1

Traffic Flow on a Freeway Network

Peter Bickel[1] Chao Chen[2] Jaimyoung Kwon[1] John Rice[1] Pravin Varaiya[2] Erik van Zwet[1]

1.1 Introduction

Traffic congestion is an unpleasant fact of modern life. Although difficult to quantify precisely, congestion must cost Californians millions of dollars per day. Since further extensive construction of freeways is unlikely, information technology is being increasingly looked to for amelioration by providing information allowing more efficient use of existing freeways. Statistics plays a major role in such efforts.

A large interdisciplinary team of faculty, postdocs, graduate students, and undergraduates at the University of California, Berkeley, has been working on a host of problems of this kind. Researchers come from Computer Science, Electrical Engineering, Statistics, and Transportation Engineering.

This paper gives an overview of some of our activities, focusing on gathering statistics on traffic flow over the network of freeways in Los Angeles and on the prediction of travel times over this network. The paper is organized as follows: The freeway system of Los Angeles is equipped with a densely deployed array of sensors, loop detectors, which we describe in the next section. Information from these sensors is captured in real time, displayed, and archived by the Freeway Performance Measurement System, as described in a section 1.3. In section 1.4 we describe briefly our attempts to globally model the evolution of the fascinating spatial-temporal field of traffic flow. Ultimately, however, rather than trying to fit and update such comprehensive models, we found it preferable to use simpler, direct methods. These are described in section 1.5 for the purpose of predicting the particular functional of interest, travel time. Section 1.6 contains final remarks.

[1]Department of Statistics, University of California, Berkeley
[2]Department of Electrical Engineering and Computer Science, University of California, Berkeley

Figure 1.1. A set of double loop recorders.

1.2 Loop Detectors

Inductive loop detectors are the basic sensors monitoring the state of a freeway. A detector consists of a wire buried beneath the roadway. An alternating current generates an electromagnetic field, resulting in a change of inductance when an engine passes by on the surface. Such loops are located fairly densely on many freeway systems, with loops in each lane located in banks every half mile or so. Figure 1.1 shows a set of double loop recorders. Data from loops is usually sampled at rates ranging from 30 seconds to five minutes.

The fundamental variables that can be deduced from loops are flow (the number of vehicles per second) and occupancy (the percentage of time that vehicles are over the loops). The latter is essentially the density of vehicles. With assumptions about average vehicle length, these measurements can be converted to average velocity:

$$v(t) = g(t) \times \frac{c(t)}{o(t) \times T}. \tag{1.1}$$

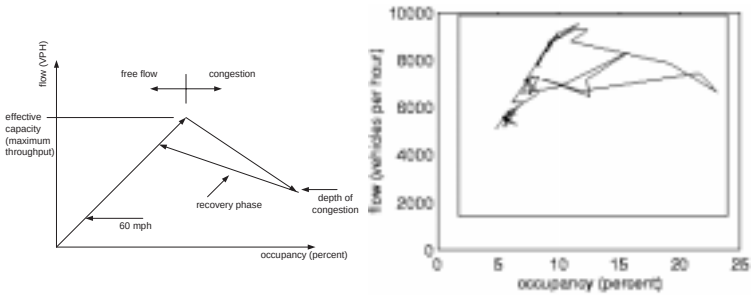
Here $c(t)$ is the flow, $o(t)$ is the occupancy, and $g(t)$ is the effective vehicle length during a time period of duration T. The effective vehicle length depends upon the mix of traffic (trucks and cars) and thus upon the lane and the time of day an also on the electronics of an individual loop. If loops are spaced in nearby pairs, velocity can be measured directly, but single loop detectors are more common.

Because of transmission problems, data from banks of loops are often missing. Furthermore loops may malfunction due to a number of causes including stuck sensors, hanging (on or off), chattering, and cross-talk, the mutual coupling of magnetic fields of two or more neighboring detectors.

1.3 Freeway Performance Measurement

The Freeway Performance Measurement System (PeMS) is an experimental project conducted by researchers at the University of California at Berkeley, with the cooperation of California Department of Transportation. The intent of this project is to collect historical and real-time data from freeways in the State of California, in order to compute freeway performance measures, thus providing managers with a comprehensive assessment of freeway performance. It also provides a wide variety of tools for transportation researchers to examine historical loop detector data. PeMS receives 30 second loop data in real time from Caltrans districts in California; about 1 gigabyte per day comes in from District 7, Los Angeles. From flow and occupancy records of single loop detectors, velocity is derived by using an adaptive algorithm [3] to estimate the effective vehicle length in 1.1 for each loop. Los Angeles has 4,000 detector at 1,300 locations, on 400 conventional highway miles. 70 % or more of the detectors are usually functional.

As an example of the kind of information available from PeMS, consider the following schematic representation of the classic traffic theory of the relationship of flow, density, and velocity, Figure 1.2. At low values of occupancy, traffic flows freely, as shown by the arrow emanating from the origin, and during this phase we have constant velocity (which is proportional to the slope of the line joining the point to the origin). Beyond some point of maximum efficiency, occupancy is sufficiently high so that the velocity decreases, and flow, the throughput of the system, decreases. Empirical versions of this diagram can be constructed from PeMS loop data as in Figure 1.2. Maximal efficiency of the system depends on keeping occupancy below the critical level, and this is the central aim of ramp metering.



(a) Classical flow-occupancy diagram.

(b) Empirical values of flow and occupancy at a particular loop.

Figure 1.2. The relationship between flow an occupancy.

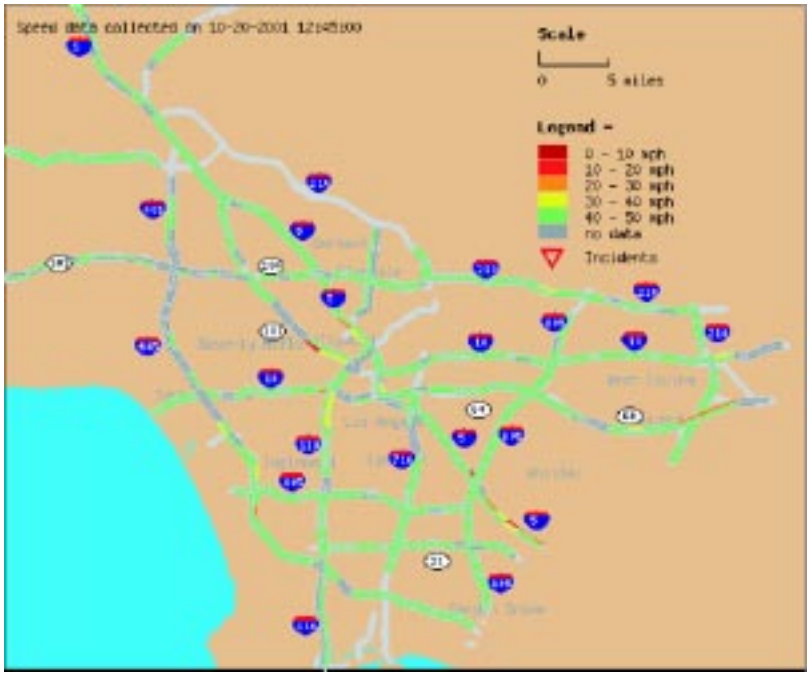


Figure 1.3. The Los Angeles freeways as currently displayed at the PeMS website.

PeMS can be interactively queried via a web browser.[1] A map of the entire freeway network can be displayed and updated every five minutes, Figure 1.3. These maps can also be played back in an animation, providing a vivid visualization of the propagation and dissipation of congestion.

In this paper, we focus particularly on one aspect of PeMS: travel time prediction. We describe the statistical methodology underlying an interface which allows a user to query the system for estimated travel times between locations selected by mouse-clicks, leaving at arbitrary times in the future. PeMS is also working on user interfaces via cell phones and via direct voice inquiry. In the not-too-distant future, continuously updated information on the state of the entire freeway system will be available to drivers as they negotiate it.

1.4 Global Models

Global models are inspired by the pattern of onset, propagation and dissipation of traffic congestion seen in Figures 1.4. Note in particular the

[1] htpp://transacct.eecs.berkeley.edu

characteristic wedge shapes reflecting the course of congestion from onset to dissipation. There are various theories [8] that try to explain such traffic dynamics, based on models of fluid flow, cellular automata, and microscopic computer simulations among others. Using such models for control and prediction is non-trivial, due to the complexity of the phenomenon.

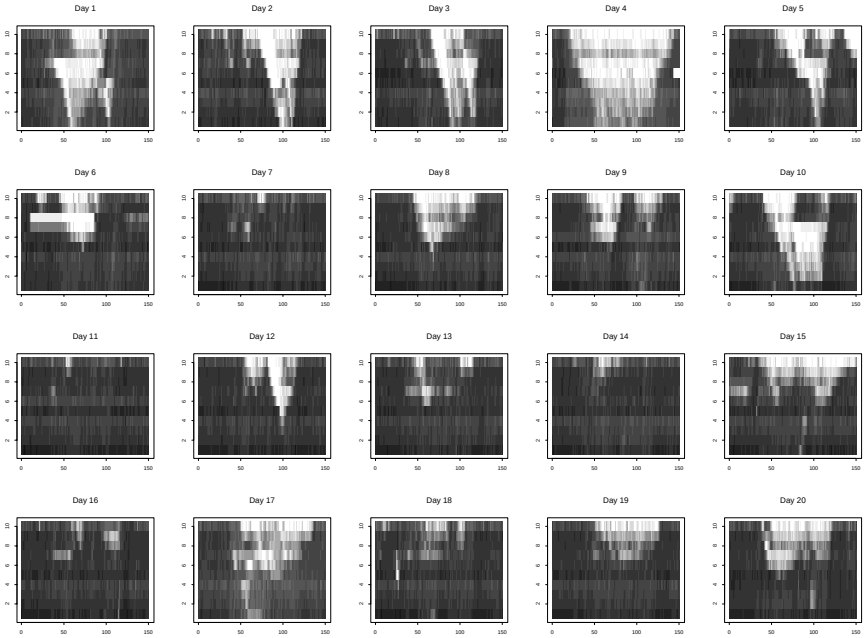


Figure 1.4. The velocity field for 20 weekdays, 2-7pm, between February 22 and March 19, 1993. The measurements come from 10 loop detectors (0.6 miles apart) in the middle lane of I-880 near Hayward, California, with 1 measurement per 2 minutes. The x-axis corresponds to time, and the y-axis to space. Vehicles travel upward in this diagram. The darkest gray-scale corresponds to the average velocity of 20 miles per hour (mph) and the lightest to 70 mph. The horizontally stretched bright blob on the left of the sixth day is due to a sensor failure.

1.4.1 A Coupled Hidden Markov Model

The model we consider, a coupled hidden Markov model (CHMM), is proposed as a phenomenological model for how the macroscopic dynamics of the freeway traffic arise from local interactions. This model views the velocity at each location as a noisy representation of the underlying binary traffic status (free flow or congestion) at that location and assumes the unobserved state vector is a Markov chain with a special structure of local dependencies. Such binary state assumption is justified from the clear

distinction between free flow and congestion regime shown in Figure 1.5, although a richer state space can be incorporated.

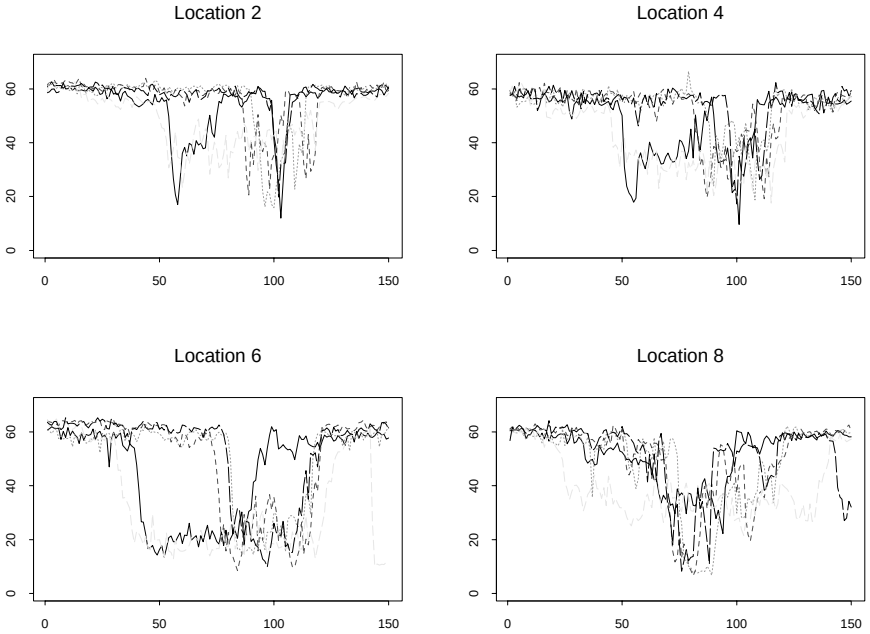


Figure 1.5. The velocity measurements for the first 5 days of Figure 1.4 at locations 2, 4, 6 and 8. The x-axis is in units of 2-minutes (0 to 150), the y-axis is mph and each line corresponds to a different day.

Consider a fixed day, d. Given all other variables, the observed velocity $y_{l,t}$ (mph) at location l ($l = 1, ..., L$, from upstream to downstream) and time t ($= 1, ..., T$) has a distribution depends only on the underlying state variable $x_{l,t} \in \{0, 1\}$. The two states 0 and 1 correspond to 'congestion' and 'free flow' each, by setting $E(Y_{l,t}|X_{l,t} = 0) < E(Y_{l,t}|X_{l,t} = 1)$. In particular, assume that the observed velocity is Gaussian whose mean and variance depends on the underlying state at the location, or

$$P_\lambda(y_{l,t}|x_{l,t} = s) \sim N(\mu_l^{(s)}, \sigma_l^{(s)2}), \quad s = 0, 1, \tag{1.2}$$

where $\lambda = (\mu_l^{(s)}, \sigma_l^{(s)2}, s = 0, 1)$ are parameters for the emission probability.

We assume the hidden process of $x_t = (x_{1,t}, \cdots, x_{L,t}) \in \{0, 1\}^{\otimes 2}$ is not only Markovian, i.e. $P(x_{t+1}|x_1, \cdots, X_t) = P(x_{t+1}|x_t)$, but also its transition probability allows the following decomposition.

$$P(x_{t+1}|x_t) = \prod_{l=1}^{L} P(x_{l,t+1}|x_t) = \prod_{l=1}^{L} P_\phi(x_{l,t+1}|x_{l-1,t}, x_{l,t}, x_{l+1,t}), \tag{1.3}$$

where ϕ specifies the transition probability and initial probability. This implies that the traffic condition at a location at time $t+1$ is affected only by the conditions at the neighboring locations at time t. This local decomposability assumption in space-time should be reasonable for certain spatial and temporal scales.

Then the complete likelihood of (x, y) is

$$P_\theta(x, y) = P_\phi(x)P_\lambda(y|x) = \prod_{i=1}^{n} P_\phi(x_{t+1}|x_t)P_\lambda(y_t|x_t).$$

where $\theta = (\phi, \lambda)$ is the parameter to estimate, and the likelihood of y is $P_\theta(y) = \sum_x P_\theta(x, y)$. This is the model for a single day, say d, and the extension to multiple days is apparent by viewing each day as an iid realization of this model. This is a hidden Markov model with a special structure and is called a *Coupled hidden Markov model* (CHMM; see [9]). See Figure 1.6 for the graphical model representation of the CHMM.

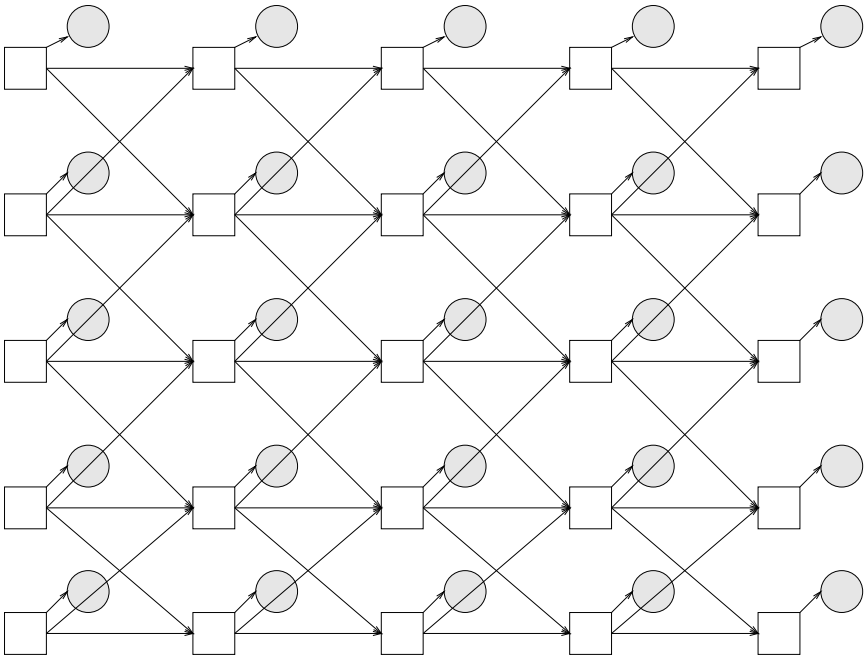


Figure 1.6. A coupled hidden Markov model represented as a dynamic Bayesian network. Square nodes represent discrete random variables (rv's) with multinomial distributions, round nodes represent continuous rv's with Gaussian distributions. Clear nodes are hidden, shaded nodes are observed. Here we show $L = 5$ chains and $T = 5$ timeslices.

1.4.2 Computation

For hidden Markov models, the parameters are usually estimated by the maximum likelihood method via the expectation-maximization (EM) algorithm which tries to find a local maximum of the likelihood. Even though the CHMM has moderately large number (of order $O(L)$) parameters, the model is still hard to fit because of the dimensionality. In particular, the E-step is in general computationally intractable while the M-step is straightforward. Sequential importance sampling with resampling (SISR; [6]) has been tried, leading to Monte Carlo EM (MC-EM) estimate of θ in which the exact E-step is replaced by an E-step approximated by the Monte Carlo sample from SISR. An alternative computational scheme, iterated conditional modes (ICM; [1]) has also been tried. For details about the computations, see [5] and [4].

1.4.3 Results and Conclusions

The approaches were applied to data collected from the I-880 freeway shown and explained in Figure 1.4. The ICM algorithm ran much faster than SISR (with 100 Monte Carlo samples), single iterations taking 1.63 and 417.73 seconds respectively on a 400MHz PC. Both algorithm seemed to stabilize after a few iterations, though the ICM algorithm, which is non-stochastic, seemed to fluctuate less. The latter however poses some conceptual difficulties.

Figure 1.7 shows the final estimates of μ and σ for the ten locations from the algorithms. Parameter estimates from the two algorithms are quite similar with each other. We can also observe that: (1)While mean free flow speed is similar (about 60 mph) for all locations, mean congestion speed varies greatly over locations, ranging from 20 to 50 mph. Location 8 where the mean congestion speed is the smallest, corresponds to the San Mateo Bridge, a notorious congestion spot, and (2) Standard errors of vehicle velocity are much larger for the congestion period than for the free flow period.

We simulated (binary) Markov chains using the parameters estimated by the algorithms to see qualitatively how well the fitted model reproduces the traffic dynamics. We can observe that inverted triangular regions of congestion are reproduced to some extent as bright patches in simulations from the SIS-estimated parameters, shown in Figure 1.8. Simulations using ICM results show similar behavior. Our initial hope for the CHMM model was to capture the behavior of the congestion and free flow regime and to reproduce pattern of propagation and dissipation of congestion. These goals have been achieved to a certain degree as illustrated above.

Given the success of such a global model, one might hope that it would prove useful in predicting future traffic patterns, including travel times which are only one of many possible functions of the predicted velocity

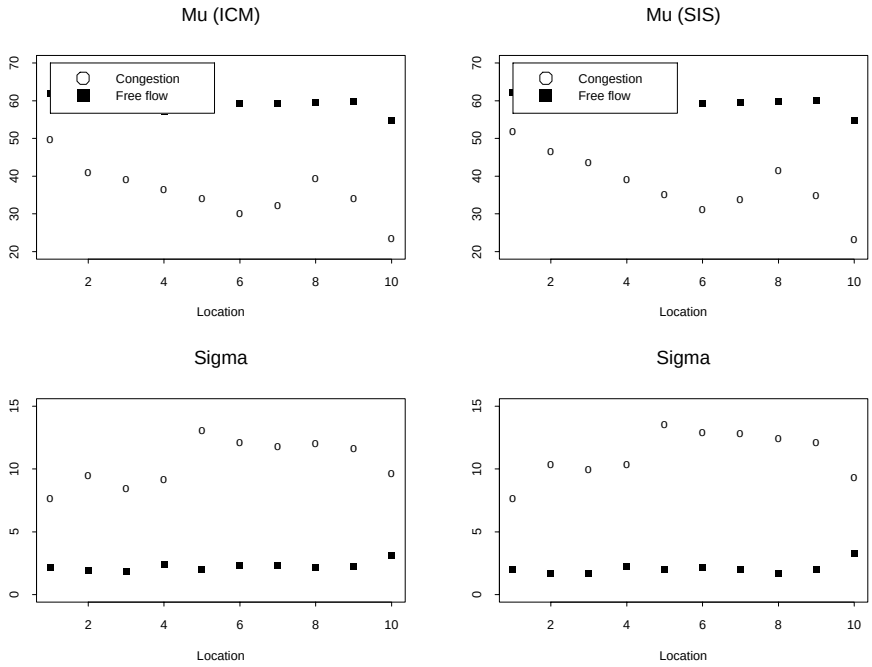


Figure 1.7. Conditional means and standard deviations of vehicle velocities for 10 locations estimated by the two algorithms.

vector. However, preliminary work on using the model for short term (5-30 minutes in the future) prediction of the velocity vector and travel time, produced disappointing initial results, comparable only to a naive predictor like the current velocity vector/travel time. This performance for prediction may be due to many possible reasons including: (1) The model does not capture the dynamics of incident propagation/dissipation in full; (2) The model fails to accommodate certain important features of the dynamics like a time-of-day factor, which would requires an inhomogeneous hidden Markov chain, (3) Having too many parameters, the model is subject to overfit, and (4) The model may be inherently too complicated and general for the specific task of travel time prediction.

To sum up, even though the proposed global model captures some interesting characteristic of the macroscopic dynamics of the freeway traffic, it is not very useful for travel time prediction. Further work is required to show whether it can be improved to outperform naive predictors or whether such a limitation is inherent in this kind of a global model.

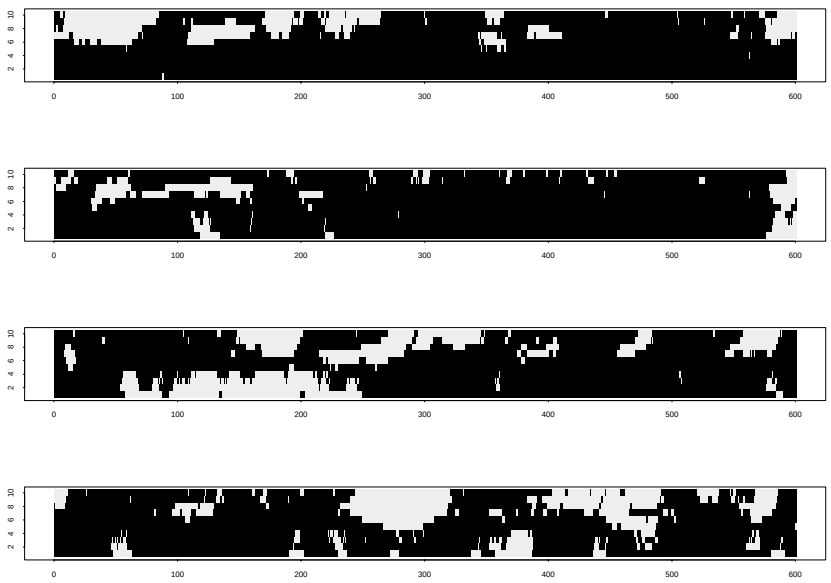


Figure 1.8. Four X fields simulated using the ϕ parameters estimated by the MC-EM algorithm using SIS. Light patches correspond to congestion and dark ones to free flow.

1.5 Travel Time Prediction

In this section we state the exact nature of our prediction problem. We then describe our prediction method and two alternative methods which will be used for comparison. This comparison is made in section 1.5.4 with a collection of 34 days of traffic data from a 48 mile stretch of I-10 East in Los Angeles, CA. Finally, in section 1.5.5, we summarize our conclusions, point out some practical observations and briefly discuss several extensions of our new method.

The data available for prediction can be represented as a matrix V with entries $V(d, l, t)$ $(d \in D,\ l \in L,\ t \in T)$ denoting the velocity that was measured on day d at loop l at time t. From V we can compute travel times $X_d(a, b, t)$, for all $d \in D$, $a, b \in L$ and $t \in T$. This travel time is to approximate the time it took to travel from loop a to loop b starting at time t on day d.

Suppose we have observed $V(d, l, t)$ for a number of days $d \in D$ in the past. Suppose a new day e has begun and we have observed $V(e, l, t)$ at times $t \leq \tau$. We call τ the 'current time'. Our aim is to predict $X_e(a, b, \tau+\delta)$ for a given (nonnegative) 'lag' δ. This is the time a trip that departs from a at time $\tau + \delta$ will take to reach b. Note that even for $\delta = 0$ this is not

trivial. In forming the prediction we have historical data on travel times and for the trip to be predicted we have data up to time τ. The predictor is some function of $V(d, l, t)$, $t \leq \tau$: the problem is to select such a function from this very high dimensional space.

We can compute a proxy for these travel times which is defined by

$$X_d^*(a, b, t) = \sum_{i=a}^{b-1} \frac{2d_i}{V(d, i, t) + V(d, i+1, t)}, \tag{1.4}$$

where d_i denotes the distance from loop i to loop $(i + 1)$. We call X^* the current status travel time (a.k.a. the snap-shot or frozen field travel time). It is the travel time that would have resulted from departure from loop a at time t on day d when no significant changes in traffic occurred until loop b was reached. It is important to notice that $X_d^*(a, b, t)$ can be computes at time t, whereas computation of $X_d(a, b, t)$ requires information of later times.

We fix an origin and destination of our travels and drop the arguments a and b from our notation. Define the historical mean travel time as

$$\mu(t) = \frac{1}{|D|} \sum_{d \in D} X_d(t). \tag{1.5}$$

Two naive predictors of $X_e(\tau + \delta)$ are $X_e^*(\tau)$ and $\mu(\tau + \delta)$. We expect—and indeed this is confirmed by experiment—that $X_e^*(\tau)$ predicts well for small δ and $\mu(\tau + \delta)$ predicts better for large δ. We aim to improve on both these predictors for all δ.

1.5.1 Linear Regression with Time Varying Coefficients

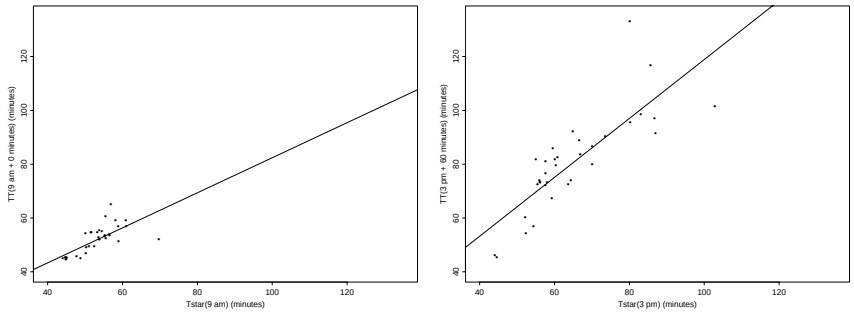
Our main result is the discovery of an empirical fact: that there exist linear relationships between $X^*(t)$ and $X(t + \delta)$ for all t and δ. This empirical finding has held up in all of numerous freeway segments in California that we have examined. This relation is illustrated by Figure 1.9, which shows scatter plots of $X^*(t)$ versus $X(t + \delta)$ for a 48 mile stretch of I-10 East in Los Angeles. Note that the relation varies with the choice of t and δ. With this in mind we propose the following model

$$X(t + \delta) = \alpha(t, \delta) + \beta(t, \delta)X^*(t) + \varepsilon. \tag{1.6}$$

where ε is a zero mean random variable modeling random fluctuations and measurement errors. Note that the parameters α and β are allowed to vary with t and δ. Linear models with varying parameters are discussed in [2].

Fitting the model to our data is a familiar linear regression problem which we solve by weighted least squares. Define the pair $(\hat{\alpha}(t, \delta), (\hat{\beta}(t, \delta))$ to minimize

$$\sum_{\substack{d \in D \\ s \in T}} (X_d(s) - \alpha(t, \delta) - \beta(t, \delta)X_d^*(t))^2 K(t + \delta - s), \tag{1.7}$$



(a) $X(9\text{ am} + 0\text{ min's})$ vs. $X^*(9\text{ am})$. Also shown is the regression line with slope α (9 am, 0 min)=0.65 and intercept β (9 am, 0 min)=17.3.

(b) $X(3\text{ pm} + 60\text{ min})$ vs. $X^*(3\text{ pm})$). Also shown is the regression line with slope α (3 pm, 60 min)=1.1 and intercept β (3 pm, 60 min)=9.5.

Figure 1.9. Scatter plots of actual travel times, $X(t)$, versus "frozen field" travel times, $X^*(t)$.

where K denotes the Gaussian density with mean zero and a certain variance which must be specified. The purpose of this weight function is to impose smoothness on α and β as functions of t and δ. We assume that α and β are smooth in t and δ because we expect that average properties of the traffic do not change abruptly. The actual prediction of $X_e(\tau + \delta)$ becomes

$$\widehat{X}_e^{\alpha\beta}(\tau + \delta) = \hat{\alpha}(\tau, \delta) + \hat{\beta}(\tau, \delta) X_e^*(\tau). \tag{1.8}$$

Writing $\alpha(t, \delta) = \alpha'(t, \delta) \times \mu(t + \delta)$ we see that (1.6) expresses a future travel time as a linear combination of the historical mean and the current status travel time—our two naive predictors. Hence our new predictor may be interpreted as the best linear combination of our naive predictors. From this point of view, we can expect our predictor to do better than both. In fact, it does, as is demonstrated in section 1.5.4.

Another way to think about (1.6) is by remembering that the word "regression" arose from the phrase "regression to the mean." In our context, we would expect that if X^* is much larger than average—signifying severe congestion—then congestion will probably ease during the course of the trip. On the other hand, if X^* is much smaller than average, congestion is unusually light and the situation will probably worsen during the journey.

Besides comparing our predictor to the historical mean and the current status travel time, we subject it to a more competitive test. We consider two other predictors that may be expected to do well—one resulting from principal component analysis and one from the nearest neighbors principle. Next, we describe these two methods.

1.5.2 Principal Components

Our predictor $\widehat{X}^{\alpha\beta}$ only uses information at one time point; the 'current time' τ. However, we do have information prior to that time. The following method attempts to exploit this by using the entire trajectories of X_e and X_e^* which are known at time τ.

Formally, let us assume that the travel times on different days are independently and identically distributed and that $\{X_d(t) : t \in T\}$ and $\{X_d^*(t) : t \in T\}$ are jointly multivariate normal. We estimate the covariance of this multivariate normal distribution by retaining only a few of the largest eigenvalues in the singular value decomposition of the empirical covariance of $\{(X_d(t), X_d^*(t)) : d \in D,\ t \in T\}$. We have experimented informally with the number of eigenvalues to retain, but one could also use cross-validation. Define τ' to be the largest t such that $t + X_e(t) \leq \tau$. That is, τ' is the latest trip that we have seen completed before time τ. With the estimated covariance we can now compute the conditional expectation of $X_e(\tau + \delta)$ given $\{X_e(t) : t \leq \tau'\}$ and $\{X_e^*(t) : t \leq \tau\}$. This is a standard computation which is described, for instance, in [7]. The resulting predictor is $\widehat{X}_e^{\mathrm{PC}}(\tau + \delta)$.

1.5.3 Nearest Neighbors

As an alternative to principal components, we now consider nearest neighbors, which is also an attempt to use information prior to the current time τ. This method is nonlinear and makes fewer assumptions (such as joint normality) on the relation between X^* and X.

The method of nearest neighbors aims to find those days in the past which are most similar to the present day in some appropriate sense. The remainder of those past days beyond time τ are then used to form a predictor of the remainder of the present day.

The critical choice with nearest neighbors is in specifying a suitable distance m between days. We suggest two possible distances:

$$m(e, d) = \sum_{l \in L,\ t \leq \tau} |V(e, l, t) - V(d, l, t)| \tag{1.9}$$

and

$$m(e, d) = \left(\sum_{t \leq \tau} (X_e^*(t) - X_d^*(t))^2 \right)^{1/2}. \tag{1.10}$$

Now, if day d' minimizes the distance to e among all $d \in D$, our prediction is

$$\widehat{X}_e^{NN}(\tau + \delta) = X_{d'}(\tau + \delta). \tag{1.11}$$

Sensible modifications of the method are 'windowed' nearest neighbors and k-nearest neighbors. Windowed-NN recognizes that not all information prior to τ is equally relevant. Choosing a 'window size' w it takes the above summation to range over all t between $\tau - w$ and τ. So-called k-NN is basically a smoothing method, aimed at using more information than is present in just the single closest match. For some value of k, it finds the k closest days in D and bases a prediction on a (possibly weighted) combination of these. Alas, neither of these variants appear to significantly improve on the 'vanilla' $\widehat{X}^{NN}$.

1.5.4 Results

We gathered flow and occupancy data from 116 single loop detectors along 48 miles of I-10 East in Los Angeles. Measurements were done at 5 minute aggregation at times t ranging from 5 am to 9 pm for 34 weekdays between June 16 and September 8 2000. We used the method described in [3] to convert flow and occupancy to velocity. Fortunately, the quality of our I-10 data is quite good and we used simple interpolation to impute wrong or missing values. The resulting velocity field $V(d, l, t)$ is shown in figure 1.10 where day d is June 16. The horizontal streaks typically indicate detector malfunction.

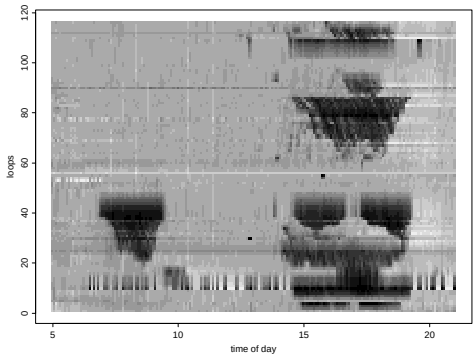


Figure 1.10. Velocity field $V(d, l, t)$ where day d = June 16, 2000. Darker shades indicate lower speeds. Note the typical triangular shapes indicating the morning and afternoon congestions building and easing. The horizontal streaks are most likely due to detector malfunction.

From the velocities we computed travel times for trips starting between 5 am and 8 pm. Figure 1.11 shows these $X_d(t)$ where time of day t is on the horizontal axis. Note the distinctive morning and afternoon congestions and the huge variability of travel times, especially during those periods. During afternoon rush hour we find travel times ranging from 45 minutes

to up to two hours. Included in the data are holidays July 3 and 4 which may readily be recognized by their very fast travel times.

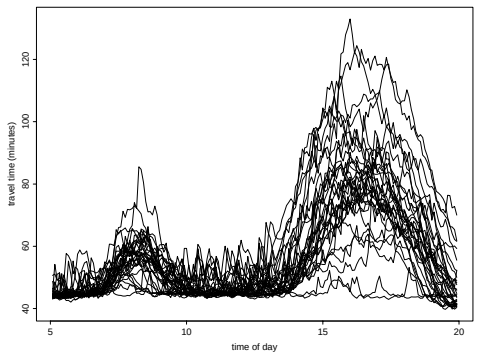


Figure 1.11. Travel Times $X_d(\cdot)$ for 34 days on a 48 mile stretch of I-10 East.

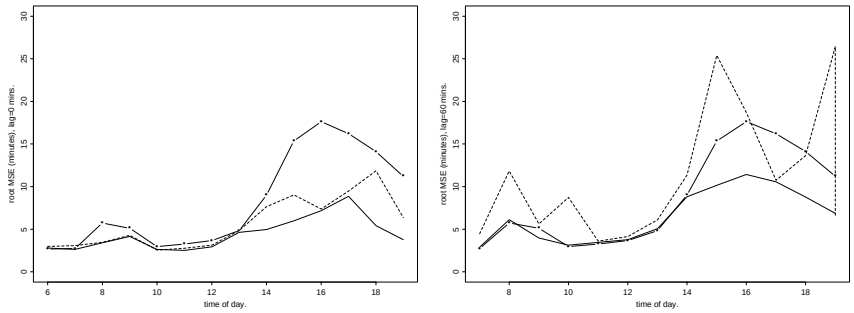
We have estimated the root mean squared error of our various prediction methods for a number of 'current times' τ (τ =6am, 7am,...,7pm) and lags δ (δ =0 and 60 minutes). We evaluated the estimation error by leaving out one day at a time, performing the prediction for that day on the basis of the remaining other days and averaging the squared prediction errors.

The prediction methods all have parameters that must be specified. For the regression method we have chosen the standard deviation of the Gaussian kernel K to be 10 minutes. For the principal components method we have chosen the number of eigenvalues retained to be 4. For the nearest neighbors method we have chosen distance function (1.10), a window w of 20 minutes and the number k of nearest neighbors to be 2.

Figure 1.12 shows the estimated root mean squared (RMS) prediction error of the historical mean $\mu(\tau + \delta)$, the current status predictor $X_e^*(\tau)$ and our regression predictor (1.8) for lag δ equal to 0 and 60 minutes, respectively. Note how $X_e^*(\tau)$ performs well for small δ ($\delta = 0$) and how the historical mean does not become worse as δ increases. Most importantly, however, notice how the regression predictor beats both hands down, except during the time period of 10am-1pm.

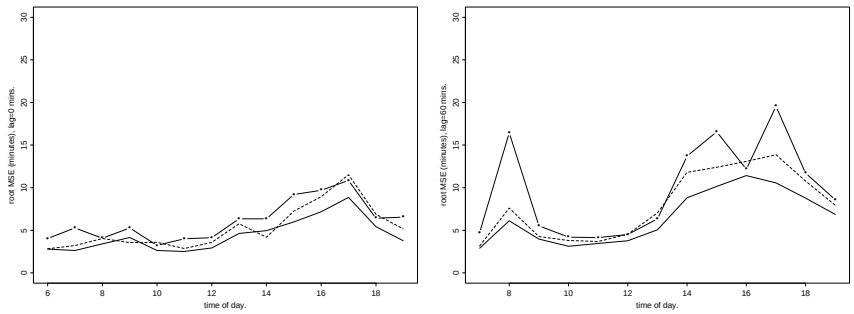
Figure 1.13 again shows the RMS prediction error of the regression estimator. Here it is compared to the principal components predictor and the nearest neighbors predictor (1.11). Again, the regression predictor comes out on top, although the nearest neighbors predictor shows comparable performance.

The RMS error of the regression predictor stays below 10 minutes even when predicting an hour ahead. We feel that this is impressive for a trip of 48 miles right through the heart of L.A. during rush hour.



(a) Estimated RMSE, lag=0 min. (b) Estimated RMSE, lag=60 min.

Figure 1.12. Comparison of Historical mean (– · –), current status (- - -) and linear regression (—).



(a) Estimated RMSE, lag=0 min. (b) Estimated RMSE, lag=60 min.

Figure 1.13. Comparison of Principal Components (– · –), nearest neighbors (- - -) and linear regression (—).

1.5.5 Conclusions and loose ends

We stated that the main contribution towards predicting travel times is the discovery of a linear relation between $X^*(t)$ and $X(t+\delta)$. But there is more. Comparison of the regression predictor to the principal components and nearest neighbors predictors unearthed another surprise. Given $X^*(\tau)$, there is not much information left in the earlier $X^*(t)$ $(t < \tau)$ that is useful for predicting $X(\tau+\delta)$. Some earlier attempts [5] at prediction using more complex models also turned out to be inferior to the regression method. In fact, we have come to believe that for the purpose of predicting travel

times all the information in $\{V(l,t),\ l \in L,\ t \leq \tau\}$ is well summarized by one single number: $X^*(\tau)$.

It is of practical importance to note that our prediction can be performed in real time. Computation of the parameters $\hat{\alpha}$ and $\hat{\beta}$ is time consuming but it can be done off-line in reasonable time. The actual prediction is trivial. In addition to making predictions available on the internet, it would also be possible to make them available for users of cellular telephones—and in fact we plan to do so in the near future.

It is also important to notice that our method does not rely on any particular form of data. In this paper we have used single loop detectors, but probe vehicles or video data can be used in place of loops, since all the method requires is current measurements of X^* and historical measurements of X and X^*. It is straightforward to make the method robust to outliers by replacing the least squares refleast-squares criterion by a robust one.

We conclude this paper by briefly pointing out two extensions of our prediction method.

1. For trips from a to c *via* b we have

$$X_d(a,c,t) = X_d(a,b,t) + X_d(b,c,t + X_d(a,b,t)). \tag{1.12}$$

We have found that it is sometimes more practical or advantageous to predict the terms on the right hand side than to predict $X_d(a,c,t)$ directly. For instance, when predicting travel times across networks (graphs), we need only predict travel times for the edges and then use (1.12) to piece these together to obtain predictions for arbitrary routes. This is precisely what is done in forming a prediction for a trip over the LA freeway network. In this way, a complex non-linear predictor is formed by composition of simpler linear ones.

2. We regressed the travel time $X_d(t+\delta)$ on the current status $X_d^*(t)$, where $X_d(t+\delta)$ is the travel time departing at time $t+\delta$. Now, define $Y_d(t)$ to be the travel time *arriving* at time t on day d. Regressing $Y_d(t+\delta)$ on $X_d^*(t)$ will allow us to make predictions on the travel time subject to *arrival* at time $t+\delta$. The user can thus ask what time he or she should depart in order to reach his or her intended destination at a desired time.

1.6 Final Remarks

The complexity of traffic flow over the Los Angeles network is bewitching. In some intriguing fashion, microscopic interactions give rise to global patterns. Tempting as it is to model this process, it is not at all clear *a priori* that doing so is the best way to achieve accurate prediction of a particular

functional, travel time in our case. Indeed, simpler, empirical methods such as those we have discussed in this paper appear to be more effective.

Our results are hardly free of blemishes and warts. Probably the greatest problem we face is the variable quantity and quality of data from loop detectors. At any given time 20% or so of the loops may not report at all. Some of those that do report give erroneous results. This poses fascinating challenges for development of statistical methodology: stated abstractly, we have a large array of often faulty or non-reporting sensors and we wish to reconstruct the random field that is driving them. We are actively working on this problem, trying to take advantage of the high correlations between nearby sensors induced by the fact that they are measuring related aspects of a common random environment. A second problem we face is the lack of ground truth in District 12. We have tested our prediction method on a smaller, higher quality data set that includes dense coverage by probe vehicles, with good results [10].

The segmentation of trips into journeys over edges of the freeway network poses some interesting problems. First, there is the question of how to segment. We have taken taken the pragmatic approach of choosing to segment using nodes formed by the major freeway intersections, but there are certainly other possibilities, given that there are loops every half mile or so. Second, there are issue of internal consistency. Consider a trip from point A to point C passing through point B. We could use the regression method to predict the travel time from A to B and then the subsequent time from B to C, or we could use the method to directly predict the travel time from A to C. There is no guarantee that the two predictions will be identical. Adding to this the possibility of using the regression method to predict backwards in time, predicting when to leave A in order to arrive at C at a desired time, and there is no guarantee that we have not created wormholes in the space-time fabric of the Los Angeles freeway network.

Our group is engaged in other interesting activities. Foremost among these is a project in which an array of video cameras will survey a stretch of freeway near the San Francisco Bay Bridge in order to study the behavior of individual drivers and the microscopic causes of congestion. The first problem encountered in this effort is identifying vehicles on the videos in order to extract their trajectories—a non-trivial challenge in computer vision. Beyond that is the challenge of formulating effective statistical procedures to study the large quantity of data from measurements of this complex spatio-temporal stochastic process.

References

[1] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

[2] T. Hastie and R. Tibshirani. Varying coefficient models. *Journal of the Royal Statistical Society Series B*, 55(4):757–796, 1993.

[3] Z. Jia, C. Chen, B. Coiffman, and P. Varaiya. The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. In *Fourth International IEEE Conference on Intelligent Transportation Systems*.

[4] J. Kwon and K. Murphy. Modeling freeway traffic with coupled HMMs. 2000.

[5] Jaimyoung Kwon, B. Coifman, and Peter J. Bickel. Day-to-day travel time trends and travel time prediction from loop detector data. In *Transportation Research Record*, 2000. Accepted for publication.

[6] J. S. Liu and R. Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[7] K. V. Mardia, J. T. Kent, and S. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[8] A. D. May. *Traffic Flow Fundamentals*. Prentice-Hall, 1990.

[9] L. K. Saul and M. Jordan. Boltzman chains and hidden markov models. In G. Tesauro, D. S. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*. MIT Press, 1995.

[10] X. Zhang and J. Rice. Short term travel time prediction. *Transportation Research C*, 2002. to appear.