PERGAMON

Transportation Research Part C 10 (2002) 303–321

TRANSPORTATION
RESEARCH
PART C

www.elsevier.com/locate/trc

Comparison of parametric and nonparametric models for traffic flow forecasting

Brian L. Smith [a,*], Billy M. Williams [b], R. Keith Oswald [c]

[a] *Department of Civil Engineering, Thornton Hall, University of Virginia, Charlottesville, VA 22903-2442, USA*
[b] *School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0355, USA*
[c] *Department of Systems Engineering, Olsson Hall, University of Virginia, Charlottesville, VA 22903-2442, USA*

Received 12 July 1999; accepted 5 July 2001

Abstract

Single point short-term traffic flow forecasting will play a key role in supporting demand forecasts needed by operational network models. Seasonal autoregressive integrated moving average (ARIMA), a classic parametric modeling approach to time series, and nonparametric regression models have been proposed as well suited for application to single point short-term traffic flow forecasting. Past research has shown seasonal ARIMA models to deliver results that are statistically superior to basic implementations of nonparametric regression. However, the advantages associated with a data-driven nonparametric forecasting approach motivate further investigation of refined nonparametric forecasting methods. Following this motivation, this research effort seeks to examine the theoretical foundation of nonparametric regression and to answer the question of whether nonparametric regression based on heuristically improved forecast generation methods approach the single interval traffic flow prediction performance of seasonal ARIMA models.

Keywords: Traffic forecasting; Nonparametric regression; ARIMA models; Motorway flows; Short-term traffic prediction; Statistical forecasting

1. Introduction

As intelligent transportation systems (ITS) are implemented widely throughout the world, managers of transportation systems have access to large amounts of “real-time” status data. While this data is key, it has been recognized by researchers and practitioners alike that the full

[*] Corresponding author. Tel.: +1-434-243-8585; fax: +1-434-982-2951.
E-mail address: briansmith@virginia.edu (B.L. Smith).

0968-090X/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0968-090X(02)00009-8

benefits of ITS cannot be realized without an ability to anticipate traffic conditions in the short-term (less than one hour into the future). The development of traffic condition forecasting models will provide this ability, and support proactive transportation management and comprehensive traveller information services. What distinguishes traffic condition forecasting from the more traditional forecasts of transportation planning is the length of the prediction horizon. While planning models use socioeconomic data and trends to forecast over a period of years, traffic condition forecasting models predict conditions within the hour using data from roadway sensors.

Without a predictive capability, ITS will provide services in a reactive manner. In other words, there will be a lag between the collection of data and the implementation of traffic control strategies. Therefore, the system will be controlled based on old information. In order to control the system in a proactive manner, ITS must have a predictive capability. "The ability to make and continuously update predictions of traffic flows and link times for several minutes into the future using real-time data is a major requirement for providing dynamic traffic control" (Cheslow et al., 1992). Furthermore, traffic condition forecasting is important in traveller information systems. This is illustrated in the statement, "the rationale behind using predictive information (for route guidance) is that traveller's decisions are affected by future traffic conditions expected to be in effect when they reach downstream sections of the network" (Kaysi et al., 1993). In fact, current traveller information activities are being hampered by the lack of a capability to predict future traffic conditions.

1.1. Transportation condition forecasting

Predicting future traffic conditions is a general concept. In fact, there are a wide variety of needs for condition forecasts depending on particular applications. For example, forecasts of traffic flow, speed, travel time, and queue length are required for specific transportation management applications. Furthermore, it is certainly true that future traffic conditions will be dependent on transportation management strategies employed, and that traffic flow can be tracked, given a sufficient number and placement of sensors, as it progresses through a network. For these reasons, an ideal approach to traffic condition forecasting is a network-based model that simulates the system as it responds to transportation management strategies. In other words, a model that takes advantage of spatial information and traffic flow dynamics.

However, every network must have boundary points, locations that serve as entry "nodes" to the network that do not have the advantage of upstream detectors to use in predicting their state. In fact, any dynamic traffic network models, such as commonly used simulation models (for example, CORSIM), are "driven" by the traffic flow rate (usually in vehicles per hour) at these boundary points. In essence, the boundary points define the short-term demand for travel through the network. For this reason, an important subset of traffic condition forecasting is the ability to predict future traffic flow at a location (a boundary point), given only data describing past conditions at this point. The research described in this paper addresses this specific challenge of *single point* traffic flow prediction.

Based on this definition of the problem, one will note that single point traffic flow prediction can be described as a time series problem. Given a series of past flow readings measured in regular intervals, a model is desired to predict the flow at the next interval. Thus, we consider the demand for travel (measured by flow) at a point to be a dynamic system, which by definition consists of two parts: a state, or the essential information describing the system, and a dynamic, which is the rule that governs how the state changes over time (Mulhern and Caprara, 1994). A dynamic system evolves towards an attractor, which is the set of values to which the system eventually settles (Thearling, 1995). Dynamic systems are often complex due to chaotic attractors, which have a bounded path that is fully determined, yet never recurs (Mulhern and Caprara, 1994). The presence of chaotic attractors tends to cloud the distinction between deterministic and stochastic behavior, and makes the selection of an ideal modeling approach difficult. A considerable amount of research has addressed dynamic system modeling in areas such as fluid turbulence (Takens, 1981), disease cases (Sugihara and May, 1990), and marketing (Mulhern and Caprara, 1994). In addition, this area has seen a large level of interest in the transportation research community of late.

When evaluating dynamic systems models for traffic flow forecasting, it is certainly important to consider the accuracy of the model. However, it is equally important to consider the environment in which the model will be developed, used, and maintained. Intelligent transportation systems deployed in metropolitan regions include thousands of traffic detectors. For example, in the Hampton Roads region of Virginia, a moderate-sized metropolitan area, the state department of transportation is in the process of installing over 1200 sensor stations on the freeways alone. In addition, the department, like most others, faces severe personnel shortages. As a result, models that require significant expertise and time to calibrate and maintain, on a site-by-site basis, are simply infeasible for use in this environment. Therefore, one must explicitly consider the requirements for dynamic system model calibration and maintenance.

1.2. Modeling approaches

Predicting traffic flow at a point decomposes into a time-series modeling problem in which one attempts to predict the value of a variable based on a series of past samples of the variable at regular intervals. Time series models have been studied in many disciplines, including transportation. In particular, transportation researchers have developed time series traffic prediction models using techniques such as autoregressive integrated moving average (ARIMA), nonparametric regression, neural networks, and heuristics. A complete review of transportation applications in this area can be found in Smith (1995) and Williams (1999).

Recent research by Williams (1999) has applied traditional parametric Box-Jenkins time series models to the dynamic system of single point traffic flow forecasting. This work has addressed most of the parametric model concerns for traffic condition data by establishing a theoretical foundation for using seasonal ARIMA forecast models (see Section 3.3). ARIMA forecasting is founded in stochastic system theory. ARIMA processes are nondeterministic with linear state transitions and can be periodic with the periodicity accounted for through seasonal differencing.

There are, however, some significant concerns with the ability to fit and maintain seasonal ARIMA models on a "production" basis in ITS systems, given the level of expertise and amount of time required to do so. For example, the researchers recently developed a number of seasonal ARIMA models for the Hampton Roads system described above. It was found that the necessary outlier detection and parameter estimation for the seasonal ARIMA models was quite time-consuming. Using only three months of ten-minute time series data, it took more than six days to detect and model the outliers and estimate the parameters for a seasonal ARIMA model at a single sensor location. Estimating parameters sequentially for each location in a moderately sized system with 200 detector locations would take more than 3 years. Clearly, this is an extreme example resulting from a non-time critical research environment. In fact, the researchers are currently exploring methods to automate and expedite the seasonal ARIMA "fitting" process. However, it does illustrate a significant practical concern regarding the wide-scale use of seasonal ARIMA models in traffic condition forecasting.

The alternative modeling approach that we will focus upon in this research is nonparametric regression. Nonparametric regression was selected as a high potential alternative to seasonal ARIMA modeling due to the fact that the authors have demonstrated the advantages of nonparametric regression over other approaches, such as neural networks, in previous research efforts (Smith, 1995). Nonparametric regression forecasting is founded on chaotic system theory. By definition, chaotic systems are defined by state transitions that are deterministic and non-linear. Furthermore, the state transitions of a chaotic system are ergodic, not periodic.

Given that ARIMA models are founded on stochastic system theory, it may seem inappropriate on the surface to simultaneously consider both nonparametric regression and ARIMA as candidates for modeling and forecasting of the same date type. However, although an argument that traffic condition data is characteristically stochastic may be more easily asserted, the presence of "chaos like" behavior cannot be completely dismissed, especially during congestion when traffic flow is unstable and a stronger causative link may be operating in the time dimension. For example, Disbro and Frame presented evidence that traffic flow exhibits chaotic behavior in a 1989 paper (Disbro and Frame, 1989). Furthermore, the characteristic weekly pattern of traffic condition data supports the assertion that historical neighbor states should yield effective forecasts in the same way that past chaotic orbits passing through nearby points provide "good" short-term forecasts. Coupling this expectation of reasonably accurate forecasts with the attractive implementation characteristics of a data driven approach provides ample motivation to investigate nonparametric regression performance.

2. Nearest neighbor nonparametric regression

Nonparametric regression relies on data describing the relationship between dependent and independent variables. The basic approach of nonparametric regression is heavily influenced by its roots in pattern recognition (Karlsson and Yakowitz, 1987). In essence, the approach locates the state of the system (defined by the independent variables) in a "neighborhood" of past, similar states. Once this neighborhood has been established, the past cases in the neighborhood are used to estimate the value of the dependent variable.

Clearly, this approach assumes that the bulk of the knowledge about the relationship lies in the data rather than the person developing the model (Eubank, 1988). Of course, the quality of the database, particularly in storing past cases that represent the breadth of all possible future conditions, dictates the effectiveness of a nonparametric regression model. To put it simply, if similar cases are not present in the database, an informed forecast cannot be generated.

2.1. Implementation challenges

There are four fundamental challenges when applying nonparametric regression: definition of an appropriate state space, definition of a distance metric to determine nearness of historical observation to the current conditions, selection of a forecast generation method given a collection of nearest neighbors, and management of the potential neighbors database.

2.1.1. State space

For data that is time series in nature, a state vector defines each record with a measurement at time $t, t-1, \ldots, t-d$ where d is an appropriate number of lags. For example, a state vector $x(t)$ of flow rate measurements collected every 10 min with $d = 3$ can be written as

$$x(t) = [V(t), V(t-1), V(t-2), V(t-3)], \tag{1}$$

where $V(t)$ is the flow rate during the current time interval, $V(t-1)$ is the flow rate during the previous 10-min interval, and so on. Note that an infinite number of possible state vectors exist. Furthermore, they are not restricted to incorporating only lagged values but may also be supplemented with aggregate measures such as historical averages.

2.1.2. Distance metric

A common approach to measuring "nearness" in nonparametric regression is to use Euclidean distance in the independent variable space. Such an approach "weights" the value of each independent variable equally. In other words, it considers the information content of each to be of equal value. In many cases, knowledge of the problem domain may make such an assumption unreasonable. In that case, a weighed distance metric where the "dimension" of variables with higher information content would be weighed more heavily may be appropriate. While this makes intuitive sense, it is clear that such an approach is heuristic in nature, and requires careful consideration by the modeler.

2.1.3. Forecast generation

The most straightforward approach to generating the forecast for the dependent variable is to compute a simple average of the dependent variable values of the neighbors that have fallen within the nonparametric regression neighborhood. The weakness of such an approach is that it ignores all information provided by the distance metric. In other words, it is logical to assume that past cases "nearer" the current case have higher information content and should play a larger role in generating the forecast.

To address this concern, a number of weighting schemes have been proposed for use within nonparametric regression. In general these weights are proportional to the distance between the neighbor and the current condition. An alternative to the use of weights is to fit a linear or nonlinear model to the cases in the neighborhood, and then use the model to forecast the value of the dependent variable. For example, Mulhern and Caprara (1994) use a regression model in the selected neighborhood to generate marketing forecasts with promising results. However, we chose not to investigate this approach due to the fact that it introduces a parametric model to a nonparametric technique.

Finally, it should be clear that there are an infinite number of approaches to forecast generation. As we will discuss later in the paper, we were able to improve the effectiveness of our models by weighing our forecast directly with historical data.

2.1.4. Management of potential neighbor database

As stated earlier, the effectiveness of nonparametric regression is directly dependent on the quality of the database of potential neighbors. Clearly, it is desirable to possess a large database of cases that span the likely conditions that a system is expected to face. However, while as large a database as possible is desirable for increasing the accuracy of the model, the size of the database has significant implications on the timeliness of model execution.

When considering how nonparametric models work, one will see that the majority of effort at runtime is expended in searching the database for neighbors. As the database grows, this search process grows accordingly. For real-time applications, such as traffic flow forecasting, this is a significant issue. Steps must be taken to keep the database size manageable, while ensuring that the database has the depth and breadth necessary to support forecasting.

Again, the number of approaches to accomplish this is infinite. One approach would be to cluster the database, and only search those cases in the cluster in which the current state falls. Another approach would be to periodically delete records from the database. This process would involve searching for multiple records that are nearly identical. Such an approach would require the use of a distance metric such as those discussed earlier. While examining this issue fully is beyond the scope of this paper, it is important to realize that the management of the database is a critically important issue, particularly in real-time applications of nonparametric regression.

2.2. Nonparametric regression theoretical foundation

It is important to recognize that nonparametric regression has a strong theoretical basis. The statistical properties of the nearest neighbor approach are attractive. Using such an approach will result in an asymptotically optimal forecaster (Karlsson and Yakowitz, 1987). This means that for a state space of size m values, the nearest neighbor model will asymptotically be at least as good as any mth order parametric model. This property indicates that if one has access to a large, high-quality database, the nearest neighbor approach should yield results comparable to parametric models.

While the state definition for a nonparametric model is flexible, and can be defined in an infinite number of ways, a common approach for time-series problems is to define the state as the series of system values (in our case, traffic flow measurements) measured during the past d intervals. In dynamic systems, an attractor is defined as a value to which a system eventually settles as $t \rightarrow \infty$ (Takens, 1979). In other words, the system dynamic is driven by the attractor. In his work, Takens found that for a D dimension attractor, it is necessary to define the number of lags (d) in the state as $2D + 1$ (1979). Therefore, given that D is not known for single point traffic flow prediction, there is theoretical justification to consider multiple values of d when defining the state.

Furthermore, there is justification to investigate the inclusion of other values in the state definition beyond lagged time series values. For example, since nearest neighbor models "geometrically attempt to reconstruct whatever attractor is operating in a time series" (Mulhern and Caprara, 1994), including historical averages in the state vector further clarifies the position of each observation along the cyclical flow-time curve, which may improve forecast accuracy by finding neighbors that are more similar to the current conditions.

Thus, based on dynamic systems literature, there is theoretical justification for considering multiple system state definitions when applying nonparametric regression. This highlights a promising opportunity to improve the accuracy of nonparametric regression forecasting.

3. Experimental approach

3.1. Research question

Earlier work by Smith (1995) found that a simple implementation of the nearest neighbor forecasting approach provides reasonably accurate traffic condition forecasts. Subsequent research by Williams (1999) demonstrated that seasonal ARIMA models outperformed the straight average *k*-nearest neighbor method. However, the advantages associated with a data-driven nonparametric forecasting approach motivate further investigation of refined nonparametric regression forecasting methods. Following this motivation, the research effort presented in this article sought to answer the question of whether nonparametric regression based on heuristically improved forecast generation methods would approach the single interval traffic flow prediction performance of seasonal ARIMA models.

3.2. The data

The data were collected by the United Kingdom Highways Agency using the motorway incident detection and automatic signaling (MIDAS) system. The pilot project for the MIDAS system included extensive instrumentation of the southwest quadrant of the London Orbital Motorway, M25. The instrumentation included four travel lane specific loop detectors at 500-m intervals. Traffic condition data from these detectors is collected on a one-minute interval, 24 h a day. The data were collected from 4 September to 30 November, 1996. The data are 15-min traffic flow rates from two loop detector locations, detector number 4757A and 4762A, which lie between the M3 and M23 interchanges. The 15-min data were formed by averaging the raw 1-min data over 15-min intervals and aggregating across all travel lanes. The data are for traffic flow in the clockwise direction on the outer loop of the M25. Fig. 1 shows the general location of the data sites.

For this research the data were split into two independent samples. The data collected between 4 September and 18 October provided the basis for model development. The models were then evaluated using the data collected between 19 October and 30 November.

Table 1 presents descriptive statistics for the data sets. The data contain very few missing observations with only approximately 1% of the data missing for each development data set and no missing observations for the evaluation data. The PM peak flow rates fall in the range of 7000–8000 vehicles per hour (vph).

3.3. Benchmark models

As stated in Section 3.1, this research focuses on improving the performance of nonparametric regression. To assess the quality of the forecasts, two benchmark models were used to "frame" the

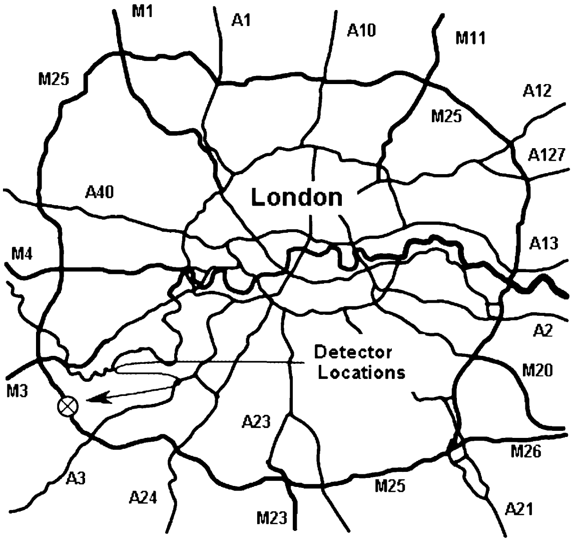


Fig. 1. M25 detector locations.

Table 1
Descriptive statistics—M25 data

Data series	Series length (number of observations)	Missing values	Percent missing (%)	Series mean (veh/h)	Mean absolute one-step change
Detector 4757A Sep/Oct 1996	4320	44	1.02	3106	309
Detector 4757A Oct/Nov 1996	4128	0	0.0	2916	282
Detector 4762A Sep/Oct 1996	4320	51	1.18	3100	315
Detector 4762A Oct/Nov 1996	4128	0	0.0	2915	287

results. First, a simple naïve model was used to serve as a worst-case approach. Any complex model must certainly outperform a naïve model. Secondly, a seasonal ARIMA model was used in the context of a best case. In other words, the goal of modifying the nonparametric regression models was to approach the performance of parametric seasonal ARIMA models. The benchmark models utilized in the work are described more fully below.

3.3.1. Naïve forecast

Naïve forecasts were calculated in a way that takes into account both the current conditions and the historical pattern. Average flow rates were calculated for each time-of-day and day-of-the-week using the development data for each site. For the test data, naïve forecasts were calculated using these historical average flow rates by the equation

$$\hat{V}(t+1) = \frac{V(t)}{V_{\text{hist}}(t)} V_{\text{hist}}(t+1), \tag{2}$$

where $\hat{V}(t+1)$ is the forecast for the next 15 min time interval, $V(t)$ is the traffic flow rate at the current time interval, and $V_{\text{hist}}(t)$ is the historical average of the traffic flow rate at the time-of-day and day-of-the-week associated with time interval t.

This forecast approach follows the intuition that the ratio of current conditions to historic conditions will be a good indicator of how the traffic conditions in the next interval will deviate from the historic conditions. A proposed forecast method that could not be shown to consistently produce more accurate forecasts than this naïve method would be of little value.

3.3.2. Seasonal ARIMA parametric model

Recent work by Williams (1999) put forth a strong theoretical foundation for modeling traffic condition data as seasonal ARIMA processes and demonstrated the forecast accuracy of fitted seasonal ARIMA models. The theoretical foundation rests on two fundamental time series concepts, namely the definition of time series stationarity and the theorem known as the *Wold Decomposition*.

3.3.2.1. Seasonal ARIMA definition. The formal definition of a seasonal ARIMA process follows:

Seasonal ARIMA process

A time series $\{X_t\}$ is a seasonal ARIMA $(p,d,q)(P,D,Q)S$ process with period S if d and D are nonnegative integers and if the differenced series $Y_t = -(1-B)^d(1-B^s)^D X_t$ is a stationary autoregressive moving average process defined by the expression

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)e_t, \tag{3}$$

where B is the backshift operator defined by $B^a W_t = W_{t-a}$,

$$\phi(z) = 1 - \phi_1 - \cdots - \phi_p z^p, \Phi(z) = 1 - \Phi_1 Z - \cdots - \Phi_P Z^P,$$

$$\theta(z) = 1 - \theta_1 - \cdots - \theta_q Z^q, \Theta(z) = 1 - \Theta_1 Z - \cdots - \Theta_Q Z^Q,$$

and e_t is identically and normally distributed with mean zero, variance σ^2 and $\mathrm{cov}(e_t, e_{t-k}) = 0\ \forall k \neq 0$, i.e., $\{e_t\} \sim WN(0,\sigma^2)$.

In the definition above, the parameters p and q represent the autoregressive and moving average order, respectively, and the parameters P and Q represent the autoregressive and moving average order at the model's seasonal period length, S. The parameters d and D represent the order of ordinary and seasonal differencing, respectively. The reader is referred to a comprehensive time series text, such as Brockwell and Davis (1996) or Fuller (1996) for additional background on ARIMA time series models.

3.3.2.2. Theoretical justification. Formal ARIMA model definitions apply to time series that are weakly stationary or that can be made weakly stationary through differencing. By definition, a time series $\{X_t\}$ is said to be weakly stationary if

(i) the unconditional expected value of X_t is the same for all t and
(ii) the covariance between any two observations in the series is dependent only on the lag between the observations (independent of t).

Given the characteristic cyclical pattern of traffic condition data, it is obvious that neither raw data series nor data series that have been transformed by a first ordinary difference are stationary. Traffic flow, for example, rises to and falls from characteristic peaks. Although the daily patterns for the weekdays are quite similar to each other, the weekday patterns do vary from day to day to some degree, and Saturday and Sunday patterns are quite different from the weekdays. Therefore, it is intuitive that the transformation resulting from differencing traffic condition data at a 1-week lag could be considered stationary for modeling purposes. In time series analysis terminology, such a transformation is called a first seasonal difference at a one-week seasonal period or, more concisely, a weekly seasonal difference. The notion that stationarity could be achieved by a weekly seasonal difference was first recognized in the literature by Okutani and Stephanedes (1984).

If we allow that time series of traffic condition data can be made stationary by a weekly seasonal difference then the theory known as the Wold Decomposition comes into play. The Wold Decomposition states that

If $\{X_t\}$ is a nondeterministic stationary time series, then

$$X_t = \sum_{j=0}^{\infty} \psi_j e_{t-j} + V_t, \tag{4}$$

where $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, $\{e_t\} \sim \mathrm{WN}(0, \sigma^2)$, $\{V_t\}$ and $\{e_t\}$ are uncorrelated, e_t is the limit of linear combinations of $X_s, s \leqslant t$ and V_t is deterministic (Brockwell and Davis, 1996) and (Fuller, 1996).

In practical terms, the Wold Decomposition says that any stationary time series can be decomposed into a deterministic series and a stochastic series and further that the stochastic series can be represented as an infinite moving average time series. ARIMA models are effective at modeling just such stochastic series. The vector $\Psi = \{\psi_0, \psi_1, \ldots\}$ is referred to as a causal time-invariant linear filter.

3.3.2.3. Seasonal ARIMA model fitting. Over the last two decades, application of ARIMA forecast models has been primarily guided by the methodology put forth by Box and Jenkins (1976). The procedures in this methodology are organized in the iterative steps of model identification, model estimation, and model diagnosis. The Box–Jenkins approach relies heavily on visual assessment of plots of the training data correlation structure. Although the basic elements of the Box–Jenkins technique continue to be relevant and useful, the combined effect of increased computing power, improved software, and the development and widespread use of information criteria for model selection has removed much of the subjectivity. Increased computing power and improved software have made it possible to efficiently search an extensive list of candidate models. Information criteria such as the Akaike information criterion (AIC) and the Schwarz Bayesian criterion (SBC) provide a model selection metric that directly addresses the potential for overfilling.

Applying this information criterion search procedure to the M25 model development data leads to the selection of ARIMA $(1,0,1)(0,1,1)_{672}$ models for both detector locations based on SBC. The forecasting performance of these seasonal ARIMA models is reflected in the model comparisons below.

3.4. Nonparametric model definitions

Application of nonparametric regression involves specification of the input space, distance metric, neighbor selection criteria, and the forecast generation method. Euclidean distance was used as the distance metric for each of the nearest neighbor models. The state space, forecast calculation methods, and neighbor selection criteria are described in detail below.

3.4.1. State definition

In order to examine Takens' theories regarding the number of necessary lags in a dynamic system state definition, this research investigated different state vector definitions, each using the current flow rate plus one to eight lagged observations. Using the *straight average* forecast function described in Section 3.4.2, none of the state vectors produced forecasts that were more accurate than the naïve forecast shown in Figs. 2 and 3. In the figures, notice the large decrease in forecast error for state vector definitions that include more than one lagged observation. These results indicate that state vectors with only lagged observations may not contain enough information

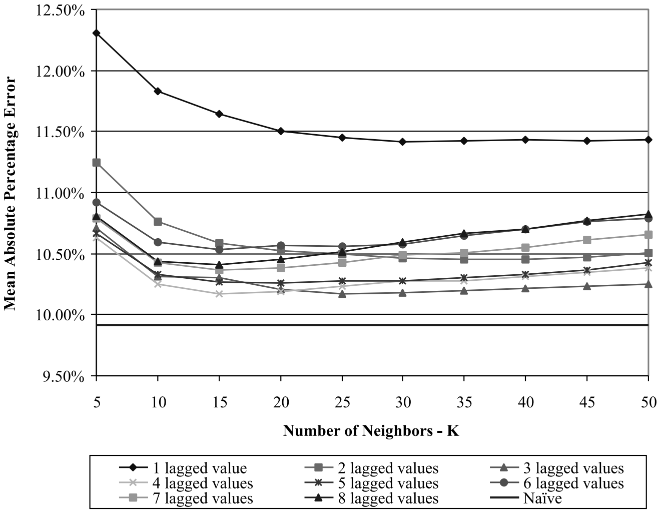


Fig. 2. Detector 4757A state length analysis.

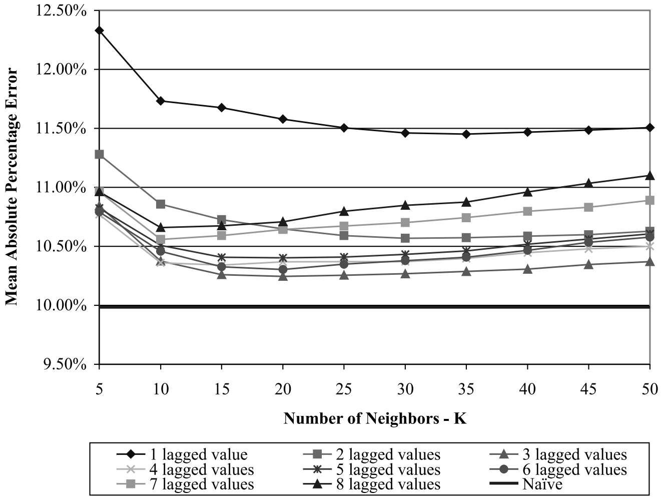


Fig. 3. Detector 4762A state length analysis.

about the state of the system to find neighbors capable of adequately estimating future demand. Therefore, the state vector requires other values to more accurately describe conditions.

Because nearest neighbor models “geometrically attempt to reconstruct whatever attractor is operating in a time series” (Mulhern and Caprara, 1994), including historical averages in the state vector further clarifies the position of each observation along the cyclical flow-time curve, which may improve forecast accuracy by finding neighbors that are more similar to the current conditions. Therefore, a hybrid state vector, $x(t)$, was defined as

$$x(t) = [V(t), V(t-1), V(t-2), V_{\text{hist}}(t), V_{\text{hist}}(t+1)], \tag{5}$$

where $V(t)$ is the traffic flow rate at time interval t and $V_{\text{hist}}(t)$ is the historical average volume at the time-of-day and day-of-the-week associated with time interval t. The associated output vector, $y(t)$, for each state observation in the potential neighbor database is, of course, the volume at time interval $t+1$ relative to the historical state. As mentioned above, the discrete time interval for the M25 data is 15 min. Therefore time interval $t+1$ is 15 min in the future relative to time interval t.

Clearly, there are an infinite number of possible definitions of $x(t)$. However, the state definition given in Eq. (5) is reasonable both in terms of sufficiency and simplicity. The state, as defined, weighs the current conditions (through the elements $V(t)$, $V(t-1)$ and $V(t-2)$ slightly more than the “normal” historical pattern at the current time-of-day and day-of-the-week, which include only two elements ($V_{\text{hist}}(t)$ and $V_{\text{hist}}(t+1)$).

3.4.2. Forecast calculation

Assuming that the neighbor search procedure identifies k neighbors, $x_i(t)$ where $i = 1, 2, \ldots, k$, for a forecast point, $x_c(t)$ the selected neighbor set and forecast point will be as follows with the elements defined as for Eq. (5):

Selected neighbor set ($x_i(t)$ where $i = 1, 2, \ldots, k$)

i	⟨.................. Input Vector $x_i(t)$ ....⟩						Output	Euclidean distance from $x_c(t)$
1	$V_1(t)$	$V_1(t-1)$	$V_1(t-2)$	$V_{\text{hist},1}(t)$	$V_{\text{hist},1}(t+1)$	→	$V_1(t+1)$	Dist_1
2	$V_2(t)$	$V_2(t-1)$	$V_2(t-2)$	$V_{\text{hist},2}(t)$	$V_{\text{hist},2}(t+1)$	→	$V_2(t+1)$	Dist_2
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮
k	$V_k(t)$	$V_k(t-1)$	$V_k(t-2)$	$V_{\text{hist},k}(t)$	$V_{\text{hist},k}(t+1)$	→	$V_k(t+1)$	Dist_k

Forecast point (the current state, $x_c(t)$)

$$V_c(t), V_c(t-1), V_c(t-2), V_{\text{hist},c}(t), V_{\text{hist},c}(t+1).$$

The input vector elements for the selected neighbors, i.e., the elements of $x_i(t)$ are the nonparametric regression independent variables while the output elements (the historical next flow rate for each neighbor) provide the basis for calculating the dependent variable (i.e., the forecast).

Given the selected neighbors and forecast point as defined above, the following forecast calculation methods were used for each nonparametric model.

Straight average

The *straight average* approach applies equal weight to each of the selected neighbor output elements and is calculated by the following equation:

$$\hat{V}(t+1) = (1/k)\sum_{i=1}^{k} V_i(t+1). \tag{6}$$

Weighted by inverse of distance

The *weighted by inverse of distance* approach assumes that selected neighbors that are relatively closer to the current state provide a better indication of the future condition. This assumption is incorporated in the calculation by weighting the output elements by the inverse of the corresponding Euclidean distance as follows:

$$\hat{V}(t+1) = \sum_{i=1}^{k} \frac{V_i(t+1)}{\mathrm{Dist}_i} \Bigg/ \sum_{i=1}^{k} \frac{1}{\mathrm{Dist}_i}. \tag{7}$$

Adjusted by $V(t)$

The *adjusted by $V(t)$* approach assumes that the output elements will provide better information if they are adjusted by the ratio of the current condition to the corresponding element in the selected neighbor input vectors prior to averaging.

$$\hat{V}(t+1) = (1/k)\sum_{i=1}^{k} V_i(t+1)(V_c(t)/V_i(t)). \tag{8}$$

Adjusted by $V_{\mathrm{hist}}(t+1)$

The *adjusted by $V_{\mathrm{hist}}(t+1)$* approach assumes that the output elements will rate provide better information if they are adjusted by the ratio of the historical flow rate at the forecast interval to the corresponding element in the selected neighbor input vectors prior to averaging.

$$\hat{V}(t+1) = (1/k)\sum_{i=1}^{k} V_i(t+1)\big(V_{\mathrm{hist},c}(t+1)/V_{\mathrm{hist},j}(t+1)\big). \tag{9}$$

Adjusted by both $V(t)$ and $V_{\mathrm{hist}}(t+1)$

The *adjusted by $V(t)$* and $V_{\mathrm{hist}}(t+1)$ approach combines the previous two adjustments by averaging the ratios. This approach assumes that a composite adjustment that considers both the current condition and the historical condition at the forecast interval will provide a better forecast than either adjustment alone.

$$\hat{V}(t+1)(1/k)\sum_{i=1}^{k} V_i(t+1)\big[\big(V_c(t)/V_i(t) + V_{\mathrm{hist},c}(t+1)/V_{\mathrm{hist},j}(t+1)\big)/2\big]. \tag{10}$$

Adjusted by both $V(t)$ and $V_{\mathrm{hist}}(t+1)$ and weighted by inverse of distance

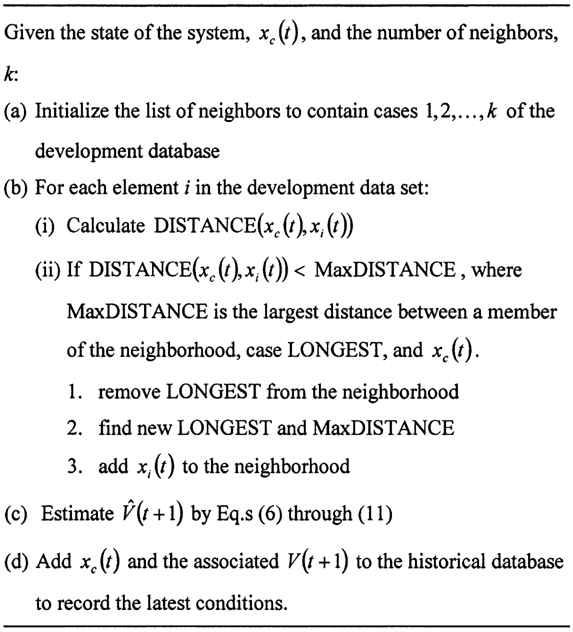
The adjusted by both $V(t)$ and $V_{\mathrm{hist}}(t+1)$ and weighted by inverse of distance approach applies the inverse of Euclidean distance weighing to the output elements after applying the composite adjustment used in the *adjusted by $V(t)$ and $V_{\mathrm{hist}}(t+1)$* approach. This approach assumes that the forecast can be further improved by applying Euclidean distance weighing to the adjusted output elements.

$$\hat{V}(t+1) = \sum_{i=1}^{k} \frac{V_i(t+1)[(V_c(t)/V_i(t) + V_{\text{hist},c}(t+1)/V_{\text{hist},i}(t+1))/2]}{\text{Dist}_i} \Bigg/ \sum_{i=1}^{k} \frac{1}{\text{Dist}_i}. \tag{11}$$

In summary, the latter five calculation methods are heuristic attempts to improve upon the straight average forecast by taking into account the relative distance of the neighbors from the forecast point and/or the relation of key neighbor state elements to the corresponding forecast point elements. These methods are founded on the notion that the forecast can be further informed by considering the overall nearness of the neighbors to the forecast state (weighing by Euclidean distance), the correlation of each neighbor to the current conditions (adjusting by $V(t)$), and the correlation of the neighbors to the historic conditions at the time interval of the forecast (adjusting by $V_{\text{hist}}(t+1)$).

3.5. Measures of predictive effectiveness

The forecasting performance of the various models was evaluated using three summary statistics: root mean square error, mean absolute deviation, and mean absolute percentage error (MAPE). Since traffic flow observations vary from a few hundred vehicles per hour in the off peak to several thousand vehicles per hour during the peak periods, absolute percentage error provides the most useful basis for comparison. For this reason and for clarity, only the MAPE results are presented in this paper.



```
Given the state of the system, $x_c(t)$, and the number of neighbors, $k$:
(a) Initialize the list of neighbors to contain cases $1,2,\ldots,k$ of the development database
(b) For each element $i$ in the development data set:
    (i) Calculate DISTANCE$(x_c(t), x_i(t))$
    (ii) If DISTANCE$(x_c(t), x_i(t)) <$ MaxDISTANCE, where MaxDISTANCE is the largest distance between a member of the neighborhood, case LONGEST, and $x_c(t)$.
        1. remove LONGEST from the neighborhood
        2. find new LONGEST and MaxDISTANCE
        3. add $x_i(t)$ to the neighborhood
(c) Estimate $\hat{V}(t+1)$ by Eq.s (6) through (11)
(d) Add $x_c(t)$ and the associated $V(t+1)$ to the historical database to record the latest conditions.
```

Fig. 4. Pseudo-code k-nearest neighborhood algorithm.

Nonparametric repeated measures tests were used to test whether the differences in MAPE for the various forecast models are statistically significant. For these nonparametric tests, the various forecast methods are ranked by absolute percentage error at each forecast point, and the tests operate on these ranks. The Friedman test was used for testing groups of three or more models. The Friedman test evaluates the null hypothesis that three or more related samples are from the same population. The Wilcoxon signed-rank test was used to test for statistically significant difference between two models. The Wilcoxon signed-rank test evaluates the null hypothesis that two related samples have the same distribution.

3.6. Forecasts

Forecasts were generated using k-nearest neighbors for the test data from each site using the development data and development data historical averages as the potential neighbor database. Forecasts were generated using k-nearest neighbor forecasts for values of k between 5 and 40 inclusive in increments of five. Pseudo-code for the k-nearest neighbor algorithm is given in Fig. 4.

4. Forecast performance test results

Model forecast performance in terms of absolute percentage error is presented below. The results indicated that a value of $k = 20$ yields the best forecasts considering both detector locations. Therefore, statistical significance tests were conducted for the k-nearest neighbor model forecasts with $k = 20$ in comparison to the naïve forecasts and the ARIMA $(1,0,1)(0,1,1)_{672}$ model forecasts.

4.1. Mean absolute percentage error

The k-nearest neighbor model performance is presented in Figs. 5 and 6 in terms of MAPE. The *adjust by* $V(f)$ method is the best performing k-nearest neighbor forecast method for both detectors. Conversely, the *adjust by* $V_{\text{hist}}(t+1)$ is the poorest performer in both cases, being outperformed by the naïve forecasts at every value of k investigated. Although outperforming the naïve forecasts, the best performing k-nearest neighbor forecast methods did not match the performance of the seasonal ARIMA forecasts. At the closest approach of the *adjust by* $V(t)$ forecasts to the seasonal ARIMA performance, the difference in MAPE was approximately 0.6% for detector 4757A and 0.7% for detector 4762A.

In addition to the *adjust by* $V(t)$ method, only the two most complex methods, *adjust by both* $V(t)$ *and* $V_{\text{hist}}(t+1)$ and *adjust by both and weight by distance*, consistently outperformed the naïve forecasts. For detector 4757A, the *straight average* forecasts have a lower MAPE than the naïve forecasts for $k = 10$ and $k = 15$, and the *weight by distance* forecasts outperform the naïve for $k = 10$ through $k = 25$. For detector 4762A, the *straight average* forecasts are outperformed by the naïve forecasts for all values of k, and the *weight by distance* forecasts match the naïve forecast MAPE only at $k = 10$.

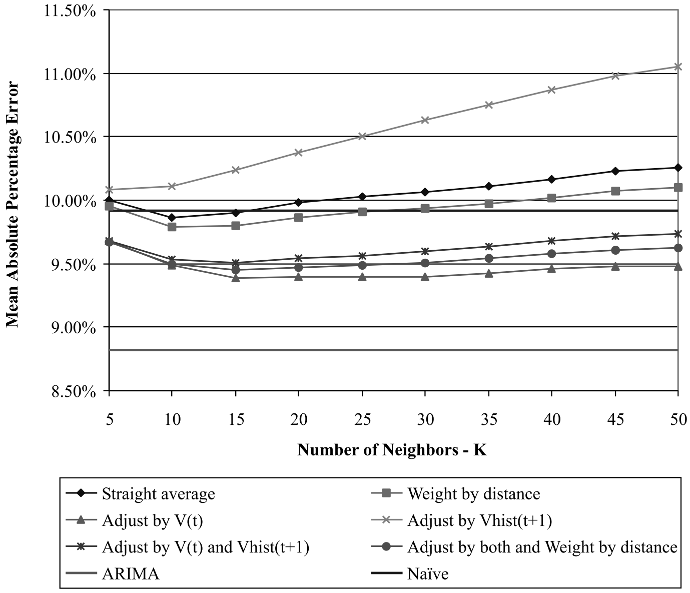


Fig. 5. Detector 4757A *k*-nearest neighbor results.

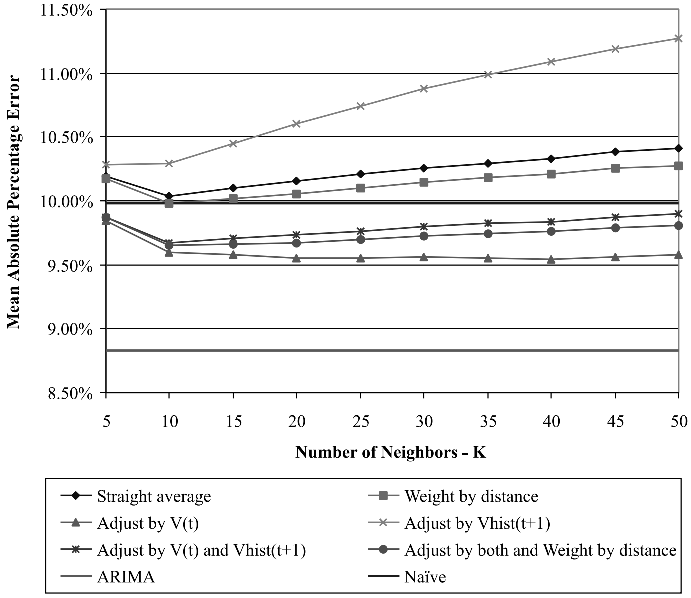


Fig. 6. Detector 4762A *k*-nearest neighbor results.

4.2. Statistical significance test results

The results summarized above indicate that *k*-nearest neighbor nonparametric regression with three of the new forecast calculation methods, *adjust by* $V(t)$, *adjust by both* $V(t)$ *and* $V_{\text{hist}}(t+1)$, and *adjust by both weight and by distance*, outperform the naïve forecasts with respect to MAPE. However, these best-performing k-nearest neighbor forecast models did not approach the predictive performance of seasonal ARIMA.

However, the foregoing analysis does not answer the question of how much statistical confidence we have in saying one method is "better" than another on the basis of MAPE. Therefore, the absolute percentage errors from each of the $k = 20$ forecasts were analyzed, along with the absolute percentage errors of the naïve and seasonal ARIMA forecasts, as related samples using the Friedman test and Wilcoxon signed-rank test. An $\alpha = 0.05$ significance level was used for hypothesis testing. The results of these related-samples analyses are presented in Tables 2 and 3.

The Friedman and Wilcoxon signed-rank tests operate on the ranks of the related measures (in this case absolute percentage error) instead of the actual measures. In other words, the tests convert the lowest absolute percentage error at each forecast point to rank 1, the second lowest to rank two, and so forth. The eight forecast methods included in Tables 2 and 3 are listed in ascending order of average rank across all forecasts. This ordering does not necessarily coincide with MAPE, which is also given in the tables. Dashed lines are shown in the table where the null

Table 2
Detector 4757A—statistical significance test results

Forecast model	Mean rank	MAPE (%)
ARIMA $(1,0,1)(0,1,1)_{672}$	4.13	8.82
Adjust by $V(t)$ and $V_{\text{hist}}(t+1)$ and weight by distance	4.26	9.47
Adjust by $V(t)$	4.37	9.39
Adjust by $V(t)$ and $V_{\text{hist}}(t+1)$	4.37	9.54
Weight by distance	4.57	9.87
Straight average	4.69	9.98
Naïve	4.71	9.91
Adjust by $V_{\text{hist}}(t+1)$	4.91	10.4

Table 3
Detector 4762A—statistical significance test results

Forecast model	Mean rank	MAPE (%)
ARIMA $(1,0,1)(0,1,1)_{672}$	4.09	8.83
Adjust by $V(t)$ and $V_{\text{hist}}(t+1)$ and weight by distance	4.30	9.67
Adjust by $V(t)$	4.37	9.55
Adjust by $V(t)$ and $V_{\text{hist}}(t+1)$	4.40	9.73
Weight by distance	4.54	10.05
Straight average	4.66	10.15
Naïve	4.68	9.98
Adjust by $V_{\text{hist}}(t+1)$	4.96	10.60

hypothesis of either the Friedman test or Wilcoxon signed-rank test, as applicable, could be rejected at the $\alpha = 0.05$ significance level. Groups of three not separated by a dashed line are groups either for which the Friedman test null hypothesis could not be rejected or for which the Wilcoxon signed-rank test did not support further ordered segregation even though the Friedman test null hypothesis could be rejected. The group of two shown in Table 2 could not be segregated based on the Wilcoxon signed-rank test.

As shown in Tables 2 and 3, the related samples tests for both detectors support the ARIMA $(1,0,1)(0,1,1)_{672}$ model as the best performer, followed by the group of three best performing k-nearest neighbor forecast methods. These three models, *adjust by* $V(t)$, *adjust by both* $V(t)$ *and* $V_{\text{hist}}(t+1)$, and *adjust by both weight and by distance*, are statistically indistinguishable at the $\alpha = 0.05$ significance level. However, a strong case could be made for extending preference to the *adjust by* $V(t)$ method since it is the simplest approach of the three in addition to having the lowest MAPE. This group of three models is followed in both cases by the *weight by distance* method which is in turn performs better statistically than the remaining three models.

The three poorest performing models in terms of absolute percentage error converted to ranks were, *straight average*, naïve, and *adjust by* $V_{\text{hist}}(t+1)$. No distinction can be made between the *straight average* and naïve forecasts at the $\alpha = 0.05$ significance level for either detector. For detector 4757A the *adjust by* $V_{\text{hist}}(t+1)$ model can be labelled as the worst performer. For detector 4762A this distinction cannot be made at the $\alpha = 0.05$ significance level (the critical Wilcoxon signed-rank test statistic for the comparison of the naïve and *adjust by* $V_{\text{hist}}(t+1)$ forecasts equates to $\alpha = 0.054$).

5. Conclusion

While the heuristic forecast generation methods examined in this research did significantly improve the performance of nonparametric regression, they did not equal the performance of seasonal ARIMA models. This result lends strong evidence to the argument that traffic condition data is characteristically stochastic, as opposed to chaotic.

However, the results did indicate that in cases when the implementation requirements of seasonal ARIMA models cannot be met, using nonparametric regression coupled with heuristic forecast generation methods is preferred to using a naïve forecasting approach. As demonstrated in Tables 2 and 3, nonparametric regression with a sufficiently large value of k and adjusting the forecast by the current traffic volume will provide performance significantly better than a naïve forecast.

Finally, the results of this work point to high potential areas for future research. First, given the strong performance of the parametric models, there is a need to investigate ways to decrease the effort required to "fit" seasonal ARIMA models at different sensor station locations. Furthermore, the improvement of nonparametric regression due to the introduction of heuristic forecast generation methods illustrates that there are other opportunities to further improve the performance of nonparametric regression models. For example, larger databases conceptually will provide a better set of neighbors to use in producing forecasts. In addition, different state definitions and/or distance metrics may lead to better results.

Acknowledgements

The authors would like to acknowledge the support of the Virginia Transportation Research Council, the Virginia Department of Transportation, and the Federal Highway Administration through the Mid-Atlantic Universities Transportation Center. The authors would also like to acknowledge the Telematics Division of the United Kingdom Highways Agency for their graciousness in providing the data used in this study.

References

Box, G.E., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Brockwell, P.J., Davis, R.A., 1996. Introduction to Time Series and Forecasting. Springer-Verlag, New York.

Cheslow, M., Hatcher, S.G., Patel, V.M., 1992. An Initial Evaluation of Alternative Intelligent Vehicle Highway Systems Architectures. MITRE Report 92W0000063. McLean, Virginia.

Disbro, J.E., Frame, M., 1989. Traffic flow theory and chaotic behavior. In: Transportation Research Record, TRB, vol. 1225. National Research Council, Washington, DC, pp. 109–115.

Eubank, R.L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker Inc., New York.

Fuller, W.A., 1996. Introduction to Statistical Time Series, second ed. John Wiley and Sons Inc., New York.

Karlsson, M., Yakowitz, S., 1987. Rainfall runoff forecasting methods old and new. Stochastic Hydrology and Hydraulics 1, 303–318.

Kaysi, I., Ben-Akiva, M., Koutsopoulos, H., 1993. An integrated approach to vehicle routing and congestion prediction for real-time driver guidance. In: Transportation Research Record, TRB, vol. 1408. National Research Council, Washington, DC, pp. 66–74.

Mulhern, F.J., Caprara, 1994. A nearest neighbor model for forecasting market response. International Journal of Forecasting 10, 191–207.

Okutani, I., Stephanedes, Y.J., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B 18B, 1–11.

Smith, B.L., 1995. Forecasting Freeway Traffic Flow for Intelligent Transportation System Applications. Doctoral dissertation. Department of Civil Engineering, University of Virginia, Charlottesville.

Sugihara, G., May, R.M., 1990. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 344, 734–741.

Takens, F., 1981. Detecting Strange Attractors in Turbulence Dynamical Systems and Turbulence. Springer-Verlag, Berlin.

Thearling, K., 1995. Massively parallel architectures and algorithms for time series analysis. 1993 Lectures in Complex Systems, Addison-Wesley.

Williams, B.M., 1999. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal Stochastic Time Series Process. Doctoral dissertation. Department of Civil Engineering, University of Virginia, Charlottesville.