Composite Nearest Neighbor Nonparametric Regression to Improve Traffic Prediction

Matthew D. Kindzerske and Daiheng Ni

The ability to predict traffic conditions accurately is of paramount importance in effective management of a highway network. A more accurate prediction will allow for better allocation of resources, which may reduce experienced travel times. This paper introduces a composite approach to the already popular nonparametric regression used in predicting traffic conditions. The composite approach performs a nearest neighbor search for each loop detector station using only data that are in proximity to the detector’s position on the roadway. This method accommodates every detector station individually to minimize the forecast error on the entire roadway. A case study using data from the Next Generation Simulation program recorded on US Highway 101 demonstrates that the composite approach significantly mitigates forecast error and performs the forecast in a reasonable amount of computational time. The case study also shows the ability of the composite approach to predict the onset and propagation of traffic shock waves.

Motor vehicle traffic congestion is an everyday reality for millions of commuters. Development and sprawl continue to test the capacities of roadway infrastructure and put pressure on politicians to allocate funds to increase those capacities. It is the job of traffic engineers to efficiently manage the roadways as a resource, and recently this has meant implementation of an intelligent transportation system (ITS) to maximize efficiency and safety. The ultimate goal in the transportation industry is to integrate detection, management, and control systems across large roadway networks and transmit pertinent information to travelers at any stage of a trip. The keystone of this integration is the ability to predict traffic conditions (i.e., flow, speed, and occupancy) on a road network. Numerous methods for short-term prediction of traffic conditions have been researched and implemented. Smith and Demetsky (1) state that the “nearest neighbor (k-NN) formulation of nonparametric regression holds considerable promise for application to traffic flow forecasting.” Clark (2) recently extended the nonparametric regression approach into a multivariate method for robust short-term traffic condition predictions. This paper seeks to define and support a composite approach to nearest neighbor nonparametric regression with the goal of increasing its performance and reducing prediction error.

MEASURING TRAFFIC

Cities are continuously installing new hardware to measure and detect traffic. The two types of point sensors currently in use are loop detectors and video cameras. Both have advantages and disadvantages. Video cameras have the capacity to supply much more information with greater detail than loop detectors. With advancements in optical content recognition algorithms, video cameras can identify individual vehicles and plot accurate microscopic trajectories. Their disadvantage is that sometimes the rate of failed reads is high as a result of external conditions such as time of day or solar glare.

Loop detectors tend to be easier and cheaper to install, and hence they are favored for mass use over video cameras. Loop detectors are not as informative as video cameras, but instead provide data describing the traffic condition (i.e., flow, speed, and occupancy). Given the wide adoption of loop detectors, methods for traffic prediction focus on traffic condition and they attempt to predict the flow, speed, and occupancy on the short-term horizon.

PREDICTING TRAFFIC CONDITION

Efforts to predict traffic conditions can fall under four main categories: historical averaging, time series, neural networks, and nonparametric regression.

Historical averaging is close to a naive method. Because it relies on the repetition of traffic from day to day, historical averaging “has no way to react to dynamic changes” (1).

Traffic condition data are a classic application of time series analysis, given the seemingly random behavior of traffic. A significant number of papers have been published with an analysis and case study on the autoregressive integrated moving average (ARIMA) technique. ARIMA’s major disadvantages include an inability to deal with incomplete data and a potentially large amount of time to calculate parameters. In addition, time series is a good technique for recurrent traffic conditions, but it lacks the ability to deal with non-recurrent traffic conditions or random events (such as those caused by incidents).

Neural networks have recently been introduced to intelligent transportation systems. Neural networks do not model traffic. Instead, they are trained with historical data to determine the weights between numerous processing elements in a network, including any elements in hidden layers. To make a short-term prediction, current traffic condition information is fed into the input layer of the neural net-

M. D. Kindzerske, University of Massachusetts, 1235L Lederle Tower, 710 North Pleasant Street, Amherst, MA 01003. D. Ni, Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 219 Manzan Hall, 130 Natural Resources Road, Amherst, MA 01003. Corresponding author: D. Ni, ni@ecs.umass.edu.


30
work and then the respective weights are propagated down the network until the information reaches the output layer, which should represent a prediction. Although the neural network paradigm lends itself to modeling complex relationships such as traffic, its unavoidable disadvantage is the lengthy amount of time required to train the network, which makes the paradigm an unlikely candidate for large-scale traffic prediction (3).

Nonparametric regression received attention when Davis and Nihan (4) realized the applicability of Yakowitz's work in "extending the k-[nearest neighbor] method to time-series data" (5). Davis and Nihan (4) concluded that the k-NN method would lend itself to traffic condition forecasting because of its adaptability to nonlinear relationships that traffic conditions tend to exhibit.

Oswald et al. (6) addressed the issue of the amount of computational time required to implement a k-NN method. Because the method is only as good as the historical data from which it is choosing nearest neighbors, a large, varied, and detailed historical data set is ideal. Oswald et al. (6) investigated the use of approximate nearest neighbors with kernel nonparametric regression as well as K-dimensional (KD) trees. An approximate nearest neighbor search sacrifices accuracy in a neighbor match to deliver a prediction in a time-critical situation. KD trees initially partition historical data using a given key. When a search is performed, a query will descend the KD tree according to the key, and as the search descends, the size of the historical data space is reduced. "KD trees were used to reduce execution time by as much as 1,000 times without sacrificing forecast accuracy," concluded Oswald et al. (6), who went on to say that approximate nearest neighbors "further reduced execution time but at the expense of forecast accuracy."

Oswald et al. (6) also defined a traffic state vector to include data from time $t$, $t - 1$, ..., $t - d$, where $d$ is the number of lags chosen to maximize the efficacy of the k-NN method to make accurate predictions. In Nikovski et al. (3) experimentation showed that "the current level of congestion has a major effect on the future level of congestion, but most often it is not important how the current level of congestion was reached."

Clark (2) defined a closeness statistic to allow the k-NN method to match multiple variables such as flow, speed, and occupancy. The statistic is a "total sum of squares," which uses specific weights to specify how a match should relate. The case study in Clark (2) showed that the least error occurred when all three parameters were included in the search.

Recently, research has been conducted to identify the number $k$ (from k-NN) that optimizes a nonparametric prediction. Of course, when $k$ is greater than one, a scheme must be used to combine $k$ results into one prediction. Turoch (7) concluded that schemes that combined nearest neighbors with an inverse proportion to their relative distances performed better with more complex procedures. Further trials showed that possible optimal values for $k$ exist between 20 and 25.

NONPARAMETRIC REGRESSION—COMPOSITE METHOD

A multivariate nonparametric regression method is intended to make traffic condition predictions, that is, to predict the data a traffic management center will observe in the short term. Smith et al. (8) defined $V(t)$ as a vector representing the flow rate ($q$) over an entire section at time $t$. This can be extended into a multivariate approach by letting $V(t)$ be a two-dimensional array representing the flow rates ($q$), speeds ($v$), and occupancies ($o$) over an entire section. A nonparametric regression method compares the current traffic condition with an archive of historical traffic conditions to find a closest match to make a prediction. Previous studies have used the studied roadway's entire section to make forecasts for the entire section. This aggregated approach can be counterintuitive. Conditions at one end of a section of roadway will most likely have little or no effect on the short-term future conditions at the opposite end of a roadway. This is motivation for a composite approach for nonparametric regression.

If $n$ loop detectors on a section of roadway are numbered with respect to order, then $V_j(t)$ can represent the traffic condition ($q$, $v$, and $o$) at the $j$th loop detector. To represent the predicted traffic condition at time $t + 1$, the notation $V_j(t + 1)$ is used. To predict the traffic condition at detector $i$ on the next time interval, the current conditions at detector $i$ will certainly be taken into account. Then the question arises as to what else contributes to the traffic condition at detector $i$. The condition of the traffic downstream (in the positive direction of travel) of detector $i$ will determine how vehicles in detector $i$'s vicinity will exit that vicinity. Now $V_j(t + 1)$ can be forecast using $V_j(t)$ and several downstream detectors' data: $V_j(t)$, $V_{j+1}(t)$, $V_{j+2}(t)$, ..., $V_n(t)$, where detector $(i + c)$ is the farthest detector that a vehicle can travel from detector $i$ within the time defining the short term at the designated maximum speed allowed for the section of roadway.

The same reasoning can be applied for the upstream direction (in the negative direction of travel). The condition of traffic entering the vicinity of detector $t$ will depend on the condition of traffic upstream of detector $i$. Now the data $V_{i-1}(t)$, $V_{i-2}(t)$, ..., $V_{i-d}(t)$ will be incorporated to predict $V_j(t + 1)$, where detector $(i - d)$ is the farthest upstream point that a vehicle can start and still arrive at detector $i$ within the time defining the short term and at the designated maximum speed allowed for the section of roadway.

If detector $i$, the highlighted loop detector, is the last detector on a section of roadway, there will of course be no downstream detector data to contribute to a prediction and vice versa for the first detector in a section of roadway. As a result, predictions may not be as precise on the ends of a roadway, but forecast quality always hinges on the quality of the historical database with nonparametric regression.

A diagram of a sample roadway section and its detectors at two different times ($t$ and $t + 1$) is shown in Figure 1. The arrows...
demonstrate which detector data will be used to compute a prediction for $V_{oa}(t + 1)$.

As seen in Figure 1, the data from Detector 006 are excluded from the forecast of $V_{oa}(t + 1)$. An incident at 006 has little chance of affecting the quantitative value of $V_{oa}(t + 1)$ within the short term, given its distance from Detector 003. If there was an incident at 006 and it was included in the forecast for the entire line (as the aggregate method does), it would affect the values of the distance metric in the search and potentially choose a nearest neighbor that satisfies the entire road section’s condition, while predicting overly inaccurate conditions for Detector 003. The composite method accommodates every detector individually to minimize the error on the entire section of roadway.

**TECHNICAL DEFINITION**

Figure 2 illustrates the method used to decide which data are used to form a prediction for a single loop detector with regard to proximity. All loop detector data within $d$ ft downstream and upstream of detector $i$ will be used in the predictions for $Q_i(t + 1)$, $S_i(t + 1)$, and $O_i(t + 1)$. When $(t + 1) - t$ is the amount of time in seconds defining the short-term horizon and $L$ is the observed speed limit in miles per hour at detector $i$, $d$ can be defined as follows:

$$d = ((t + 1) - t) \times L \times \frac{5.280}{3.600}$$

The number of detectors within $d$ feet upstream of detector $i$ will be denoted $m_u$, and similarly $m_d$ will be the number of detectors within $d$ feet downstream of detector $i$. For the example in Figure 2, $m_u$ and $m_d$ both have a value of 2.

If the posted speed limit is constant for an entire section of roadway and loop detectors are uniformly spaced, $m_u$ and $m_d$ will be constant except at the beginning and the end of the roadway where the number of upstream and downstream detectors is respectively limited. Nonuniform spacing of loop detectors will result in varying values for $m_u$ and $m_d$.

The state vector, $x(t)$, can be defined to describe the traffic condition at detector $i$ at time $t$:

$$x(t) = [Q_{i-m_u}(t) \ldots Q_i(t) \ldots Q_{i+m_d}(t)]$$

$$[S_{i-m_u}(t) \ldots S_i(t) \ldots S_{i+m_d}(t)]$$

$$[O_{i-m_u}(t) \ldots O_i(t) \ldots O_{i+m_d}(t)]$$

This state vector is compared with historical data in a database to find a closest match that minimizes the following statistic:

$$MAPE = \sum_{j=m_u}^{m_d} \left( \frac{|Q_j(t) - Q_j(t-c)|}{|Q_j(t)|} + \frac{|S_j(t) - S_j(t-c)|}{|S_j(t)|} + \frac{|O_j(t) - O_j(t-c)|}{|O_j(t)|} \right)$$

where $c = 1, 2, 3 \ldots$ and data from the time index $(t - c)$ is present in the historical database.

Once $c$ is determined, predictions can be made for $\hat{Q}_i(t + 1)$, $\hat{S}_i(t + 1)$, and $\hat{O}_i(t + 1)$:

$$\hat{Q}_i(t + 1) = Q_i(t - c + 1)$$

$$\hat{S}_i(t + 1) = S_i(t - c + 1)$$

$$\hat{O}_i(t + 1) = O_i(t - c + 1)$$

This approach will be repeated for all detectors, 1 through $n$, to form a prediction for $V(t + 1)$, the traffic condition on the entire section of roadway.

**CASE STUDY**

Backward propagating shock waves on congested roadways are well-documented phenomena. They are a product of not only the design of the roadway but also the behavior and reactions of the people in control of their respective vehicles. It is in the best interest of traffic engineers to mitigate the effects that these shock waves have on travel time. To manage a shock wave, it becomes absolutely necessary for an ITS to be able to detect and predict its propagation. The case study in this paper aims to show that the improved performance of nonparametric regression with the composite method will allow for accurate predictions of a backward propagating shock wave.

A data set developed by the Next Generation Simulation (NGSIM) program will be used for this study. The data set consists of detailed vehicle trajectory data on a 2,100-ft, six-lane section of US Highway 101 in Los Angeles, California. The data set was developed from measurements taken on June 15, 2005, between 7:50 a.m. and 8:35 a.m. A map of the studied highway can be seen in Figure 3.

The conventional uses of nonparametric regression are not readily applicable to vehicle trajectory data; therefore single loop detector data needed to be simulated using the trajectory data. Standardized
definitions of flow, speed, and occupancy from Edie (9) were used in the calculations of the traffic conditions. Entirely because of the high quality of the trajectory data, it was determined that the simulated loop detectors could be spaced every 100 ft and measurements recorded every 10 s. Such spacing and timing would almost never be used with actual loop detectors, but it serves the purposes of this study. Figure 4 is a 15-min space–time trajectory plot from the data set looking at a single lane. Two backward-propagating shock waves are clearly distinguishable.

A backward propagating shock wave was identified in the final 10 min of the data set. The time interval that contained the first manifestation of the shock wave was used as the experimental query. An implementation of the composite method used data previous to the experimental query to find the nearest neighbor for each loop detector and build a composite prediction for the next immediate time interval. [To eliminate the effect that the magnitude of a measurement would have on a distance metric, the distance metric used was mean absolute percent error (MAPE).] An aggregate nearest neighbor method as well as an ARIMA implementation was tested with the same data for comparison purposes. The composite method showed a MAPE reduction of more than 33% over the traditional aggregate method. The composite method’s MAPE was nearly half of the forecast’s MAPE from ARIMA. Trials made with different experimental queries showed similar reductions in error.

All three prediction methods were used to forecast traffic conditions for $V(t+1)$ through $V(t+5)$. Figure 5 displays the observed errors for each method. Although the composite method offers significant improvement over the aggregate method for the first two time intervals, it appears to converge to the aggregate method (on average) after the fourth time interval. The composite method did select nearest neighbors in the data set that were part of historical shock waves. Figure 6 shows the similarity between the experimental shock wave and the historical shock waves from which the composite method selected nearest neighbors.

**COMPUTATIONAL TIME**

The composite method was performed on a consumer computer. The amount of time required to form a composite prediction never exceeded 5 s on trials performed in the case study. The minimal computational time required by the composite method shows that it is viable for an online implementation.

**DATABASE SIZE**

The same experiment was performed with varying amounts of historical data to demonstrate the effect that the size and quality of the database has on the traffic condition forecast. The following table shows the results from four database sizes and their respective MAPE on the short-term horizon:

<table>
<thead>
<tr>
<th>Database size</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>0.1605</td>
</tr>
<tr>
<td>90</td>
<td>0.2341</td>
</tr>
<tr>
<td>60</td>
<td>0.2962</td>
</tr>
<tr>
<td>30</td>
<td>0.4374</td>
</tr>
</tbody>
</table>

**FURTHER RESEARCH**

During the implementation of the composite method, various forms of the state vector, $x(t)$, were considered for the search algorithm. A modified state vector with weighted values for detector data with an inverse proportion to the distance from the highlighted detector could improve performance.

**IMPLEMENTATION STRATEGY**

A composite nearest neighbor nonparametric regression can be incorporated into an ITS and could require no additional hardware. The implementation would call for changes in the management of loop detector data.

**Implementation Environment**

Any infrastructure of roadway monitored by single loop detectors is a candidate for composite nonparametric regression. Double loop detectors would also qualify provided that recordings could be translated into similar traffic condition characteristics monitored by single loop detectors (i.e., flow, speed, and occupancy). Thus a monitoring system comprising a mix of single and double loop detectors could equally qualify to benefit from composite nonparametric regression.

A central computing hub with real-time access to the data of every pertinent detector would house the historical database and perform all computational tasks.
Data with the quality of the NGSIM data used in the case study are not required. The detailed microscopic trajectory data from NGSIM were used for demonstration purposes.

**Historical Database**

The quality of the traffic condition prediction will depend on the quality and depth of the historical database. A trade-off has to be made in regard to the size of the database. On the one hand, it is desirable to have a large database (a) to contain a diverse assortment of traffic condition scenarios that a roadway will typically experience and (b) to contain anomalies, which are desirable for the prediction method to perform well. On the other hand, a light-weighted database is preferable because of constraints on memory storage and computation time. As a good scheme to begin with, the authors recommend three continuous days of historical data to create the database. Three days is a fairly large sample time, and 3 days of data for typical single loop detectors would have a pessimistic memory storage estimate of 10 megabytes, which most consumer computers could easily analyze in a reasonable amount of time to make a prediction on the short-term horizon. After midnight of any given day, the previous day’s data could replace the oldest day in the historical database to create a rolling update scheme to ensure that the assortment of traffic condition scenarios stays current in the historical database. Another potentially beneficial historical database scheme is to maintain databases for the individual days of the week.

**Composite Algorithm**

Once the data for the current or most recent traffic condition have reached the central computing hub, individual state vectors will be constructed for each detector. The section on technical definition describes the composite algorithm in detail.

**Expectations**

As the case study has shown, a reduction in error and improved prediction traffic forecasts are expected benefits. The study also showed the ability of the composite method to tailor itself to an individual roadway to detect and predict scenarios typical of the individual roadway, such as shock waves.

**Precautions and Limitations**

The composite method is still a form of nonparametric regression and does not model traffic conditions. Anomalies never before experienced may confuse any regression scheme and ultimately result in
inaccurate predictions. However, clever management of the historical database could allow the composite method to "learn" from any new anomaly and never make the same mistake twice.

When a nonparametric regression search finds a nearest neighbor, there is always the potential that the nearest neighbor is an undesirable representation of the traffic state. A simple tolerance check could prevent such an occurrence, but then other means would be required to form a prediction.

CONCLUSIONS

Nonparametric regression continues to warrant further research given its proven performance in numerous research papers written by traffic engineers and scientists. As advanced traffic management and information systems come closer every day to being realized, it is imperative to develop the algorithms those centers will use and the theory on which those algorithms are based.

Knowledge of the past is important to anticipate the direction of the future. That is how nonparametric regression works. The composite approach to nonparametric regression has demonstrated effectiveness at increasing accuracy for predictions over the short-term horizon. The forecast is based on the quality of the historical database, but advances in hardware manufacturing are trivializing concerns of memory as a resource. The case study showed that the computational time for the composite method is reasonable for an online implementation.

Adaptability is possibly nonparametric regression’s greatest attribute. An implementation need not be restricted to a stretch of roadway. Loop detectors near intersections could include detector data from another roadway, provided the roadway falls within range during the time interval defined by the short term. This strategy would allow compatibility between nearly any roadway and composite nonparametric regression.

ACKNOWLEDGMENTS

The first author thanks the University of Massachusetts Amherst Transportation Center for its support and resources. Additional thanks are due to the NGSIM program for its work in traffic data set development.

REFERENCES


The Statewide Transportation Data and Information Systems Committee sponsored publication of this paper.