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
July 30, 2013

3300 words + 8 figure(s) + 0 table(s) ⇒ 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 one of the most important traffic performance measures. Accurate travel-time information enables drivers to understand the traffic conditions, and hence to choose routes or manage 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 to monitor and control traffic with signal lights, ramp metering, etc. (*1*)

However, travel-time is difficult to predict. Since it is affected by many different kinds of traffic parameters: flow, density, speed, route length, geometry, etc. These parameters are obtained through various sources, which carry different kinds of uncertainties, making the prediction more challenging. There have been many methods for predicting travel-time. One method involves computing travel-time from historical data using data mining techniques. Methods have been developed based on linear regression (*2*), time series (*3*), Kalman filter (*4*) and (*5*), artificial neural networks (*6*). However, these data-based prediction methods requires large amount of traffic data, which can be missing or incorrect for some road sections due to missing or bad sensors. Solving this problem requires large investment in new sensors, which is often infeasible. Another prediction technique is to build a traffic flow model and to obtain travel-time forecasts through simulation. Model-based prediction does not depend as much on real time measurements as data-based techniques, and the data it needs is easy to access. To the authors' 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, these models require large amounts of computation and are often difficult to calibrate. In this paper, a macroscopic model is used, which formulates the relationships among aggregate traffic quantities.

One of the main challenges is to predict a distribution of travel-time rather than a deterministic value. There are various uncertainties in the traffic model. Hence the travel-time is not a deterministic quantity but a probabilistic one. The range of the distribution grows with the size of the uncertainties, and with the length of the predicting horizon. 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. Therefore it is difficult to evaluate how reliable the predicted travel-time is. Rather than giving one sample travel-time, this paper aims to provide the travel-time probability distribution. Such a distribution has important uses in traveler information as well as traffic control systems.

Another challenge is the multidimensionality of the problem. Travel-time is affected by various factors, each of which may be generated by different kinds of distributions (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. The Online Expectation Maximization clustering method is selected in part for its economy of the resources.

The rest of the paper is organized as follows: First, the model of the simulator is introduced, then we describe the computation of travel-time samples and the estimation of a travel-time distribution. Simulation results and analysis are discussed next. Discussion and conclusion are shown at last.

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 (*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 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.

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 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-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 $\{r_i\}, i = 1, 2, ...n$. The driver starts at the 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:

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

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

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 and 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 $t = t_{start}$.

2. We use the technique described in [ref33] for computing travel-time on a single link. Take the ith link in the route and denote its incoming and outgoing flow with $f^i_{in}(k)$ and $f^i_{out}(k)$ at

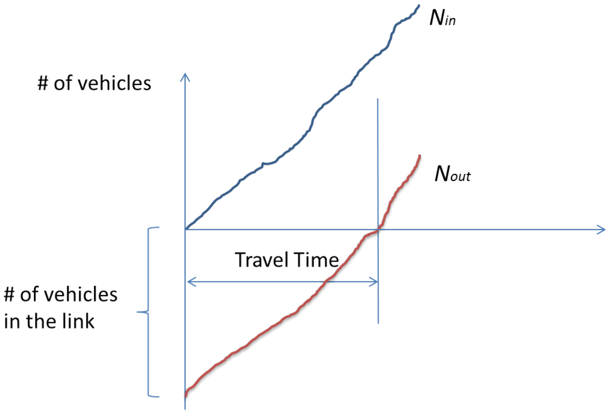


FIGURE 1 Cumulative Counts for a Link

time step k. Then the "cumulative counts" $N^i_{in}(k)$ and $N^i_{out}(k)$ are:

$$N^i_{in}(k) = \sum_{\alpha=0}^{k} f^i_{out}(\alpha) \cdot \Delta t + N^i_{in}(0) \tag{2}$$

$$N^i_{out}(k) = \sum_{\alpha=0}^{k} f^i_{out}(\alpha) \cdot \Delta t + N^i_{out}(0) \tag{3}$$

where Δt is the length of time step, and $N^i_{in}(k)$ and $N^i_{out}(k)$ 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:

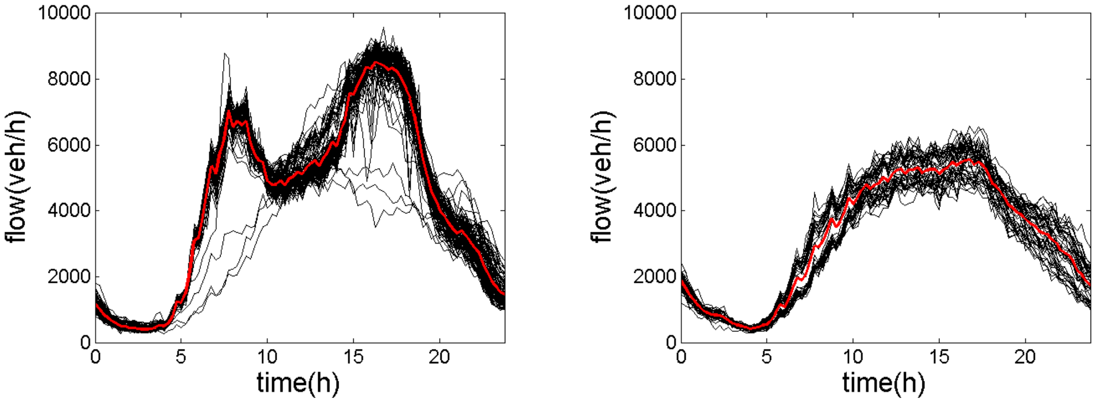
$$\rho_i(t_{in}) \cdot l_i = N^i_{out}(t_{in} + \tau) - N^i_{out}(t_{in}) \tag{4}$$

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 $N^i_{out}(k)$ vector for the value $\rho_i(t_{in}) \cdot l_i + N^i_{out}(t_{in})$. 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^i_{in} and linear interpolation is used to calculate accurate t_{in}.

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 Method

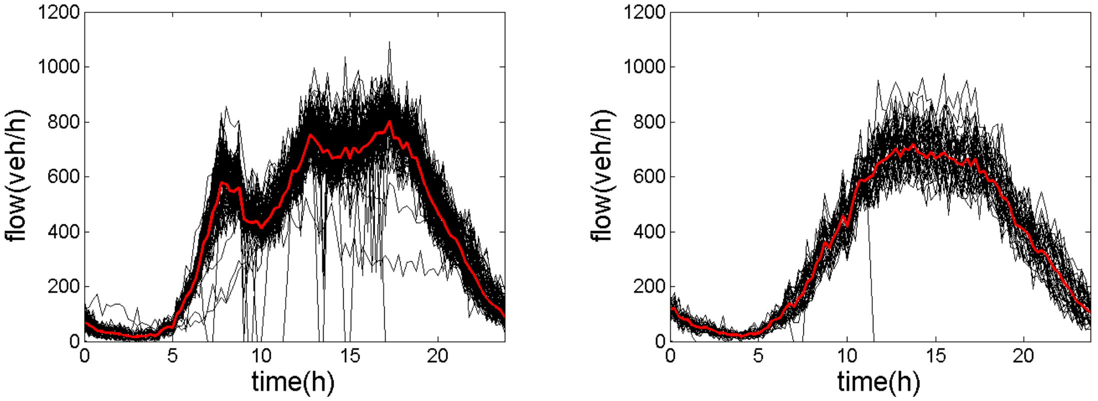
As mentioned before, travel-time is affected by factors such as capacity and demand. Considering that they are themselves non-Gaussian random quantities, and the system is inherently nonlinear, the travel-time estimation problem becomes analytically intractable. Therefore, in this paper, the Monte Carlo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



(a) Mainline, Weekdays

(b) Mainline, Weekends

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



(a) On-ramp, Weekdays

(b) On-ramp, Weekends

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, which makes 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.

EM Algorithm and Bayesian Inference Criterion

The shape of the travel-time probability distributions is not known a-priori. Based on the 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 represented as:

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

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,

$$\sum\nolimits_{i=1}^{K} \pi_i = 1 \tag{6}$$

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*). 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 $\{\gamma_i(x)\}, i = 1, 2, .., K$, where $\gamma_i(x)$ represents the conditional probability that the data x came from the ith cluster , given the current parameters $\{\mu_i, \Sigma_i, \pi_i\}$. That is,

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

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,

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

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

$$\pi_i = \frac{1}{N} \sum\nolimits_{j=1}^{N} \gamma_i(x_j) \tag{10}$$

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*).

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:

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

where $\ln P(D|\mu, \Sigma)$ is the log-likelihood function, D represents the samples, Q is the number degrees of of freedom (here since the travel-time has one freedom, $Q = 1$), K is the number of clusters, and N is the sample size. The optimal cluster number would generate maximum the 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 when doing Monte Carlo sampling, more samples provide more accurate results. However, the computation time of the simulation has 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 hading 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 the 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 considered to be GMM, the Online Expectation Maximization (On-line 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 newly coming samples x_i with $i = 1, 2, .., n$ 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 $y_i = L_0^{-1} x_i, i = 1, 2, .., n$, then the W statistic is represented as:

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

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

$$\frac{nWd}{2} \sim \chi^2_{d(d+1)/2} \tag{13}$$

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 newly coming samples $x_i, i = 1, 2, .., n$ have a mean μ_x, and a given target mean μ_0. The T^2 is defined as:

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

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

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

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

Since the 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 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, stop the simulation and return the parameters, otherwise go to step 2.

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 withs 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 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.

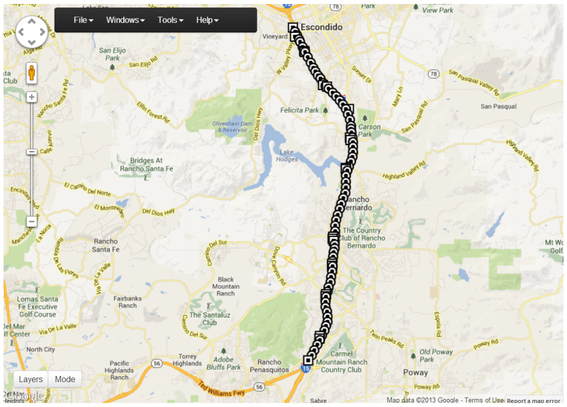


FIGURE 4 A Section of I-15 South

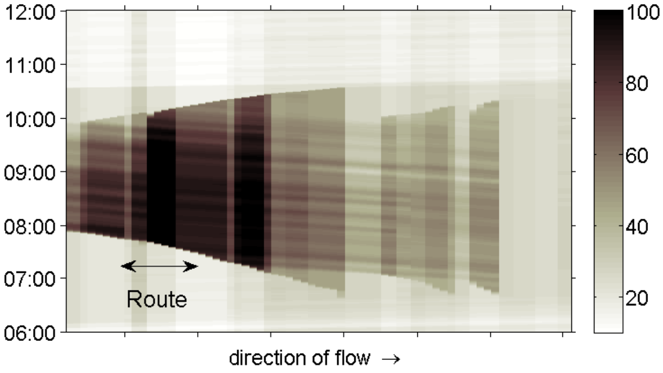


FIGURE 5 The Density Contour Plot

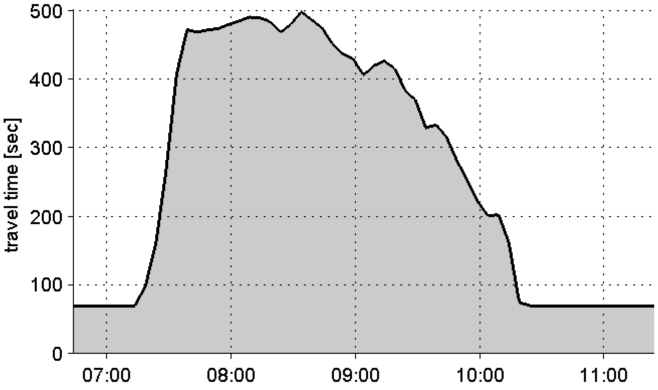
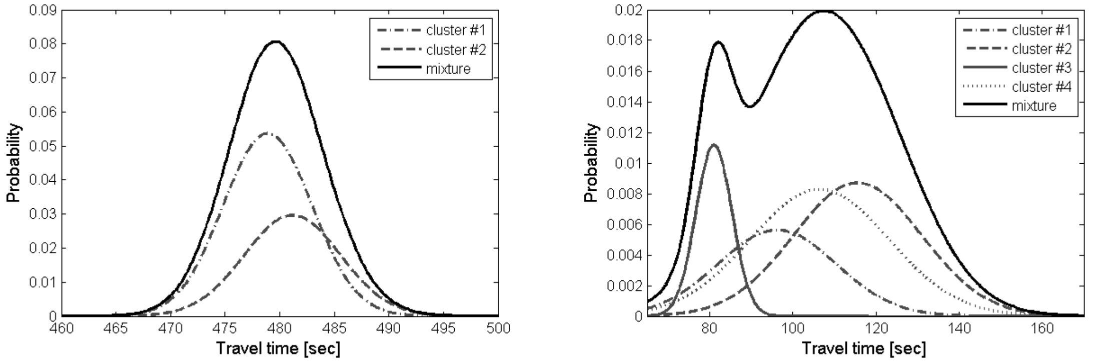


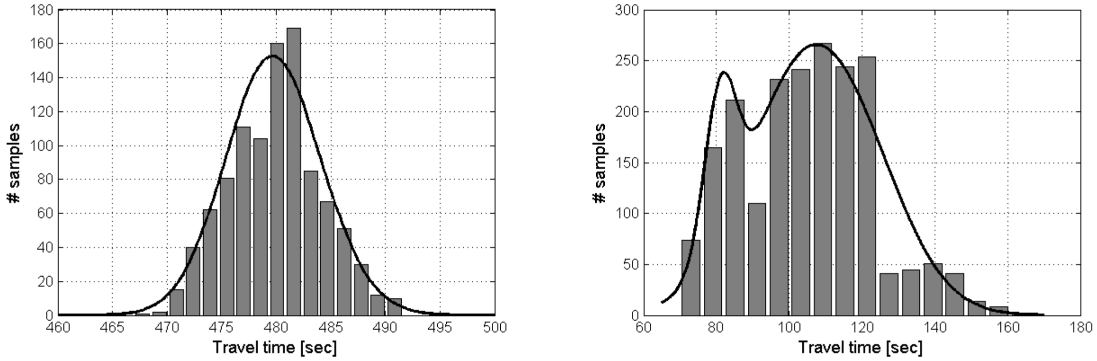
FIGURE 6 Deterministic Travel-time



(a) The first case (starting time=8:00:08 AM)

(b) The second case (starting time=7:18:23 AM)

FIGURE 7 Travel-time Distributions and Their Components



(a) The first case (starting time=8:00:08 AM)

(b) The second case (starting time=7:18:23 AM)

FIGURE 8 Travel-time Distributions and the Sample Histogram

The travel-time curve resulting form 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.

Results

Two starting times were selected. The first one is 8:00:08 AM. At this time, the route is heavily congested. Figure 7(a) shows the results. Although it looks like a Gassian 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 7:18:23 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 resents 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 stops 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 works well and requires 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 sware that at 7:20 AM the situation is less reliable and a wider range of outcomes are possible.

DISCUSSION

The travel-time prediction method is not only suitable for freeways, but also is suitable for arterials. since this method is based on LN-CTM, which also models arterial traffic. In arterials, because of the signal lights, the travel-time distribution will usually have multi-modal shape. In that case, the problem of minimizing the number of samples becomes essential. One of the direction of the work is to model more complicated arterial traffic, and predict travel-time distributions for urban drivers.

Another future work of this method is to calibrate the model states. Since a lot of ramp detector data are found to be partially or entirely missing, when modeling the traffic, estimation and imputation techniques have been used to solve the missing data problem. Currently it is difficult to calibrate the imputed data. On the other hand, increasing mobile technology allows us to obtain travel-time easier and more accurate by learning from probe data. In that case, suppose we could calibrate model states (density) through the travel-time, we could compare the predicted travel-time through this method with the real travel time through probe data. This 'closed-loop' mechanism definitely could help us to model the traffic more accurately.

EM v.s Variational Bayesian method: As the distribution becomes multimodal, EM may require longer time to be converged. In this case, alternative clustering methods such as Variational Bayesian (VB) method might be employed to reduce computation time. Early experimentation with VB suggests that it performs better than EM for larger number of clusters.

CONCLUSION

This paper developed a method to predict the travel-time distributions for freeways. We used the BeATS simulator, an implementation of the Link-Node Cell Transmission Model, to model traffic. Due to uncertainties on the input demands, the problem is multidimensional and could not be solved analytically. Monte Carlo method is introduced to obtain travel-time samples, and the Online Expectation Maximization algorithm is used to compute the Gaussian Mixture Model cluster parameters and decide when to stop the simulation. Experiments with data from I-15 southbound showed that the method could provide the travel-time distribution and requires small number of simulations. This method could be used for arterials as well as for freeways in the future.

ACKNOWLEDGEMENT

Any acknowledgments can be put here.

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 freeways. *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.