PREDICTION INTERVALS FOR TRAFFIC TIME SERIES

By

Hongyu Sun

Department of Civil and Environmental Engineering

University of Wisconsin at Madison, 1415 Engineering Drive, Madison, WI 53706

Tel: (608) 262-2524, Fax: (608) 262-5199, E-mail: sun@cae.wisc.edu

Chunming Zhang

Department of Statistics, University of Wisconsin at Madison,

4357 Computer Sciences and Statistics Building,

1210 W. Dayton Street, Madison, WI 53706

Tel: (608) 262-0084, E-mail : cmzhang@stat.wisc.edu

Bin Ran

Department of Civil and Environmental Engineering

University of Wisconsin at Madison

1415 Engineering Drive, Madison, WI 53706

Tel: (608) 262-0052, Fax: (608) 262-5199, E-mail: bran@engr.wisc.edu

Keechoo Choi, Ph.D

Associate Dean of Graduate School of ITS

Professor, Dept. of Transportation Eng./College of Eng., Ajou University

Woncheon-Dong, Paldal-Ku Suwon, 442-749 KOREA (Republic Of)

Tel: 82-(31) 219-2541 Office, Fax: 82-31-215-7604, Email: keechoo@ajou.ac.kr

Word Count: 5,419

Resubmitted on Nov 11th, 2003

ABSTRACT

Traffic forecasting systems could be improved significantly by the development of interval prediction at a given probability (confidence level) as well as point prediction which seems to have drawn most attention from researchers in this field so far. Not only can the provision of interval prediction increase the user comfort by reducing error risk associated with the information, it can also be used to assess the predictor forwards (not afterwards) for model selection and preemption in an adaptive or cooperative setting.

Few concrete research results on interval prediction seem to be presented so far in this area, partially because of the increasing complexity of the prediction models. Although closed-form expressions of asymptotic estimation for prediction intervals are not usually available for nonparametric models such as neural networks, they can be derived for the local polynomial predictor. This paper addresses the issue of the interval forecasting (constructing prediction intervals for future observations) of the traffic data time series.

Two approaches will be investigated. First, an asymptotic result will be presented. Closed-form equations of prediction intervals for the multivariate local linear regression model have been derived under certain assumptions. Second, a bootstrap approach to generate prediction intervals with bias corrections will be proposed. The bootstrap procedure relies on few assumptions and is easy to implement for many prediction models. Finally, a case study using real-world traffic data will be presented for both approaches, along with the results compared with each other. The results coincide with expectations and have validated the proposed methods.

KEYWORDS

Interval Forecasting, Prediction Interval, Prediction Bias, Bootstrap, Local Linear Regression, Short-Term Traffic Prediction, Nonparametric Model

1 INTRODUCTION

Short-term prediction by means of regression consists of two steps: The first step uses training data (historical and/or real-time) to approximate the conditional mean regression function (relationship) between inputs (data at one or more time instants) and the output (data at future time instants, with respect to the input data). Once this function is established, the second step uses it to estimate future data relative to current input. Usually, the value of the mean function for the given input data is used as the prediction.

The regression model can be parametric (such as linear regression, AR model, Kalman filter, multifractal) or nonparametric (such as neural network, local polynomial, wavelet, chaos). For traffic prediction, Sun et al. (2003) studied the local linear regression model, which belongs to the family of the local polynomial regression in nonparametric models. The conditional mean function was used to obtain the prediction.

Estimating bias, variance and prediction interval is not new for most parametric models. This is especially true for the linear regression predictor and for Gaussian data for which strategies have been well established. Although such measurements are in increasing demand in transportation application, the usually low accuracy of the parametric models for dealing with highly nonlinear, nonstationary traffic data, and the usually high complexity of the nonparametric models, have led to little research in this area. One related literature found so far, is a presentation by Rilett (2001) in which the LOESS[1] is used to analyze historical traffic patterns such as mean, variance and confidence limit.

Prediction intervals often can be constructed if the prediction of mean, bias and intervals can be calculated. Besides the conditional mean function mentioned above, the conditional variance function is needed to obtain intervals. The derivatives of the mean function can be used to obtain bias. All these results can be derived for the local polynomial regression by extending the results from global models to local models. For example, the prediction interval in linear models can be converted to local linear models (Schaal et al., 1994).

[1] a software program for smoothing multivariate scattered data by locally weighted least square criteria.

On the other hand, such derived expressions based on certain assumptions may not perform very well for real-world traffic data. For example, the mean function for traffic prediction is usually not smooth enough to have the second derivative. In that case, the bias cannot be estimated using the equations. Also, the residual error distribution for the small sample is often unknown instead of being assumed as normal or *t*-distribution. In this context, the more general bootstrap method may be proposed.

Bootstrap is a simple resampling procedure which generates samples by randomly resampling the original training set with replacements (Enfron et al., 1993). The idea borrows the spirit similar to Monte Carlo simulation. It has found wide application in solving all kinds of "hard" real-world problems, in estimating bias and interval when the data distribution is unknown or sample size is small. Care must be taken in applying the bootstrap in nonparametric situations and dependent data such as time series. Several papers by Freedman and Peters (1984; 1987) documented the fact that the bootstrap does not give the correct answer for multivariate regression situations where the number of variables is of a similar order as the number of observations. This paper will propose a method of applying bootstrap to time series prediction when implemented by the local linear model. A bootstrap procedure that is plausibly similar for prediction interval was described by Cho et al. (2003), however, the prediction interval in that paper refers to the interval of prediction error which was calculated based on feedback errors of prediction in order to verify the accuracy of the forecast, which could be treated as a measure of uncertainty of the forecast. That is, the interval was calculated after the observations. Instead, the prediction interval in this paper is constructed for future observations before they arrive.

This article is organized as follows. Section 2 will describe the closed-form equation and the bootstrap scheme of interval prediction, both for the local linear regression model. Section 3 will be devoted to numerical study. Discussion and future research directions will be provided in Section 4.

2 METHODOLOGY

2.1. Asymptotic Prediction Interval for the Local Linear Regression Model

This section will, first, briefly review the mean prediction of the local linear model and address the variance prediction. Then the estimator bias and variance will be introduced. Finally, interval prediction will be derived.

2.1.1. MEAN AND VARIANCE PREDICTION

Given the observations $\{(\mathbf{X}_i^T, Y_i)$: $i = 1, \ldots, n\}$ of the multivariate covariate $\mathbf{X}$ and a univariate response Y, the relationship between $\mathbf{X}$ and Y can be modeled as:

$$Y = m(\mathbf{X}) + s(\mathbf{X})e, \tag{1}$$

where $\mathbf{X}$ and e are not necessarily independent, e is the additive error term with

$$\mathrm{E}(e|\mathbf{X}) = 0 \tag{2}$$

and

$$\mathrm{Var}(e|\mathbf{X}) = 1. \tag{3}$$

Here n is the number of the observations,

$$\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_d)^T \tag{4}$$

And

$$\mathbf{X}_i = (\mathrm{X}_{i1}, \ldots, \mathrm{X}_{id})^T \tag{5}$$

with d the dimension of $\mathbf{X}$.

It is of interest to estimate the mean regression function

$$m(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x}) \tag{6}$$

and the possible heteroscedasticity (conditional variance function)

$$s^2(\mathbf{x}) = \mathrm{Var}\,(Y|\mathbf{X} = \mathbf{x}), \tag{7}$$

where

$$\mathbf{x}^T = (x_1, .., x_d) \tag{8}$$

is a point in R^d.

Once the estimated mean regression function (denoted as $\hat{m}(\ )$) is obtained, the fitted regression is used as a mechanism for prediction of response values. That is, if the prediction of Y at $\mathbf{X} = \mathbf{x}$ is denoted as $\hat{y}(\mathbf{x})$, then

$$\hat{y}(\mathbf{x}) = \hat{m}(\mathbf{x}). \tag{9}$$

$\hat{y}(\mathbf{x})$ may be viewed as the estimated response. It is the estimated mean response at $\mathbf{X} = \mathbf{x}$.

2.1.1.1. *Mean*

A local polynomial model is formed at the query point **x**, much as a Taylor series model, a function in the neighborhood of a point. In the local linear model, the Taylor expansion terms up to the first (linear) order are used to make the local approximation. That is, the function *m* is estimated locally by a linear model:

$$m(\mathbf{X}) \approx m(\mathbf{x}) + \mathbf{g}^T(\mathbf{X} - \mathbf{x}), \tag{10}$$

where $\mathbf{g} = (b_1, \ldots, b_d)^T$. For the convenience of a matrix expression, redefine the vectors taking into account the constant term. Write

$$\mathrm{b} = (b_0, b_1, \ldots, b_d)^T, \tag{11}$$

where

$$b_0 = m(\mathbf{x}), \tag{12}$$

and

$$\tilde{\mathbf{X}} = (1, (\mathbf{X} - \mathbf{x})^T)^T, \tag{13}$$

Then

$$m(\mathbf{X}) \approx \mathrm{b}\,\tilde{\mathbf{X}} \tag{14}$$

The observations $\{(\mathbf{X}_i^T, Y_i): i = 1, \ldots, n\}$ are used as training data to estimate b. The weighted least square criterion is used to obtain the fit (Fan et al., 1996). The local model will fit nearby training points well with less concern for distant points by the weighting functions (kernels) with the bandwidth defining the validity region of the local model. The effective number of training points is modulated by the bandwidth of the kernels.

The estimation of b

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}_d\boldsymbol{\beta})^T \mathbf{W}(\mathbf{y} - \mathbf{X}_d\boldsymbol{\beta}) \tag{15}$$

is

$$\hat{\boldsymbol{\beta}} = (\hat{b}_0, \ldots, \hat{b}_d)^T = \left(\mathbf{X}_d^T\mathbf{W}\mathbf{X}_d\right)^{-1}\mathbf{X}_d^T\mathbf{W}\mathbf{y}, \tag{16}$$

where

$$\mathbf{X}_d = \begin{pmatrix} 1 & X_{11} - x_1 & \cdots & X_{1d} - x_d \\ 1 & X_{21} - x_1 & \cdots & X_{2d} - x_d \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} - x_1 & \cdots & X_{nd} - x_d \end{pmatrix}, \quad (17)$$

$$\mathbf{W} = \text{diag}\{K_B(\mathbf{X}_i - \mathbf{x})\}, \quad (18)$$

$$\mathbf{y} = (Y_1, \ldots, Y_n)^T \quad (19)$$

And

$$K_B(\mathbf{u}) = \tfrac{1}{|B|} K(B^{-1}\mathbf{u}), \quad (20)$$

where $K(\bullet)$ is a multivariate probability density function with mean zero and the covariance matrix of $m_2(K)\mathbf{I}_d$, with $\mathbf{I}_d$ the $d \times d$ identity matrix. $\boldsymbol{B}$ is called bandwidth matrix and $|\boldsymbol{B}|$ denotes its determinant. In this study, the weighting kernel K is chosen as a Gaussian function and $\boldsymbol{B} = h\mathbf{I}_d$, where h is called bandwidth. That is,

$$K(\mathbf{u}) = e^{-(dis(\mathbf{u}))^2} \quad (21)$$

The distance function $dis(\bullet)$ used in this study is the Euclidean distance.

$$dis(\mathbf{u}) = \sqrt{\mathbf{u}^T\mathbf{u}} \quad (22)$$

Thus

$$\hat{m}(\mathbf{x}) = \hat{b}_0, \quad (23)$$

and

$$(\frac{\partial \hat{m}}{\partial x_j})(\mathbf{x}) = \hat{b}_j, j = 1, \ldots, d. \quad (24)$$

The prediction value $\hat{y}$ is equal to $\hat{b}_0$. That is,

$$\hat{y}(\mathbf{x}) = \hat{m}(\mathbf{x}) = \hat{b}_0 = \mathbf{q}^T\hat{b} = \mathbf{q}^T\left(\mathbf{X}_d^T\mathbf{W}\mathbf{X}_d\right)^{-1}\mathbf{X}_d^T\mathbf{W}\mathbf{y}$$

$$= \mathbf{p}_\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} p_i(\mathbf{x})Y_i \quad (25)$$

where

$$\mathbf{q} = (1, 0, \ldots, 0)^T. \quad (26)$$

The vector $\mathbf{p}_\mathbf{x}$, also written as $\mathbf{p}(\mathbf{x})$, will be useful for calculating the bias and variance of the local model.

$$\mathbf{p}_\mathbf{x} = \mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_n(\mathbf{x}))^T$$

$$= \left(\mathbf{q}^T \left(\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d\right)^{-1} \mathbf{X}_d^T \mathbf{W}\right)^T \quad (27)$$

It is easy to find that

$$\mathbf{p}_x^T \mathbf{p}_x = \mathbf{q}^T \left(\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d\right)^{-1} \mathbf{X}_d^T \mathbf{W} \left(\mathbf{q}^T \left(\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d\right)^{-1} \mathbf{X}_d^T \mathbf{W}\right)^T$$

$$= \mathbf{q}^T \left(\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d\right)^{-1} \mathbf{q} \quad (28)$$

and

$$\Sigma p_i(\mathbf{x}) = 1. \quad (29)$$

2.1.1.2. ***Variance***

Atkeson et al. derived (1997) an estimate of $s^2(\mathbf{x})$, the local noise variance, that is, the variance for the traffic data in our study. First, some additional quantities in terms of the weighted variables are defined. A locally weighted linear regression centered at a point $\mathbf{x}$ produces local model parameters $b(\mathbf{x})$. It also produces errors (residuals) at all training points. The weighted residual $r_i(\mathbf{x})$ is given by:

$$r_i(\mathbf{x}) = w_i(\mathbf{x}) \mathbf{X}_i^T b(\mathbf{x}) - w_i(\mathbf{x}) Y_i \quad (30)$$

where

$$w_i(\mathbf{x}) = \sqrt{K(d(\mathbf{X}_i, \mathbf{x}))}\,. \quad (31)$$

The training criteria which are the weighted sum of the squared errors is:

$$C(\mathbf{x}) = \sum_i r_i^2(\mathbf{x})$$

$$= \left(\mathbf{y} - \mathbf{X}_d \mathbf{ß}\right)^T \mathbf{W} \left(\mathbf{y} - \mathbf{X}_d \mathbf{ß}\right) \quad (32)$$

A reasonable estimator for the local value of the noise variance is

$$\hat{s}^2(\mathbf{x}) = \frac{\sum r_i^2(\mathbf{x})}{n_{LL}(\mathbf{x})} = \frac{C(\mathbf{x})}{n_{LL}(\mathbf{x})} \quad (33)$$

where n_{LL} is a modified measure of how many data points there are:

$$n_{LL}(\mathbf{x}) = \sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} K\left(\frac{d(X_i, \mathbf{x})}{h}\right) \quad (34)$$

In an analogy to unweighted regression (Myers, 1990), the bias of the estimate $\hat{s}^2(\mathbf{x})$ can be reduced by taking into account the number of parameters in the local linear regression:

$$s^2(\mathbf{x}) = \hat{s}^2(\mathbf{x}) \frac{n_{LL}(\mathbf{x})}{n_{LL}(\mathbf{x}) - p_{LL}(\mathbf{x})}$$

$$= \frac{\sum r_i^2(\mathbf{x})}{n_{LL}(\mathbf{x}) - p_{LL}(\mathbf{x})} \quad (35)$$

where $p_{LL}(\mathbf{x})$ is a measure of the local number of the free parameters in the local model:

$$p_{LL}(\mathbf{x}) = \sum_i w_i^4 X_i^T (\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d)^{-1} X_i \quad (36)$$

2.1.2. ESTIMATOR VARIANCE AND BIAS

For any estimator $\hat{m}(\mathbf{x})$, define

$$\text{Bias}(\hat{m}(\mathbf{x})) = \text{E}\{\hat{m}(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\} - m(\mathbf{x}) \quad (37)$$

$$\text{Var}(\hat{m}(\mathbf{x})) = \text{Var}\{\hat{m}(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\} \quad (38)$$

$$= \text{E}\{[\hat{m}(\mathbf{x}) - \text{E}\{\hat{m}(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\}]^2 \mid \mathbf{X} = \mathbf{x}\}$$

Mean Squared Error:

$$\text{MSE}(\mathbf{x}) = \text{E}\{[\hat{m}(\mathbf{x}) - m(\mathbf{x})]^2 \mid \mathbf{X} = \mathbf{x}\}$$

$$= \text{Bias}^2(\mathbf{x}) + \text{Var}(\mathbf{x}) \quad (39)$$

An important aspect of local polynomial learning is that it is possible to estimate the prediction error and derive confidence intervals (bounds) on the predictions. In order to develop the intervals, the parameter $\text{Var}(\hat{y}(\mathbf{x}))$ must be determined. A standard error $s_{\hat{y}(\mathbf{x})}$ of $\hat{y}(\mathbf{x})$ can be interpreted as the standard error of the estimator of mean response, conditional on $\mathbf{x}$. The notion standard error, of course, evokes the image of precision or variation. In this case, it reflects the variation of $\hat{y}$ at $\mathbf{x}$, if repeated regressions were conducted, based on the same $\mathbf{X}$-levels and new observations on Y each time.

The variance and bias of the multivariate local linear estimator are shown below as given by Atkeson et al. (1997).

$$\text{E}(\hat{y}(\mathbf{x})) = \text{E}(\mathbf{q}^T (\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d)^{-1} \mathbf{X}_d^T \mathbf{W} \mathbf{y}) = \text{E}(\mathbf{p}_\mathbf{x}^T \mathbf{y})$$

$$= m(\mathbf{x}) + \mathbf{p}_\mathbf{x}^T (\mathbf{m} - \mathbf{X}_d \mathbf{b})$$

$$= m(\mathbf{x}) + \mathbf{p}_\mathbf{x}^T \mathbf{t} \quad (40)$$

where

$$\mathbf{m} = [m(\mathbf{X}_1), \ldots, m(\mathbf{X}_n)]^T \quad (41)$$

and

$$\mathbf{t} = \mathbf{m} - \mathbf{X}_d\mathbf{b} \ . \tag{42}$$

$$\begin{aligned}\mathrm{Var}(\hat{y}(\mathbf{x})) &= \mathrm{Var}(\hat{m}(\mathbf{x})) \\ &= \mathrm{E}[\hat{y}(\mathbf{x}) - \mathrm{E}(\hat{y}(\mathbf{x}))]^2 \\ &= s^2(\mathbf{x})\,\mathbf{p}_\mathbf{x}{}^T\mathbf{p}_\mathbf{x}\end{aligned} \tag{43}$$

From (40), it is easy to get

$$\begin{aligned}\mathrm{Bias}(\hat{y}(\mathbf{x})) &= \mathrm{Bias}(\hat{m}(\mathbf{x})) = \mathrm{E}(\hat{y}(\mathbf{x})) - y_{true}(\mathbf{x}) \\ &= \mathrm{E}(\hat{y}(\mathbf{x})) - m(\mathbf{x}) = \mathbf{p}_\mathbf{x}{}^T\,\mathbf{m} - m(\mathbf{x}) \\ &= \mathbf{p}_\mathbf{x}{}^T\,\mathbf{t} = \mathbf{p}_\mathbf{x}{}^T(\mathbf{m} - \mathbf{X}_d\mathbf{b})\end{aligned} \tag{44}$$

Since the estimate of $s^2(\mathbf{x})$ can be given by (35), the estimator variance can be computed by using (43). If the estimator $s^2(\mathbf{x})$ is substituted by $s^2(\mathbf{x})$, from (43) the standard error of prediction can be defined as

$$s_{\hat{y}(\mathbf{x})} = s(\mathbf{x})\sqrt{\mathbf{p}_\mathbf{x}^T\mathbf{p}_\mathbf{x}} \tag{45}$$

Assessing the bias requires making assumptions about the underlying form of the true function, and the data distribution. The local linear model exactly matches any linear trend in the data.

Using Taylor's expansion (Fan et al., 1996) of $m(\mathbf{X})$, the bias (44) can be estimated.

$$\begin{aligned}m(\mathbf{X}) &= m(\mathbf{x}) + \mathbf{g}^T(\mathbf{X} - \mathbf{x}) + \text{higher terms of } (\mathbf{X} - \mathbf{x}) \\ &= \mathbf{X}_d\mathbf{b} + \text{higher terms of } (\mathbf{X} - \mathbf{x})\end{aligned} \tag{46}$$

According to (44), the bias depends only on the terms higher than the linear term. Therefore, the bias of the local linear model can only be handled if $m(\mathbf{X})$ is modeled as higher than the linear degree polynomial. That means, some degrees of smoothness of m are not used by $\hat{m}$. Denote $t_i = m(\mathbf{X}_i) - \mathbf{X}_d\mathbf{b}$ = higher terms of $(\mathbf{X}_i - \mathbf{x})$ and $t = [t_1\ t_2\ \ldots\ t_n]^T$, the estimation of Bias $(\hat{y}(\mathbf{x}))$ is

$$\hat{Bias}(\hat{y}(\mathbf{x})) = \mathbf{p}_\mathbf{x}{}^T\,t \tag{47}$$

For example, if $m(\mathbf{X})$ is a locally quadratic model, then

$$m(\mathbf{X}) = m(\mathbf{x}) + \boldsymbol{g}^T(\mathbf{X} - \mathbf{x}) + \frac{1}{2}(\mathbf{X}-\mathbf{x})^T\mathbf{H}(\mathbf{X}-\mathbf{x}), \tag{48}$$

where **H** is the Hessian matrix of second derivatives of $m(\mathbf{X})$ at **x**. The estimated bias is

$$\hat{Bias}\,(\hat{y}(\mathbf{x})) = \frac{1}{2}\sum_{i=1}^{n} p_i(\mathbf{x})(\mathbf{X}_i - \mathbf{x})^T \mathbf{H}(\mathbf{X}_i - \mathbf{x}).$$

$$= \frac{1}{2}\mathbf{p_x}^T (\mathbf{X}-\mathbf{x})^T \mathbf{H}(\mathbf{X}-\mathbf{x}), \qquad (49)$$

H can be estimated if $m(\mathbf{X})$ is modeled as a locally quadratic model. But this demands a new model and additional computation. Also, due to the roughness of the $m(\mathbf{X})$ in the data which was used in the present study (refer to Figure 1 and 2), another approach, instead, is chosen to estimate bias.

2.1.3. ASYMPTOTIC PREDICTION INTERVAL

To derive confidence intervals requires the distribution of the error. Here, the error is assumed normal, $e \sim N(0,1)$. From Equation (1), $Y \sim N(m(\mathbf{x}), s(\mathbf{x}))$.

If $m(\mathbf{x})$ is linear in **x**, the local linear estimator is unbiased (Atkeson et al. (1997)). That is,

$$\text{Bias}\,(\hat{y}(\mathbf{x})) = \text{E}\,(\hat{m}(\mathbf{x})) - m(\mathbf{x})$$

$$= \text{E}\,(\hat{y}(\mathbf{x})) - y_{true}(\mathbf{x}) = 0. \qquad (50)$$

The following section will first give the derivation of the prediction interval for the unbiased case and then will discuss the biased case. Under the condition of normal errors, $\hat{y}(\mathbf{x})$ is normal, and a confidence interval at the $100(1-\alpha)\%$ confidence level for $\text{E}(Y|\,\mathbf{x})$ can be written

$$\hat{y}(\mathbf{x}) \pm t_{\alpha/2,n-2} s_{\hat{y}(\mathbf{x})} \qquad (51)$$

The expression in (51) is, indeed, that of a confidence interval and is not to be confused with the prediction level on a new response observation at $\mathbf{X} = \mathbf{x}$. The latter reflects bounds in which the analysts can realistically expect an observation of y at $\mathbf{X} = \mathbf{x}$ to fall.

The statistics $\hat{y}(\mathbf{x})$, the point on the regression line at $\mathbf{X} = \mathbf{x}$, serves the dual purpose for the estimate of mean response and the predicted value. The standard error of prediction, given by (45), is used in constructing a confidence interval on the mean

response. However, it is not appropriate for establishing any form of inference on a future single observation. Suppose the mean response at a fixed $\mathbf{X} = \mathbf{x}$ is not of interest. Rather, one is interested in some type of bound on a single response observation at $\mathbf{x}$. Consider a single observation at $\mathbf{X} = \mathbf{x}$ denoted symbolically by $y_{new}(\mathbf{x})$, independently of $\hat{y}(\mathbf{x})$. A prediction interval on y can be constructed by beginning with $y_{new}(\mathbf{x}) - \hat{y}(\mathbf{x})$.

One way to derive confidence intervals for the predictions from local linear learning is to assume a local constant variance $s^2(\mathbf{x})$ at the prediction point $\mathbf{x}$ and to use the result of Var ($\hat{y}(\mathbf{x})$) (Equation 43).

Considering

$$\begin{aligned} \text{Var}\,(y_{new}(\mathbf{x}) - \hat{y}(\mathbf{x})) &= s^2(\mathbf{x}) + s^2(\mathbf{x})\,\mathbf{p}_\mathbf{x}{}^T\mathbf{p}_\mathbf{x} \\ &= s^2(\mathbf{x})\left(1+\mathbf{p}_\mathbf{x}^T\mathbf{p}_\mathbf{x}\right) \end{aligned} \tag{52}$$

This reflects both the additive noise in sampling at the new point ($s^2(\mathbf{x})$) and the prediction error of the estimator ($s^2(\mathbf{x})\,\mathbf{p}_\mathbf{x}{}^T\mathbf{p}_\mathbf{x}$).

Under the assumption (50),

$$\begin{aligned} \text{E}\,[y_{new}(\mathbf{x}) - \hat{y}(\mathbf{x})] &= \text{E}\,(Y|\mathbf{X} = \mathbf{x}) - \text{E}\,(\hat{y}(\mathbf{x})) \\ &= m(\mathbf{x}) - \text{E}(\hat{m}(\mathbf{x})) \\ &= -\text{Bias}\,(\hat{y}(\mathbf{x})) = 0, \end{aligned}$$

then,

$$\frac{y_{new}(\mathbf{x}) - \hat{y}(\mathbf{x})}{s(\mathbf{x})\sqrt{1+\mathbf{p}_\mathbf{x}^T\mathbf{p}_\mathbf{x}}} \sim N(0,1) \tag{53}$$

under the normal theory assumptions. Since in many nonparametric regression situations, there may be only a few local data points in the neighborhood of $\mathbf{x}$ and the asymptotic normality is not close enough. Replace s by s (see Graybill (1976)) and

$$\frac{y_{new}(\mathbf{x}) - \hat{y}(\mathbf{x})}{s(\mathbf{x})\sqrt{1+\mathbf{p}_\mathbf{x}^T\mathbf{p}_\mathbf{x}}} \sim t_{n-2} \tag{54}$$

From (54) a probability bound or prediction interval can be placed on y_{new}, i.e., an interval in which y_{new} is contained with a fixed probability $(1-\alpha)$.

This prediction interval is given by

$$\hat{y}(\mathbf{x}) \pm t_{\alpha/2,n-2}\,s(\mathbf{x})\sqrt{1+\mathbf{p}_\mathbf{x}^T\mathbf{p}_\mathbf{x}} \tag{55}$$

This expression of the prediction intervals is independent of the output values of the training data Y_i, and reflects how well the data is distributed in the input space (see Equation (27)).

When the bias is not zero, however, the variance only reflects the difference between the prediction and the mean prediction, and not the difference between the prediction and the true value, which requires knowledge of the predictor's bias. Only when the local model structure is correct will the bias be zero.

Under certain regularity conditions, extending the univariate case by Fan et al. (1996), it can be shown that asymptotically

$$\frac{\hat{y}(\mathbf{x}) - \hat{Bias}(\hat{y}(\mathbf{x}))}{s_{\hat{y}(\mathbf{x})}} \rightarrow N(0,1) \tag{56}$$

Therefore, the prediction interval can be estimated as

$$\hat{y}(\mathbf{x}) - \hat{Bias}(\hat{y}(\mathbf{x})) \pm t_{a/2,n-2} s(\mathbf{x}) \sqrt{1 + \mathbf{p}_{\mathbf{x}}^T \mathbf{p}_{\mathbf{x}}} \tag{57}$$

Since the t-distribution may not be valid for our data, the bootstrap method is proposed.

2.2. Bootstrap Prediction Interval

The bootstrap is a method for estimating the distribution of an estimator or a test statistic by resampling one's data or a model estimated from the data. The bootstrap principle is that the distribution of (resampled – sample), which can be computed directly from data, approximates the distribution of (sample – true). Often, the bootstrap provides approximations that are more accurate than those of the first-order asymptotic theory.

The bootstrap method provides a direct computational way of assessing uncertainty in settings where no formulas are available, by sampling from the training data. Bootstrap is a popular method despite its disadvantage of being time consuming. In some sense, bootstrap is a versatile approach. In terms of obtaining prediction intervals, it could be applied to many prediction models and needs few assumptions. Bootstrap can provide a reliable solution and it is easy to implement when the asymptotic equations are not available or not valid. This may occur due to a small sample size, or because of the limitations set by the problem characteristics such as smoothness of the mean function.

From an original sample

$$\Psi_n = (Y_1, Y_2, \ldots, Y_n) \overset{iid}{\sim} F$$

draw a new sample of n observations among the original sample with replacements, each observation having the same probability of being drawn (= $1/n$). A bootstrap sample is often denoted

$$\Psi_n^* = (Y_1^*, Y_2^*, \ldots, Y_n^*) \overset{iid}{\sim} F_n \text{ the empirical distribution}$$

The behavior of a random variable $\hat{q} = q(\Psi_n, F)$ can be studied by considering the sequence of B new values obtained through computation of B new bootstrap samples. An approximation of the distribution of the estimate $\hat{q} = q(\Psi_n, F)$ is provided by the distribution of

$$\hat{q}^{*b} = q(\Psi_n^{*b}, F_n),\ b = 1, \ldots, B$$

2.2.1. BOOTSTRAP BIAS

In general, let q be a parameter and $\hat{q}$ an estimate. Let $\hat{q}^*$ be the bootstrap estimate calculated in the same way as $\hat{q}$. Then the bootstrap assessment of the bias is

$$\text{Bias} = \text{Mean of } (\hat{q}^*) - \hat{q} \tag{58}$$

The bias-corrected estimate of q is then

$$\bar{q} = \hat{q} - \text{Bias} = 2\hat{q} - \text{Mean of } (\hat{q}^*) \tag{59}$$

2.2.2. BOOTSTRAP PREDICTION INTERVAL

A residual-based bootstrap with bias correction is used to compute the prediction interval based on the percentile method (Efron et al., 1993).

Denote the bootstrap distribution of $\hat{q}^*$ by $G_n^*(t) = P_{F_n}(\hat{q}^* \le t)$, approximated by

$$\hat{G}_n^*(t) = \#\{\hat{q}^* \le t\} / B$$

The percentile method consists in taking the 1-2a confidence interval for q as being

$$[\hat{G}_n^{*-1}(a), \hat{G}_n^{*-1}(1-a)] \quad (60)$$

Theoretically this is equivalent to the replacement of the unknown distribution $G(t,F) = P_{F_n}(\hat{q}^* \le t)$ by the estimate $G(t, F_n)$.

The bootstrap interval prediction procedure can be divided into three steps:

1. Given training data $\{(\mathbf{X}_i^T, Y_i): i = 1, \ldots, n\}$ of size n (n = 14 for our case study), fit the local linear model $m(\mathbf{X})$ and calculate the corresponding residuals $\hat{e}_i = Y_i - \hat{y}_i = Y_i - \hat{m}(\mathbf{X}_i)$, $i = 1, \ldots, n$. Since Ee =0 and Var e = 1 are assumed by our model (Equation 2 and 3), $\hat{e}_i$ needs to be divided by the square root of Var($\hat{y}(\mathbf{x})$) (Equation 45) before standardization to avoid a system error in the bootstrap. The standardization includes centering by subtracting the average (Freedman, 1981)

$$\tilde{e}_i = \hat{e}_i - \frac{1}{n}\sum_{k=1}^{k} \hat{e}_k \text{ , } k = 1, \ldots, n. \quad (61)$$

2. Then, draw B bootstrap errors $\{e_i^*(b), i = 1, \ldots, n; b = 1, \ldots, B\}$ each of size n with replacement from the sample distribution given by the centered residuals. Finally B bootstrap outputs are formed as $Y_i^*(b) = \hat{y}_i + e_i^*(b)$ to get B bootstrap training datasets $(X_i^T, Y_i^*(b))$, $i = 1, \ldots, n$; $b = 1, \ldots, B$. To each bootstrap dataset a local linear model is fitted as $\hat{m}^{(b)}(\mathbf{X})$ and the prediction $\hat{m}^{(b)}(\boldsymbol{x})$ for the testing data query point $\boldsymbol{x}$ is computed. To get the final prediction $\hat{y}^{(b)}(\boldsymbol{x})$, bias is estimated by using average of $\hat{m}^{(b)}(\boldsymbol{x})$. A bias corrected prediction

$$\hat{y}^{(b)}(\boldsymbol{x}) = 2\,\hat{m}^{(b)}(\boldsymbol{x}) - \frac{1}{B}\sum_{b=1}^{B} \hat{m}^{(b)}(x). \quad (62)$$

3. The prediction interval for $\hat{y}(\boldsymbol{x})$ with the confidence level of 100(1- a) percent is obtained as [$\hat{y}(\boldsymbol{x})^{*(y)}$, $\hat{y}(\boldsymbol{x})^{*(1-y)}$], where $\hat{y}(\boldsymbol{x})^{*(y)}$ is the 100y-th percentile of the bootstrap distribution $\{\hat{y}^{(b)}(\boldsymbol{x})\}(b = 1, \ldots, B)$ and $y = 0.5a$. For B = 400 bootstrap samples, a 95% pointwise confidence bound can be formed from the percentiles of prediction values for each $\boldsymbol{x}$ by finding the 2.5% $\times$400 = tenth smallest and largest values at each $\boldsymbol{x}$: $\hat{y}(\boldsymbol{x})^{*\_\text{low}}$ and $\hat{y}(\boldsymbol{x})^{*\_\text{up}}$. The prediction interval will be [$\hat{y}(\boldsymbol{x})^{*\_\text{low}}$ and $\hat{y}(\boldsymbol{x})^{*\_\text{up}}$].

3 NUMERICAL STUDY

The detailed process for obtaining data, preparing the data and providing the preliminary data analysis and experimental design is basically the same as in the study by Sun et al., (2003) for the point prediction performance of the local linear predictor. The differences between this numerical study from that of the point prediction study should be pointed out.

First is the performance index or evaluation criterion. Since the main concern of the interval prediction is the predicted bounds instead of the predicted value, therefore, the relative mean error (RME) won't be used as the performance index. The expected result is that the predicted bounds should include the predicted values. That is, the predicted values should fall within the intervals formed by predicted upper bounds and lower bounds. Provided this premise is satisfied, the narrower the prediction intervals, the better performance. Obviously, under given confidence levels, the narrower prediction intervals reduce uncertainty and give better results. Also, when comparing different methods of interval prediction, results over multiple experimental runs may not require to be averaged. If consistent results for each run are observed and one-run results may be enough to infer comparison conclusions. In this paper, one-day results are given to illustrate the comparison since all other days show the same trends. It is expected that the bootstrap method is better in terms of giving narrower prediction intervals.

Secondly, results using 16-day data instead of 32-day data as the data set are given. Since it is found that the two data sets give the same comparison results for the two interval prediction approaches. But a smaller data set has less total computation time.

In the same way, this study is based on Houston's US-290 Northwest freeway eastbound traffic speed data collected from February 2002 to July 2002 every five minutes. The selected road segment for study is US-290 from the cross street Sam Houston toll way to the cross street Fairbanks. Two days form the testing set and the rest form the training set. For each day, the first two points were deleted because certain days missed those two points. The other missing data are replaced by the most recent time data for that day. Thus, each day has 286 data points.

Since the predictor uses two time instants data in covariates, the 3D figure visualizing the relationship between the future data and the current and past data is reproduced here (in Figure 1) for review. To some extent, the figure indicates that $m(\boldsymbol{x})$ is not very smooth.

A resampling size (B) of 500 is used in the bootstrap procedure. Figure 2 and 3 show one-day prediction upper-bounds and lower-bounds at a confidence level of 95% computed by asymptotic equations and the bootstrap method, respectively. Figure 3.2 is a zoom-in view of Figure 3.1 for a clearer display. Figure 4 compares the one-sided prediction intervals obtained by these two methods.

It is observed that the prediction intervals given by both methods can include the prediction values. That is, the predicted upper-bounds and lower-bounds can cover the predicted data. This is self-evident for the asymptotic method since the bounds are derived after computing one-sided intervals. So the results validate the proposed bootstrap procedure using the percentile method, where the bounds are computed directly from the data. On the other hand, the similarity of the trends of both results validates the asymptotic equations.

The asymptotic equation approach provides a one-sided prediction interval with an average of approximately ± 10Mph and a maximum of around ± 16Mph.

The bootstrap approach results in a prediction interval with an average of approximately ± 0.3Mph and a maximum of about ± 2.5Mph. It is much smaller and thus better than that of the asymptotic results. This is in accordance with the expected outcome of the experimental design.

From Figure 4, it is shown that both approaches have a larger interval when entering and leaving peak hours than at other times. But the bootstrap method gives much more stable intervals than the asymptotic method. This again proves the bootstrap is advantageous.

4 DISCUSSION AND FUTURE WORK

The work of this study has evolved from theoretical methodologies to software implementations in order to achieve more informational traffic forecasting by providing interval predictions. This paper has explored two approaches to compute prediction

intervals. First the asymptotic equations of prediction intervals were derived for the local linear predictor. Second a bootstrap method which does not rely on specific models or assumptions was proposed to obtain the prediction intervals. Then these two computational approaches were implemented in an experimental test, using a set of real-world data. Note that the bootstrap method is very adaptable to many predictors. The experimental results for the local linear predictor were given and have validated the proposed methodologies. Four figures in this paper illustrate results visually, to compare the results of the two methods. The case study results are consistent with what are expected. That is, both methods are valid and the bootstrap method gives better results.

Although the asymptotic prediction intervals, i.e., $\pm$ 10 to 16 Mph, seem somewhat large in this initial study, there may be reasons to account for this. One possibility could be the effect of ridge parameters, which also induce bias. This was not considered in the formula derivation. Also, the small sample size and the roughness of the real-world data (shown in Figure 1) may violate the *t*-distribution assumption and the second-order smoothness assumption in the local linear model. Another aspect is the bandwidth parameter, which is very important in both point prediction and interval prediction. More choices of bandwidth should be tested in the numerical study to obtain optimal results. Such factors could affect the outcome of an asymptotic prediction. Remarkably, the bootstrap method does not rely on those assumptions and therefore it has yielded very promising results which are much better than the asymptotic method.

Since the bootstrap method is both accurate and easy to implement, it is a promising method and strongly recommended for traffic forecasting practice. A cautious view to such an excellent result suggests that further investigation would be prudent before this approach can be extended to predictors other than the local linear predictor. This is because the refined bootstrap method, as used in this paper, requires the computation of predictor variances, which may not be available for other predictors. Future work also includes more study on the asymptotic results of the local linear predictors due to the directness of its computational method as well as its theoretical potential. It does not need the resampling required in the bootstrap method. Additionally, the model selection/preemption achieved by interval prediction should also be included in future research.

REFERENCES

Atkeson, C., Moore, A. and Schaal, S. (1997). "Locally Weighted Learning". *Artificial Intelligence Review*, 11, 1-5, pp.11-73.

Cho., H and Rilett, L. (2003). "Forecasting Train Travel Time". TRB 2003 Annual Meeting CD-ROM.

Enfron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Fan, J. and I. Gijbels. *Local Polynomial Modelling and its Applications*. London: Chapman & Hall, 1996.

Freedman, D.A. (1981). "Bootstrapping regression models". Annals of Statistics, 9, 1218-1228.

Freedman, D., Peters, S. (1984). "Bootstrapping a Regression Equation: Some Empirical Results (in Theory and Methods) ". *Journal of the American Statistical Association*, Vol. 79, No. 385. (Mar., 1984), pp. 97-106.

Härdle, W., Horowitz, J. and Kreiss, J. (2001). "Bootstrap Methods for Time Series". SFB 373 (HU Berlin) Discussion Paper 59 (2001)

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. Springer.

Neumann, M.H., and J.-P. Kreiss (1998). "Regression-type Inference in Nonparametric Autoregression". Annals of Statistics, 26, 1570-1613.

Peters, S. and Freedman, D. (1987). "Better Bootstrap Confidence Intervals: Comment (in Theory and Methods)". *Journal of the American Statistical Association*, Vol. 82, No. 397. (Mar., 1987), pp. 186-187.

Rilett, L. & Eisele, B. (2001). "Investigation of Travel Time Reliability Estimates for ITS Data". Presentation in May 2001, Texas A&M University and Texas Transportation Institute.

Schaal, S., & Atkeson, C. G. (1994). "Assessing the quality of learned local models". In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems* 6 (pp. 160-167). San Mateo, CA: Morgan Kaufmann.

Sun, H., Liu, H., Xiao, H., He, R. and Ran, B. (2003). "Short-Term Traffic Forecasting Using the Local Linear Regression Model". To appear in *Transportation Research Record, Journal of the Transportation Research Board*.

LIST OF FIGURES

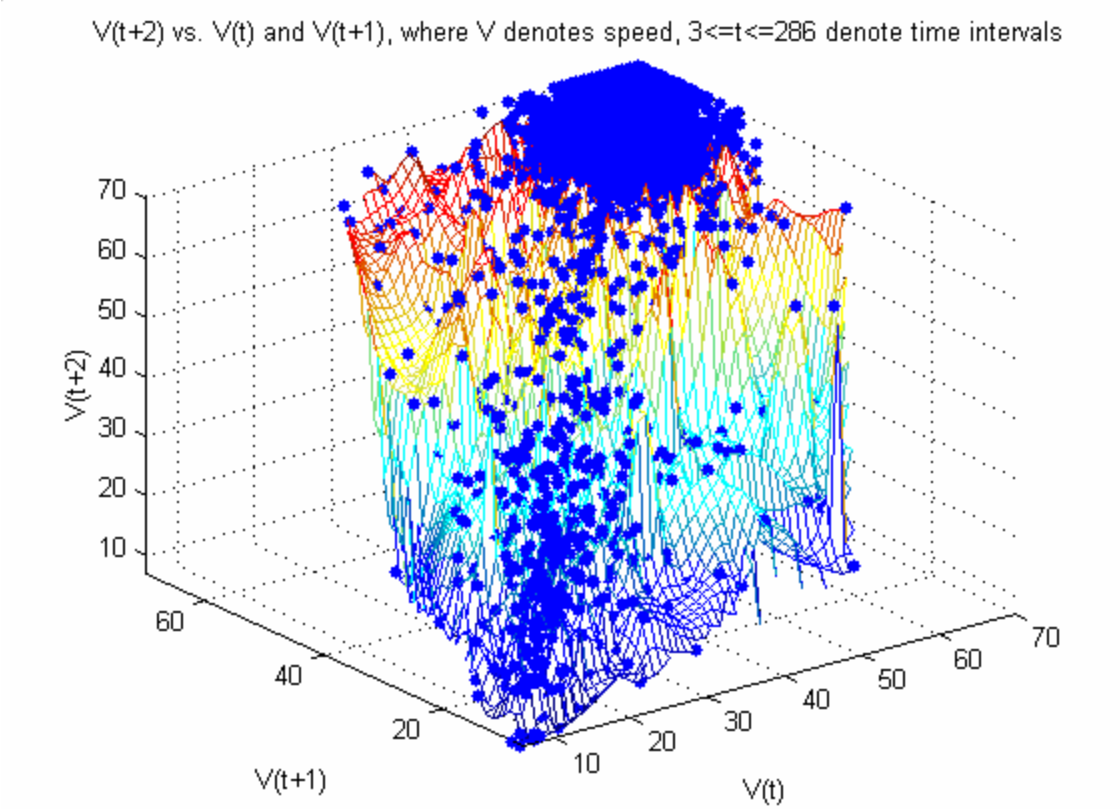

Figure 1. The Relationship of One-step Future Traffic Speed Data versus Current and Past Traffic Speed Data for 32 days

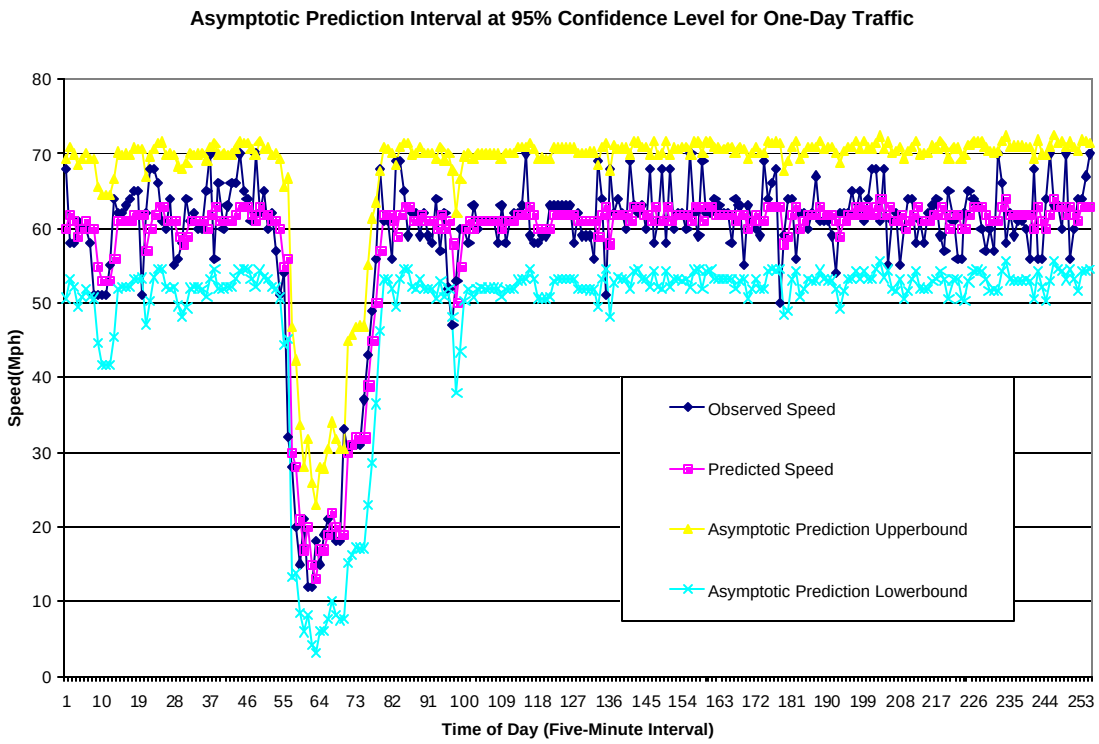

Figure 2. 95% Prediction Upper and Lower Bounds for One-Day Traffic Time Series, Computed by Asymptotic Equations of the Local Linear Predictor

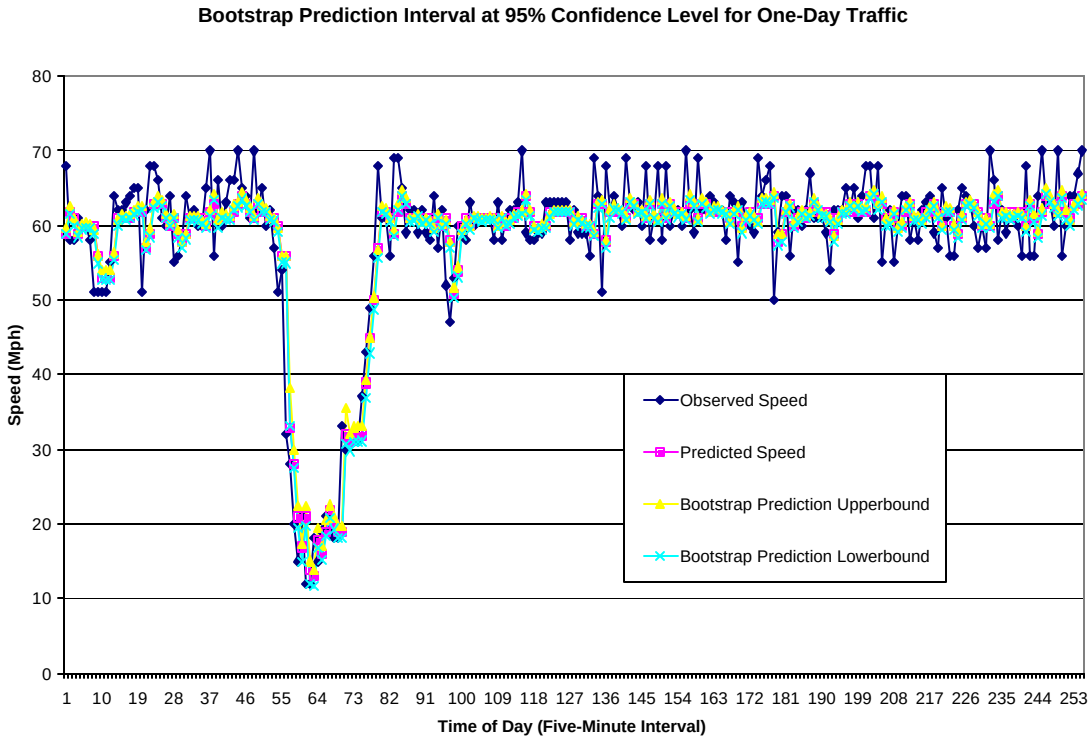

Figure 3.1. 95% Prediction Upper and Lower Bounds for One-Day Traffic Time Series, Computed by the Bootstrap Procedure for the Local Linear Predictor

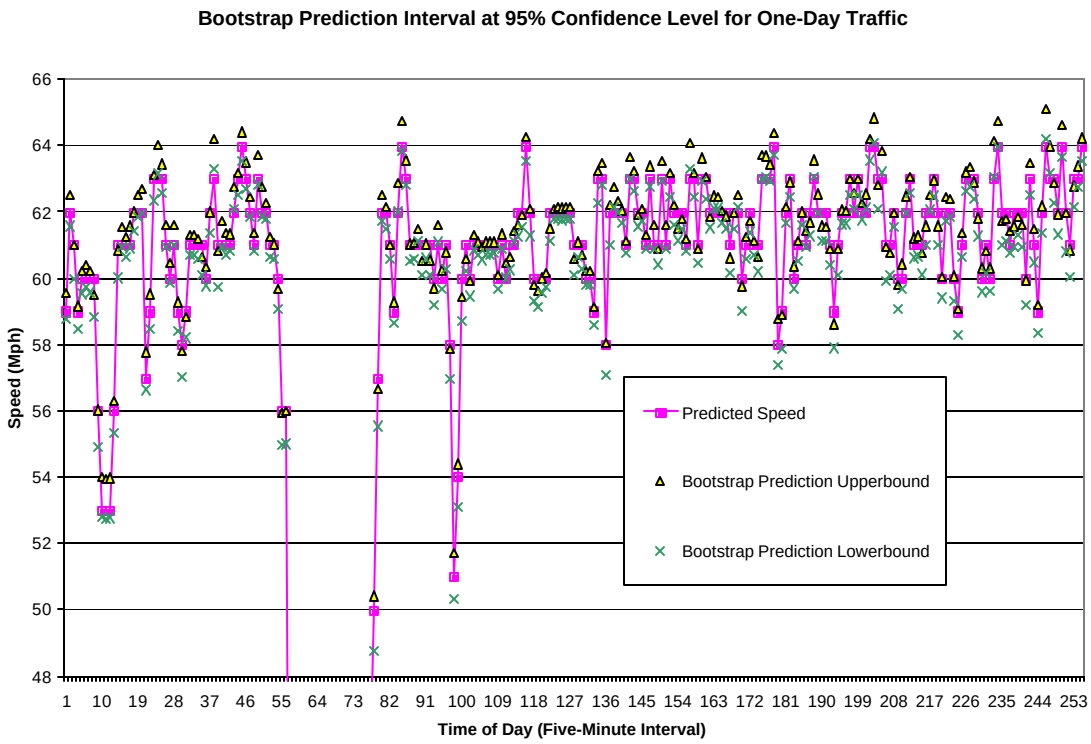

Figure 3.2. Zoom-In Results for 95% Prediction Upper and Lower Bounds for One-Day Traffic Time Series, Computed by the Bootstrap Procedure for the Local Linear Predictor

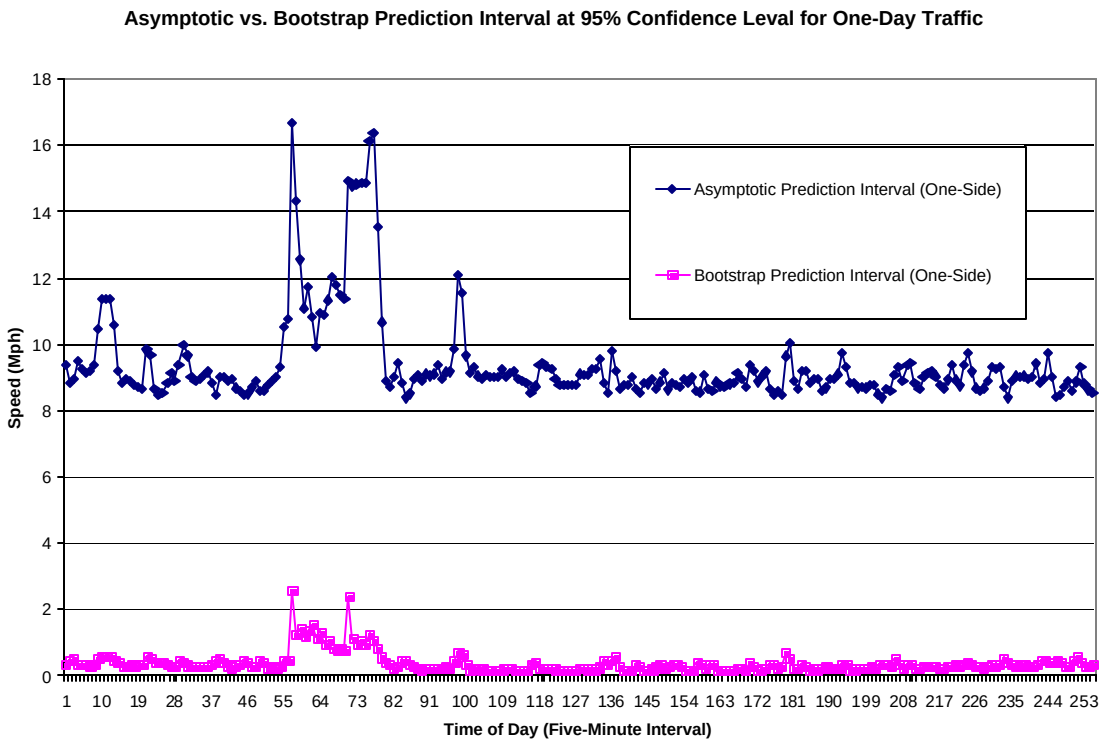

Figure 4. Comparison of One-Sided Prediction Intervals at 95% Confidence Level for One-Day Traffic Time Series, Computed by Asymptotic Equations and the Bootstrap Procedure for the Local Linear Predictor, Respectively.