A Simple and Effective Method for Predicting Travel Times on Freeways

John Rice and Erik van Zwet

Abstract—**We present a method to predict the time that will be needed to traverse a given section of a freeway when the departure is at a given time in the future. The prediction is done on the basis of the current traffic situation in combination with historical data. We argue that, for our purposes, the current traffic situation of a section of a freeway is well summarized by the *current status travel time*. This is the travel time that would result if one were to depart immediately and no significant changes in traffic would occur. This current status travel time can be estimated from single- or double-loop detectors, video data, probe vehicles, or any other means. Our prediction method arises from the empirical observation that there exists a linear relationship between any future travel time and the current status travel time. The slope and intercept of this relationship may change subject to the time of day and the time until departure, but linearity persists. This observation leads to a prediction scheme by means of linear regression with time-varying coefficients.**

Index Terms—**Intelligent transportation systems (ITS), linear regression, prediction, travel time, varying coefficients.**

I. Introduction

THE PERFORMANCE measurement system (PeMS) [1] is a large-scale freeway data collection, storage, and analysis project. It involves the Electrical Engineering and Computer Science (EECS) and Statistics Departments and the Institute of Transportation Studies, University of California, Berkeley, in cooperation with the California Department of Transportation (CalTrans). PeMS's goal is to facilitate traffic research and assist CalTrans by quantifying the performance of California's freeways. Useful information in various forms is to be distributed among traffic managers, planners, and engineers; freeway users; and researchers. In real time, PeMS obtains loop-detector data on flow (count) and occupancy at selected locations, aggregated over 30-s intervals. For all of California, this amounts to 2 GB/d. In its raw form, this data is of little use.

In this paper, we focus our attention on travel-time prediction between any two points on a freeway network for any future departure time. Besides being useful *per se*, travel-time prediction serves as an input or aid to dynamic route guidance [2], congestion management, and control [3]; optimal routing and dispatching [4], [5]; and incident detection [6].

Manuscript received January 4, 2002; revised August 16, 2003 and February 9, 2004. The performance measurement system (PeMS) project is supported by grants from the National Science Foundation and from the California Department of Transportation (CalTrans) through the California PATH program. The Associate Editor for this paper was M. Papageorgiou.

J. Rice is with the Department of Statistics, University of California, Berkeley, CA 94720 USA (e-mail: rice@stat.berkeley.edu).

E. van Zwet is with the Mathematical Institute, University of Leiden, Leiden 2300 RA, The Netherlands (e-mail: evanzwet@math.leidenuniv.nl).

Digital Object Identifier 10.1109/TITS.2004.833765

We want our methods to be simple, fast, and scalable. We are currently developing an Internet application that will give the commuters of CalTrans District 7 (Los Angeles) the opportunity to query the prediction algorithm that is described in this paper. The user will access our Internet site and state the origin, destination, and time of departure (or desired time of arrival). He or she will then receive a prediction of the travel time and the best (fastest) route to take. To make useful predictions in a rapidly changing environment, it is clear that we need to be able to very quickly process very large amounts of data. Complex algorithms will not do for our purpose.

In Section II, we state the exact nature of our prediction problem. We then describe our new prediction method ("linear regression") and two alternative methods that will be used for comparison. This comparison is made in Section III with a collection of 34 d of traffic data from a 48-mi stretch of I-10 East in Los Angeles, CA. Finally, in Section IV, we summarize our conclusion, note some practical observations, and briefly discuss several extensions of our new method.

II. Methods of Prediction

Consider an array $v(d,l,t)$ ($d \in D,\ l \in L,\ t \in T$) denoting the velocity that was measured on day d at loop l at time t. In Fig. 3, we see an example of a velocity field for one day. From v, we can approximate the time $T(d,t)$ needed to travel from loop 1 to loop L starting on d at time t. This travel time can be thought of as belonging to a path through the velocity field. It is important to note that in order to actually compute $T(d,t)$ we need information after time t. Using information available at time t, we can compute a proxy for the travel time defined as

$$T^*(d,t) = \sum_{l=1}^{L-1} \frac{2d_l}{v(d,l,t) + v(d,l+1,t)} \tag{1}$$

where d_l denotes the distance from loop l to loop $(l+1)$. We call T^* the instantaneous or current status travel time. It is the travel time that would have resulted from the departure from loop 1 at time t on day d when no significant changes in traffic occurred until loop L was reached. If we have computed $T(d,t)$ for a collection D of days in the past, we then can also compute the average historical travel time as

$$\bar{T}(t) = \frac{1}{|D|} \sum_{d \in D} T(d,t). \tag{2}$$

Our goal is to predict $T(d, t+\delta)$ for $\delta \geq 0$ on the basis of the available information on day d at time t. We call δ the "time lag" and note that even for $\delta = 0$ this problem is not trivial. We are certainly not the first to study this problem.

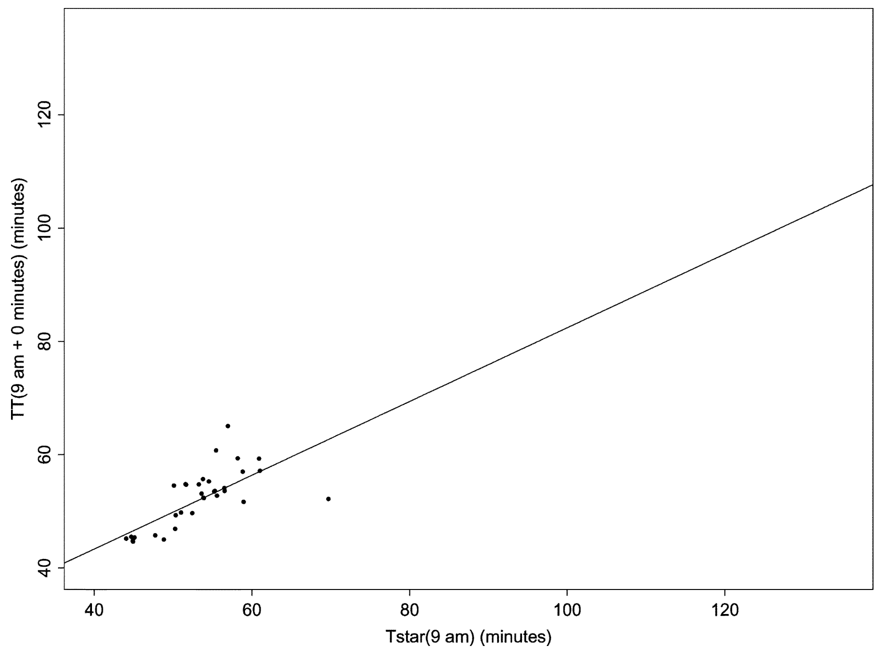

Fig. 1. $T^*(9\text{AM})$ versus $T(9\text{AM})$. Also shown is the regression line with intercept $\alpha = 17.3$ and slope $\beta = 0.65$.

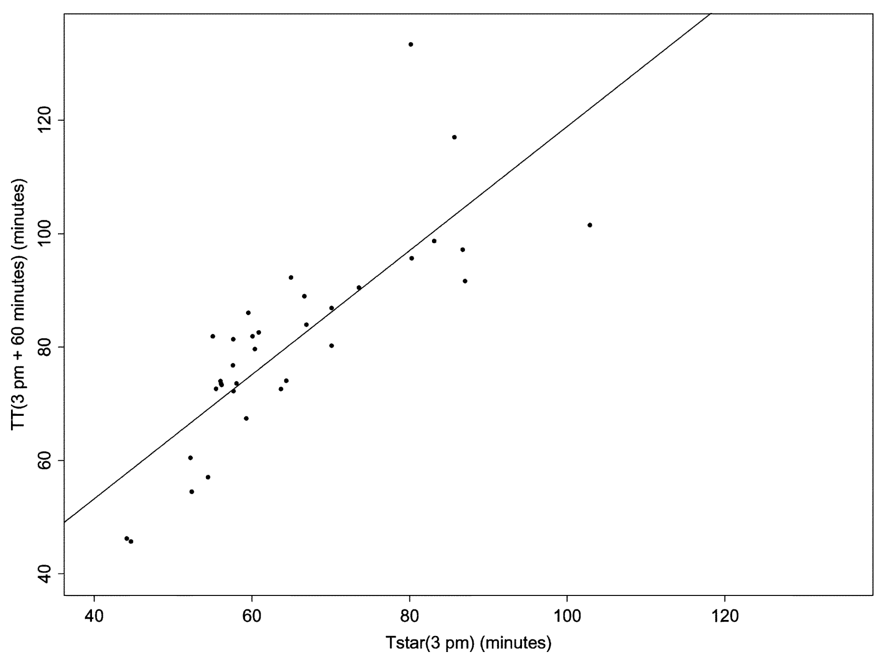

Fig. 2. $T^*(3\text{PM})$ versus $T(4\text{PM})$. Also shown is the regression line with intercept $\alpha = 9.5$ and slope $\beta = 1.1$.

In the literature, we find spectral analysis [7], Kalman filtering [8], [9], linear models [10], [11], and autoregressive-integrated moving average (ARIMA) models [12], [13]. Clustering techniques have been applied [10], [12] and, in recent years, also neural networks [14], [15]. We refer to [15] for additional references. Many of these efforts are aimed at forecasting traffic flow rather than travel time. Also, several make use of automated vehicle identification (AVI), which is not available in many situations. Finally, and most importantly, our *focus* differs from the abovementioned papers. We do not aim for sophistication or statistical optimality, but for ease of implementation and computational efficiency. We want our method to be fully scalable to the very large amounts of data that we need to process very quickly to be able to advise motorists in real time.

Two naïve predictors of $T(d, t + \delta)$ are the instantaneous travel time $T^*(d, t)$ and the historical average $\bar{T}(t + \delta)$. We expect—and, indeed, this is confirmed by an experiment—that $T^*(d, t)$ predicts well for small δ and $\bar{T}(t + \delta)$ predicts better for large δ. We will improve on both these predictors for all δ.

A. Linear Regression

The main point of this paper is what appears to be an empirical fact: that there exist linear relationships between $T^*(d, t)$

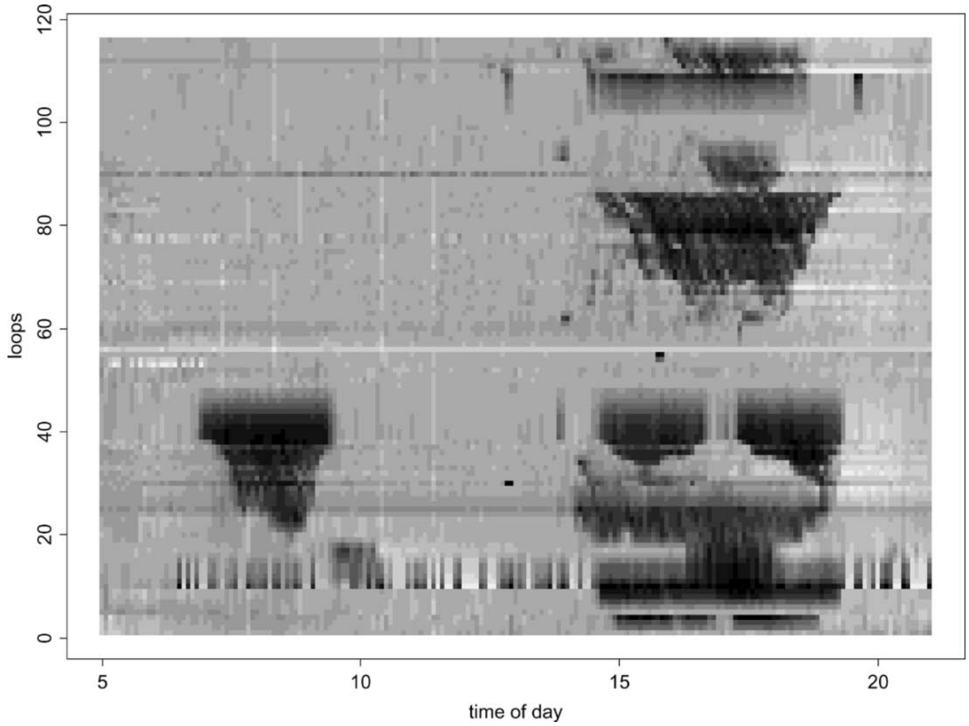

Fig. 3. Velocity field $v(d, l, t)$ where day $d = $ June $16, 2000$. Darker shades refer to lower speeds. Note that the typical triangular shapes indicate the morning and afternoon congestions building and easing. The horizontal streaks are most likely due to detector malfunction.

and $T(d, t+\delta)$ for all t and δ. This observation has held up in all of numerous freeway segments in CA that we have examined. This relation is illustrated by Figs. 1 and 2, which are scatter plots of $T^*(d,t)$ versus $T(d, t+\delta)$ for a 48-mi 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 model

$$T(d, t+\delta) = \alpha(t, \delta) + \beta(t, \delta) T^*(d, t) + \varepsilon \tag{3}$$

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 by Hastie and Tibshirani [16].

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

$$\sum_{\substack{d \in D \\ s \in T}} (T(d,s) - \alpha(t,\delta) - \beta(t,\delta) T^*(d,t))^2 K(t+\delta-s) \tag{4}$$

where K denotes the Gaussian density with mean zero and a certain variance σ^2, which the user needs to specify.

$$K(x) = \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-x^2/2\sigma^2}. \tag{5}$$

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 traffic do not change abruptly. The actual prediction of $T(d, t+\delta)$ becomes

$$\hat{T}(d, t+\delta) = \hat{\alpha}(t, \delta) + \hat{\beta}(t, \delta) T^*(d, t). \tag{6}$$

Writing $\alpha(t,\delta) = \alpha'(t,\delta)\bar{T}(t+\delta)$, we see that (3) expresses a future travel time as a linear combination of the historical mean and the current status travel time—our two naïve predictors. Hence, our new predictor may be interpreted as the best linear combination of our naïve predictors. From this point of view, we can expect our predictor to do better than both. In fact, it does, as demonstrated in Section III.

Another way to think about (3) is by remembering that the word "regression" arose from the phrase "regression to the mean." In our context, we would expect that if $T^*(d,t)$ is much larger than average, signifying severe congestion, then congestion will probably ease during the course of the trip. On the other hand, if it 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.

B. Principal Components

Our predictor only uses information at one time point: the current time t on day d. However, we do have information prior to that time. The following method attempts to exploit this by using the trajectories $T(d,s)$ and $T^*(d,s)$ for all $s \leq t$.

Let $\vec{T}(d)$ denote the vector of travel times for all departure times t on day d. Let us assume that the travel times on different days are independently and identically distributed (i.i.d.) and that for a given day d, the vectors $\vec{T}(d)$ and $\vec{T}^*(d)$ are 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 $\vec{T}(d)$ and $\vec{T}^*(d)$. The number of the eigenvalues retained must be specified by the user.

For given d and t, define τ to be such that $\tau + T(d, \tau) = t$. That is, τ is the departure time of the last trip that was completed before time t. With the estimated covariance, we can now compute the conditional expectation of $T(d, t+\delta)$ given $T(d, s)$ for all $s \le \tau$ and $T^*(d, s)$ for all $s \le t$. This is a standard computation that is described; for instance, in [17]. We call the resulting predictor $\hat{T}^{PC}(d, t+\delta)$.

C. Nearest Neighbors

As an alternative to principal components, we now consider nearest neighbors, which also is an attempt to use information prior to the current time t. Similar to principal components, it is a nonparametric method, but makes fewer assumptions (such as joint normality) on the relation between T^* and T.

Nearest neighbors aims to find that day in the past that is most similar to the present day in some appropriate sense. The remainder of that past day beyond time t is then taken as a predictor of the remainder of the present day.

The trick with nearest neighbors is in finding a suitable distance m between two days d_1 and d_2. We suggest two possible distances

$$m(d_1, d_2) = \sum_{l \in L,\ s \le t} |v(d_1, l, s) - v(d_2, l, s)| \tag{7}$$

and

$$m(d_1, d_2) = \left(\sum_{s \le t} (T^*(d_1, s) - T^*(d_2, s))^2 \right)^{1/2}. \tag{8}$$

Now, if day d' minimizes the distance to d among all previous days, our prediction is

$$\hat{T}^{\mathrm{NN}}(d, t+\delta) = T(d', t+\delta). \tag{9}$$

Sensible modifications of the method are "windowed" nearest neighbors (NN) and k-NN. Windowed-NN recognizes that not all information prior to t is equally relevant. Choosing a "window size" w, it takes the above summation to range over all s between $t - w$ and t. So-called k-NN basically is 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" $\hat{T}^{\mathrm{NN}}$.

III. Results

We have gathered flow and occupancy data from 116 single-loop detectors along 48 miles of I-10 East in Los Angeles (between postmiles 1.28 and 48.525). Measurements were done at 5-min aggregation at times t ranging from 5 AM to 9 PM for 34 weekdays between June 16 and September 8, 2000. We have used the so-called g-factor method to convert flow and occupancy to velocity using the well-known formula

$$\text{velocity} = g \times \frac{\text{flow}}{\text{occupancy}}. \tag{10}$$

Here, g is the unknown average length of vehicles, which has to be estimated. Often, a fixed car length is assumed at all loops and all times and inserted into the formula. PeMS research, carried out by Jia *et al.* [18], indicates that this is far too naïve. Jia *et al.* have studied data from 20 double-loop detectors from Orange County, CA (CalTrans District 12) for a 3-mo period starting in January 1998. Double-loop detectors consist of two single loops in close proximity, allowing for direct measurement of velocity. Jia *et al.* conclude that

> The analysis shows that the g-factor for any two randomly picked detectors in this network differ on average by 26 percent, and with probability 0.12 that they will differ by 50 percent. Furthermore, the g-factor for the same detector can vary by as much as 50 percent over a 24-hour period. [18, p. 536]

The differences between two loops might be due to different calibration or sensitivity. It could also be due to the fact the ratio of trucks versus cars will differ at different locations. This ratio also changes during the day from mostly trucks at night to compacts during rush hour. This explains why the g-factor at the same loop varies with time.

Jia *et al.* [18] propose a method that essentially works as follows. If occupancy is below a certain threshold, say 10%, then it is assumed that the freeway is in a state of freeflow with a nominal speed of 60 mi/h. Keeping the velocity fixed, (10) allows us to track the g-factor. As soon as the occupancy threshold is exceeded, a state of congestion is assumed and the g-factor is kept fixed at the last computed value. We again turn to (10), but this time to track velocity as we keep the g-factor fixed. The method appears to work well in practice.

Another difficulty in computing the velocity field is that loop detectors often do not report correct values or do not report at all. Fortunately, the quality of our I-10 data is quite good and we have used simple interpolation to impute incorrect or missing values. The resulting velocity field $v(d, l, t)$ is shown in Fig. 3, where day d is June 16. The horizontal streaks typically indicate detector malfunction.

From the velocities, we computed travel times for trips starting between 5 AM and 8 PM. Fig. 4 shows the $T(d, t)$ as functions of the time of day t. 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 of 45 min to up to 2 h. Included in the data are holidays July 3 and 4, which may be readily recognized by their very fast travel times.

We have estimated the root-mean-squared (rms) error of our various prediction methods for a number of "current times" t ($t = $ 6AM, 7AM..., 7PM) and lags δ ($\delta = 0$ and 60 min). We did the estimation by leaving out one day at a time, performing the prediction for that day on the basis of the other remaining days and averaging the squared prediction errors.

The prediction methods all have parameters that must be specified by the user. For the regression method, we have chosen the standard deviation of the Gaussian kernel K to be 10 min. For the principal components method, we have chosen the number of eigenvalues retained to be 4. For the NN method, we have chosen distance function (8), a window w of 20 min, and the number k of nearest neighbors to be 2.

Figs. 5 and 7 show the estimated rms error of $\bar{T}(t+\delta)$, $T^*(d, t)$, and $\hat{T}(d, t+\delta)$ as predictors of $T(d, t+\delta)$, for lag δ

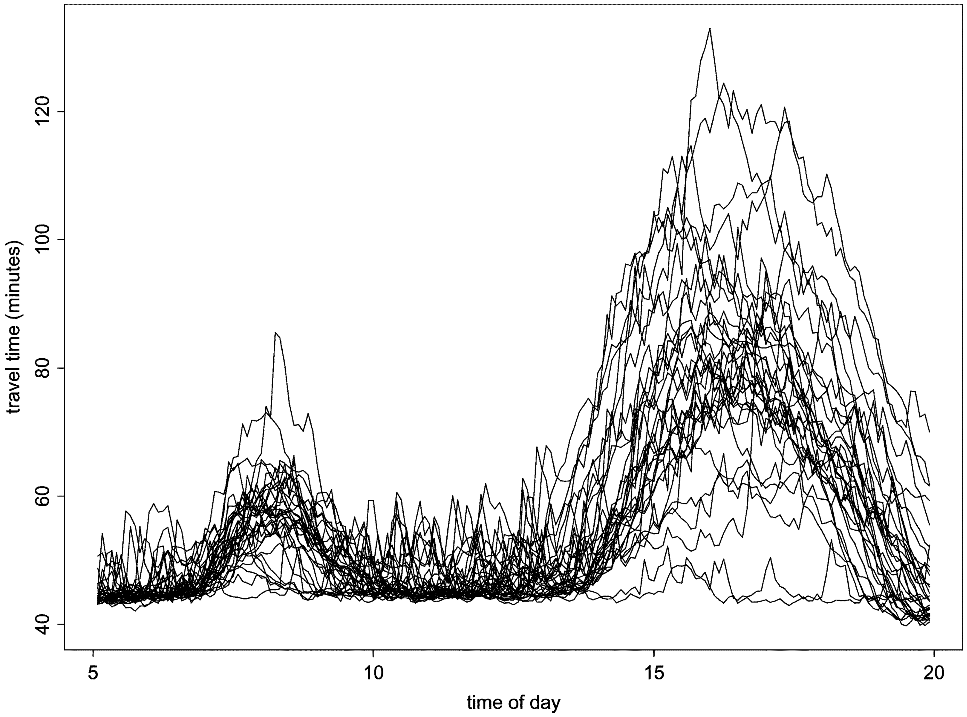

Fig. 4. Travel times $T(d, \cdot)$ for 34 d on a 48-mile stretch of I-10 East.

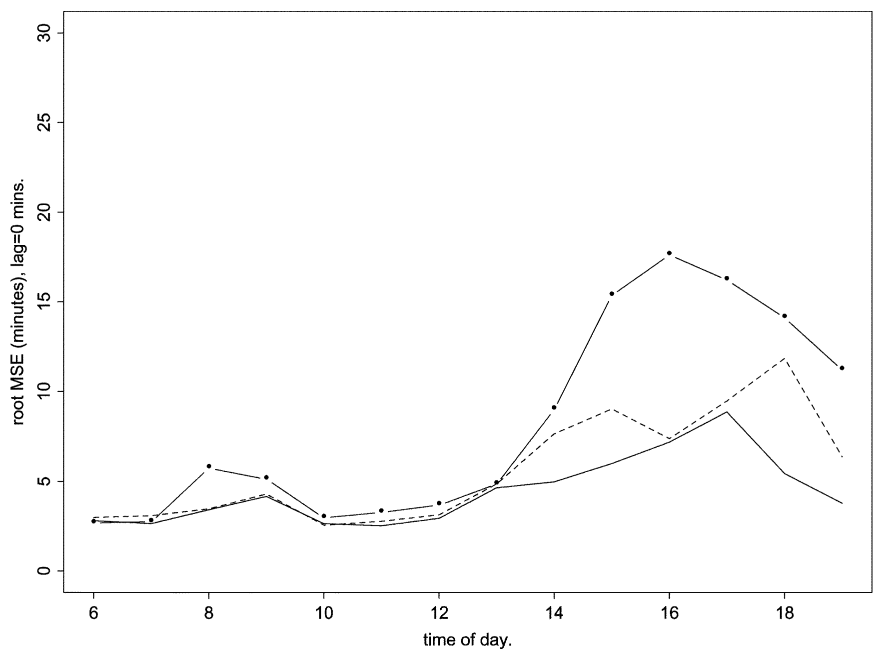

Fig. 5. Estimated root-mean-square error (rmse), lag $= 0$ min. Historical mean (– · –), current status (– – –), and linear regression (—).

equal to 0 and 60 min, respectively. Note how the current status $T^*(d,t)$ performs well for small δ ($\delta = 0$) and how the historical mean $\bar{T}(t+\delta)$ does not become worse as δ increases. Most importantly, however, notice how the regression predictor $\hat{T}(d, t+\delta)$ beats both uniformly.

Figs. 6 and 8 again show the rms prediction error of the regression predictor. This time, its performance is compared to the principal components predictor $\hat{T}^{\mathrm{PC}}(d, t+\delta)$ and the NN predictor $\hat{T}^{\mathrm{NN}}(d, t+\delta)$. Again, the regression predictor comes out on top, although the NN predictor shows comparable performance.

The rms error of the regression predictor stays below 10 min even when predicting 1 h ahead. As the average travel time is multiplied by 3, the rms error is still less than 10%. We feel that this is particularly impressive given the simplicity of the predictor.

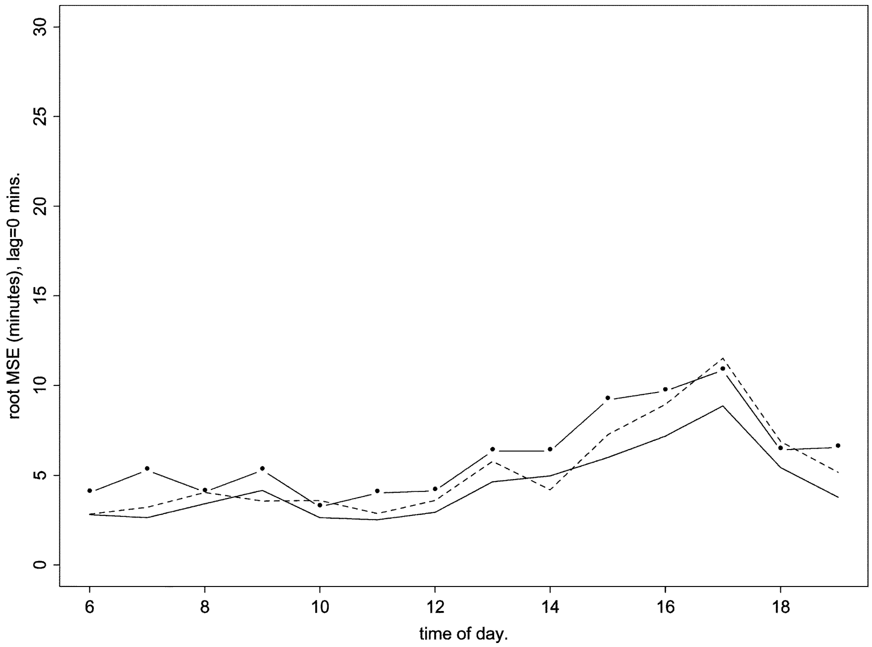

Fig. 6. Estimated rmse, lag $= 0$ min. Principal components (– · –), nearest neighbors (– – –), and linear regression (—).

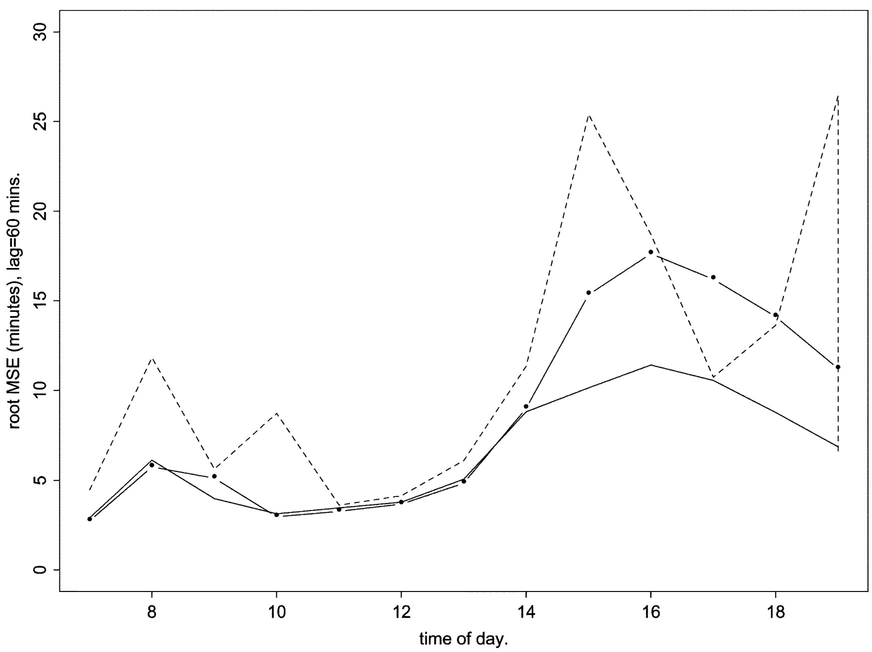

Fig. 7. Estimated rmse, lag $= 60$ min. Historical mean (– · –), current status (– – –), and linear regression (—).

IV. Conclusion and Loose Ends

The main contribution of this paper is the discovery of a linear relation between $T^*(d,t)$ and $T(d,t+\delta)$, which we put to use to predict travel times. Comparison of the regression predictor to the principal components and NN predictors yields another surprise. Given $T^*(d,t)$, there is not much information left in the earlier $T^*(d,s)$ $(s < t)$ that is useful for predicting $T(d,t+\delta)$. In fact, we have come to believe that for the purpose of predicting travel times all the information in $\{v(d,l,s),\ l \in L,\ s \leq t\}$ is well summarized by one single number: $T^*(d,t)$.

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 can be done offline in reasonable time. The actual prediction is trivial. As this paper was submitted, we were in the process of making our travel-time predictions and associated optimal routings available through the Internet for the network of freeways of California District 7 (Los Angeles). It would also be possible to make our service available for users of cellular telephones—in fact, we plan to do so in the near future.

It also is 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

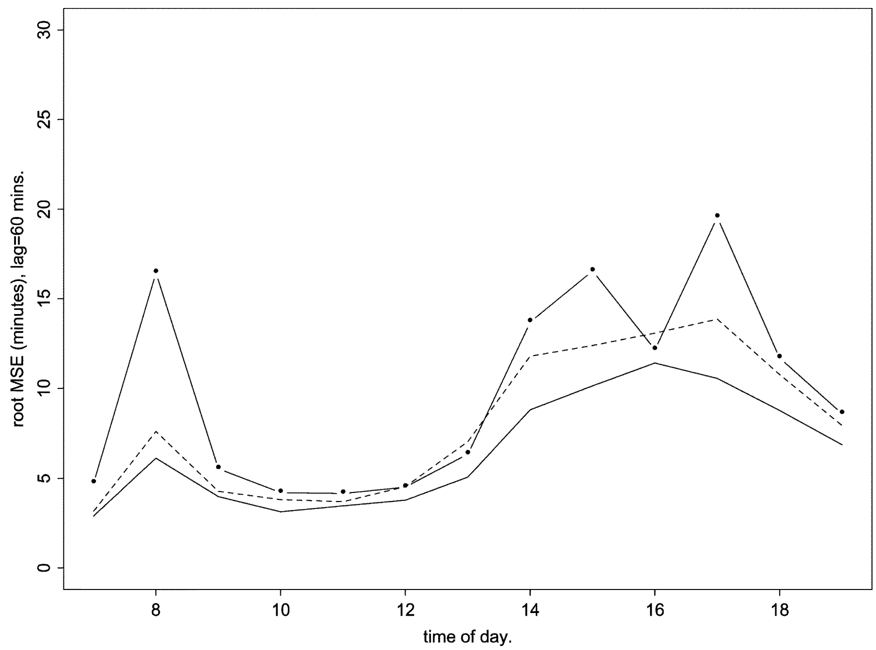

Fig. 8. Estimated rmse, lag $= 60$ min. Principal components (– · –), NN (– – –), and linear regression (—).

used in place of loops, since all the method requires is the current value of $T^*(d,t)$ and historical measurements of T and T^*. Earlier work with the method [19] used a short stretch of freeway with very-high-quality data, including probe vehicles. In that work, however, the regression method was not compared to methods such as NN and principal components, which take more information into account.

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

1) For trips from A to B to C, we have

$$T_{AC}(d,t) = T_{AB}\big(d,t\big) + T_{BC}(d, t + T_{AB}(d,t)\big). \quad (11)$$

We have found that it is sometimes more practical or advantageous to predict the terms on the right-hand side than to predict $T_{AC}(d,t)$ directly. For instance, when predicting travel times across networks (graphs), we need only predict travel times for the edges and then use (11) to piece these together to obtain predictions for arbitrary routes.

2) We regressed the travel time $T(d,t+\delta)$ on the current status $T^*(d,t)$. Now, define $T'(d,t+\delta)$ to be the travel time for a trip *arriving* at time $t+\delta$ on day d. Regressing $T'(d,t+\delta)$ on $T^*(d,t)$ 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.

Acknowledgment

The authors would like to acknowledge the help (alphabetically) from P. Bickel, C. Chen, J. Kwon, X. Zhang, and the other members of the PeMS project. Also, we would like to thank the engineers of CalTrans Headquarters and Districts 7 and 12.

References

[1] [Online]. Available: http://pems.EECS.Berkeley.EDU/Public/
[2] H. S. Mahmassani and S. Peeta, “System optimal dynamic assignment for electronic route guidance in a congested traffic network,” in *Urban Traffic Networks: Dynamic Flow Modeling and Control*, Gartner and Improta, Eds. Berlin, Germany: Springer-Verlag, 1995, pp. 2–27.
[3] F. Ho and P. Ioannou, “Traffic flow modeling and control using artificial neural networks,” *IEEE Control Syst. Mag.*, vol. 16, pp. 16–27, Oct. 1996.
[4] I. Chabini, “Discrete dynamic shortest path problems in transportation applications: Complexity and algorithms with optimal run time,” *Transport. Res. Rec.*, vol. 1645, pp. 170–175, 1998.
[5] A. Orda and R. Rom, “Minimum weight paths in time-dependent networks,” *Networks*, vol. 21, no. 3, pp. 295–320, 1991.
[6] Y. J. Stephanedes and A. P. Chassiakos, “Freeway incident detection through filtering,” *Transport. Res. C*, vol. 1, no. 3, pp. 219–233, 1993.
[7] H. Nicholson and C. Swann, “The prediction of traffic flow volumes based on spectral analysis,” *Transport. Res.*, vol. 8, pp. 533–538, 1974.
[8] I. Okutani and Y. Stephanedes, “Dynamic prediction of traffic volume through Kalman filtering theory,” *Transport. Res. B*, vol. 18, no. 1, pp. 1–11, 1984.
[9] P. C. Vythoulkas, “Alternative approaches to short term traffic forecasting for use in driver information systems,” in *Proc. Int. Symp. Transportation Traffic Theory*, 1993, pp. 485–506.
[10] M. Danech-Pajouh and M. Aron, “ATHENA, a Method for Short-Term Inter-Urban Traffic Forecasting,” INRETS, Paris, France, Rep. 177, 1991.
[11] J. Kwon, B. Coifman, and P. Bickel, Day-to-day travel time trends and travel time prediction from loop detector data, in *Transport. Res. Rec.1717*, Transportation Research Board, pp. 120–129, 2000.
[12] M. Van der Voort, M. Dougherty, and S. Watson, “Combining KOHONEN maps with ARIMA time series models to forecast traffic flow,” *Transport. Res. C*, vol. 4, no. 5, pp. 307–318, 1996.
[13] T. Oda, “An algorithm for prediction of travel time using vehicle sensor data,” in *Proc. 3rd Int. Conf. Road Traffic Control*, 1990, pp. 40–44.
[14] D. Park and L. Rilett, “Multiple-period freeway link travel times usingmodular neural networks,” *Transport. Res. Rec.*, vol. 1617, pp. 163–170, 1998.
[15] H. R. Kirby, S. M. Watson, and M. S. Dougherty, “Should we use neural networks or statistical models for short-term motorway traffic forecasting?,” *Proc. Int. J. Forecasting*, vol. 13, no. 1, pp. 43–50, 1997.
[16] T. Hastie and R. Tibshirani, “Varying coefficient models,” *J. R. Stat. Soc.*, ser. B, vol. 55, no. 4, pp. 757–796, 1993.
[17] K. V. Mardia, J. T. Kent, and S. M. Bibby, *Multivariate Analysis*. London, U.K.: Academic, 1979.

[18] Z. Jia, C. Chen, B. Coifman, and P. Varaiya, “The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single loop detectors,” in *Proc. 4th Int. ITSC Conf.*, 2001, pp. 536–541.
[19] X. Zhang and J. Rice, “Short term travel time prediction,” *Transport. Res.*, ser. C, vol. 11, pp. 187–210, 2003.

John Rice was born in New York on June 14, 1944. He received the B.A. degree in mathematics from the University of North Carolina, Chapel Hill, in 1966 and the Ph.D. degree in statistics from the University of California, Berkeley, in 1972.

He was with the Department of Mathematics, University of California, San Diego, from 1973 to 1991. Since 1991, he has been with the University of California, Berkeley, where he currently is Professor of statistics. His research interests include applied and theoretical statistics.

Dr. Rice is a Member of the Institute of Mathematical Statistics, the American Statistical Association, and the International Statistical Institute.

Erik van Zwet was born in The Hague, The Netherlands, on November 10, 1970. He received the *doctorandus* (M.Sc.) degree from the University of Leiden, Leiden, the Netherlands, in 1995 and the Ph.D. degree from the University of Utrecht, Utrecht, The Netherlands, in 1999, both in mathematics.

After three years as a post-doctorate with the University of California, Berkeley, he is now an Assistant Professor with the University of Leiden.