremental GMM estimation algorithm merges Gaussian components that are statistically equivalent, and maintains other components.

The W statistic test is used for equality of covariance. Let the newly coming samples x_i with $i = 1, 2, \dots, n$ have a covariance matrix Σ_x , and a given target covariance matrix Σ_0 . The null hypothesis is $\Sigma_x = \Sigma_0$. Define L_0 as a lower triangular matrix by Cholesky decomposition of Σ_0 , that is, $\Sigma_0 = L_0 L_0^T$. Let $y_i = L_0^{-1} x_i, i = 1, 2, \dots, n$, then the W statistic is represented as:

$$W = \frac{1}{d} \text{tr}[(S_y - I)^2] - \frac{d}{n} \left[\frac{1}{d} \text{tr}(S_y) \right]^2 + \frac{d}{n} \quad (12)$$

Where S_y is covariance of y_i , d is the dimension, n is the sample size, and $\text{tr}(\cdot)$ is the trace of the matrix. From (18), $\frac{nWd}{2}$ has χ^2 distribution, that is:

$$\frac{nWd}{2} \sim \chi_{d(d+1)/2}^2 \quad (13)$$

Once we set a significance value for the χ^2 distribution, we can decide whether the test has passed or failed.

The Hotellings T^2 statistic is used for equality of mean. Let the newly coming samples $x_i, i = 1, 2, \dots, n$ have a mean μ_x , and a given target mean μ_0 . The T^2 is defined as:

$$T^2 = n(\mu_x - \mu_0)^T S^{-1} (\mu_x - \mu_0) \quad (14)$$

Where S is covariance of x_i . From (19), $\frac{n-d}{d(n-1)}T^2$ has F distribution, that is:

$$\frac{n-d}{d(n-1)}T^2 \sim F_{d,n-d} \quad (15)$$

Once we set a significance value for the F distribution, we can decide whether the test has passed or failed.

Since it is essential to stop the Monte Carlo simulation properly to avoid too many samples, several rules are added:

1. When a cluster can be merged with more than one other clusters, the one with the highest weight is chosen;
2. A cluster is eliminated whenever its weight falls below a threshold;
3. The clustering algorithm is stopped if a certain number of iterations pass without new clusters being created.

Simulation Flow Process

In summary, the process of generating a distribution of travel-time from stochastic simulation is as follows:

1. The simulator advances to the given starting time.
2. (Monte Carlo step) The simulator applies uncertainties to the model, and runs a certain number of times to get travel-time samples. Each travel-time is calculated through the method mentioned at the start of this section.
3. (Online EM step) The simulator clusters the incoming samples, and calculates the parameters using the EM algorithm.
4. (Merging or Maintaining step) The simulator merges qualified new clusters to the old ones, and maintains the rest.
5. (Eliminating step) If the clusters have less weight than a threshold, the simulator merges them into the nearest cluster.
6. If there is no more new cluster in several steps, the simulation is stopped and returns the parameters, otherwise step 2-6 are repeated.

EXPERIMENTAL SETUP AND RESULTS

A section of I-15 southbound was used to test the algorithm. The stretch is located between Escondido and San Diego in California. Figure 4 shows the section in Google Maps. The section contains more than 120 nodes and 100 links. A route with 9 consecutive links was created. The total length of the route is 1.47 miles. It contains two on-ramps and no off-ramp. The demand profiles and the split ratio data was obtained from PeMs for Monday, January 7, 2013.

Figure 5 shows the density contour plot of one simulation sample. The vertical axis is the time, the horizontal axis is the spatial dimension, and the color represents the amount of density on each link. The stretch over which travel-time was computed is highlighted. The contour plot shows that on the route the congestion begins around 7:15 AM and ends at about 9:45 AM, which illustrates the Monday morning rush hour on I-15. With reasonable uncertainties, the boundary of the congestion changes and the travel-time changes too.

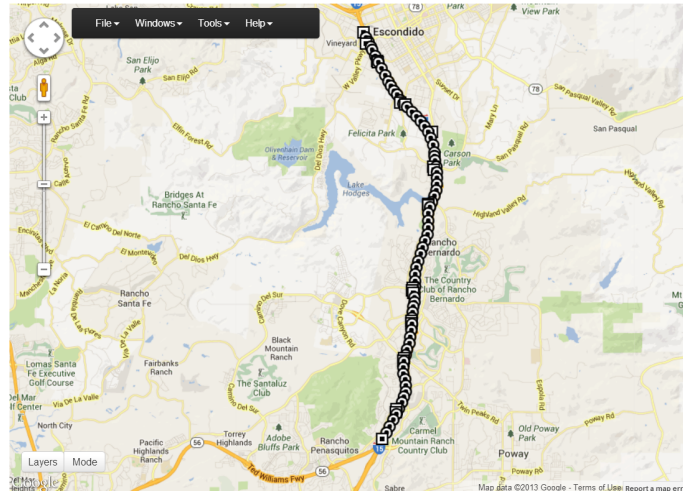


FIGURE 4 A Section of I-15 South

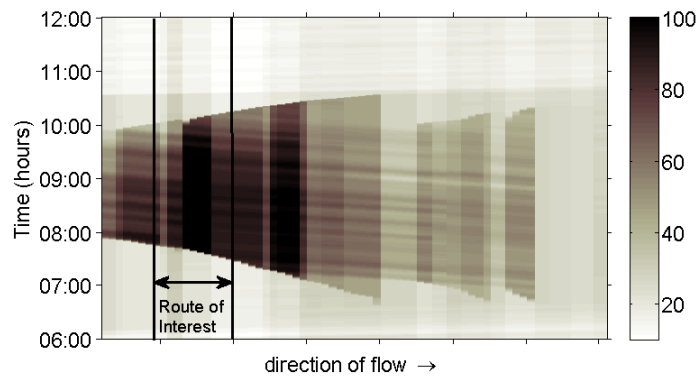


FIGURE 5 The Density Contour Plot

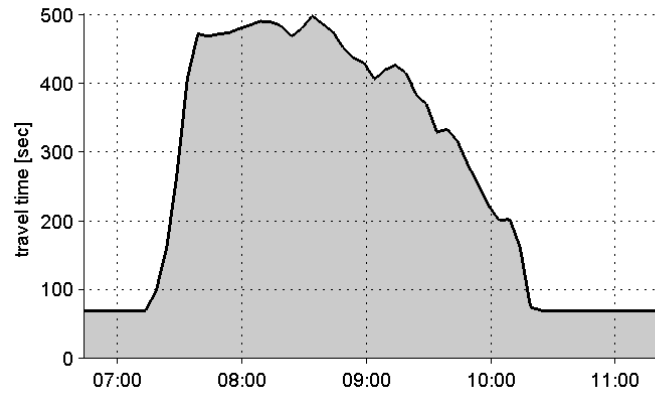


FIGURE 6 Deterministic Travel-time

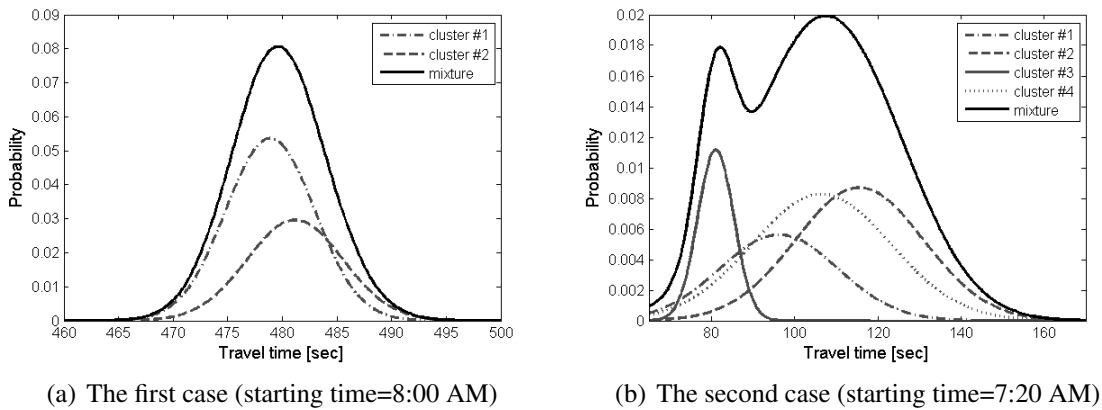


FIGURE 7 Travel-time Distributions and Their Components

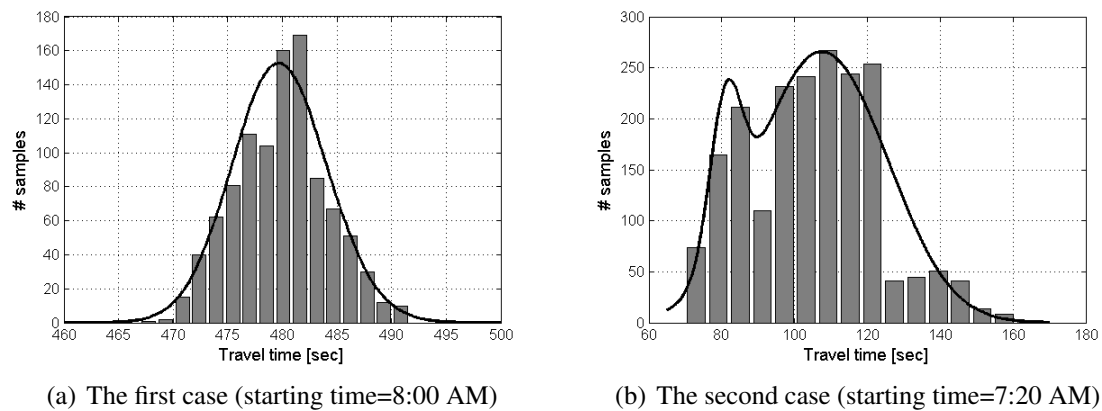


FIGURE 8 Travel-time Distributions and the Sample Histogram

The travel-time curve resulting from a simulation with mean values of demands is shown in Figure 6.

The significance level for the Online EM algorithm was set to 0.05. The threshold for eliminating cluster was set to 0.05, and the number of steps for convergence to 3.

Two starting times were selected. The first one is 8:00 AM. At this time, the route is heavily congested. Figure 7(a) shows the results. Although the distribution seems to be a Gaussian distribution, the Online EM clustering algorithm found two clusters. The GMM distribution result is more accurate. In this heavily congested example, the travel-time distribution is unimodal. The second starting time is 7:20 AM. At this time, the route is on the edge of congestion. With variable uncertainties, in some cases it is congested while in others it is in freeflow. Since vehicles cannot travel faster than freeflow speed, the minimum travel-time is the freeflow speed travel-time, which is 69 seconds. Figure 7(b) shows that there are two well separated modes in the distribution. In the first mode, the travel-time stays around freeflow speed travel-time, which indicates that there is no congestion or only a small number of links in the route are congested while others are in freeflow. In the second mode, more links become congested. In that case the travel-time distribution represents multiple modes.

By using the On-line EM algorithm, the clusters are eliminated, merged, or maintained and after several steps, the clusters number and parameters become stable. In the first case, the simulation stopped with 170 samples. In the second one, 290 samples were needed. Figure 8 compares the travel-time distribution prediction results with the histogram of 1000 travel-time samples, which more precisely capture the shape of the distribution. The comparison shows good agreement, suggesting that the Online EM algorithm and the stopping criterion work well and require fewer samples.

Travel-time distributions can be used by drivers and traffic managers to make more informed decisions about expected traffic patterns. For example, from Figure 7, drivers could expect with a high degree of certainty to take between 470 seconds and 490 seconds to travel the given route if they start at 8:00 AM. On the other hand, they will be aware that at 7:20 AM the situation is less reliable and a wider range of outcomes are possible.

CONCLUSION

This paper introduced a model-based approach for predicting probability distribution of travel-times for freeways. We used the BeATS simulator, an implementation of the Link-Node Cell Transmission Model, to model vehicular traffic flow and to estimate link travel times for a given set of demand profiles. Uncertainty in demand was handled by Monte Carlo sampling and an Online Expectation Maximization algorithm that estimated a Gaussian mixture probability distribution for travel times. Simulations with data from a freeway in California showed that the method could provide a robust estimate of travel time probability distribution with a relatively small number of simulations.

The proposed approach is not only suitable for freeways, but also applicable to arterial roads. Travel time for arterial roads is expected to have a multi-modal distribution due to the stop-and-go pattern induced by traffic signals at intersections. Therefore optimizing the sampling process becomes more critical for arterial roads and is the subject of future work.

Our model-based travel-time estimator runs open-loop and therefore is sensitive to accuracy of its parameters. Combining model predictions and sparse estimates of travel times from probe vehicle data, in a closed-loop estimator, can generate more accurate estimates of travel-time distribution, and will be investigated in our future work.

REFERENCES

- [1] Ho, F.-S. and P. Ioannou. Traffic flow modeling and control using artificial neural networks. *Control Systems, IEEE*, Vol. 16, No. 5, 1996, pp. 16–26.
- [2] Rice, J. and E. van Zwet. A simple and effective method for predicting travel times on free-ways. *Intelligent Transportation Systems, IEEE Transactions on*, Vol. 5, No. 3, 2004, pp. 200–207.
- [3] Billings, D. and J.-S. Yang, Application of the ARIMA Models to Urban Roadway Travel Time Prediction - A Case Study. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, 2006, Vol. 3, pp. 2529–2534.
- [4] Yang, J.-S., Travel time prediction using the GPS test vehicle and Kalman filtering techniques. In *American Control Conference, 2005. Proceedings of the 2005. IEEE*, 2005, pp. 2128–2133.
- [5] Chu, L., S. Oh, and W. Recker, Adaptive Kalman filter based freeway travel time estimation. In *84th TRB Annual Meeting, Washington DC*, 2005.
- [6] Park, D. and L. R. Rilett. Forecasting multiple-period freeway link travel times using modular neural networks. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1617, No. 1, Trans Res Board, 1998, pp. 163–170.
- [7] Chen, M. and S. I. Chien. Dynamic freeway travel-time prediction with probe vehicle data: Link based versus path based. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1768, No. 1, Trans Res Board, 2001, pp. 157–161.
- [8] Hollander, Y. and R. Liu. Estimation of the distribution of travel times by repeated simulation. *Transportation Research Part C: Emerging Technologies*, Vol. 16, No. 2, 2008, pp. 212 – 231.
- [9] Daganzo, C. F. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, Vol. 28, No. 4, Elsevier, 1994, pp. 269–287.
- [10] Muralidharan, A., G. Dervisoglu, and R. Horowitz, Freeway traffic flow simulation using the Link Node Cell transmission model. In *American Control Conference, 2009. ACC'09. IEEE*, 2009, pp. 2916–2921.
- [11] Dervisoglu, G., G. Gomes, J. Kwon, R. Horowitz, and P. Varaiya, Automatic calibration of the fundamental diagram and empirical observations on capacity. In *Transportation Research Board 88th Annual Meeting*, 2009, 09-3159.
- [12] Koehler, E., E. Brown, and S. J.-P. Haneuse. On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, Vol. 63, No. 2, Taylor & Francis, 2009, pp. 155–162.
- [13] Ko, J. and R. L. Guensler, Characterization of congestion based on speed distribution: a statistical approach using Gaussian mixture model. In *Transportation Research Board Annual Meeting*. Citeseer, 2005.

- [14] Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, 1977, pp. 1–38.
- [15] Meilă, M. and D. Heckerman, An experimental comparison of several clustering and initialization methods. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 386–395.
- [16] Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, Vol. 6, No. 2, Institute of Mathematical Statistics, 1978, pp. 461–464.
- [17] Song, M. and H. Wang, Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. In *Defense and Security*. International Society for Optics and Photonics, 2005, pp. 174–183.
- [18] Ledoit, O. and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, JSTOR, 2002, pp. 1081–1102.
- [19] Hotelling, H. *The generalization of Students ratio*. Springer, 1992.