# Day-to-Day Travel Time Trends and Travel Time Prediction from Loop Detector Data

Jaimyoung Kwon
University of California
Department of Statistics
367 Evans Hall #3860
Berkeley, CA 94720-3860

Benjamin Coifman
Ohio State University
Department of Civil and Environmental Engineering and Geodetic Science
Department of Electrical Engineering
470 Hitchcock Hall, 2070 Neil Ave
Columbus, OH 43210-1275

Peter Bickel
University of California
Department of Statistics
367 Evans Hall #3860
Berkeley, CA 94720-3860

## ABSTRACT

This paper presents an approach to estimate future travel times on a freeway using flow and occupancy data from single loop detectors and historical travel time information. The work uses linear regression with stepwise variable selection method and more advanced tree based methods. The analysis considers forecasts ranging from a few minutes into the future up to an hour ahead. Leave-a-day-out cross-validation was used to evaluate the prediction errors without under-estimation. The current traffic state proved to be a good predictor for the near future, up to 20 minutes, while historical data is more informative for longer-range predictions. Tree based methods and linear regression both performed satisfactorily, showing slightly different qualitative behaviors for each condition examined in this analysis. Unlike preceding works that rely on simulation, this study uses real traffic data. Although the current implementation uses measured travel times from probe vehicles, the ultimate goal of this research is an autonomous system that relies strictly on detector data. In the course of presenting the prediction system, the paper examines how travel times change from day-to-day and develops several metrics to quantify these changes. The metrics can be used as input for travel time prediction, but they should be also beneficial for other applications such as calibrating traffic models and planning models.

Keywords: loop detectors, travel time prediction, advanced traveler information systems (ATIS), regression, cross-validation.

## INTRODUCTION

As congestion increases on urban freeways, more and more journeys are impacted by delays. Unless a traveler routinely traverses a given route, the extent of possible delays are unknown before departing on a journey and the uncertainty must be addressed by allocating extra time for traveling. Advanced Traveler Information Systems (ATIS) attempt to reduce the uncertainty by providing the current state of the system and sometimes a prediction of future states. In this context, travel time is an important parameter to report to travelers. From the user's perspective, accurate predictions (and an estimate of their precision) are more beneficial than the current travel time since conditions may change significantly before a traveler completes their journey.

To this end, we study a travel time prediction methodology using flow and occupancy data from single loop detectors and historical travel time information. By restricting the detector data to flow and occupancy, the methodology should be applicable to most automated traffic surveillance systems. The raw data for this study come from the I-880 database (*1*), which includes detector data and probe vehicle travel times. In practice, the probe data could be replaced with estimated travel times from the detector data (e.g., (*2*-*3*)). In contrast, preceding works tend to rely on simulation (e.g., (*4*-*9*)) or automatic vehicle identification (e.g., (*10*-*11*)) The former may capture biases intrinsic to the given traffic simulator, while the latter requires a large investment in new detector infrastructure.

Using the I-880 database, we present a methodology for generating and evaluating travel time prediction models. Such models are likely to be site-specific, so the paper should be viewed as a tool for generating prediction models suitable for a specific site rather than for developing a universal model applicable to all roadways. Although the I-880 database used in this study is relatively small, it is the most detailed set of real traffic data currently available. This fact allows for detailed analysis and verification in the examples presented herein.

Under the current setup, the problem can be abstracted as one of fitting the response variable (probe vehicle travel time) on the explanatory variables (detector measurements, probe departure time, etc.) in various ways and thus, can be considered a *regression* problem. In this work, linear regression with stepwise variable selection and advanced methods such as the tree based method and neural networks are investigated.

Traditionally, a database would be divided into separate training and testing subsets to evaluate these methods since the prediction error will be underestimated if the two subsets intersect. Unfortunately, the small database employed in this study does not allow for such a simple division. To make efficient use of the sample while avoiding the intersection problem, we employ *cross-validation* (CV), i.e., repeatedly taking a different subset of the data as a training set and using the remaining observations as a test set. We will use a form of cross-validation to correctly estimate the prediction error.

### Overview

The first section discusses the data and notation used throughout this paper. The next section provides empirical and exploratory analysis of the data, highlighting implications for travel time prediction. Then, the regression methods are studied in detail. Finally, the paper closes with conclusions and future directions.

## DATA AND NOTATION

### The I-880 Database

The I-880 database includes loop data and probe vehicle travel times over a 10 km (6.2 mi) segment of freeway, south of Oakland, California. There are approximately 19 loop detector stations in each direction (northbound and southbound) and up to five probe vehicles circling the 20 km round trip on a given day. The probe vehicle drivers were instructed to stay in the middle lane (third lane out from the median). The I-880 study used dual loop speed traps to measure flow, occupancy and velocity. The data were collected during the morning and evening peak periods.

This work uses data from the middle lane in each direction over 20 weekdays (all weekdays between February 22 and March 19, 1993). The loop data are aggregated in 30-second samples and the velocity measurements are not used, thus, replicating the data that would be available from single loops. Treating each direction and each shift (5:00- 10:00 AM and 2:00-7:00 PM) independently, there are four distinct samples or scenarios (South/AM, South/PM, North/AM and North/PM) each day. Because there are significant differences between the various scenarios, they will be treated as independent populations in the subsequent analysis.

### Loop Data Processing

Although the I-880 database is a rich source of information, it does not have data from every detector station for every day. Missing data at one station are estimated by interpolating data from adjacent stations. The distance between successive stations is irregular and to prevent the part of the segment with more densely located loops from being over-represented as predictor covariates, this work interpolates the original detector data to 10 equidistant points along the freeway. Henceforth, the discussion treats these virtual detectors as if they were real and they will simply be referred to as detector stations. One could also consider travel time prediction from these virtual stations as fitting travel time to a function of the true detector data.

Finally, because 30-second detector samples are inherently noisy, the analysis uses a simple low pass filter to eliminate transients. Specifically, for flow and occupancy time series at each detector, we apply:

$$z_t = \sum_{i=0}^{4} a_i y_{t-i} \ , \ a_i = \frac{e^{-i^2/5}}{\sum_{j=0}^{4} e^{-j^2/5}}, i = 0,1,...,4 \tag{1}$$

to get filtered series $z_t$ from the original series $y_t$. The exponential weights $a_i$ 's for the filter were selected for convenience and no attempt was made to optimize the filter for the given application.

After filtering, a given sample reflects traffic conditions over the previous 2.5 minutes. With 600 samples per shift, the net result of this processing is a pair of 10 by 600 matrices for each scenario, corresponding to flow and occupancy, respectively.

### Notation

As noted above, there are ten detector stations, or loops, in this study. These stations are indexed by $x$ =1,…,10 (we will assume that a smaller index is upstream of a larger one, without loss of generality.) and each scenario spans 600 time points, indexed by $t$ =1,…,600. Occupancy and flow at time $t$ and location $x$ on day $d$ will be denoted as $o_d(x,t)$ and $f_d(x,t)$ respectively. So, for a given scenario and day $d$, the data can be summarized as vector time series of occupancy $\{\mathbf{o}_d(t)\}$ and flow $\{\mathbf{f}_d(t)\}$ evolving over time $t$, where $\mathbf{o}_d(t) = (o_d(1,t),...,o_d(10,t))$ and $\mathbf{f}_d(t) = (f_d(1,t),..., f_d(10,t))$. The data are observed over 20 days, $d = 1,2,...,20$, yielding three-dimensional arrays of size 20 by 10 by 600, $\{o_d(x,t)\}$ and $\{f_d(x,t)\}$.

The travel time between station $x$ and $x+1$ for a probe vehicle that departs from the upstream loop $x$ at time $s$ on day $d$ in a given direction is denoted $\tau_d(x,s)$. For each scenario, successive probe vehicle runs are indexed by $i$. For a given probe vehicle run, $\tau_d(s)$ denotes the total travel time over all 10 detector stations, with the corresponding departure time $s$ and day $d$.

## EMPIRICAL AND EXPLORATORY DATA ANALYSIS

With the three dimensional arrays of flow and occupancy, dimension reduction and visual representation of traffic state are important tasks in preliminary analysis of the data. To this end, the following subsections examine the I-880 data using empirical and exploratory data analysis, introducing three measures of freeway traffic status for a given day in the form of a field, a series, and a scalar.

### Occupancy and Flow Field for a Given Day

The time-space field of occupancy $o_d(x,t)$ conveys all of the available loop occupancy information for day $d$. See Figure 1 for contour plots of occupancy field for an entire day, selected at random from each of the four scenarios. As can be seen from the plots, the fields vary widely for the different scenarios.

Variability over different days in the same scenario is also quite large. To summarize such day-to-day variability, we propose the field of historical median, $\tilde{o}(x,t)$ , and the field of historical MAD (Median Absolute Deviation), $\tilde{\sigma}(x,t)$ , as measures for 'location' and 'scale,' respectively. For generic sample values $y_i = 1,...,n$, MAD is defined as $MAD(y_i) = med_i(| y_i - med_j y_j |)/.6745$. The median and MAD are robust measures of center and dispersion of a distribution, respectively. These measures are insensitive to outliers, unlike mean and standard deviation (SD).

Formally, $\tilde{o}(x,t)$ and $\tilde{\sigma}(x,t)$ are denoted

$$\tilde{o}(x,t) = med\{o_d(x,t) : d = 1,..., D\} \tag{2}$$

$$\tilde{\sigma}(x,t) = MAD\{o_d(x,t) : d = 1,..., D\}. \tag{3}$$

Figures 2 and 3 show plots of these two fields for each of four scenarios. Among other features, recurrent congestion can be easily identified from Figure 2. Consider the northbound AM traffic, the figure clearly shows recurring congestion at station 2 around time sample 400 (8 AM); and again in the northbound PM traffic at station 7 around time sample 300 (4:30 PM). Note that the freeway segment is exhibiting different characteristics during different time periods. Looking at Figure 2 and 3 side by side, we can also observe that day-to-day variation of occupancy is small when the average occupancy (over days) is light and conversely. Although we considered the occupancy field for simplicity, the flow field, the velocity field if available, or the field of any function of flow and occupancy can be analyzed similarly.

## Evolution of Probe Vehicle Travel Time

Though the occupancy field gives much information about the traffic condition for a day, a more informative aspect of the data for travel time prediction would be the daily pattern of observed travel times. We will define one such measure based on the probe vehicle travel times and examine time-of-day dependency of this parameter.

For a day $d$ , we have a set of probe vehicle travel times $\{\tau_d(s), s = s_{d,1},..., s_{d,R(d)}\}$ where $s_{d,i}, i = 1,..., R(d)$ are the departure times of each probe vehicle run on day $d$ . We linearly join $\tau_d(s)$ 's which are adjacent in time to get a regularly spaced time series for each day. We denote the interpolated travel times as the *travel time evolution* for the day $d$ and write it as $\{\hat{\tau}_d(t), t = 1,...,T\}$ . These series are plotted in Figure 4 for each scenario.

As a summary of these series, we calculate the historical median $\tilde{\tau}(t) = med_d \hat{\tau}_d(t)$ and the historical 1st and 3rd quartiles, $Q_1(t)$ and $Q_3(t)$ , of $\{\hat{\tau}_d(t), d = 1,..., D\}$ . In Figure 4, $\tilde{\tau}(t)$ , $Q_1(t)$ and $Q_3(t)$ are plotted as solid lines. The length of the vertical lines joining $Q_1(t)$ and $Q_3(t)$ , forming the shaded area is the interquartile range (IQR; the difference between $Q_1(t)$ and $Q_3(t)$ ), which serves as a measure of day-to-day variation of the evolution.

The plot in Figure 4 for South/AM shows that few days have unusually heavy traffic throughout this scenario. Although not as apparent in the other scenarios, clearly there are a few days in each scenario that show a significant level of congestion during a large portion of the time shift (1 to 2 hours). We will refer to these outlying days (as well as days with outlying occupancy fields) as *unusual* days. Because $\tau_d(t)$ is bounded by the minimum free flow travel time, the unusual days typically reflect unusually *bad* days, which can be observed from Figure 4. In other words, the distribution of travel time is skewed towards larger values.

Related to this property is the observation that the evolution of travel times on a given day is relatively smooth, implying dependence between successive probe vehicle travel times within a particular day. This property will be discussed more in the next section.

Finally, note that Figure 4 clearly shows the difference between the various scenarios, e.g., the PM travel exhibits greater variability than the AM. The directional and temporal differences suggest that location and time-of-day factors cannot be excluded.

**Unusualness Measure**

As shown above, most days are similar in terms of the travel time evolution, but the outliers can be significantly different. Further, probe vehicle travel time evolution appears auto-correlated within each day, meaning congestion is dispersed over time and space, rather then being localized in one or both dimensions. It will be useful for prediction if we can tell early in a day whether it will be an 'unusual' day or not. To measure how bad the traffic is compared to the 'average', we propose an *unusualness measure* $U_d$ for a day $d$,

$$U_d = \frac{1}{n_X n_T} \sum_{t=1}^{n_T} \sum_{x=1}^{n_X} | \xi_d(x,t) - \tilde{\tilde{\xi}}(x,t) |, \tag{4}$$

where $\xi_d(x,t) = f_d(x,t) / o_d(x,t)$ and $\tilde{\tilde{\xi}}(x,t) = med\{\xi_d(x,t), d = 1,..., D\}$. For this study, there are $n_X = 10$ loops and $n_T = 600$ time samples.

Figure 5 shows the plots of $U_d$ for 20 weekdays for each scenario. We can visually pick days when $U_d$ is substantially larger than the other days, e.g., the 4th day of South/PM or the 8th day of South/AM. Clearly, the "day of the week" is not sufficient to quantify these relationships. We expect $U_d$ can provide additional explanatory power.

It is plausible that the days with large $U_d$ in Figure 5 are likely to have $\hat{\tau}_d(t)$ outside the $Q_1(t) - Q_3(t)$ band for most of the day in Figure 4. This hypothesis is confirmed in Figure 6, where the travel time evolutions corresponding to days with relatively large $U_d$ are shown with bold lines. Although the relationship is not perfect (especially for North/PM), this unusualness measure appears to detect the unusual travel time evolutions well. In view of this observation, it makes sense to classify some days as usual ones and assume they have similar traffic flow patterns, while treating other days as outliers based on the value of $U_d$.

This work developed an unusualness measure using data for the entire day for illustration, but the approach is not useful for prediction. In practice, the analysis would be modified to use all data up to the most recent measurements to have a *predictive unusualness measure*. It is plausible that incorporating such a parameter in a travel time prediction, which is equivalent to modeling the travel time as a mixture of 'normal days' and 'bad days', will increase predictive power. For outliers or unusual days, historical information is not likely to be useful, while current condition can be informative. For 'usual' days, both historical and current information could be of some value. We simply note this idea but we do not include this analysis in our regression predictors, which will be introduced in the next section.

Another possibility is to use different $\xi_d(x,t)$ fields. The field we used, $\xi_d(x,t) = f_d(x,t) / o_d(x,t)$, has the property that $\xi_d(x,t) \approx v_d(x,t) / \hat{l}$, where $v$ ='true velocity for the given sample' and $\hat{l}$ ='assumed average vehicle length' which is held constant over all samples. Though this feature is reasonable, there are other possibilities such as $o_d(x,t) / f_d(x,t)$ or even $\tau_d(x,t)$.

## REGRESSION

The goal of this section is to predict travel time $\tau$ (sec) a fixed amount of time, $\Delta$ (min), in the future using available information. We will call $\Delta$ *prediction headway* or *lag*. For each scenario, this work fits the response variable, $\tau$, to the following covariates: departure data, $(s, w)$ where $s$ is the departure time and $w$ is the day of the week; and loop data, $(\mathbf{o}, \mathbf{f})$ at time $s - \Delta$.

Note the different nature of the loop covariate and departure covariate. The former can be thought of as *current information,* representing what is happening right now, while the latter can be though of as *historical information,* representing what has happened at this time of the day and this day of the week in the past. Presumably additional information, such as incidents or weather conditions, would improve the fit; but with only 20 days in this initial study, the decision was made to exclude these covariates for risk of over-fitting the data.

**Regression Methods**

Each scenario in the I-880 database has about 300 observations of $\tau_d(s)$ for $d = 1,...,20$ and $s = t_{d,1},...,t_{d,R(d)}$ for given $d$. We will identify $\tau_d(s)$ by a single index $i$ to write them as $\tau_i, i = 1,...,n$, where $n$ is total number of probe vehicle runs for all 20 days. Departure time, day and day of the week corresponding to $\tau_i$ are identified as $s_i$, $d_i$, and $w_i$. We then define $\mathbf{o}_i = \mathbf{o}_d(s_i - \Delta)$ and $\mathbf{f}_i = \mathbf{f}_d(s_i - \Delta)$. Thus, $X_i = (\mathbf{o}_i, \mathbf{f}_i, s_i, w_i)$ and $\tau_i$ become the covariate vector and the response variable, respectively.

We will treat these vectors $(X_i, \tau_i), i = 1,...,n$ as a *training sample* (an independent sample from an identical distribution) from a population for which we wish to construct a rule for predicting $\tau$ from $(\mathbf{o}, \mathbf{f}, s, w)$. As noted in the previous section, the sample is not a genuine training sample since there are dependencies within each day of the week, which are only partially eliminated by knowledge of the departure time. However, with this assumption we can apply linear, nonlinear or nonparametric regression to construct predictors.

Thus, the model is

$$\tau_i = g(X_i) + \varepsilon_i, \quad i = 1,...,n, \tag{5}$$

where $E(\varepsilon_i \mid X_i) = 0$ for all $i$ and $g$ is some function. Our aim is to find a function or 'predictor' $\hat{g}$ that can be calculated from the sample and is close to $g$. As a method to construct a predictor $\hat{g}$, *linear regression* and *tree methods* are considered.

Linear regression is a standard model where $g$ is assumed to be a linear function of covariates. Since using all variables in $X_i$ will lead to over-fitting, the stepwise method is used for variable selection (see (*12*) for description of linear regression and stepwise methods). Tree methods include many varieties, and we consider the one provided by S-plus (version 4.3). In the tree method, the model is fitted using binary recursive partitioning whereby the data are successively split along coordinate axes of the predictor variables. The split is done so that at any node the response variable is maximally distinguished in the left and the right branches. The splitting continues until data are too sparse for each node and then the tree is pruned using cross-validation. Terminal nodes are called leaves, while the initial node is called the root (for details, see (*13*)).

In addition to the tree method and linear regression model, neural networks (feed-forward neural networks as in (*13*)) with various numbers of hidden layers were tested. The neural networks did not perform as well as the other two methods in any situation, and so the results are omitted. However, one should note that neural nets are highly tunable and further adjustments may provide better results. More refined use of neural networks or other computer intensive prediction or learning methods such as support vector machines could be tried. One may well also try modern methods for improving a given prediction scheme such as Boosting, ARCing, etc.. These prediction schemes give highly non-linear (and non-structural) predictors and are known to give better prediction than traditional methods under some conditions. Since it is not our aim in this paper to find the best predictor for general situations we leave the application of these sophisticated methods to future studies.

**Measure for Prediction Error and Cross-Validation**

The prediction error is usually measured by Mean Squared Prediction Error (MSPE), defined as

$$MSPE = E(\tau - \hat{g}(X))^2(\sec^2), \tag{6}$$

where the expectation is taken with respect to both the distribution of the training set used to construct $\hat{g}$ and that of the test set $(X, \tau)$, which are independent with each other. Since MSPE (or other measures of prediction errors) cannot be calculated from the sample, it needs to be estimated. In estimating MSPE, a naive 'plug-in' estimate

$$MSPE^{plug-in} = \frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{g}(X_i))^2 \tag{7}$$

is overly optimistic, since it uses the same data set both for training and for testing. A partial remedy is to use the leave-one-out cross-validation MSPE estimate defined by

$$MSPE^{CV1} = \frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{g}_{(i)}(X_i))^2 \tag{8}$$

where $\hat{g}_{(i)}$ is a predictor constructed using all the data except the $i$ 'th observation. The cross-validation scheme we employ here is slightly more complex, which is done as follows. We leave a day out, construct a predictor using the remaining 19 days, and predict the travel times for the day we left out. It is repeated for each of 20 days and average of the squared empirical prediction errors can now be calculated. This method might be called 'leave a day out' cross-validation and formally, the estimate of prediction error can be written as

$$MSPE^{CV} = \frac{1}{n}\sum_{i=1}^{n}(\tau_i - \tilde{g}_{(i)}(X_i))^2 \tag{9}$$

where $\tilde{g}_{(i)}$ is a predictor constructed using $(X_j, \tau_j)$ 's for all $j$ 's belonging to different days from the day $i$ belongs to. Leave a day out cross-validation makes our estimate of prediction error more realistic since it partially takes into account the effect of the dependence between travel times on the same day that in fact makes our data not a genuine training sample.

As a relative measure of error, the cross-validation estimate of Mean Absolute Percentage Prediction Errors (MAPPE) defined by

$$MAPPE^{CV} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\tau_i - \tilde{g}_{(i)}(X_i)}{\tau_i}\right| \times 100(percent) \tag{10}$$

will be used.

**Results**

The cross-validation estimates of root-MSPE and MAPPE for various prediction headway $\Delta$ =5, 10, 20, 30, 60 (min) are summarized in Table 1 and Figure 7 for both tree and linear regression predictors. The root-MSPE values are given in seconds and should be considered relative to the median travel times, which are respectively 358, 397, 405 and 446 (sec) for South/AM, South/PM, North/AM and North/PM. The "No Loop Info" entry corresponds to a predictor that does not use the current loop information $(\mathbf{o}, \mathbf{f})$, but instead, relies strictly on the historical information $(s, w)$. These estimates serve as a baseline against which we can ask- "how much do we gain by using the current loop information?"

The tree method does not perform better than linear regression, both in terms of MSPE and MAPPE. Except for a few occasions, its errors are consistently larger than those of linear regression. Still, the differences are slight and their behavior differs from scenario to scenario and for different $\Delta$ 's, so neither of the two predictors is a clear winner. Performance of the individual method may differ from freeway to freeway and it would be more reasonable to use the method that performs best for a given situation rather than try to find a single regression method that performs well for all situations. In our situation, for example, it is perfectly reasonable to use tree method for $\Delta$ =0 while using regression for $\Delta$ =30 min for South/AM.

Still there are some patterns common to many situations, the most important of which are the benefits of current information in contrast to only using the baseline historical data. For all four scenarios, the root-MSPE for $\Delta = 0$ is 30%-40% smaller than the root-MSPE for the predictor that only relies on historical information. Likewise, the MAPPEs are reduced by about 30-40% as well. Even if we allow for the fact that estimates of prediction errors contain errors themselves, we can still consider $\Delta = 20$ (min) as maximum lags that give similar prediction error as lag 0 prediction for all scenarios. Thus, making use of the current loop information is desirable for prediction of travel time up to about 20 min in the future, which is consistent with Dynamic Trip Assignment research (*4-10*).

The stepwise method we employed usually chooses 8 to 10 variables out of 22 input explanatory variables. For shorter prediction headway up to $\Delta$=20 min, current loop information are chosen as significant variables at early stages. Especially, occupancies at those stations where traffic conditions vary significantly, i.e., stations with high variability in $\{o_d(x,t), d=1,...,20, t=1,...,600\}$ are chosen first (see Figure 2 and 3). In contrast, historical information, particularly the departure time, is chosen as most significant for longer prediction headway. The tree method shows similar behavior, i.e. similar sets of variables appear as nodes closer to the tree root.

Again, there is a lot of variation from one scenario to the next. For example, the PM travel time predictions are significantly worse than the AM in each direction in terms of both MSPE and MAPPE.

## CONCLUSIONS

Travel time prediction is an important task for ATIS and ATMS; however, the ability to predict future traffic conditions is non-trivial. To facilitate the analysis, this paper has introduced a number of metrics to characterize the spatial-temporal variations along the roadway: the scalar “unusualness” measure; the vector travel time evolution; and the two-dimensional occupancy field. As one would expect, these metrics are sensitive to location and time-of-day. Under some scenarios the metrics exhibited large day-to-day variations, while under other scenarios, they did not. The metrics also indicate a strong correlation between travel times during a given day.

Exploiting these phenomena, we found that simple prediction methods (such as linear regression on the current flow and occupancy measurements, departure time and day of the week) are beneficial for short-term travel time forecasts (up to 20 minutes), while historical data are better predictors for longer-range travel time predictions. Of course true travel time will not be observable under most situations; however, earlier works have shown that it can be estimated reliably from single loop detectors. Naturally, other relevant parameters should be included in the model when they are available, e.g., major events, incidents, or weather conditions.

With these factors in mind, the models presented in the final section are clearly site-specific and they should be viewed as such. It was not our intent to produce a single prediction model for all roadways; rather, we have provided a set of tools to develop and evaluate models that capture the relevant phenomena local to any site under study. Some notable features of our strategy are the use of cross-validation for efficient data utilization and digital filtering to smooth out measurement transients. Perhaps more importantly, the rigorous statistical analysis and verification presented in this work could be used independently from our models, providing consistent measures of effectiveness from one study to another.

Finally, note that the methodology presented in this work is relatively simple to implement, but as a result, the work is only applicable to short stretches of roadway (up to 15 km). Over longer distances, more complex models would be necessary to account for changing traffic conditions during a journey.

## ACKNOWLEDGMENTS

The authors would like to thank Joe Palen from Caltrans, as well as John Rice, Pravin Varaiya and Alex Skabardonis from the University of California, Berkeley for their helpful comments.

## REFERENCES

1. Skabardonis, A., Petty, K., Noeimi, H., Rydzewski, D. and Varaiya, P. I-880 Field Experiment: Data-Base Development and Incident Delay Estimation Procedures. In *Transportation Research Record* 1554, TRB, National Research Council, Washington D.C., 1996, pp. 204-212.
2. Petty, K., Bickel, P., Jiang, J., Ostland, M., Rice, J., Ritov, Y., and Schoenberg, F. Accurate Estimation of Travel Times from Single-Loop Detectors. *Transportation Research, Part A*, Vol. 32A, No. 1, 1998, pp. 1-17.
3. Coifman, B. Vehicle Reidentification and Travel Time Measurement in Real-Time on Freeways Using the Existing Loop Detector Infrastructure. In *Transportation Research Record* 1643, TRB, National Research Council, Washington D.C., 1998, pp. 181-191.
4. Ben-Akiva, M., de Palma, A. and Kaysi, I. Dynamic Network Models and Driver Information Systems. *Transportation Research, Part A*, Vol. 25A, No. 5, September 1991, pp. 251-266.

5. Ben-Akiva, M., Cascetta, E., Gunn, H., Smulders, S. and Whittaker, J. DYNA: A Real-Time Monitoring and Prediction System for Inter-Urban Motorways. *Proc. of the First World Congress on Applications of Transportation Telematics and Intelligent Vehicle- Highway Systems*, Vol 3, ERTICO, 1994, pp. 1166-1180.
6. Ben-Akiva, M., Cascetta, E. and Gunn, H. An On-Line Dynamic Traffic Prediction Model for and Inter-Urban Motorway Network. *Urban Traffic Networks - Dynamic Flow Modeling and Control*, Springer, 1995, pp. 83-122.
7. Mahmassani, H. and Peeta, S. System Optimal Dynamic Assignment for Electronic Route Guidance in a Congested Traffic Network. *Urban Traffic Networks - Dynamic Flow Modeling and Control*, Springer, 1995, pp. 3-37.
8. Wild, D. Pattern-Based Forecasting. *Proceedings of the Second DRIVE-II Workshop on Short Term Traffic Forecasting*, TNO Institute for Policy Studies, Delft, The Netherlands, 1994, pp. 49-63.
9. Uerlings, U. The Prediction System within the Socrates Information Centre. *Proceedings of the Second DRIVE-II Workshop on Short Term Traffic Forecasting*, TNO Institute for Policy Studies, Delft, The Netherlands, 1994, pp. 27-47.
10. Hoffman, G. and Janko, J. Travel Times as a Basic Part of the LISB Guidance Strategy. *Proc. of the Third International Conference on Road Traffic Control*, IEE Conference Publication Number 320, 1990, pp. 6-10.
11. Park, D. and Rilett, L. Forecasting Multiple-Period Freeway Link Travel Times Using Modular Nural Networks. In *Transportation Research Record* 1617, TRB, National Research Council, Washington D.C., 1998, pp. 163-170.
12. Draper, N. R. and Smith, H. *Applied Regression Analysis*. John Wiley and Sons, Inc., New York, 1981.
13. Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York, 1994.

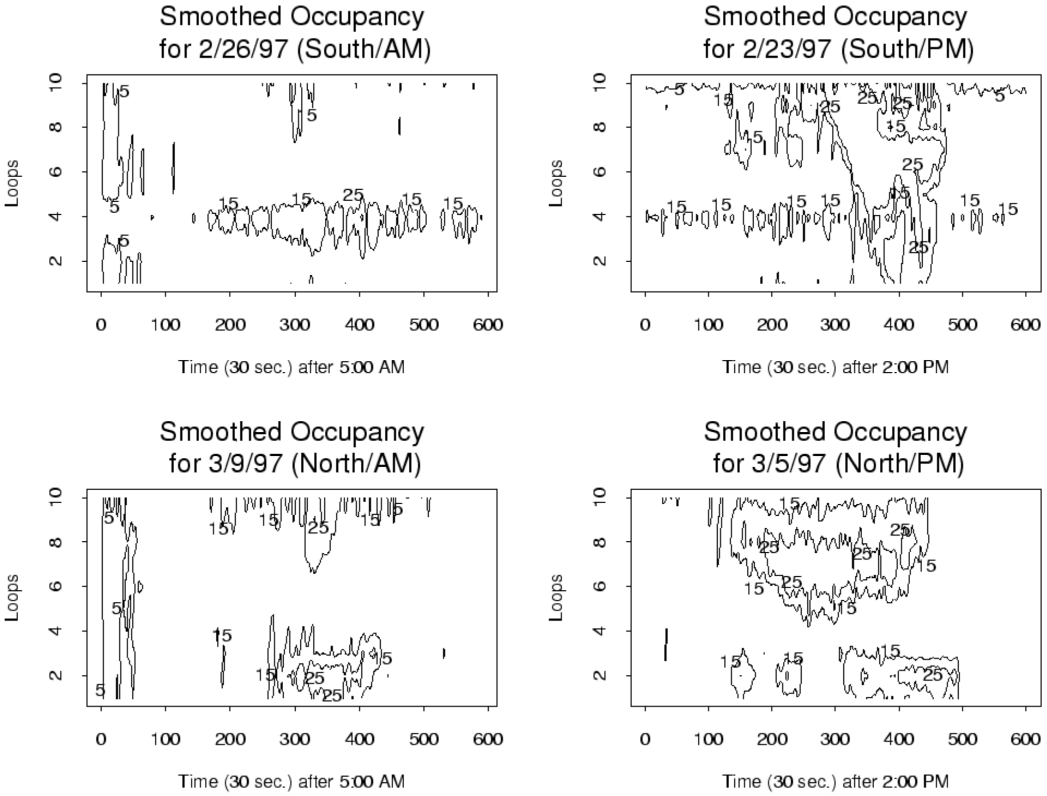
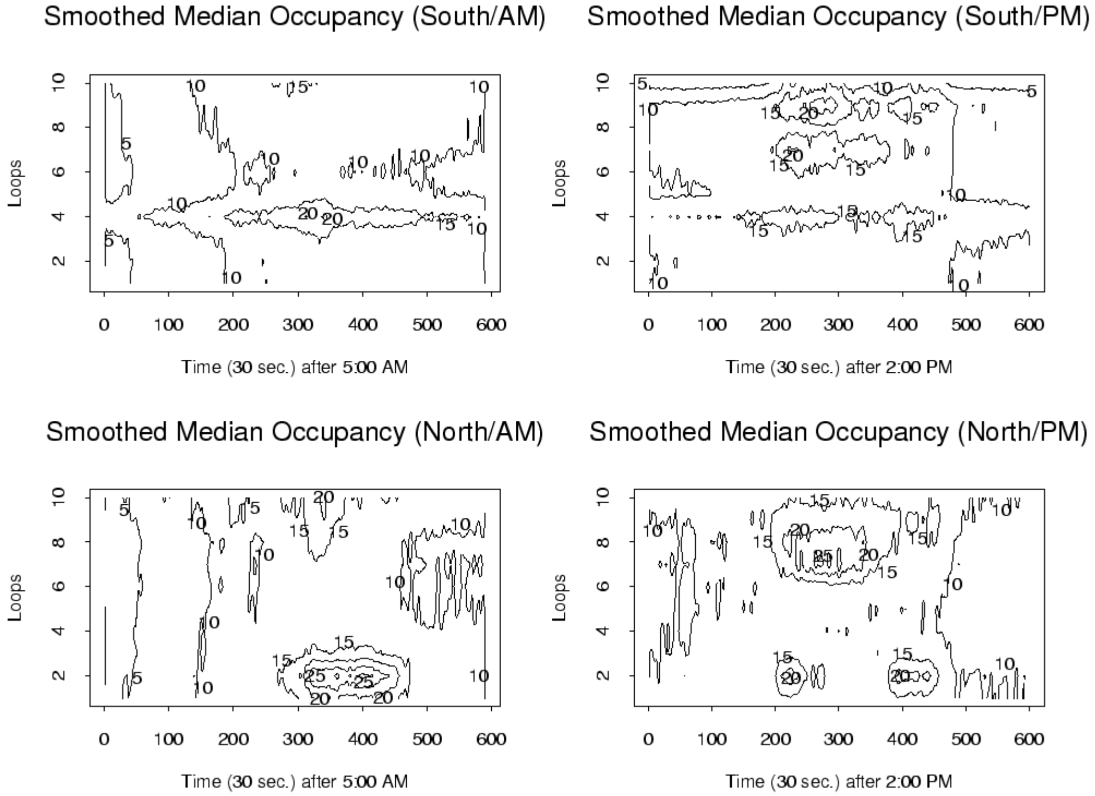


FIGURE 1. Contour plots of smoothed occupancy field for a random day from each of the four scenarios. The axes represent the time-space plane and curves indicate contours of the occupancy levels 5, 15, 20 and 25.


Smoothed Median Occupancy (South/AM)

Smoothed Median Occupancy (South/PM)

Smoothed Median Occupancy (North/AM)

Smoothed Median Occupancy (North/PM)

FIGURE 2. Contour plots of smoothed median occupancy field for each scenario.

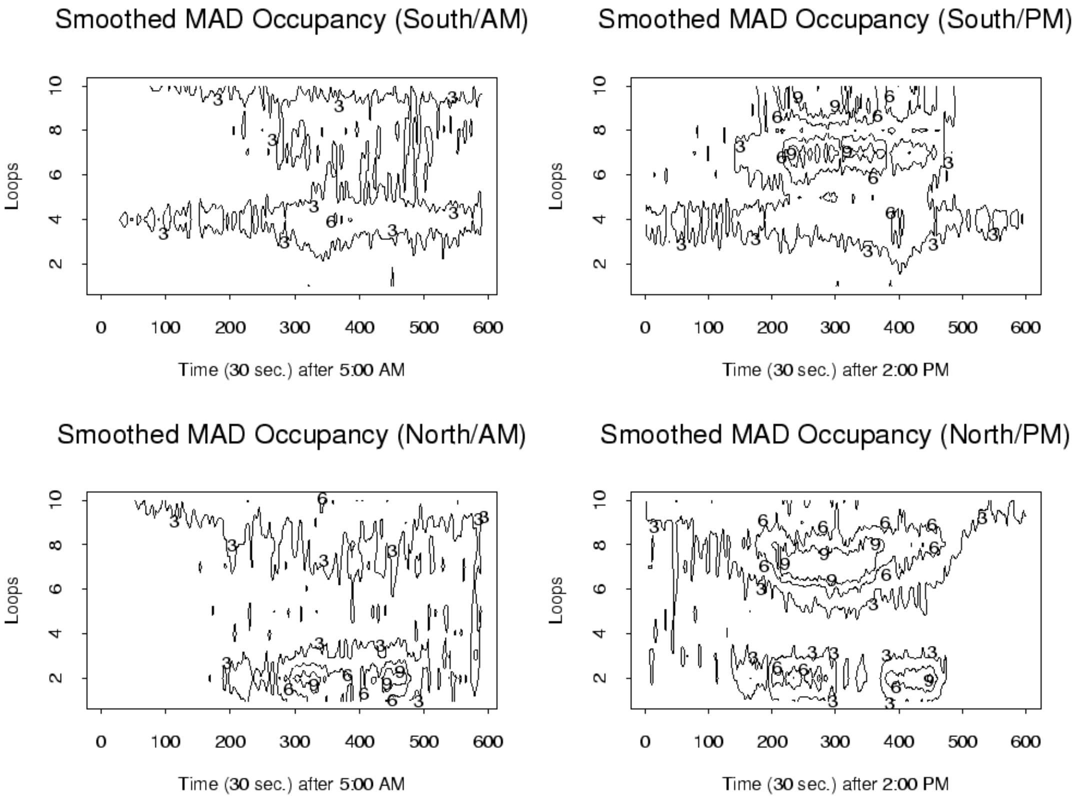


FIGURE 3. Contour plots of smoothed MAD occupancy field for each scenario.

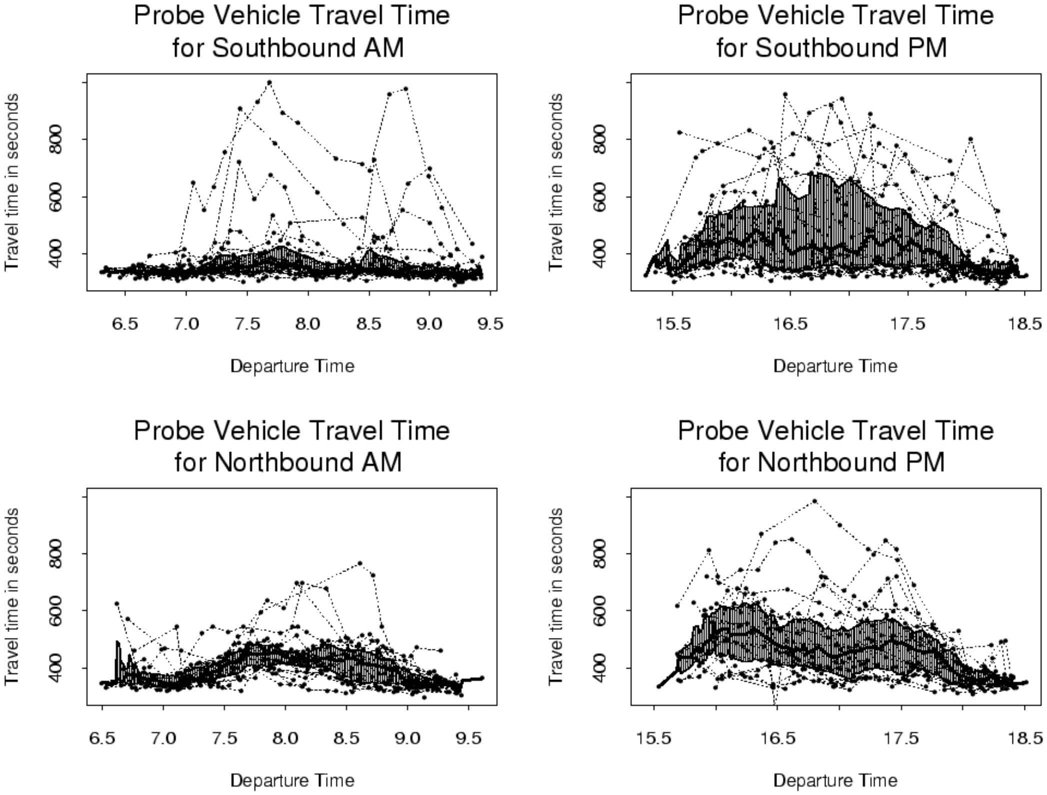


FIGURE 4. Plots of probe vehicle travel time evolutions for 24 days for each scenario. Each dotted line corresponds to a single day's evolution. The first and third quartile and median of these series are plotted in solid lines and vertical lines join the former two. Each point represents a single probe vehicle run over the freeway segment.

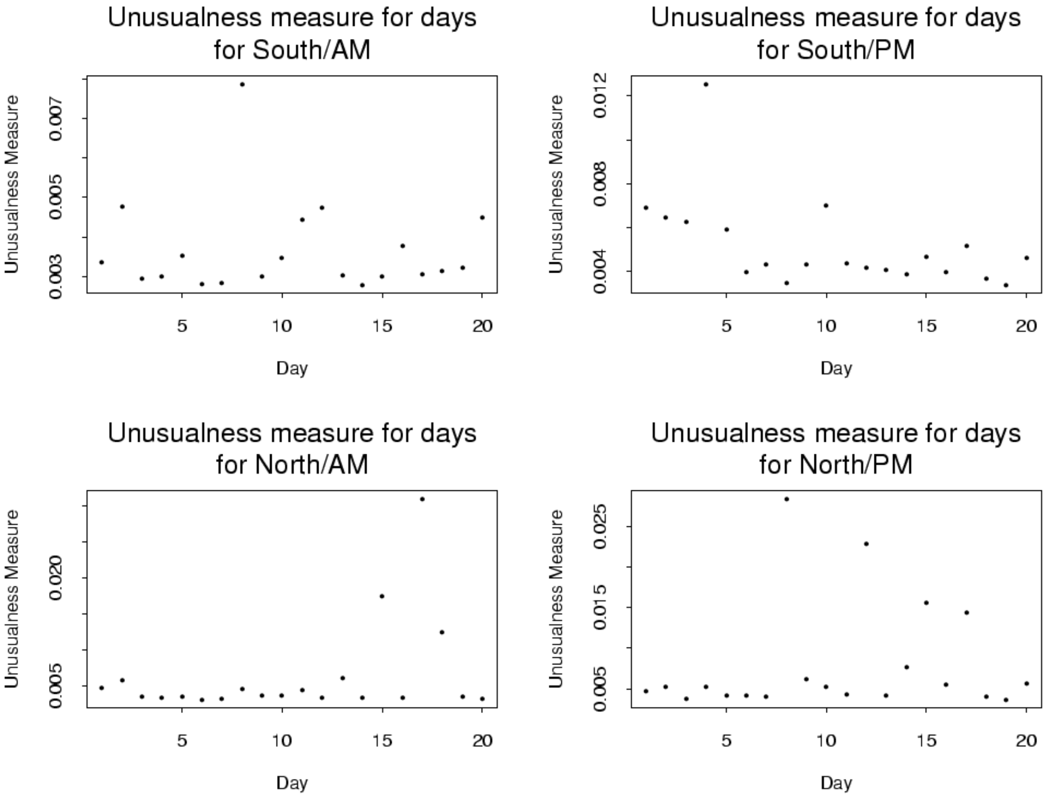


FIGURE 5. Plots of unusualness measure $U_d$ for 20 days from each scenario. Day 1 corresponds to 2/22/1993 and day 20 to 3/19/1993. Day 1, 6, 11, 16 are Mondays, 2, 7, 12, 17 are Tuesdays, etc.

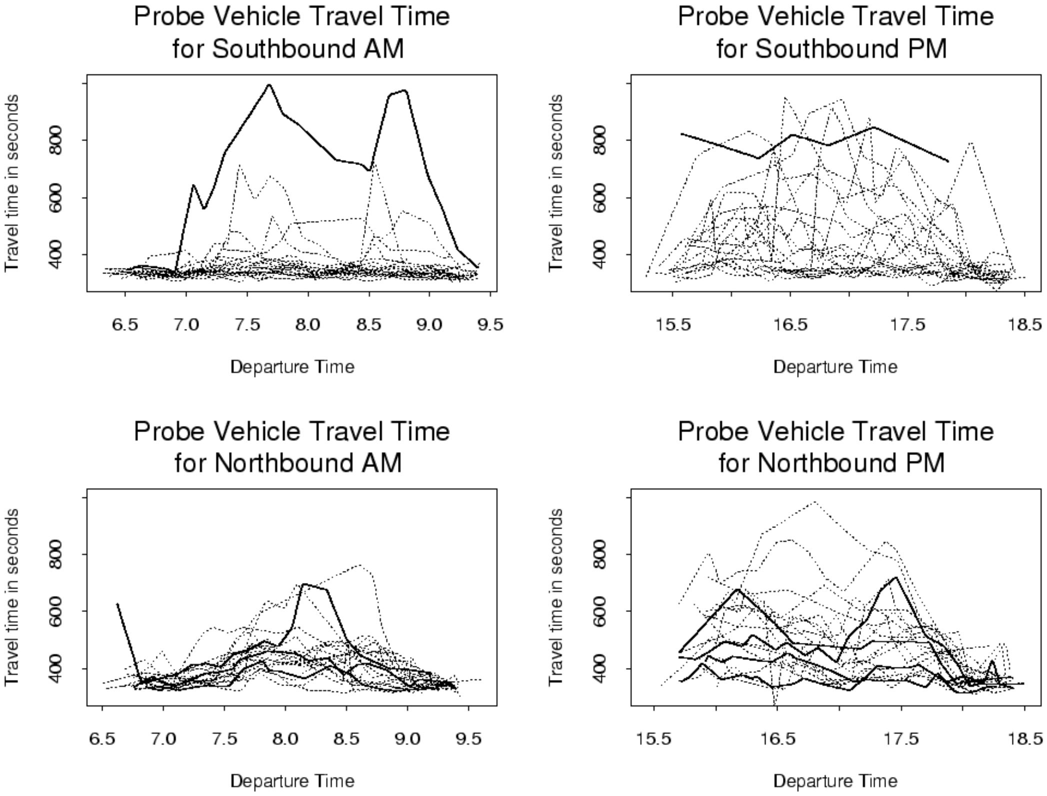


FIGURE 6. The probe vehicle travel time evolutions for each scenario, drawn for 20 days for which loop data is also available. Evolutions corresponding to days with high $U_d$ are drawn in thick solid lines.

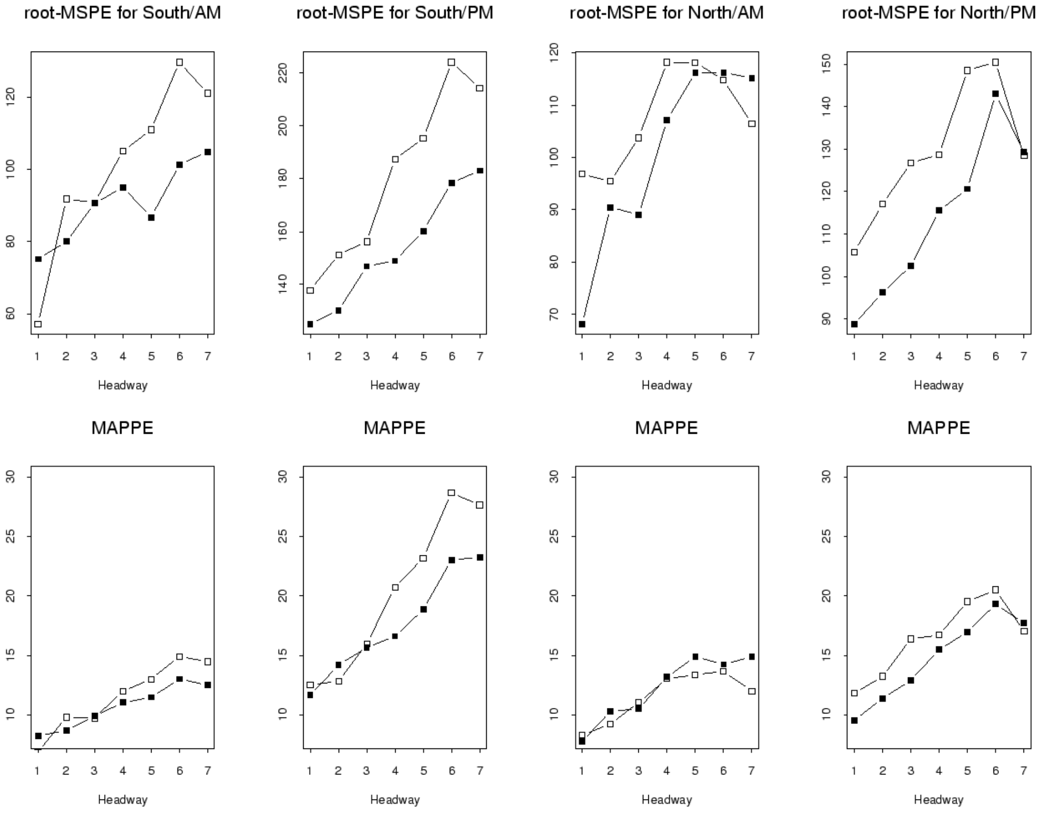


FIGURE 7. Cross-validation estimates of prediction errors of linear regression and tree predictor for South/AM, South/PM, North/AM and North/PM, from left to right. Top plots are root-MSPE and bottoms are MAPPE. Black boxes correspond to linear regression and white boxes to tree method. Each headway index means: 1= 0min, 2= 5min, 3= 10min, 4=20min, 5=30min, 6=60min, 7= "No Loop Entry". Bottom plots are scaled across the row for readability. See Table 1 for their values.

**TABLE 1. Cross-validation Estimates of Prediction Errors of the tree method and linear regression with stepwise variable selection method.**

| Scenario | Prediction Headway (min) | Tree Method | | Linear Regression | |
|---|---|---|---|---|---|
| | | $\sqrt{MSPE}$ (Sec.) | MAPPE (%) | $\sqrt{MSPE}$ (Sec.) | MAPPE (%) |
| South/AM | 0 | 57 | 6.9 | 75 | 8.2 |
| | 5 | 92 | 9.8 | 80 | 8.7 |
| | 10 | 91 | 9.7 | 91 | 9.9 |
| | 20 | 105 | 12.0 | 95 | 11.0 |
| | 30 | 111 | 13.0 | 87 | 11.5 |
| | 60 | 130 | 14.9 | 101 | 13.0 |
| | No Loop Info | 121 | 14.5 | 105 | 12.5 |
| South/PM | 0 | 138 | 12.5 | 125 | 11.6 |
| | 5 | 151 | 12.8 | 130 | 14.2 |
| | 10 | 156 | 16.0 | 147 | 15.6 |
| | 20 | 187 | 20.7 | 149 | 16.6 |
| | 30 | 195 | 23.2 | 160 | 18.9 |
| | 60 | 224 | 28.7 | 178 | 23.0 |
| | No Loop Info | 214 | 27.6 | 183 | 23.3 |
| North/AM | 0 | 97 | 8.3 | 68 | 7.7 |
| | 5 | 95 | 9.2 | 91 | 10.3 |
| | 10 | 104 | 11.0 | 89 | 10.5 |
| | 20 | 118 | 13.0 | 107 | 13.2 |
| | 30 | 118 | 13.4 | 116 | 14.9 |
| | 60 | 115 | 13.7 | 116 | 14.2 |
| | No Loop Info | 106 | 12.0 | 115 | 14.9 |
| North/PM | 0 | 106 | 11.8 | 89 | 9.5 |
| | 5 | 117 | 13.2 | 96 | 11.4 |
| | 10 | 127 | 16.4 | 102 | 12.9 |
| | 20 | 129 | 16.7 | 116 | 15.5 |
| | 30 | 149 | 19.5 | 121 | 17.0 |
| | 60 | 150 | 20.5 | 143 | 19.3 |
| | No Loop Info | 129 | 17.0 | 129 | 17.7 |