Path and Travel Time Inference from GPS Probe Vehicle Data

Article · January 2009

4 authors, including:

Ryan Herring
University of California, Berkeley
19 PUBLICATIONS   **2,044** CITATIONS

Alexandre M. Bayen
University of California, Berkeley
335 PUBLICATIONS   **10,536** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   Control of PDE-Modeled Traffic Flow Systems View project

Path and travel time inference from GPS probe vehicle data

Timothy Hunter
Department of Electrical Engineering
and Computer Science
University of California, Berkeley
tjhunter@eecs.berkeley.edu

Ryan Herring
Department of Industrial Engineering
and Operations Research
University of California, Berkeley
ryanherring@berkeley.edu

Pieter Abbeel
Department of Electrical Engineering
and Computer Science
University of California, Berkeley
pabbeel@cs.berkeley.edu

Alexandre Bayen
Systems Engineering
Department of Civil
and Environmental Engineering
University of California, Berkeley
bayen@berkeley.edu

Abstract

We consider the problem of estimating real-time traffic conditions from sparse, noisy GPS probe vehicle data. We specifically address arterial roads, which are also known as the secondary road network (highways are considered the primary road network). We consider several estimation problems: historical traffic patterns, real-time traffic conditions, and forecasting future traffic conditions. We assume that the data available for these estimation problems is a small set of sparsely traced vehicle trajectories, which represents a small fraction of the total vehicle flow through the network. We present an expectation maximization algorithm that simultaneously learns the likely paths taken by probe vehicles as well as the travel time distributions through the network. A case study using data from San Francisco taxis is used to illustrate the performance of the algorithm.

1 Introduction

Traffic congestion affects nearly everyone in the world due to the environmental damage and transportation delay it causes. The 2007 Urban Mobility Report [3] states that traffic congestion causes 4.2 billion hours of extra travel in the United States every year, which accounts for 2.9 billion extra gallons of fuel, which cost taxpayers an additional $78 billion. Providing drivers with good traffic information reduces the stress associated with congestion and allows drivers to make informed decisions, which generally increases the efficiency of the entire road network [5].

Modeling highway traffic conditions has been well-studied by the transportation community dating back to the pioneering work of Lighthill, Whitham and Richards [8, 9]. Recently, researchers demonstrated that estimating highway traffic conditions can be done using only GPS probe vehicle data [11, 7]. Arterial roads are major urban city streets that connect population centers within and between cities. Recent studies focusing on estimating real-time arterial traffic conditions have investigated traffic flow reconstruction for single intersections [4, 10] using dedicated traffic sensors. Dedicated traffic sensors are expensive to install, maintain and operate, which limits the amount of funding that governmental agencies can spend to deploy these sensors on the secondary network. The lack of sensor coverage across the arterial network thus motivates the use of GPS probe vehicle data for estimating traffic conditions.

The specific problem addressed in this article is how to extract travel time distributions from *sparse, noisy* GPS measurements from probe vehicles (section 2). A probabilistic model of travel times through the arterial network is presented along with an expectation maximization (EM) algorithm for learning the parameters of this model (section 3). We then extend this model to the case where the paths of the vehicles are unknown and we wish to infer the path taken (section 4). The goal is to identify historical traffic patterns upon which we build a real-time model for processing incoming data into estimates and forecasts of traffic conditions. Our initial results indicate that this is a promising approach (section 5).

(a) Real-time traffic conditions displayed through a Nokia E71 smart phone.

(b) An example of how the model results are visualized in the *Mobile Millennium* visualizer.
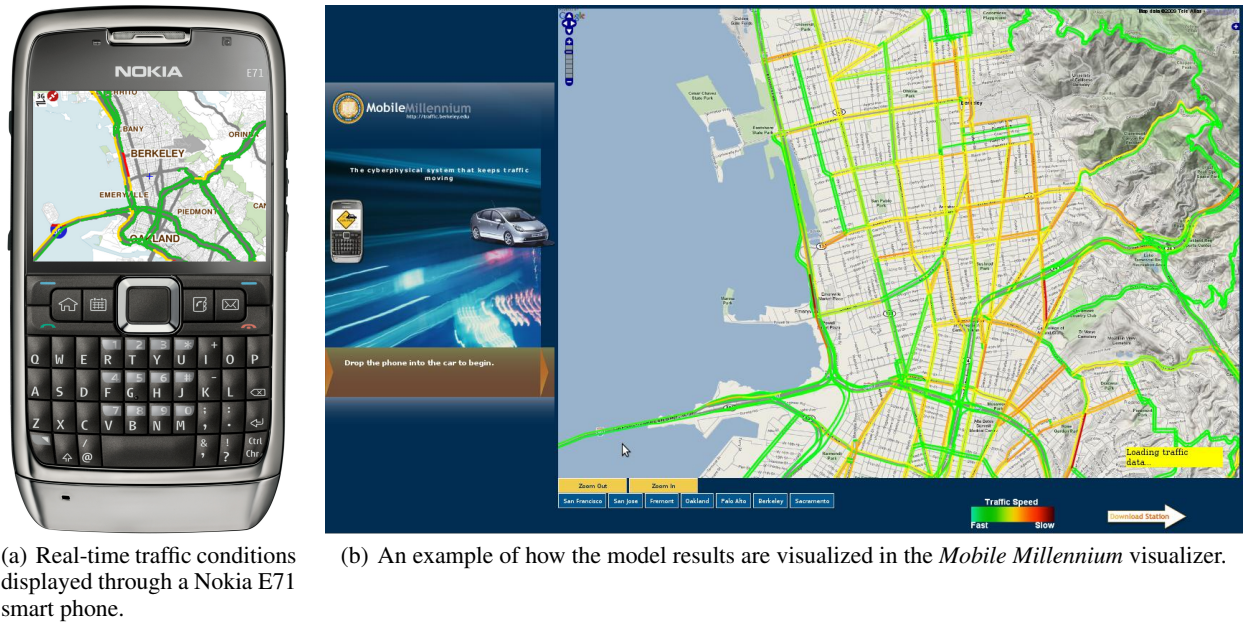
Figure 1: Visualizing traffic conditions from the *Mobile Millennium* system.

This work represents one component of the Berkeley-Nokia *Mobile Millennium* project [2]. The primary goal of the project is to assess the capability of GPS-enabled mobile phones to provide traffic data that can be used to estimate real-time conditions, forecast future conditions, and provide optimal routing in a time-varying stochastic network. The project has resulted in a real-time traffic estimation system that combines dedicated sensor data with GPS data from probe vehicles (figures 1(a) and 1(b)).

2   Problem Statement

The ability to estimate and forecast traffic conditions with limited amounts of streaming data relies on learning the general traffic patterns for a network. The goal of a "historical" model of traffic is to provide the recurring trend observed each day of the week at each time of the day. Therefore, our first objective is to develop an algorithm for processing all of the data previously collected into a form which accurately represents these trends.

Define the road network as a graph $\mathcal{D} = (\mathcal{L}, \mathcal{E})$, where the set $\mathcal{E}$ will be referred to as the "links" of the network and $\mathcal{L}$ as the "nodes" or "intersections". Let $\mathcal{C}$ be the set of day/time pairs for which we want to know the trend across the network. Our goal is then to estimate X^c, the joint link travel time distribution across all links in E, for all $c \in \mathcal{C}$. This historical estimate will then be used as a basis for estimating or forecasting the joint link travel time distribution at a current or future time.

In our framework, a vehicle observation r is defined as being two consecutively sampled locations from the same vehicle, which the two points having d^r time between them. Figure 2 illustrates the fact that some paths are easier to determine than others. Our model both infers the path and then extracts the relevant link travel times for the path chosen.

3   Probabilistic Model and Parameter Estimation Algorithm

We consider a road network as a set of directed links. The problem of interest is to predict the time it takes for a vehicle to travel across the links of the network. Four broad classes of influences on this travel time can be identified:

- Intrinsic characteristics of this link (the length of the link, the number of lanes, the presence of traffic signals, etc).
- Traffic behavior characteristics such as the congestion on this particular link or on neighboring links.
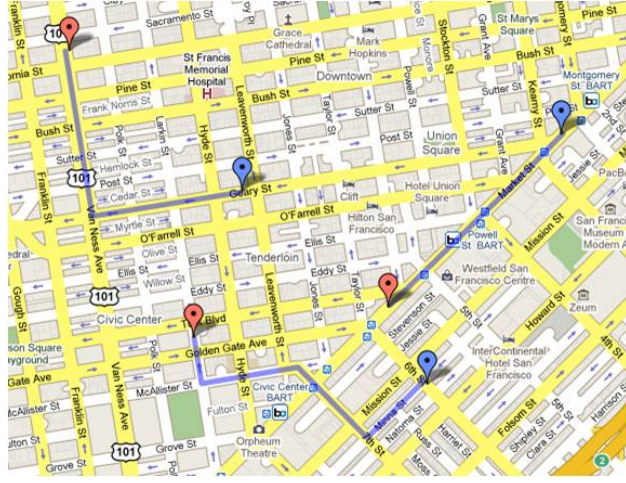
Figure 2: Probe vehicle observations with inferred paths. Each pair of points represents consecutive location samples from a vehicle and the lines represent the possible paths between those points.

- Vehicle-dependent characteristics. For example in many states in the U.S., a car making a right turn does not have to wait for the green light, and may go faster than cars going straight or making a left-turn.
- Exogenous conditions such as weather or sporting events can cause a part of the network to behave much differently than on a "typical day".

These and other influences make arterial traffic a complex stochastic process. From a probabilistic graphical model perspective, traffic estimation is a hard inference problem: if one assigns a random variable to each link that represents the travel time, then a network with tens of thousands of nodes in the case of a major city is not uncommon, and these variables have usually strong local correlation. This estimation and prediction problem should be solved in a matter of minutes for any practical purpose, which adds some additional constraints on the efficiency of the inference algorithm.

3.1 Basic generative model

In order to simplify the parameter estimation problem, we assume that each day can be decomposed into time intervals that share some common patterns on a day-by-day basis. For the rest of this section and section 4, we assume that the time interval c has been fixed and we remove the index to simplify the notation. The observations are samples of the trajectories from probe vehicles that span at most a few links per observation. In the example road network shown in figure 3(a), the nodes are intersections connected by road links, and we observe the travel time of a probe vehicle from intersection 1 to intersection 3 along the path $[1\rightarrow5\rightarrow3]$.

Formally, the road network is described as a directed graph $\mathcal{D} = (\mathcal{L}, \mathcal{E})$ where the set of vertices $\mathcal{L}$ is the set of links on the road network (i.e. each direction of the road). For our previous example, the path is represented as a list of links $[l_7, l_3]$. For each road link indexed by $l \in \mathcal{L}$, we call X_l the distribution of the travel time on this link. This joint distribution X is parameterized by a fixed set of hyperparameters P, which our goal is to estimate.

We are given R observations corresponding to this time interval. For each observation r, the travel time between the start point and the end point is denoted d^r. This observation is also associated to a particular trajectory along the links of the network $l_1^{(r)}, l_2^{(r)}, ..., l_{N_r}^{(r)}$. It can be assumed for our data set that vehicles do not have the opportunities to drive along a link twice or more times during an observation. With this hypothesis, we can represent the trajectory of a vehicle as a vector w of size L where $w_l = 1$ if the link was taken during this observation and 0 otherwise. The travel time D^r will be the random variable corresponding to the sum of the travel times on each link encountered along the path: $D^r = X_{l_1^{(r)}} + X_{l_2^{(r)}} + \ldots + X_{l_{N_r}^{(r)}}$. These variables D^r are the variables we observe. Using the vector w, the travel time over a path can be expressed in vector form: $D^r = (X)^T w^r$. The problem can be represented as a Bayesian network that includes the variables X^r, the observed variables D^r and the prior P that parameterizes the X^r. Our model is presented in figure 4(a) using a plate representation.

The joint probability of this distribution is: $p(\mathbf{X}, \mathbf{D}|P) = \prod_r f^r(d^r|X^r) g(X^r|P)$ with $\mathbf{X} = (X^r)_r$ and $\mathbf{D} = (D^r)_r$. The distribution g is based on our intuition of the traffic and models the travel times with respect to some set of

(a) Example of an observed trajectory in the road network.

(b) Example of a observed pair of start point and end point in the network. Two possible paths to consider are represented in dashed and bold lines.
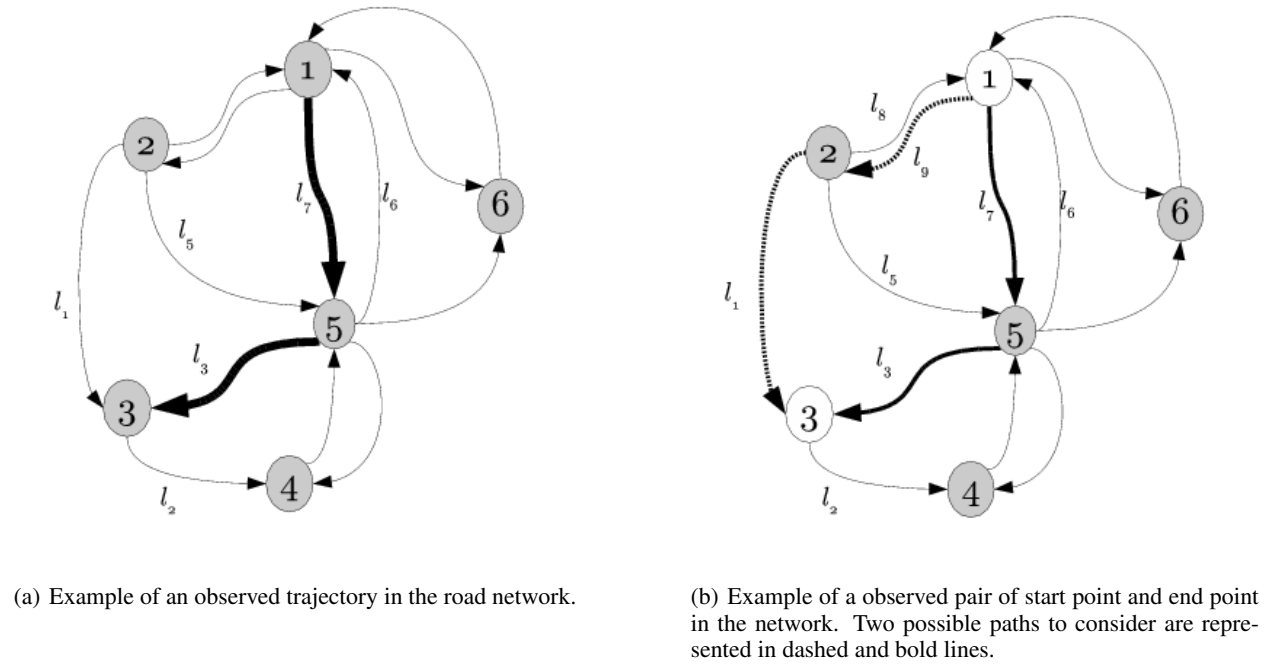
Figure 3: Illustration of known and unknown probe vehicle paths.

(a) Bayesian network representing our basic generative model.

(b) Graphical model representing a generic traffic network and our observations.

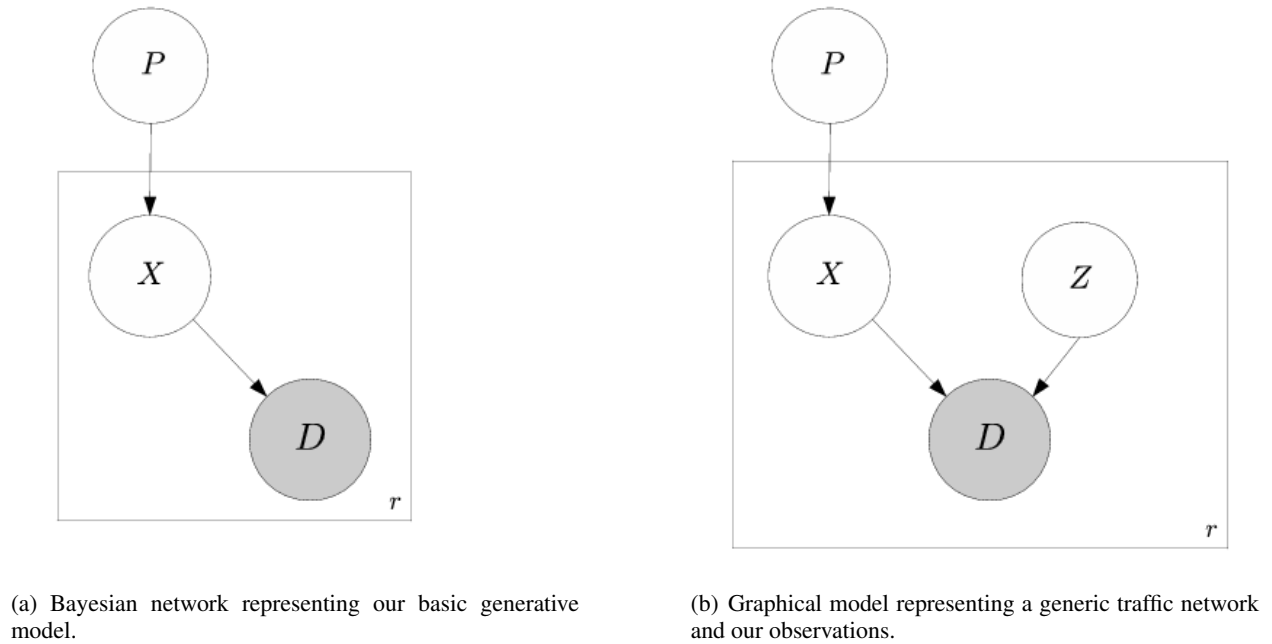Figure 4: Bayesian networks representing two models of travel times.

parameters. The functions $f^r \left(d^r | X^r \right) = \mathbf{1} \left\{ d^r = \left(X^r \right)^T w^r \right\}$ express the observations we make. One should note that with these definitions for f^r, the density function p is ill-defined (i.e. does not sum up to one) when using the usual Lebesgue measure over the space $\mathbf{X} \times \mathbf{D}$ because of the deterministic relationship. In the latter derivations, we will call ν an appropriate measure over $\mathbf{X} \times \mathbf{D}$.

Endowed with this probabilistic representation, the problem is:

$$
\begin{aligned}
\max_P ll \left(\mathbf{d}; P \right) &= \log \int_{\mathbf{X}} \nu \left(\mathrm{d}x \right) p \left(x, \mathbf{d} | P \right) \\
&= \sum_r \log \int_X g \left(x | P \right) \left[f^r \left(d^r | x \right) \nu \left(\mathrm{d}x \right) \right] \\
&= \sum_r \log \int_X g \left(x | P \right) \nu^r \left(\mathrm{d}x \right)
\end{aligned}
$$

where we introduce some measure ν^r that depends on the vector w^r and the observed travel time d^r.

3.2 Model learning for independent links

3.2.1 Optimization of a lower bound of the log-likelihood

In the general formulation presented in the previous section, we need to maximize P over an integral that is generally impractical to compute. Therefore, we propose to maximize a lower bound on the log-likelihood that will give convergence guarantees and works well in practice. To alleviate most of the computational complexity, we propose the use of a distribution g such that each link is independent from all other links. We will show how these simplifications lead to an elegant EM-like ascent algorithm.

We consider the lower bound to the maximum log-likelihood:

$$
L_b \left(\mathbf{d}; P \right) = \sum_r \int_X \log g \left(x | P \right) \nu^r \left(\mathrm{d}x \right)
$$

and we make this assumption that all the links have independent distributions parameterized by a set of features P_l: $g \left(x | P \right) = \prod_l h \left(x_l | P_l \right)$. Then L_b becomes:

$$
L_b \left(\mathbf{d}; P \right) = \sum_l \sum_r \int_X \log h \left(x_l | P_l \right) \nu^r \left(\mathrm{d}x \right)
$$

Thus an ascent algorithm can be written as follows:

- For all r and l, sample $\int_X \log h \left(x_l | P_l \right) \nu^r \left(\mathrm{d}x \right)$ and compute some sufficient statistics T_l^r for $\log h \left(x_l | P_l \right)$.
- For all l, find P_l that maximizes $\sum_r \int_X \log h \left(x_l | P_l \right) \nu^r \left(\mathrm{d}x \right)$ using the sufficient statistics computed in the step above
- Iterate the previous steps

This algorithm is guaranteed to improve L_b at each iteration, in the same way as the EM algorithm does. We detail in the next section the case of h in the exponential family.

3.2.2 The case of the exponential family

We consider the special case when h can be written $h \left(x_l | P_l \right) = s \left(x \right) \exp \left(P_l^T T \left(x_l \right) - A \left(P_l \right) \right)$ where T is the sufficient statistic and A is the cumulant function. We suppose that P_l is the canonical parameter of the distribution. Then the bound L_b becomes easy to maximize since:

$$
L_b \left(\mathbf{d}; P \right) = o \left(x \right) + \sum_l P_l^T \left(\sum_r \int_X T \left(x_l \right) \nu^r \left(\mathrm{d}x \right) \right) - R A \left(P_l \right)
$$

and the maximization step is done by solving the equation:

$$
\frac{\partial A}{\partial P_l} = \sum_r \int_X T \left(x_l \right) \nu^r \left(\mathrm{d}x \right)
$$

5

4 Extensions to the basic model - path uncertainty

In practice, we only observe the start point and the end point of the path as described in the previous section. For example, in figure 3(b) we observe a travel time from the intersection 1 to the intersection 3, for which at least two possible paths can be considered: $[l_9, l_1]$ and $[l_7, l_3]$. Suppose that for each observation, there are M_r possible paths to consider. We integrate this ambiguity in our model by introducing a latent variable Z^r that takes values in $\{1...M_r\}$ and that indicates which trajectory was taken by the vehicle. This variable is an additional prior to D^r in our Bayesian representation, as seen in figure 4(b).

Since the variable Z^r is a discrete latent variable, we can readily integrate our previous approach in an Expectation-Maximization approach. The complete log-likelihood is:

$$
\begin{aligned}
\max_P ll\left(\mathbf{d}; P\right) &= \sum_r \log \sum_{z^r=1}^{M_r} \int_X g\left(x|P\right)\left[f^r\left(d^r|x, z^r\right) p\left(z^r\right) \nu\left(\mathrm{d}x\right)\right] \\
&= \sum_r \log \sum_{i=1}^{M_r} p\left(Z^r = i\right) \int_X g\left(x|P\right) \nu^{r,i}\left(\mathrm{d}x\right)
\end{aligned}
$$

where $\nu^{r,i}$ extends our previous notation to that fact that ν depends now on the particular trajectory i followed during the observation r. The ascent algorithm derived for the case of independent links in section 2 becomes:

- For all r, i and l, sample $\int_X \log h\left(x_l|P_l\right) \nu^{r,i}\left(\mathrm{d}x\right)$ and compute some sufficient statistics T_l^r for $\log h\left(x_l|P_l\right)$. Update the distribution $\hat{p}\left(Z^r = i\right) \sim \int_X g\left(x|P\right) \nu^{r,i}\left(\mathrm{d}x\right)$.

- For all l, find P_l that maximizes $\sum_r \sum_{i=1}^{M_r} \hat{p}\left(Z^r = i\right) \int_X \log h\left(x_l|P_l\right) \nu^{r,i}\left(\mathrm{d}x\right)$ using the sufficient statistics computed in the step above

- Iterate the steps above

5 Preliminary Experimental Results

Our preliminary results show a lot of promise. We first introduce the data source used, namely samples from San Francisco taxi drivers. We validate our results by holding out a test data set and comparing our travel time estimates to those experienced in this test set.

5.1 San Francisco Taxi Data

For our study, we use the GPS data from San Francisco taxis (figure 5) as provided by the Cabspotting project [1]. The observed data are tuples of a start position on the network along with a time stamp and an end position with a time stamp. The travel time between the start and the end of each observation is around one minute, which is enough for a vehicle to travel up to five links during [low traffic hours] in our network.

The *Mobile Millennium* project has gathered two months of data, which represents about $60,000$ observations for an average of 50 taxi cabs a day. For each observation, the four shortest trajectories between the start point and the end point were selected as the the most probable ones. If the shortest trajectory is at least 50% shorter that the second one, it is decided that it was selected with probability one. Indeed, for a number of observations, it is obvious by looking at the start point and end point that the vehicle was driving along the same road. The heuristics explained above attempt to capture this fact and at the same time ensure that no valid trajectory would be discarded.

5.2 Model Results

There is precedent in the transportation community for using lognormal travel time distributions [6]. Each link distribution is assumed to be independent in first approximation. 90% of the samples were selected for training and 10% randomly held out for cross validation. We get the following preliminary results:

	Gaussian model	Lognormal model
Mean cross validation error	31sec	27sec

Figure 6 shows an example of the type of output produced by the model and illustrates the ability of the model to capture the increase in congestion during morning and afternoon peak congestion times.

6

Figure 5: One full day of GPS measurements from San Francisco taxis.
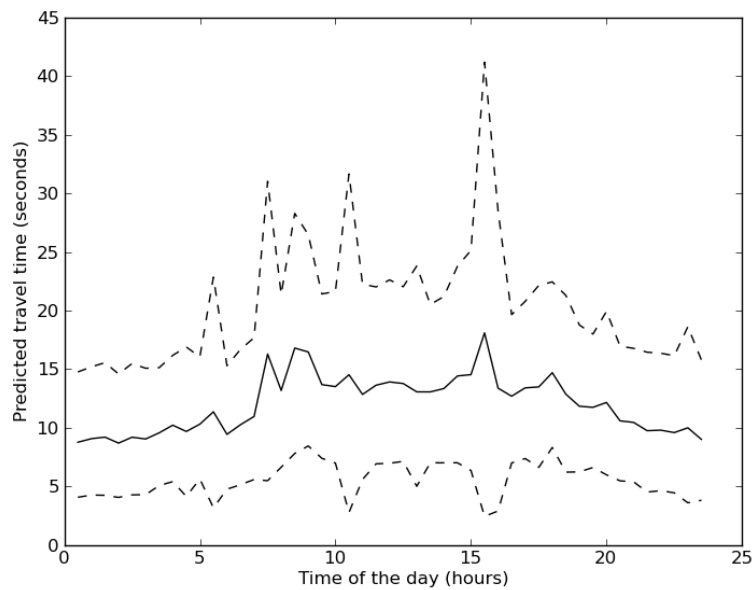

Figure 6: Example of a travel time prediction using the lognormal distribution. on a link in San Francisco (Franklin street at the intersection of Fulton street). This link is 120 meters long in a steep rise. Also plotted are the 90% quantiles. Note the traffic surge in morning and the afternoon.

6 Future Directions

This article described an EM algorithm for estimating historical link travel time distributions across an arterial road network. This is just the first step toward accomplishing the goals set out in the introduction. In particular, our future work will focus on how to use the historical model as the basis for a real-time model as well as a forecast model. We will soon implement an algorithm for estimating and forecasting traffic conditions by combining the historical model with real-time data. This represents a huge step forward in the transportation community for providing timely, accurate estimates of traffic conditions.

Additionally, our future work will focus on how to relax the assumption of independent link travel times. In reality, there are strong correlations between links and quantifying this correlation is essential for computing real-time stochastic shortest paths. This is a challenging topic with relatively little research to date.

References

[1] Cabspotting. http://www.cabspotting.org.

[2] The *Mobile Millennium* Project. http://traffic.berkeley.edu.

[3] TTI, Texas Transportation Institute: Urban Mobility Information: 2007 Annual Urban Mobility Report. http://mobility.tamu.edu/ums/.

[4] X. Ban, R. Herring, P. Hao, and A. Bayen. Delay pattern estimation for signalized intersections using sampled travel times. In *Proceedings of the 88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

[5] X. Ban, R. Herring, J.D. Margulici, and A. Bayen. Optimal sensor placement for freeway travel time estimation. In *Proceedings of the 18th International Symposium on Transportation and Traffic Theory, Hong Kong*, 2009.

[6] E. Emam and H. Al-Deek. Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board, No. 1959, Transportation Research Board of The National Academies*, pages 140–150, 2006.

[7] J.C. Herrera, D. Work, J. Ban, R. Herring, Q. Jacobson, and A. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century experiment. *Submitted to Transportation Research Part C*, December 2008.

[8] M. J. Lighthill and G. B. Whitham. On kinematic waves. II. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345, May 1955.

[9] P. Richards. Shock waves on the highway. *Operations Research*, 4(1):42–51, February 1956.

[10] A. Skabardonis and N. Geroliminis. Real-Time monitoring and control on signalized arterials. *Journal of Intelligent Transportation Systems*, 12(2):64–74, March 2008.

[11] D. Work, O.P. Tossavainen, S. Blandin, A. Bayen, T. Iwuchukwu, and K. Tracton. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 5062–5068, Cancun, Mexico, December 2008.