Traffic Prediction Using Multivariate Nonparametric Regression

Stephen Clark[1]

Abstract: The efficient control of traffic on motorways or freeways can produce many benefits, including quicker journey times, fewer pollutant emissions, and reduced driver stress. If it were possible to accurately predict the future state of traffic on a motorway, active measures could be taken to forestall congestion and its attendant negative impacts. This paper presents an intuitive method of producing these forecasts using a pattern matching technique. The technique adopted is novel in that it is a multivariate extension of nonparametric regression that exploits the three-dimensional nature of the traffic state. The application and other facets of the technique are illustrated with actual data from the London orbital motorway. The technique is able to produce forecasts for two of the three traffic state variables with reasonable accuracy and is capable of application on site.

DOI: 10.1061/(ASCE)0733-947X(2003)129:2(161)

CE Database keywords: Traffic control; United Kingdom.

Introduction

The English Highways Agency has embarked on an ambitious program of instrumentation of the London orbital motorway, the M25. This infrastructure consists of many elements. The primary input component is a series of traffic loops, approximately every 500 m, on every lane of the carriageway. These loops are able to report, at 1 min intervals, the state of traffic back to a control center. This traffic state is measured in terms of traffic flow (vehicles), speeds (vehicles/second), loop occupancies (percentage of time a loop is covered by a vehicle), and vehicle headways (seconds between vehicles). The primary use to which these data are put is to automatically control a system of variable speed signs and to help detect incidents. The purpose of the variable speed signs is to impose reduced speed limits that smooth out the flow of vehicles, enabling a more sustained and controlled throughput of vehicles along the carriageway (Nuttall 1995; Maxwell and Beck 1996). At present this system is largely reactive; i.e., responses are made to the current, or more precisely, most recent traffic state. An enhancement would need to be more proactive; i.e., responses could be made on the basis of the anticipated future traffic state. It is hypothesized that this approach may require a less drastic intervention, in terms of its severity and duration, by the operators.

Characteristic Traffic States

A large volume of literature exists on the states in which traffic may exist. The typical situation is free-flowing conditions when the demand flows are below the capacity of the road network. Here, speeds tend to be near the speed limit, the occupancies are low, and vehicle headways are comfortable. In congested conditions, the actual flows reduce, but the demand flows remain high; the vehicles slow down, the occupancies increase, and vehicles pack more closely together. During congestion the road system is operating in an inefficient manner, with increased vehicle delays, driver frustration, and greater potential for accidents. In addition to these two states, there exist two distinct transition states, where the traffic state changes from free-flowing to congested and from congested to free-flow conditions. These two states may be different from each other in their characteristics.

The behavior of the traffic can be illustrated by producing bivariate scatter plots of each measure. Figs. 1–3 show such plots using 10 min carriageway data aggregated from 1 min lane data from one week of observations in March 1998. It is important to realize that the flows reported are the actual flow over the detector in each lane and into the downstream road section and not the demand flow. When the demand flow is less than the capacity of the downstream road section, the demand and actual flows will be the same; however, when the demand flow is greater than the capacity, the demand flow usually exceeds the actual flow, resulting in delays or queues.

Examining the plot of flow versus speed shows a characteristic shape, first identified by Greenshields (1934). Along the top part of this curve, the traffic state is in free-flow conditions where the speed of traffic is relatively constant and high over a wide range of traffic flows. As the flow increases to near the capacity of the road section, the speed begins to fall. At the cusp of the curve the road section is at its maximum capacity, and additional demand forces the traffic state to deteriorate with a dramatic drop in speed and a reduction in the actual flow. The traffic is then in a congested state where the flow and speeds are low, even though the demand is high. Eventually, as the traffic demand reduces, the traffic state will begin to recover, with increases in speeds and actual flows until the traffic is restored to a free-flowing state. If efficiency is measured in terms of the number of vehicles that can pass through the road section, then the most efficient state is at the cusp of the speed-flow curve, where demand is near capacity. The

[1]Research Fellow, Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, U.K.

Note. Discussion open until August 1, 2003. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on February 27, 2001; approved on March 20, 2002. This paper is part of the ***Journal of Transportation Engineering***, Vol. 129, No. 2, March 1, 2003. ©ASCE, ISSN 0733-947X/2003/2-161–168/$18.00.

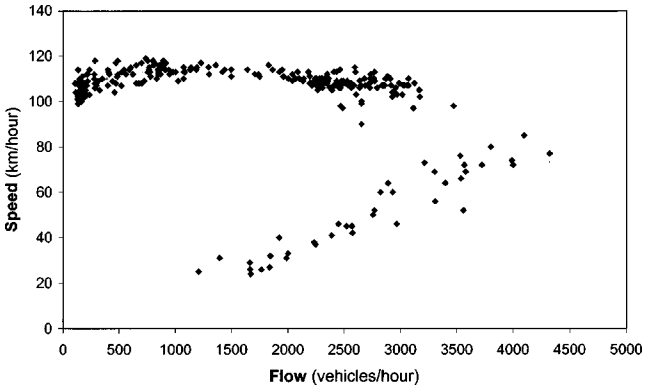

Fig. 1. Scatter plot of speed/flow relationship

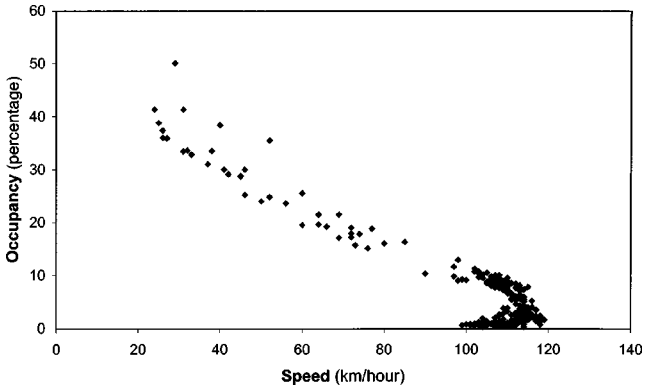

Fig. 3. Scatter plot of occupancy/speed relationship

speed of vehicles at this point is less than the limit but still greater than that experienced during congested conditions.

A similar picture emerges by looking at the other two scatter plots. As demand flows increase, the occupancy across the detector increases linearly until the capacity is reached, and with further increases in the demand flow the actual flow reduces and the occupancy increases dramatically. The increased occupancy is a result of each vehicle spending more time on the detector rather than more vehicles passing over the detector. During congestion, flows remain low and detector occupancies remain high. As seen for the speed-flow relationship, the traffic state eventually reverts to a free-flowing condition when the demand flow reduces to less than the capacity of the downstream road section. The simplest, though less revealing, relationship is evident with the speed-occupancy plot. Here there is evidence of an inverse linear relationship between speed and detector occupancy. In practice, it is not necessary to include headway in these pair-wise considerations, because it is the inverse of flow and conveys no additional information. A more mathematical treatment of these relationships can be found in Haight (1963) and Transportation Research Board (1992).

The aim is to maintain the traffic state in either the free-flowing state or the transition state from free-flowing to congestion for as long as possible, since it is in these states that the road network capacity is being used most effectively. The measures available for this involve control of traffic flows or traffic speeds. In a motorway context, the flows may be controlled by ramp metering on the junctions along the motorway or by enabling variable message speed signs. With ramp metering, if it were possible to know ahead of time when ramp metering was to be deployed, drivers could be advised to try an alternative route or mode of travel. This would avoid the congestion that can occur in the local road network adjacent to ramp metering sites. The other approach is to advise drivers on the motorway to travel at a speed less than the statutory maximum. The Highways Agency has adopted this latter approach to control the traffic state on the M25. As the traffic flow exceeds a specified threshold, the variable speed signs are used to advise a lower speed limit, and a further increase in flow above a second threshold limit advises an even lower speed limit. When the road section recovers from congestion the signs are deactivated.

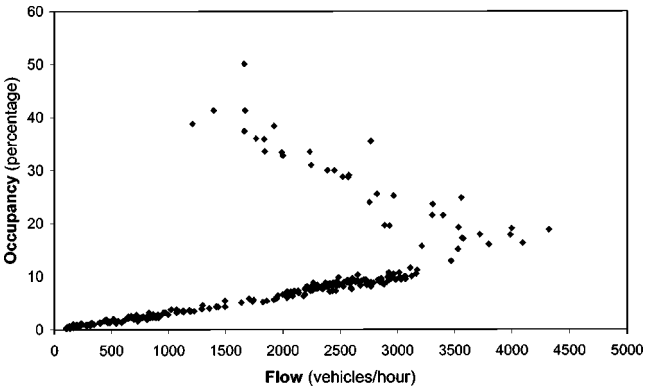

Fig. 2. Scatter plot of occupancy/flow relationship

Forecasting Problem

Currently, the system is reactive to the traffic state in the previous time period, and an enhancement is to operate the system in a proactive manner. This requires a prediction of the traffic state in the near future, if no action were to be taken. In order to define the state of traffic, two of the three measures are required. To illustrate this requirement, consider the situation where only flow data is available. The system can be activated when the flow is expected to exceed a threshold value. The system can then be deactivated when the flow is predicted to fall below another threshold. It is not possible, however, to know whether this reduction is due to reduced demand or a congestion effect, and in the latter case the deactivation is a false alarm. If speed or occupancy predictions are available, then the two situations can be distinguished.

Many attempts have been made to predict traffic flows along a motorway network. Of the more recent attempts, a study of French interurban motorway data described in Danech-Pajouh and Aron (1991) uses time period specific regression models of 30 min traffic data to produce forecasts. In Davis and Nihan (1991), a comparative study of a limited data set was conducted, consisting of 1 min traffic flows, and concluded that the nonparametric regression approach was no better than a linear time series model. In their conclusions, they suggest that the poor performance may be a result of the size of the small ''training'' database used in the study. Another comparative study by Smith and Demetsky (1997) fitted historical, time series, artificial neural network, and nonparametric models to 15 min North American freeway data to produce one-step-ahead forecasts, and in Smith and Demetsky (1996) this is extended to produce many-step-ahead

J. Transp. Eng. 2003.129:161-168.

traffic flow forecasts. In both these papers the nonparametric method is identified as the one that produces the lowest and best structured errors.

A part special issue of the *International Journal of Forecasting* contains a substantial body of work in this area. The introductory paper of this issue, Van Arem et al. (1997), contains an overview description of the issues involved in this area of study, in particular a consideration of methodologies and current practice. They also précis other papers in the issue that are mainly concerned with the application of a wide variety of modeling and forecasting techniques to primarily European traffic data.

The same data used in this study was used to forecast 15 min traffic flows using a range of naive techniques, Box-Jenkins ARIMA (Box and Jenkins 1976) models, and the composite method in Clark et al. (1999). A longer range forecasting study of monthly turnpike interstate traffic volumes was conducted by Marshment et al. (1996) using ARIMA and regression techniques. The regression approach was found to perform as well as or better than the ARIMA method, but it did require a larger database of socio-economic variables.

Other attempts have been made to forecast urban traffic flows. Stephanedes et al. (1991) describe five forecasting methods that can be used to produce traffic forecasts for inclusion in an urban traffic control system. ARIMA methods were used by Moorthy and Ratcliffe (1988) to produce forecasts of monthly traffic flows. Use of 5 min flows from an urban traffic control system to evaluate the forecasting performance of time series and artificial neural network models of traffic flow is reported in Clark et al. (1993).

A question arises from the parametric modeling approaches as to whether there is indeed an appropriate model formulation for the future behavior of the traffic state. If the models proposed are not appropriate, then there can be little expectation of a reasonable forecasting performance. A model that does not rely on any parametric assumptions is likely to be a more robust forecasting tool but may not have the forecasting accuracy when compared with a true model formulation.

Modeling Approach

This paper adopts a nonparametric regression technique described as a k nearest neighbor (k-nn) model. This model is best described as a pattern matching exercise, where recent observations are matched with those contained in a database of historical observations. From all the matches, either the k nearest matches or all the matches below a given distance threshold are located. The successive observations from these ''best'' matches are then averaged, usually by taking the arithmetic mean, to obtain the forecasts. The only parameters in the model are the number of observations to match with and the number of best matches to retain, or the distance threshold. Once a sequence of recent observations has been matched and forecasts made, the recent observations can then be added to the historical matching database for use in subsequent matching operations. This is termed a nonparametric method, because there are no distributional constraints or assumptions placed on either the input variables or the output variables. In particular, there are no restrictions on the form of the residuals from such models as are typically required for parametric regression techniques.

Case Study

For this study a site, 4989A, on the M25 was selected, and its location is given in Fig. 4. From this site a matching database of

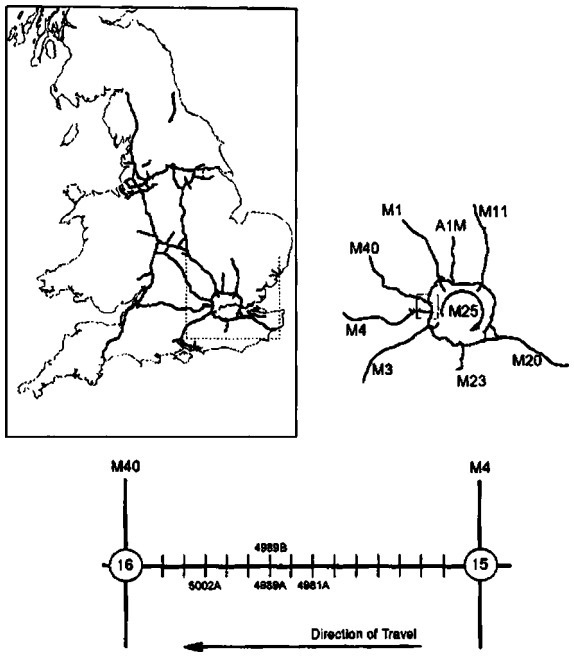

Fig. 4. Schematic diagram of case study area

three weeks of aggregate 10 min traffic statistics were assembled from February 10 to March 2, 1998, resulting in a database of 3,024 observations. Each observation consisted of the recorded actual flow, speed, and occupancy at the site. The test observations consist of one week of data from March 3 to March 9 in the same format from the same site and separately from neighboring sites. Stored as binary files, a week of data occupies just 8 k of disk storage. The dataset contains rare periods of missing data, where data for a traffic lane is not available for up to 5 min. The aggregation process across all four traffic lanes and across 10×1 min recording periods, however, helps to reduce the impact of these missing observations.

The question arises as to how to establish the closeness of a sequence of recent observations to those in the matching database. With a univariate approach, the statistic could be the sum of squares (ss):

$$ss=\sum_{i=1}^{L}(x_{ri}-x_{mi})^2$$

where x_{ri} = recent observation at lag i; x_{mi} = matching observation at lag i; and L = number of lags to match with. An alternative could be a measure based on the absolute difference.

Given the multidimensional nature of the traffic state, an extension is a multivariate match on the three measures of flow, speed, and occupancy. In this case, the closeness statistic could be the total sum of squares (tss):

$$tss=\sum_{i=1}^{L}\left[\frac{(q_{ri}-q_{mi})}{w_q}\right]^2+\sum_{i=1}^{L}\left[\frac{(v_{ri}-v_{mi})}{w_v}\right]^2+\sum_{i=1}^{L}\left[\frac{(o_{ri}-o_{mi})}{w_o}\right]^2$$

where q = 10 min aggregate flow; v = 10 min average speed; o = 10 min average occupancy; and w_q, w_v, w_o = measure specific weights.

The weights were necessary to reflect the different magnitudes at which each of the measures is recorded. The values of these weights have been set at w_q = 1,000 vehicles per 10 min; w_v = 100 km/h; and w_o = 15%. These weights specify that a match

J. Transp. Eng. 2003.129:161-168.

of 1% in the occupancy score is equivalent to a match of 6.67 km/h in the speed and 66.67 vehicles in the flow measure. The choice of these weights has been based on a comparison of the mean, spread, and maximum observed value within a day for a sample of the data set. If, however, matches against a specific measure were thought to be more critical, then these weights could be varied by reducing the weight associated with that measure. An alternative is a percentage statistic, which does not require an explicit weight.

This technique of matching and forecasting uses two measures of quality, one during the matching process and one in producing the forecast. A good match between the testing observations and those in the matching database is quantified by a low total sum of squares (*tss*). The second measure is an accurate forecast, which is measured by close agreement between the forecast and the actual observation. It should be noted that a good quality match does not necessarily imply an accurate forecast. Recently, Smith et al. (2000) have suggested the use of a weighing technique where the forecast is no longer a simple average of those selected but is a weighted average. The weights can be a function of the closeness of the match between the testing and matching data (i.e., the *tss* values) or some other relative measure.

Goodness of Fit Measures

A common statistic used for the assessment of goodness of fit is the root mean square error (RMSE). This is:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

where x_i = observed value of the measure; $\hat{x}_i$ = forecast value of the measure; and N = number of observations.

The drawbacks to this goodness of fit measure are that it has the units of the measure and its magnitude depends upon the magnitude of the observations. The mean absolute percentage error (MAPE) goodness of fit statistic is unitless and insensitive to changes in the magnitude of the forecasts:

$$\text{MAPE} = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - \hat{x}_i|}{x_i}\right) \times 100$$

Both of these statistics will be used to assess the quality of the forecasts produced. In addition, a naive method of forecasting where

$$\hat{x}_i = x_{i-1}$$

is used, and the RMSE and MAPE from this naive forecast is compared with that obtained from the method of *k-nn*.

The experience from other modeling attempts may also provide insight into the performance of the *k-nn* method. The level of aggregation of the observations may have an impact on the performance statistic. Data from a shorter time span will tend to be noisier or less smooth than that from a longer time span, making accurate predictions more of a challenge. Many of the more relevant studies have used 15 min data, while this study has adopted a 10 min span, which may produce larger RMSE and MAPE statistics. Smith and Demetsky (1997) suggested one-step-ahead flow prediction performance in the range of 7.5–11%, whereas Davis and Nihan (1991) presented RMSE statistics that roughly translate to an 11–15% error in flow forecasts. They also suggest larger errors of between 20 and 30% in forecasts of occupancy. The ATHENA method reported in Danech-Pajouh and Aron (1991) suggests MAPE errors in 30 min flow forecasts of between 8 and 13%. The study by Marshment et al. (1996) reported MAPE errors in the range of 3.9–8.2%, but these models were based on less volatile monthly traffic volume statistics. Wilde (1997) produces MAPE statistics for 4.5 min ahead flow predictions, using three prediction methods of 14.60%, 13.44%, and 12.76%, and for 15 min ahead predictions, 18.70%, 17.98%, and 17.73%. Wilde notes that the first of these methods uses a form of pattern identification and matching. It is also noted that ''it is difficult to beat the simple time-series predictors for horizons shorter than about 10–15 min.'' In Dougherty and Cobbett (1997) one-, two-, and six-step-ahead forecasts of flow, speed, and occupancy are made, but the evaluation statistic they use, the root mean square error proportional, is not directly compatible with either the MAPE or the RMSE.

Exploring Options

The first task is to establish whether the univariate or multivariate approach to prediction is better. Given an approach, the issues then are what length of recent observations to match against and what number of retained matches gives the best performance. The next task was to see whether this forecasting method was transferable to other sites. A final consideration is the composition of the matching database, in particular, whether it is more efficient to maintain smaller, day-specific, databases. For the purposes of this study, the oldest set of observations was replaced when the most recent ones became available, keeping the size of the matching database constant and contemporary.

Univariate or Multivariate

During this test, the performance of methods that match against a subset of measures to produce forecasts are compared. Here the length of matches to make (L) is set at four observations, representing 40 min of traffic data, and the number of matches to retain (M) is set at eight. Only a one-step-ahead forecasting horizon is used, representing a forecast of the traffic state for the next 10 min. Table 1 shows how well matches on speed (v), flow (q), occupancy (o), speed and flow (vq), speed and occupancy (vo), flow and occupancy (qo), and all three measures (vqo) perform. The naive forecasting method's performance was the same irrespective of the measures used in the match and is given at the bottom of the table.

Computer code to perform the matching and forecasting was written in C and is capable of producing all the forecasts for a week (1,008 forecasts) from a matching database of three weeks (3,024 observations) in 20 s on a 350 MHz Pentium II PC. The most time-consuming element of the process is the sequential search for matches. For a reasonably small dataset of this size, this is not too great a consideration, but for larger datasets where results are required in real time, more sophisticated data structures and search algorithms may be required (Oswald et al. 2000).

It is clear from Table 1 that the best forecasting performance is achieved if the measure being forecast is part of the matching criteria. It is also clear that, with the exception of the speed measure, a multivariate match performs better. The univariate speed criterion produces very poor forecasts of flow and occupancy. This is due to the situation where many very similar speed levels can be associated with a wide range of flow and occupancy levels,

J. Transp. Eng. 2003.129:161-168.

Table 1. Comparison of Univariate and Multivariate Matching Strategy

Match	RMSE			MAPE		
	q	*v*	*o*	*q*	*v*	*o*
q	74.67	14.30	5.03	10.24	12.19	24.42
v	317.90	6.25	4.35	204.35	5.06	205.61
o	86.41	7.68	3.01	14.87	6.59	16.66
vq	69.93	6.66	3.14	10.48	5.43	18.03
vo	82.97	6.59	3.01	13.11	5.44	16.67
qo	75.06	7.54	3.02	10.52	6.27	15.86
vqo	75.38	6.64	3.00	10.76	5.47	15.87
Naive	82.10	6.40	3.61	10.98	4.89	17.07

as is evident in Figs. 1 and 3. The best compromise performance is gained by matching on all three measures, although there is some evidence that, if only a forecast of speed is required, then either a match on speed alone or the naive method is worth exploring further. With the exception of speed, the performance is better than the naive method, and the results compare well with those reported in earlier studies. Of particular note is the performance in predicting occupancy, which is below the range reported in Davis and Nihan (1991).

Estimation of *L* and *M*

To explore the issue of what values to adopt for *L* and *M*, different values of *L* and *M* are applied in a three-variable model and the accuracy of the forecasts calculated. It may be the case that different values are required for accurate forecast of different traffic measures. As the value of *L* changes, the naive forecasts will change [because only $(N\text{-}L)$ forecasts are possible] but not as *M* changes.

Examining Table 2, using only a small number of matches (*M*) tends to give poor forecasting accuracy, suggesting a value of *M* of near 8 or 10. Reasonable and consistent accuracy is obtained by choosing lags (*L*) somewhere in the middle range, perhaps three or four lags, representing 30 and 40 min of observations. Forecasts of speed tend to be most accurate for large values of *M* but low values of *L*.

It may appear that, by performing this exercise of estimating *L* and *M*, the *k-nn* model is actually a parametric model, with *L* and *M* being its parameters. The term nonparametric, however, refers to the absence of assumptions on the distribution of the data and the model form, not to the absence of parameters. More sophisticated procedures for efficiently estimating the values of *L* and *M* are available; for example, genetic algorithms are suggested in Oswald et al. (2000).

Transferability of Database

Maintaining a site-specific database can be a complex task. The method outline in this paper would be much more practical if it

Table 2. Comparison of Forecast Accuracy with Differing Values of *L* and *M*

M	RMSE						MAPE					
	10	8	6	4	2	Naive	10	8	6	4	2	Naive
						(a) Flow						
L: 6	78.95	78.63	78.13	76.36	86.57	82.19	10.71	10.7	10.74	10.75	11.66	10.99
5	77.37	76.38	76.68	77.95	84.98	82.15	10.69	10.66	10.62	11.03	11.72	10.99
4	76.20	75.38	74.67	75.28	84.49	82.10	10.85	10.76	10.85	11.22	11.80	10.98
3	77.33	77.56	78.25	80.68	85.97	82.07	10.98	11.07	11.08	11.38	12.22	11.01
2	75.69	77.26	77.99	80.41	83.18	82.03	11.03	11.10	11.27	11.44	12.40	11.01
1	75.44	77.10	78.30	80.09	85.35	81.99	11.14	11.25	11.35	11.71	12.57	11.01
						(b) Speed						
L: 6	7.18	7.24	7.21	7.34	8.20	6.41	5.92	5.89	5.92	5.93	6.61	4.90
5	7.14	7.06	7.00	7.09	7.79	6.41	5.85	5.78	5.69	5.75	6.30	4.89
4	6.64	6.64	6.79	6.79	7.48	6.40	5.47	5.47	5.56	5.50	6.08	4.89
3	6.59	6.65	6.75	6.97	7.57	6.40	5.49	5.50	5.56	5.71	6.10	4.89
2	6.30	6.39	6.52	6.79	7.24	6.40	5.20	5.25	5.34	5.57	6.08	4.88
1	6.11	6.23	6.48	6.60	7.07	6.39	5.06	5.16	5.36	5.49	5.74	4.88
						(c) Occupancy						
L: 6	3.27	3.24	3.19	3.23	3.61	3.62	16.23	16.38	16.29	17.00	18.25	17.06
5	3.20	3.15	3.13	3.18	3.39	3.62	16.23	15.91	15.89	16.34	17.27	17.05
4	3.03	3.00	3.09	3.01	3.47	3.61	16.10	15.87	16.02	16.54	16.98	17.07
3	3.11	3.13	3.21	3.26	3.51	3.61	16.48	16.48	16.72	16.98	18.52	17.08
2	3.02	3.09	3.17	3.34	3.43	3.61	16.76	16.08	17.14	17.55	19.09	17.08
1	3.01	3.10	3.21	3.24	3.43	3.61	17.24	17.41	17.79	18.36	19.76	17.06

J. Transp. Eng. 2003.129:161-168.

Table 3. Comparison of Forecast Accuracy at Different Sites

Site	Model	RMSE			MAPE		
		q	*v*	*o*	*q*	*v*	*o*
4981A	*k-nn*	79.71	7.07	2.86	10.76	6.01	15.84
	Naive	92.29	6.91	3.35	11.17	5.58	16.39
4989B	*k-nn*	74.25	6.06	2.62	10.39	4.68	14.85
	Naive	78.06	5.39	2.94	10.44	3.98	14.69
5002A	*k-nn*	83.28	7.22	3.56	11.51	6.31	17.82
	Naive	84.82	7.09	4.11	11.36	5.23	16.89

were possible to maintain a general ''pool'' of historic data to match against. Before this is done, it is reasonable to ask if a matching database from a site is suitable for forecasting observations at a neighboring site. To test this, a week's data from three neighboring sites is chosen. One is at the same physical location as the matching site, but on the opposite carriageway, i.e., counterclockwise traffic (4989B), another is a site 1.5 km downstream (5002A) on the same carriageway, and the third is 1 km upstream (4981A), again on the same carriageway. When recent observations from the appropriate forecasting site became available for storage in the matching database, they replaced the oldest observation in the matching database from site 4989A. This approach means that, as forecasting continues, more observations from the forecasting site become eligible for matching, and ultimately these observations constituted a third of the matching database. The length of matches to make (L) is set to four and the number of matches to retain (M) is set at eight.

The proportion of matches that have come from the newer observations (i.e., previous observations from the testing week) is around the 20–30% mark, a similar value to that for forecasting at site 4989A. This demonstrates that there is no tendency for a site to match against its own recent historic observations rather than those from the other site.

The forecast accuracy of the measures appears to have remained at the expected levels suggested by the results in Tables 1 and 2. The performance relative to the naive method is, however, poorer. This last feature may suggest that there are site-specific characteristics that are important, and it may be necessary to maintain site-specific databases. Alternatively, if some method of clustering like sites together could be established, then sites within the cluster could share the same common database. In any case, the results in Table 3 suggest that it is reasonable to use data from a neighboring site in order to prime a new site or a site with poor quality data.

Day of the Week

It could be hypothesized that when forecasting for a particular day, there is a bias towards matching with observations from the same day in the matching database. To determine which days the observations were matched against, a record was kept of where the eight matches came from. This information is presented in Table 4.

This table shows that with the Monday testing day, 306 matches came from Mondays in the matching database, 206 came from Tuesdays, 160 from Wednesdays, etc. As would be expected, there are large values on the main diagonal, meaning that testing days are matched with the same or similar days. This was especially true for Saturdays and Sundays. This would suggest that the size of the matching database could be reduced so that it contained only the relevant day's data. This would reduce the time needed to search the matching database, while hopefully still maintaining the same degree of forecasting accuracy.

To test this accuracy hypothesis, a further forecasting exercise was conducted. Here the matching and testing databases were portioned into two databases, one containing only weekdays and a second containing only Saturdays and Sundays. This would mean, for example, that Monday testing data could only be matched against Monday–Friday matching data. Table 5 shows the results of three matching exercises. The first row is this new day of the week fit, the second is the naive equivalent, and the third is a comparison match against the full database of all days. The last two results are necessary, because the equivalent entries in Tables 1 and 2 are no longer valid when we use three initialization periods rather than one. These two additional initialization periods are 40 min blocks at the start of Saturday, March 7, and Monday, March 9, when no forecasts are possible due a day type transition.

A comparison of day of the week results with the naive equivalent are the same as those established earlier. What is of

Table 4. Days Used in Matching Process

Matched day	Testing Day						
	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	306	191	240	110	121	77	98
Tuesday	206	213	276	212	205	57	53
Wednesday	160	241	206	209	187	40	70
Thursday	226	176	249	266	259	63	17
Friday	121	248	147	247	321	82	49
Saturday	80	44	20	44	43	631	325
Sunday	53	7	14	64	16	202	540

J. Transp. Eng. 2003.129:161-168.

Table 5. Comparison of Forecasting Accuracy Using Day of Week Databases and Full Week Databases

	RMSE			MAPE		
Match	q	v	o	q	v	o
Day of week	76.14	6.75	3.01	10.89	5.59	15.88
Naive	82.27	6.42	3.62	10.99	4.91	17.05
Full	75.35	6.67	3.01	10.70	5.52	15.90

interest, however, is that in all but one instance the day of the week result is worse than a full matching database result. By restricting the content of the matching database, we have forced the method to accept poorer matches, which appears to have produced poorer forecasts. This suggests that a rich database that contains ''idiosyncratic'' observations may be preferable to a tailored database.

Conclusions

Application of the method of multivariate nonparametric regression of traffic data proposed in the paper is only possible now that we have a rich source of high quality traffic data. Traditionally, only short runs of very aggregate flow measurements were available to the traffic modeler. These measurements were seldom sufficient to build a matching database that contained the diversity of day-to-day variations required and the different measures needed to capture the diversity in the traffic state.

The nonparametric method poses a number of advantages over alternatives. First, the method is intuitive in its formulation and is thus readily understood by the practitioners in the field. Second, it does not rely on the formulation of a parametric model that requires assumptions on the transition of the traffic state from one time period to the next. Third, it is able to explicitly use the multivariate nature of the traffic state to produce forecasts. Finally, the method is simple and quick to implement, having only modest data storage requirements; thus, forecasts can be obtained well in advance of a decision being required. For all these reasons, this technique is of real practical benefit to the engineers whose role it is to control the traffic for the benefit of all, both drivers and nondrivers.

Forecasts of flow and occupancy produced by the method were more accurate than those provided by the naive method, but speed forecasts tended to be worse. Of the three measures considered in this paper, the speed measure contains the least information in terms of patterns, because for most of the day the speed measure is constant. Even with this reduced performance on the speed measure, the method is able to produce reasonably accurate short term forecasts of all three measures.

Further studies are possible into the use of alternative weights in the total sum of squares statistic. If a more accurate prediction of flows were thought desirable, then the ''weight'' associated with the flow component could be reduced. Alternatives to the total sum of squares statistic are possible, including a measure based on the percentage difference. The third area of future study is the production of many-step-ahead forecasts rather than the one-step-ahead forecasts produced here. If these many-step-ahead forecasts can be accurately produced then even longer notice of the future traffic state can be obtained.

Acknowledgments

The writer would like to thank Stuart Beale of the Highways Agency for his help in the supply of the data used in this study. Additionally, the writer would like to thank the referees who took the time and effort to consider earlier drafts of this paper and who made many helpful suggestions. Any views expressed in this paper are those of the writer alone and do not represent the views of the Highway Agency or the Department of Transport, Local Government and the Regions.

References

Box, G. E. P., and Jenkins, G. M. (1976). *Time series analysis, forecasting and control*, 2nd Ed., Holden-Day, San Francisco.

Clark, S. D., Chen, H., and Grant-Muller, S. M. (1999). ''Artificial neural network and statistical modelling of traffic flows—the best of both worlds.'' *Proc., 8th World Conf. on Transport Research*, Antwerp, Elsevier Science, U.K., 215–226.

Clark, S. D., Dougherty, M. S., and Kirby, H. R. (1993). ''The use of neural networks and time series models for short term traffic forecasting: a comparative study.'' *Transportation Planning Methods: Proc., PTRC 21st Summer Annual Meeting*.

Danech-Pajouh, M., and Aron, M. (1991). ''ATHENA: a method for short-term inter-urban motorway traffic forecasting.'' *Recherche Transports Sécurité*, 6, 11–16.

Davis, G. A., and Nihan, N. L. (1991). ''Nonparametric regression and short-term freeway traffic forecasting.'' *J. Transp. Eng.*, 117(2), 178–188.

Dougherty, M. S., and Cobbett, M. R. (1997). ''Short-term inter-urban traffic forecasts using neural networks.'' *Int. J. Forecast.*, 13, 21–31.

Greenshields, B. D. (1934). ''A study of traffic capacity.'' *Proc. Highway Research Board, Highway Research Board, Washington, D.C.*, 14, 448–477.

Haight, F. A. (1963). *Mathematical theories of traffic flow*, Academic, New York.

Marshment, R. S., Dauffenbach, R. C., and Penn, D. A. (1996). ''Short-range intercity traffic forecasting using econometric techniques.'' *Inst. Transp. Eng. J.*, 66(2), 37.

Maxwell, H. A., and Beck, I. (1996). ''Traffic control on the English motorway network.'' *Proc., 8th Int. Conf. on Road Traffic Monitoring and Control*, 136–144.

Moorthy, C. K., and Ratcliffe, B. G. (1988). ''Short term traffic forecasting using time series methods.'' *Transp. Plan. Technol.*, 12, 45–56.

Nuttall, I. (1995). ''Slow, slow, quick, quick, slow: taking the 'stop-start' out of the London Orbital.'' *Traffic Technology International*, Winter, 46–50.

Oswald, R. K., Scherer, W. T., and Smith, B. L. (2000). ''Traffic flow forecasting using approximate nearest neighbor nonparametric regression.'' *Research Project Rep. for ITS Implementation Research*, ⟨http://www.gmupolicy.net/its/papers.htm⟩ (Apr. 25, 2002).

Smith, B. L., and Demetsky, M. J. (1996). ''Multiple-interval freeway traffic flow forecasting.'' *Transportation Research Record 1554*, Transportation Research Board, Washington, D.C., 136–141.

Smith, B. L., and Demetsky, M. J. (1997). ''Traffic flow forecasting: comparison of modeling approaches.'' *J. Transp. Eng.*, 123(4), 261–266.

Smith, B. L., Williams, B. M., and Oswald, R. K. (2000). ''Parametric and nonparametric traffic volume forecasting.'' *Proc., Transportation Research Board Annual Meeting*, Transportation Research Board, Washington, D.C., Reprint 00-817.

Stephanedes, Y. J., Michalopoulos, P. G., and Plum, R. A. (1991). ''Im-

J. Transp. Eng. 2003.129:161-168.

proved estimation of traffic flow for real-time control.'' *Transportation Research Record 795*, Transportation Research Board, Washington, D.C., 28–39.
Transportation Research Board (TRB). (1992). ''Traffic flow theory.'' *Special Rep. 165*, ⟨http://www-cta.ornl.gov/cta/research/trb/tft.html⟩ (Apr. 25, 2002).
Van Arem, B., Kirby, H. R., Van Der Vlist, M. J. M., and Whittaker, J. C. (1997). ''Recent advances and applications in the field of short-term traffic forecasting.'' *Int. J. Forecast.*, 13, 1–12.
Wilde, D. (1997). ''Short-term forecasting based on a transformation and classification of traffic volume time series.'' *Int. J. Forecast.*, 13, 63–72.

J. Transp. Eng. 2003.129:161-168.