MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Univariate Short-Term Prediction of Road Travel Times

D. Nikovski N. Nishiuma Y. Goto
H. Kumazawa

TR2005-086 July 2005

Abstract

This paper presents an experimental comparison of several statistical machine learning methods for short-term prediction of travel times on road segments. The comparison includes linear regression, neural networks, regression trees, k-nearest neighbors, and locally-weighted regression, tested on the same historical data. In spite of the expected superiority of non-linear methods over linear regression, the only non-linear method that could consistently outperform linear regression was locally-weighted regression. This suggests that novel iterative linear regression algorithms should be a preferred prediction methods for large-scale travel time prediction.

This paper has been published in Proceedings of the 2005 IEEE Intelligent Transportation Systems Conference, Vienna, Austria, September 2005.

Final version submitted June 2004.

Univariate Short-Term Prediction of Road Travel Times

D. Nikovski, N. Nishiuma, Y. Goto and H. Kumazawa

Abstract— **This paper presents an experimental comparison of several statistical machine learning methods for short-term prediction of travel times on road segments. The comparison includes linear regression, neural networks, regression trees, k-nearest neighbors, and locally-weighted regression, tested on the same historical data. In spite of the expected superiority of non-linear methods over linear regression, the only non-linear method that could consistently outperform linear regression was locally-weighted regression. This suggests that novel iterative linear regression algorithms should be a preferred prediction methods for large-scale travel time prediction.**

I. Introduction

Prediction of highway and urban traffic is becoming a major application area within the field of intelligent transportation systems (ITS), and is considered to be one of the most important components of advanced travelers information systems (ATIS). Currently existing travel information systems, such as the popular MapQuest service, provide travel times only under free-flow conditions, i.e. in the absence of any congestion within the transportation network. Since most major urban areas experience heavy congestion precisely when commuters need accurate travel information the most, free-flow travel times are of little or no value to them. Although most cities already have or are currently introducing information services that inform drivers of the most up-date travel conditions, this information is of limited value either — what drivers actually need is information on the conditions along a particular stretch of road at the moment when they plan to be there. Such information is possible to obtain only by means of prediction into the future.

Travel times at a future moment are determined by the state of the network at that moment, as represented by the flows and densities of vehicles on all of its segments. In its turn, the state of the network is determined by three factors:

1) Future demand for the network, as measured by the number of vehicles in it, their intended trips, and the preferred routes these vehicles would take.
2) Future capacity of the network, as measured by the throughput of its segments, and any traffic accidents and/or road work that might reduce this capacity. Since specific future traffic accidents cannot be predicted, only currently existing accidents can be taken into account, if they are expected to persist until the prediction horizon.
3) The current state of the network, as measured by the number of vehicles currently in it, and their routes until the completion of their respective trips. Even if future demand could normally be handled by the available capacity without causing congestion, if such congestion already exists at the current moment, it would take some time for it to dissipate.

Any successful prediction method should take all of these factors into consideration. Furthermore, these factors suggest two trivial prediction methods for future travel times that can be used for comparison purposes:

1) **Current travel times**. This prediction would be exact if future demand and capacity were equal to the current ones, and were also at an equilibrium, i.e. the state of the network would remain the same.
2) **Long-term average travel times for the specific time of day of the prediction.** This prediction would be exact if future demand and capacity were equal to their historic averages, and furthermore, the current and past states of the network had no impact on its future state. Long-term average values can be obtained easily from a database of historical data, by averaging the travel times over all days at this specific hour. A more detailed prediction can use averages over the same day of the week, if enough data are available.

II. Traffic Prediction Methods

Traffic prediction methods generally fall into two broad classes: those based on dynamic traffic assignment (DTA), and those based on statistical machine learning (ML). Dynamic traffic assignment methods attempt to model exactly the physical process that governs the evolution of the state of the network. These methods rely on detailed simulations of the traffic process, which is both their biggest advantage and their biggest disadvantage. DTA methods can provide a solution for any number of scenarios and combinations of events, such as multiple accidents, lane closures, etc. However, these methods are only reliable when they can be calibrated precisely, which is typically a very hard and labor-intensive task.

In contrast to the DTA approach, ML methods largely ignore the exact nature of the physical process that governs the evolution of the network state. Instead, ML methods treat the prediction task as a supervised machine learning problem, where the current and past states of the network are used as model input, and the future state of the network is used as model output. Any available information about the physical system that generates the input-output pairs is only used as a guide to the type of parametric model that is likely to result in

D. Nikovski is with Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA, USA, nikovski@merl.com

N. Nishiuma, Y. Goto and H. Kumazawa are with Mitsubishi Electric Corporation, 8-1-1 Tsukaguchi-Honmachi, Amagasaki, Hyogo, Japan, {nishiuma.norihiro, goto.yukio, kumazawa.hiroyuki}@wrc.melco.co.jp

good prediction. Still, the knowledge of how DTA solutions operate can be very instructive about the complexity of the ML predictive models to be used.

From the point of view of statistical machine learning, the data observed at network road segments constitute a multivariate time series generated by an unknown dynamical system. Consequently, the prediction problem is amenable to the whole arsenal of machine learning methods. The ML approach to short-term traffic prediction has been researched extensively, and most major predictive algorithms have been tried at this problem, including linear regression, univariate and multivariate state-space methods (ARIMA), neural networks, k-nearest neighbors, locally-weighted regression, Kalman filtering, knowledge-based methods, etc.

However, there is a major difference between typical time series prediction problems, and the task of predicting travel times throughout the day. Most time series prediction problems assume that the exogenous factors acting upon the dynamical system either remain constant, or can be measured and accounted for in the model, if they vary in time. Under such assumptions, a single ARMA/ARIMA/ARMAX model can be fit to the entire time series.

This is not the case with predicting travel times: the demand on the transportation network, which acts as the main exogenous factor, varies widely throughout the day, and is typically hard to quantify. Consequently, it is not reasonable to expect that a single predictive model would model equally well travel times in peak periods, such as morning peak hour, and off-peak periods. This means that for the purposes of travel time prediction, separate models should be fit to different periods during the day. Since travel times are usually reported at regular intervals, e.g. 5 or 15 minutes, a separate model can be fit for each such interval.

In predictive applications, the most important modeling choice — that of predictive model — depends on the type of dependency that is believed to exist between model inputs and outputs. Traditionally, the simplest possible statistical prediction method is linear regression, where predicted values are regressed upon time-lagged past observations of the time series via linear coefficients. These models are also known as auto-regressive (AR) models. A more general case is that of ARIMA, or state-space approaches [1].

When the time series is multivariate, the modeler has the choice of splitting it into multiple independent AR models for each predicted variable, or fitting a single multivariate AR model, also known as a vector autoregressive (VAR) model. Since there are significant interactions between different road segments, it is likely to be expected that VAR models would be applicable to the traffic prediction problem. Another name for these models is space-time ARIMA (STARIMA), underlining the fact that the individual components of the time series are related spatially and temporally to each other. It can be expected that their use would result in increased accuracy, because the state of neighboring links would be taken into consideration when determining future travel times.

Moving beyond linear models, a natural modeling choice is neural networks (NN). There are numerous applications of both feed-forward and recurrent NN to the problem of short-term traffic prediction [2]. These applications hope to exploit the well known ability of neural nets to model complex non-linear relationships. However, one significant disadvantage of NN is their slow training rate, which makes them an unlikely candidate for large-scale traffic prediction.

Space-partitioning methods, such as regression trees and advanced variants such as CART and MARS, recursively split input space with the objective of reducing the inhomogeneity of training samples left in each partition [3]. These methods have the advantage of producing results that are easily interpretable by humans. However, they are designed to work mainly in batch mode, and although some algorithms allow iterative updates in real-time, the order of presenting training data has a significant impact on the shape of the final tree.

Another large group of prediction methods is that of non-parametric regression, also known as memory-based learning, instance-based learning, etc. Some example of non-parametric regression models are k-nearest neighbors, kernel averaging, and locally-linear regression. The main idea of non-parametric regression models is to delay the building of a predictive model until the actual time when a prediction must be made. At that time, a set of relevant data points is selected, and a local model is built for this specific prediction. One of the most promising methods for short-term traffic prediction is locally-weighted regression, and one of its variants, a linear-model with time-varying coefficients [4]. A comparison between parametric (ARIMA) models and non-parametric ones is given in [5].

III. Experimental Comparison

In our experiments, we used data collected from the Vehicle Information and Communication System (VICS) in Japan, collected over a main road of length 15km over 14 months. The first 12 months were used as training data, and the remaining 2 months were used as testing data. Since we used data from a single road segment, all experiments used univariate time series models. Under free-flow conditions, the usual travel time along this segment is around 16 minutes, but under heavy congestion, it can exceed an hour and a half.

Current travel times were reported by VICS every 5 minutes, or 288 times throughout the day. In these experiments, we were interested in the accuracy of the predictive algorithms over different prediction horizons, aggregated over all prediction times during the day. Consequently, for each predictive method described below, we fit 262 different models for each 5-minute interval between 2am and 11:55pm, using the training data, and measured the root mean squared error (RMSE) (in minutes) on the testing data. This error was averaged over all 262 prediction points, i.e., over all predictive models. A total of 24 different prediction horizons were considered, ranging from 5 minutes to 2 hours into the future, at 5-minute intervals. All experiments were performed in the statistical environment R, using the additional packages *nnet*, *rpart*, *class*, and *locfit* [3].

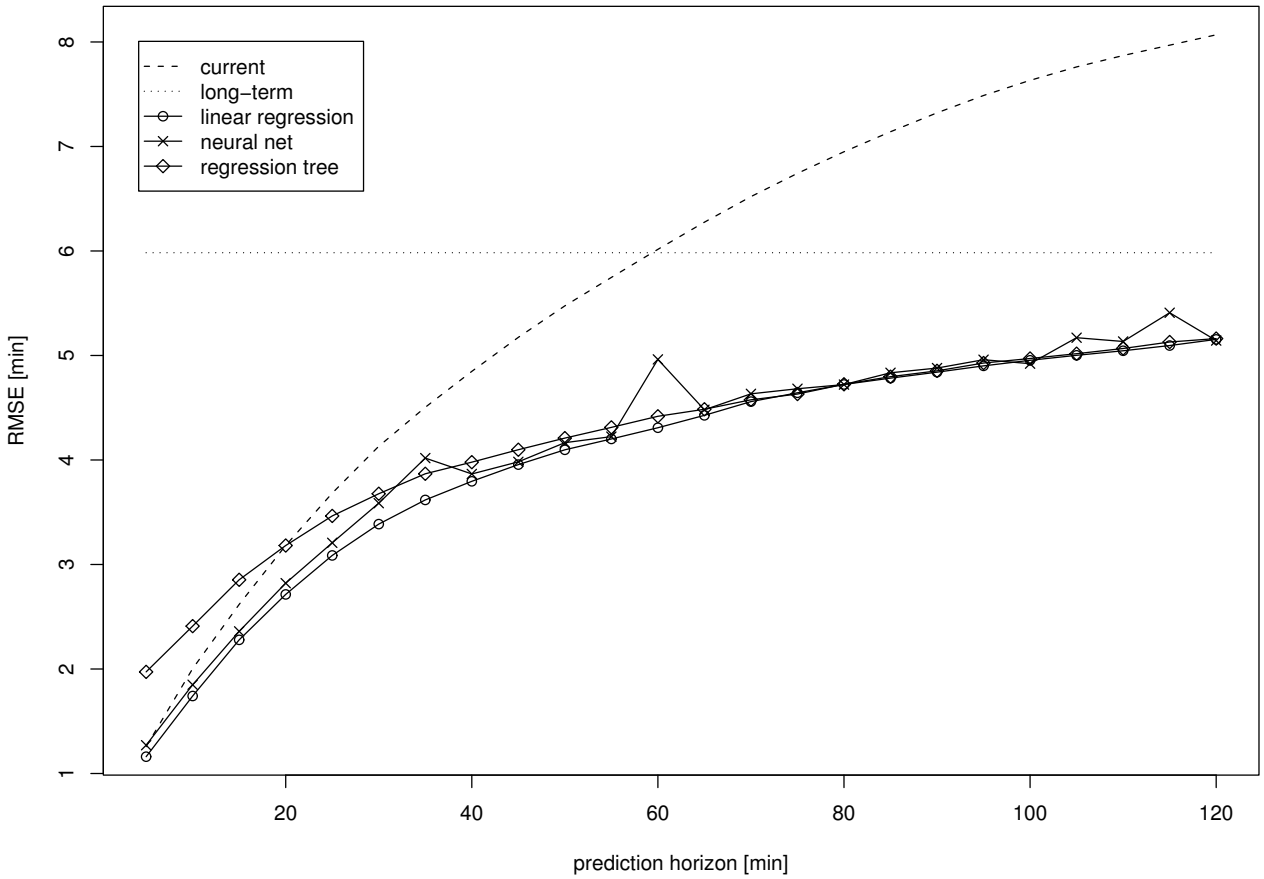

Fig. 1. Comparison between linear regression, neural networks, and regression trees. All three methods are significantly better than the trivial predictors, but still linear regression is better and more stable than the non-linear methods. The neural net does not generalize reliably, while regression trees cannot achieve better accuracy than linear regression at any prediction horizon, and are systematically inferior for short prediction horizons (less than one hour).

As described above, machine learning methods treat the prediction task as a supervised machine learning problem, where the current state of the network is used as model input, and the future state of the network is used as model output. The main modeling questions then are which variables should be used as predictors, how long in the past should the predictors go, and what kind of predictive model should be fit to the data.

The available data, collected from the VICS network, determined to a large extent the answer to the first question - the predictor variables used in prediction were the travel times on the same road segment, collected at 5-minute intervals in the past. While consistent with many other predictive models, this approach omits the information contained in other variables, such as the vehicle density along the segment, as well as the travel times on connected segments. (Reliable connectivity information was not available, in general.)

A. Dimensionality of the dynamical system

The answer to the second question, how far in the past should the predictors go, depends on the intrinsic dimensionality of the dynamical system that generates the data. From the point of view of statistical machine learning, the data observed at network road segments constitute a multivariate time series generated by an unknown dynamical system. A theorem due to Takens states that a system of dimension d needs at most $2d + 1$ past readings for successful prediction [6]. (For example, a two-dimensional dynamical system such as a linear harmonic oscillator needs at most five past readings for successful prediction.)

We were surprised to discover that the dynamical system corresponding to road traffic seemed to have a very low dimension. Experiments with linear regression suggested that only the two most recent readings were statistically significant predictors to future travel times. (And in some cases, even only the most recent reading was sufficient.) For example, in order to predict the travel time at 8:00am, it was sufficient to know only the travel times at 7:55 and 7:50, while travel times at earlier moments (7:45 and earlier) were not statistically significant predictors, as long as those at 7:55 and 7:50 were included in the model. This was true regardless of how recent the most recent reading was — if the most

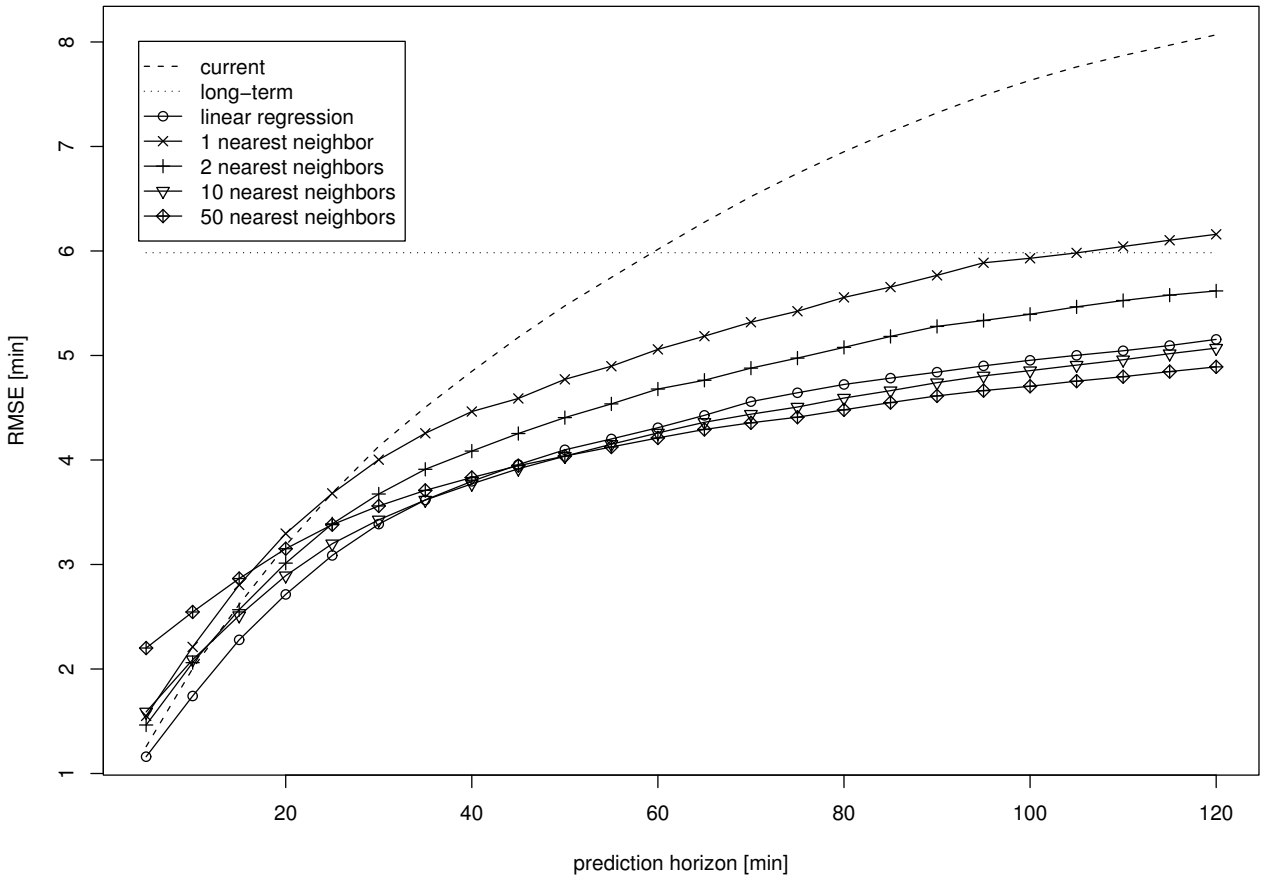

Fig. 2. Comparison between k-nearest neighbors and linear regression. There is a strong dependency on the number of nearest neighbors considered — in general, more neighbors result in higher accuracy at longer prediction horizons, while fewer neighbors are better for shorter horizons. In summary, although kNN is sometimes better than linear regression for some horizons, no single value of k can consistently outperform linear regression.

recent reading was at 7:30, then that reading and the reading at 7:25 were the only significant predictors.

This effect is illustrated in Table 1. The goal was to predict the travel time at 8am (variable T0800) from the travel times between 7:30 and 7:55, by means of linear regression. The only really statistically significant predictor was that at 7:55 (T0755), where statistical significance is beyond doubt (t-value=16.316). All of the earlier travel times are quite useless for improving the prediction, as long as T0755 is included in the model. (Of course, this is only true for linear regression — non-linear models might still pick up other significant predictors.)

In order to test this hypothesis, we varied the number of input variables of linear regression from one to three. Since statistical significance of past travel times as linear predictors decreases rapidly as they become less recent, only the three most recent travel times (at moments t, $t-5$, and $t-10$ minutes) were considered in the predictive model for time t. After averaging over all 262 prediction points t, we discovered that although the accuracy of the models on the training error increased slightly with the number of input variables, as expected for a model with more degrees of freedom, no such increase in accuracy could be observed on the testing set. This suggests once again that the most current travel time is the only input variable necessary for linear prediction. This result is consistent with the approach chosen in [4].

This effect has a perfectly rational explanation. Road traffic is a pretty inert and slowly evolving system, apparently without any harmonic components that would require high-dimensional models. The most recent reading provides information about the current level of congestion on the road, while the reading before it provides information on the direction of change in congestion (increasing or decreasing), and the rate of change. Apparently, these are the only two variables that have an effect on the future level of congestion for a specific moment in time, and sometimes even the direction of change is not important. In other words, the current level of congestion has a major effect on the future level of congestion, but most often it is not important how the current level of congestion was reached. This observation has a very positive effect on the complexity of predictive models,

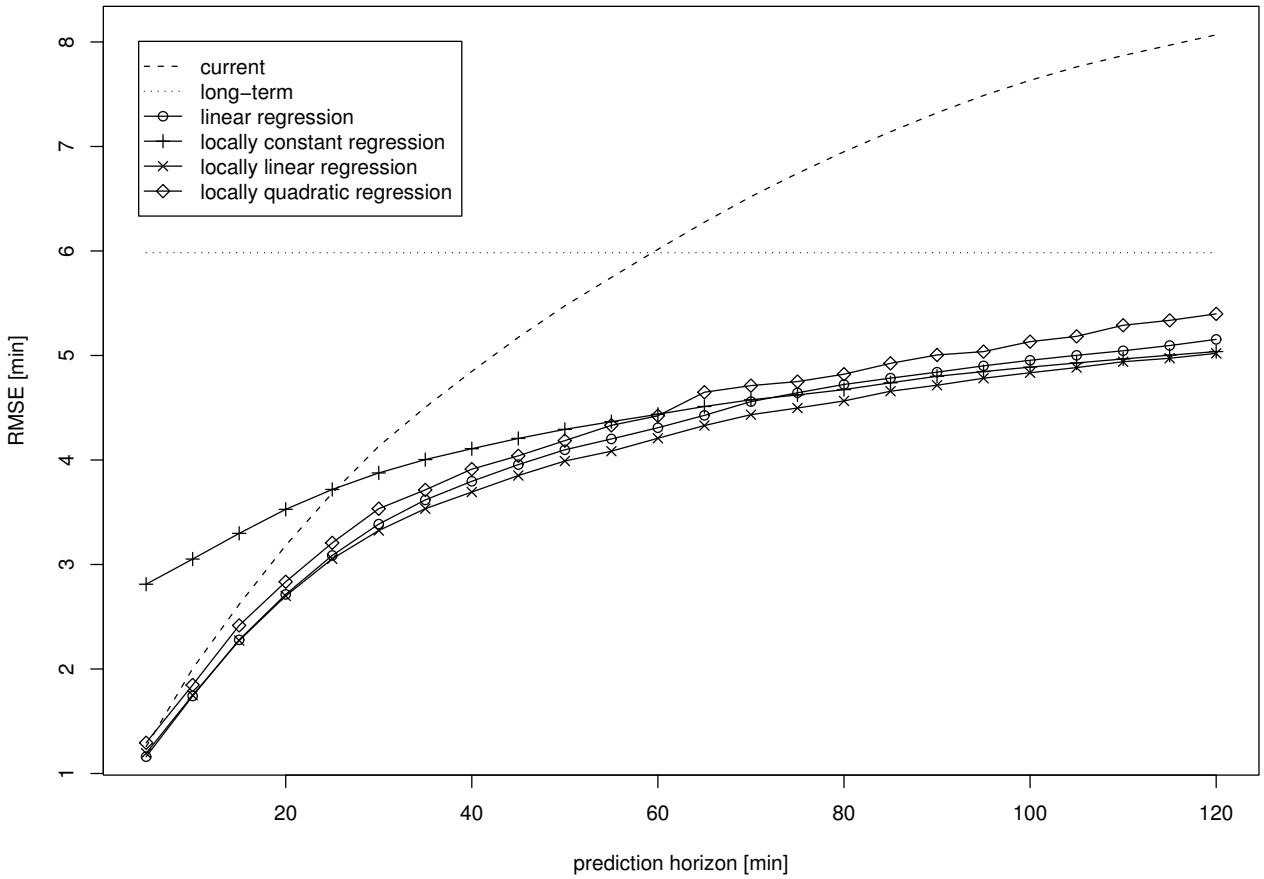

Fig. 3. Comparison between locally-weighted regression and linear regression. Locally weighted regression clearly outperforms linear regression, but not by much.

because it suggests that a real-time predictive system would have to keep in memory only the few most recent readings for the purpose of model updating, and can safely discard those from earlier times. Based on this discovery, all of our remaining experiments included only the three most recent travel times as predictors.

B. Neural networks

As possible candidates for other predictive models, we considered various non-linear machine learning methods. Moving beyond linear regression, a natural modeling choice is neural networks (NN). There are numerous applications of both feed-forward and recurrent NN to the problem of short-term traffic prediction [2]. One significant disadvantage of NN is their slow training rate, which makes them an unlikely candidate for large-scale traffic prediction. Furthermore, one has to choose carefully the complexity of the model, expressed by the number of units in the hidden layer. After several experiments, we obtained acceptable performance by a neural net with 5 units in the hidden layer, plus direct connections between the input and output layers, bypassing the hidden layer. The performance of the neural net in comparison to linear regression and the baseline predictors (current travel time and its long-term average) is shown in Fig.1. Both predictive methods are significantly better than the simple baseline predictors, but still linear regression is better and more stable than the neural network. This can be attributed to the unreliable manner in which neural nets generalize — although the training algorithm managed to train the neural net well on the training set in all cases, its performance on the training set was often systematically wrong. The graph also shows that the difference between the accuracy of the first baseline predictor (current travel time) and that of the learning methods increases with the prediction horizon, as expected.

C. Regression trees

The second non-linear machine learning method we explored was regression trees [3]. Regression trees are similar to decision trees in their operation, but differ in their application to predict continuous variables, rather than discrete ones. Fig.1 demonstrates a surprising behavior — although the regression tree can predict travel times at longer prediction horizons (more than one hour) equally well as linear

	Lin. Coeff.	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.919035	0.400073	2.297	0.0223
T0755	0.931041	0.057063	**16.316**	$< 2E-16$
T0750	0.108272	0.074738	1.449	0.1484
T0745	-0.024510	0.085604	-0.286	0.7748
T0740	0.007005	0.096753	0.072	0.9423
T0735	-0.077798	0.087802	-0.886	0.3763
T0730	0.003608	0.070759	0.051	0.9594

TABLE I

TABLE 1. STATISTICAL SIGNIFICANCE OF LINEAR COEFFICIENTS FOR SEVERAL PREDICTORS. TRAVEL TIMES BETWEEN 7:30AM (VARIABLE T0730) AND 7:55 (VARIABLE T0755) ARE USED TO PREDICT THE TRAVEL TIME AT 8:00AM IN A LINEAR MODEL. ESTIMATES FOR THE LINEAR COEFFICIENTS OF THE REGRESSION MODEL ARE SHOWN IN COLUMN LIN. COEFF. COLUMN t-VALUE IS THE t-STATISTIC FOR THE COMPUTED STANDARD ERROR OF THE LINEAR COEFFICIENT, AND $Pr(>|t|)$ IS THE PROBABILITY THAT SUCH A LARGE VALUE OF t WOULD BE OBSERVED PURELY BY CHANCE WHEN THE REGRESSION COEFFICIENT IS IN FACT ZERO. ONLY THE MOST RECENT TRAVEL TIME (T0755) HAS MAJOR STATISTICAL SIGNIFICANCE AS A LINEAR PREDICTOR; THE CONSTANT TERM (INTERCEPT) IS ALSO SIGNIFICANT AT THE 2.23% CONFIDENCE LEVEL, WHICH SIMPLY MEANS THAT TRAVEL TIME BETWEEN 7:55AM AND 8:00AM IS INCREASING.

regression, its accuracy at short horizons is quite poor, and much worse than that of linear regression. Given that fitting a linear tree is much slower than fitting a linear regression, it seems that linear regression is a surprisingly good prediction method.

D. Non-Parametric Regression

Our last prediction models were k-nearest neighbors (kNN) and locally-weighted regression (LWR), both of which are non-parametric regression models. These types of models are also known as memory-based learning, instance-based learning, etc. [7]. The main idea of non-parametric regression models is to delay the building of a predictive model until the actual time when a prediction must be made. At that time, a set of relevant data points is selected, and a local model is built for this specific prediction. For the case of kNN, the output values of the selected data points are simply averaged, while for LWR, a local regression model is fitted to the selected points. Moreover, the prediction residual of each data point is weighted proportionally to its proximity to the novel input.

A comparison between parametric (ARIMA) models and non-parametric ones is given in [5]. We explored kNN and LWR on the same data set, and the results are shown in Fig.2 and Fig.3. The simpler of the two methods, kNN, exhibits a strong dependency on the size of the local likelihood, as expressed by the number of local neighbors k. In general, more neighbors result in higher accuracy at longer prediction horizons, while fewer neighbors are better for shorter horizons. The consequence of this behavior is that no single value of k could consistently outperform linear regression.

Locally weighted regression methods can vary in the degree of the local model to be fit, and some common choices are locally-constant (similar to kNN, but using proximity weights for all data points, rather than only those in the local neighborhood), locally-linear, and locally-quadratic. Interestingly, the best results were achieved for a local model of degree one (locally linear), rather than for the more flexible locally-quadratic model. Locally weighted linear regression was also the only predictive method among all presented in this paper that systematically outperformed linear regression.

IV. CONCLUSION

The paper presented experiments comparing the predictive accuracy of five statistical machine learning methods: linear regression, neural networks, regression trees, k-nearest neighbors, and locally-weighted regression. The surprising result is that linear regression is very competitive in terms of accuracy, and given its significant advantage in terms of computational time and memory resources, it is a very strong candidate for deployment in practical systems, especially if fast iterative updating schemes are employed [8]. However, it is still possible that more complicated non-linear relationships exist between different road segments, and non-linear methods could still turn out to be better in multivariate short-term traffic prediction.

REFERENCES

[1] A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction," *Transportation Research Part C*, vol. 11, no. 3, pp. 121–135, 2003.

[2] M. S. Dougherty and M. R. Cobbett, "Short-term inter-urban traffic forecasts using neural networks," *International Journal of Forecasting*, vol. 13, no. 1, pp. 21–31, 1997.

[3] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S. Fourth Edition*. New York: Springer, 2002.

[4] J. Rice and E. van Zwet, "A simple and effective method for predicting travel times on freeways," in *Intelligent Transportation Systems*, pp. 227–232, IEEE, 2001.

[5] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C*, vol. 10, no. 4, pp. 303–321, 2002.

[6] F. Takens, "On the numerical determination of the dimension of an attractor," in *Dynamical systems and bifurcations, Groningen 1984* (B. L. J. Braaksma, H. W. Broer, and F. Takens, eds.), vol. 1125 of *Lecture Notes in Mathematics*, pp. 99–106, Berlin: Springer-Verlag, 1985.

[7] A. W. Moore and C. G. Atkeson, "Memory-based function approximators for learning control," tech. rep., MIT, MIT Artificial Intelligence Laboratory, Cambridge, MA 02139, July 1992.

[8] N. Nishiuma, Y. Goto, H. Kumazawa, D. Nikovski, and M. Brand, "Traffic prediction using singular value decomposition," in *Proceedings of the 11th World Congress on ITS, Nagoya, Oct 18-22*, 2004.