A Simple and Effective Method for Predicting Travel Times on Freeways

John Rice, Erik van Zwet

Abstract— **We present a method to predict the time that will be needed to traverse a certain stretch of freeway when departure is at a certain 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 purpose, the current situation of a stretch of 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 the traffic would occur. This current status travel time can be estimated from single or double loop detectors, video data, probe vehicles or by any other means.

Our prediction method arises from the empirical fact that there exists a linear relationship between any future travel time and the current status travel time. The slope and intercept of this relationship is observed to change subject to the time of day and the time until departure. This naturally leads to a prediction scheme by means of linear regression with time varying coefficients.

Keywords— **Prediction, Travel Time, Linear Regression, Varying Coefficients.**

I. Introduction

The Performance Measurement System (PeMS) [1] is a large-scale freeway data collection, storage and analysis project. It involves the departments of EECS and Statistics and the Institute of Transportation Studies at the University of California, Berkeley in cooperation with the California Department of Transportation (Caltrans). PeMS' goal is to facilitate traffic research and to 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, aggregated over 30 second intervals. For all of California this amounts to 2 gigabytes per day. 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 of a freeway network for any future departure time. Besides being useful per se, travel time prediction serves as input to dynamic route guidance, congestion management, optimal routing and incident detection.

We are currently developing an Internet application which will give the commuters of Caltrans District 7 (Los Angeles) the opportunity to query the prediction algorithm which is described in this paper. The user will access our Internet site and state 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.

Both authors are in the Department of Statistics of the University of California at Berkeley. Email: rice@stat.berkeley.edu and vanzwet@stat.berkeley.edu.

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 which will be used for comparison. This comparison is made in section III 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 IV, we summarize our conclusions, point out some practical observations and briefly discuss several extensions of our new method.

II. Methods of Prediction

Consider 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 $TT_d(a,b,t)$, 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. We can also compute a proxy for these travel times which is defined by

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

where d_i denotes the distance from loop i to loop $(i+1)$. We call T^* 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 the computation of $T_d^*(a,b,t)$ only requires information available at time t, whereas computation of $TT_d(a,b,t)$ requires information of later times.

We fix an origin and destination of our travels and drop the arguments a and b from our notation. We now formally state the problem that is addressed in this paper.

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 $t \le \tau$. We call τ the 'current time'. Or aim is to predict $TT_e(\tau+\delta)$ 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.

Define the historical mean travel time as

$$\mu_{TT}(t) = \frac{1}{|D|}\sum_{d \in D} TT_d(t). \tag{2}$$

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

A. Linear Regression

The main result of this paper is our discovery of an empirical fact: that there exist linear relationships between $T^*(t)$ and $TT(t+\delta)$ 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 Figures 1 and 2, which are scatter plots of $T^*(t)$ versus $TT(t+\delta)$ 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

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

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 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

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

where K denotes the Gaussian density with mean zero and a certain variance which the user needs to specify. 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 $TT_e(\tau+\delta)$ becomes

$$\widehat{TT}_e^{\alpha\beta}(\tau+\delta) = \hat{\alpha}(\tau,\delta)T_e^*(\tau) + \hat{\beta}(\tau,\delta). \tag{5}$$

Writing $\beta(t,\delta) = \beta'(t,\delta)\mu_{TT}(t+\delta)$ 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 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 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^* is much larger than average—signifying severe congestion—then congestion will probably ease during the course of the trip. On the other hand, if T^* 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 $\widehat{TT}^{\alpha\beta}$ 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 TT_e and T_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 for a given day d, $\{TT_d(t) : t \in T\}$ and $\{T_d^*(t) : t \in T\}$ 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 $\{(TT_d(t), T_d^*(t)) : d \in D,\ t \in T\}$. How many of the eigenvalues are retained must be specified by the user. Define τ' to be the largest t such that $t + TT_e(t) \leq \tau$. 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 $TT_e(\tau+\delta)$ given $\{TT_e(t) : t \leq \tau'\}$ and $\{T_e^*(t) : t \leq \tau\}$. This is a standard computation which is described, for instance, in [3]. The resulting predictor is $\widehat{TT}_e^{\mathrm{PC}}(\tau+\delta)$.

C. 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. Similar to Principal Components, it is a non–parametric method, but it makes fewer assumptions (such as joint normality) on the relation between T^* and TT.

Nearest neighbors aims to find that day in the past which is most similar to the present day in some appropriate sense. The remainder of that past day beyond time τ 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 days. We suggest two possible distances:

$$m(e,d) = \sum_{l\in L,\ t\leq\tau} |V(e,l,t) - V(d,l,t)| \tag{6}$$

and

$$m(e,d) = \left(\sum_{t\leq\tau} (T_e^*(t) - T_d^*(t))^2 \right)^{1/2}. \tag{7}$$

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

$$\widehat{TT}_e^{NN}(\tau+\delta) = TT_{d'}(\tau+\delta). \tag{8}$$

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' $\widehat{TT}^{NN}$.

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 minute 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

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

Here g is the unknown average length of vehicles, which has to be estimated. There are two problems with this method. First, the sensitivity of loop detectors differ and this difference is incorporated into g. Secondly, the average length of vehicles varies during a day from trucks late at night to compacts during rush hour. To try to counter these problems, we have used separate g factors for different detectors and allowed them to vary with the time of day. We estimate these g factors during times of freeflow when the occupancy is below 15% and known freeflow speeds occur. During congestion, when occupancy is above 15%, we keep the g constant.

Another problem with 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 wrong or missing values. The resulting velocity field $V(d, l, t)$ is shown in figure 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. Figure 4 shows these $TT_d(t)$ 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 of 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.

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 sixty minutes). We did the estimation 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 by the user. 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 (7), a window w of 20 minutes and the number k of nearest neighbors to be 2.

Figures 5 and 7 show the estimated root mean squared (RMS) prediction error of the historical mean $\mu_{TT}(\tau+\delta)$, the current status predictor $T_e^*(\tau)$ and our regression predictor (5) for lag δ equal to 0 and 60 minutes, respectively. Note how $T_e^*(\tau)$ 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.

Figures 6 and 8 again show the RMS prediction error of the regression estimator. This time, it is compared to the Principal Components predictor and the nearest neighbors predictor (8). 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.

IV. Conclusions and loose ends

We stated that the main contribution of this paper is the discovery of a linear relation between $T^*(t)$ and $TT(t+\delta)$. But there is more. Comparison of the regression predictor to the Principal Components and nearest neighbors predictors unearthed another surprise. Given $T^*(\tau)$, there is not much information left in the earlier $T^*(t)$ ($t < \tau$) that is useful for predicting $TT(\tau+\delta)$. In fact, we have come to believe that for the purpose of predicting travel times all the information in $\{V(l,t),\ l \in L,\ t \le \tau\}$ is well summarized by one single number: $T^*(\tau)$.

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. As this paper is submitted, we are 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—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 T^* and historical measurements of TT and T^*.

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

$$T_d(a,c,t) = T_d(a,b,t) + T_d(b,c,T_d(a,b,t)). \qquad (9)$$

We have found that it is sometimes more practical or advantageous to predict the terms on the right hand side than to predict $T_d(a,c,t)$ directly. For instance, when predicting travel times across networks (graphs), we need only predict travel times for the edges and then use (9) to piece these together to obtain predictions for arbitrary routes.

2. We regressed the travel time $T_d(t+\delta)$ on the current status $T^*_d(t)$, where $T_d(t+\delta)$ is the travel time departing at time $t+\delta$. Now, define $S_d(t)$ to be the travel time *arriving* at time t on day d. Regressing $S_d(t+\delta)$ 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.

Acknowledgments

The PeMS project is supported by grants from the National Science Foundation and from Caltrans through the California PATH program. The authors would like to acknowledge the help from (alphabetically) Peter Bickel, Chao Chen, Jaimyoung Kwon and Xiaoyan Zhang and the other members of the PeMS project. Also, we express our thanks to the engineers of Caltrans Headquarters and districts 7 and 12.

References

[1] `http://transacct.EECS.Berkeley.EDU/Public/`

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

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

John A. Rice was born in New York on June 14, 1944. He received his B.A. degree in mathematics from the University of North Carolina at Chapel Hill in 1966 and his Ph.D. in statistics in 1972 from the University of California, Berkeley. He was in the Department of Mathematics at the University of California, San Diego from 1973–1991 and since 1991 he has been at the University of California, Berkeley where is is professor of statistics. His research interests include applied and theoretical statistics. He is a member of the Institute of Mathematical Statistics, the American Statistical Association, and the International Statistical Institute.

Erik W. van Zwet was born in The Hague, The Netherlands on November 10, 1970. He received his *doctorandus* (M.Sc.) degree in mathematics from the University of Leiden in 1995 and his Ph.D. degree, also in mathematics, from the University of Utrecht in 1999. He is currently a post doctoral visitor at the Department of Statistics of the University of California at Berkeley.

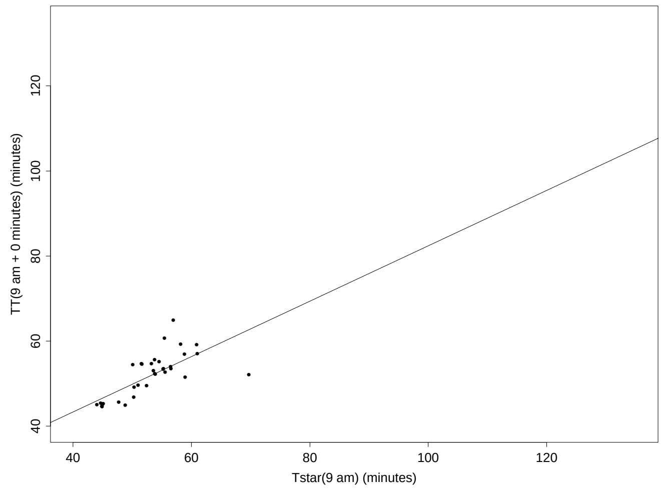

Fig. 1. T^*(9 am) vs. TT(9 am + 0 min's). Also shown is the regression line with slope α(9 am, 0 min's)=0.65 and intercept β(9 am, 0 min's)=17.3.

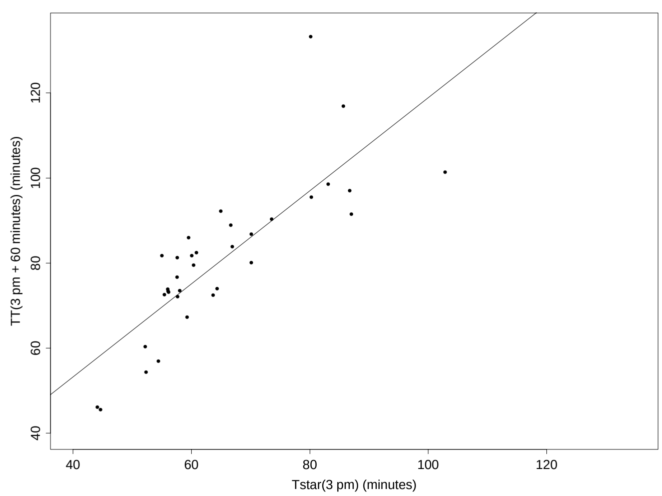

Fig. 2. T^*(3 pm) vs. TT(3 pm + 60 min's). Also shown is the regression line with slope α(3 pm, 60 min's)=1.1 and intercept β(3 pm, 60 min's)=9.5.

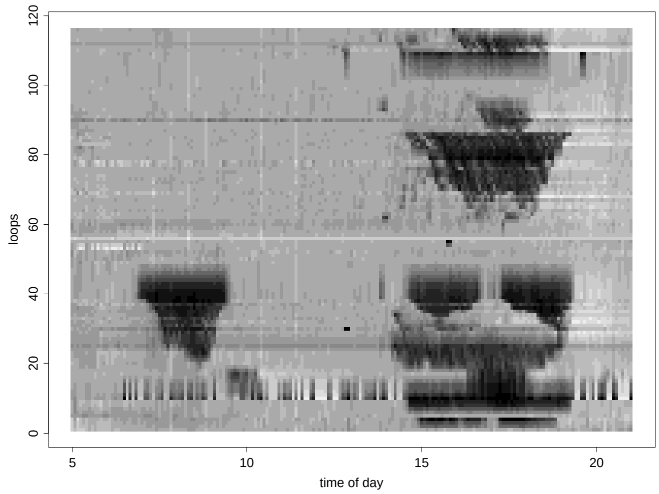

Fig. 3. Velocity field $V(d, l, t)$ where day d = June 16, 2000. Darker shades refer to lower speeds. Note the typical triangular shapes indicating the morning and afternoon congestions building and easing. The horizontal streaks are most likely due to detector malfunction.

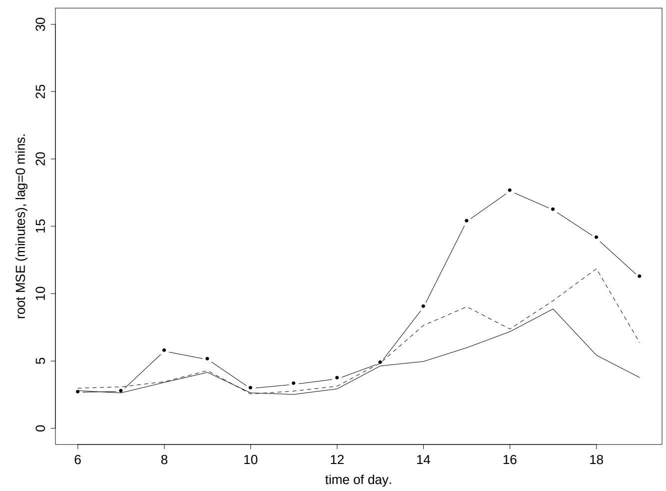

Fig. 5. Estimated RMSE, lag=0 minutes. Historical mean (– · –), current status (- - -) and linear regression (—).

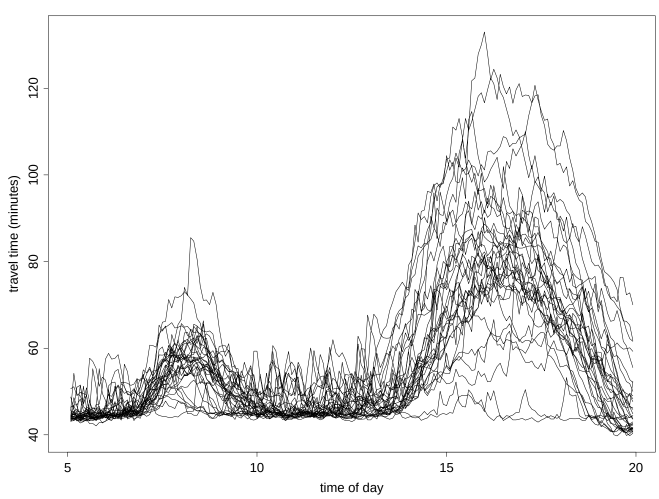

Fig. 4. Travel Times $TT_d(\cdot)$ for 34 days on a 48 mile stretch of I-10 East.

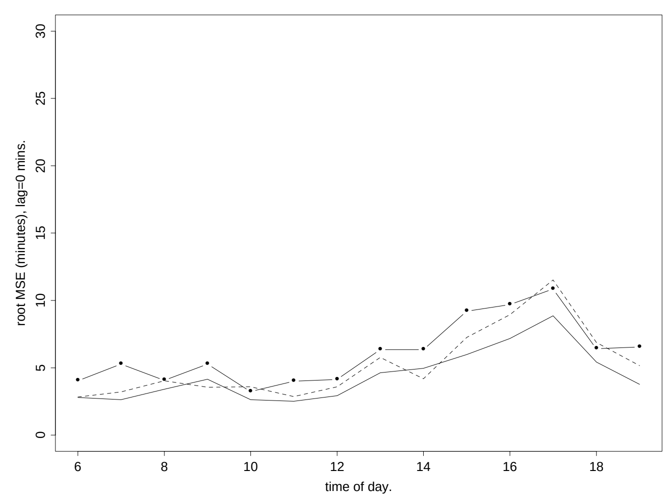

Fig. 6. Estimated RMSE, lag=0 minutes. Principal Components (– · –), nearest neighbors (- - -) and linear regression (—).

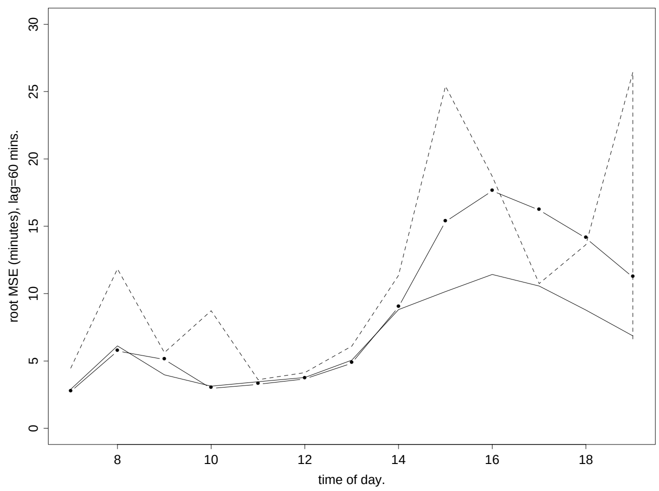

Fig. 7. Estimated RMSE, lag=60 minutes. Historical mean (– · –), current status (- - -) and linear regression (—).

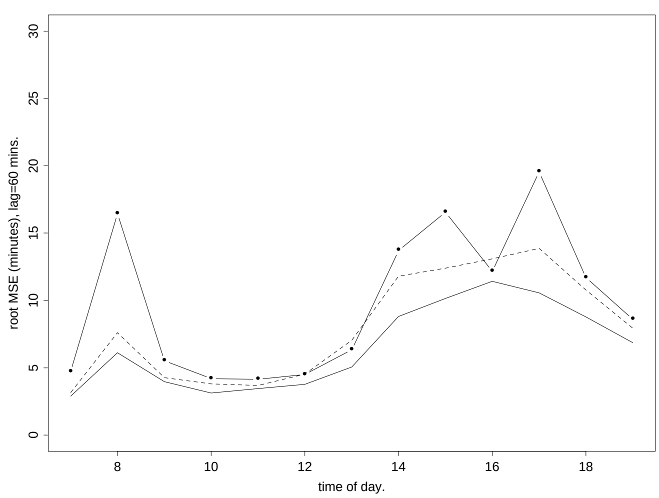

Fig. 8. Estimated RMSE, lag=60 minutes. Principal Components (– · –), nearest neighbors (- - -) and linear regression (—).