Traffic Flow Forecasting Based on Pattern Recognition to Overcome Memoryless Property

Taehyung Kim[1], Hyoungsoo Kim[2], Cheol Oh[3] and Bongsoo Son[4]

[1]*Advanced Transportation Technology Research Center, The Korea Transport Institute, Goyang-si, Gyeonggi-do 411-701, Korea*
[2]*Department of Civil and Environmental Engineering, University of Maryland, College Park, MD 20742, USA*
[3]*Department of Transportation Engineering, Hanyang University, Ansan-si, Gyeonggi-do 425-791, Korea*
[4]*Department of Urban Planning and Engineering, Yonsei University, Seodaemun-gu, Seoul 120-749, Korea*
[1]douloservant@gmail.com, [2]hsookim@umd.edu, [3]cheolo@hanyang.ac.kr, [4]sbs@yonsei.ac.kr

Abstract

A variety of methods and techniques have been developed to forecast traffic flow. Current nearest neighbor non-parametric traffic flow forecasting models treat the dynamic evolution of traffic flows at a given state as a memoryless process; the current state of traffic flow entirely determines the future state of traffic flow, with no dependence on the past sequences of traffic flow patterns that produced the current state. Since traffic flow is not completely random in nature, there should be some patterns in which the past traffic flow repeats itself. In this paper, we proposed a pattern recognition technique, which enables us to consider the past sequences of traffic flow patterns to predict the future state. It was found that the pattern recognition model is capable of predicting the future state of traffic flow reasonably well compared with the k*-nearest neighbor non-parametric regression model.*

1. Introduction

The capability to forecast traffic volume has been identified as a critical need for a proactive and dynamic traffic control system. Cheslow et al. [1] concluded in an early report on intelligent transportation systems (ITS) architecture that 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.

A variety of methods and techniques have been developed to forecast traffic flow and these have been continuously refined up to the present. Linear regression is perhaps the most well-known method but other techniques such as non-linear regression, time-series analysis, neural networks, and Kalman filtering are commonly used in forecasting traffic flow. Each method has strengths and weaknesses, and each might be said to be designed to handle a specific class of problems. However, during the modeling process, assumptions about the data are made, which may or may not be appropriate, thus affecting forecasting performance. For example, "parametric algorithms assume that the data to be modeled takes on a structure that can be described by a known mathematical expression with a few free parameters" [2] and "If the assumptions are flagrantly violated, any inferences derived from the regression are suspect" [3].

These types of conclusions are frequently used to motivate the use of non-parametric regression, which is a data-driven heuristic forecasting technique, for forecasting traffic flow or travel time using large traffic flow data sets. Non-parametric regression does not require any prior knowledge about the process being modeled, only sufficiently large quantities of data representing the underlying system. It relies on past data to describe the relationship between input and output states rather than a (possibly incorrect) model upon the data. Hence, it is useful in situations where a well-defined theory does not exist but large amounts of data are readily available.

2007 International Conference on Multimedia and Ubiquitous Engineering(MUE'07)
0-7695-2777-9/07 $20.00 © 2007 IEEE

Davis and Nihan [4] used a *k*-nearest neighbor (*k*-NN) formulation of non-parametric regression to estimate short-term freeway traffic flows. They focused on estimating the transitions from the uncongested traffic regime to the congested regime. An empirical study using actual freeway data was conducted to test the *k*-NN approach and to compare it to simple univariate linear time-series forecasts. The *k*-NN method performed comparably to, but not better than, the linear time-series forecasts. Smith *et al*. [5] also used a nearest neighbor non-parametric regression to develop a traffic flow forecasting model for two sites on Northern Virginia's Capital Beltway. They showed that the non-parametric regression model significantly outperformed other models such as historical average, time-series, and neural network. Subsequently, Smith *et al*. [6] and Smith and Oswald [7] used nearest-neighbor techniques to forecast traffic flow based on real-time traffic data. Clark [8] recently examined relationships between flow, occupancy, and speed in order to generate short-term predictions of traffic flow. He employed a *k*-nearest-neighbor regression and relied on high-quality loop detector data from England. You and Kim [9] also used this technique to forecast travel time using traffic flow data on highways in Korea.

It is noteworthy that previous nearest neighbor non-parametric traffic flow forecasting models treat the dynamic evolution of traffic flows at a given state as a memoryless process; i.e., the current state of traffic flow entirely determines the future state of traffic flow, with no dependence on the past sequences of traffic flow patterns that produced the current state (in existing nearest neighbor non-parametric models, the state includes only instantaneous conditions, not historic ones). In fact, for certain instantaneous traffic state, the most natural future states might differ, depending on how those states were reached. As a simple example, if a traffic condition is in transition state from uncongested traffic to congested traffic, the future state of traffic flow might be decreased compared with the current state. However, if it is clear that the state is moving to a more free flow condition from the congested traffic, the future state might be increased, whereas current simple state-based models (with a myopic definition of the state) would not distinguish these two situations.

Hence, the purpose of this study is to propose a methodology that considers the past sequences of traffic flow patterns to predict the future state, and that overcomes the memoryless property of previous nearest neighbor non-parametric regression models. This methodology, which adopts a pattern recognition technique to represent a series of traffic flow patterns, follows in the second section. The third section describes the comparison of performance between the proposed pattern recognition model and the nearest neighbor non-parametric regression with actual freeway time-series data. Finally, some conclusions and future studies are mentioned.

2. Model development

It has been shown that drivers tend to retain their personalities, in the sense that each driver tends to maintain his driving attributes, and in some instances, drivers return to their nominal attributes after being forced by a traffic disturbance to alter them temporarily [10]. Hence, we might assume that driving patterns such as acceleration and deceleration profiles are expected to remain the same for each driver. Furthermore, observations have shown that typical hourly variation patterns related to highway type and day of the week exist and these basic patterns repeat among the days [11].

We recognize that traffic flow is not completely random in nature. We hypothesize that there should be some patterns in which the past traffic flow repeats itself. We expect that the future state of traffic flow is affected by the past sequence of traffic flow patterns. Therefore, it may be possible to represent the traffic flow state that unfolds over a sequence of time and the existence of common patterns through a pattern recognition technique that correlates current time series states with historical data for making future predictions. A pattern recognition technique is based on the premise that current structures may be matched with old structures to generate a future prediction. This technique is one of the newer methods of forecasting in the area of traffic and there are also a number of other applications which are still being explored, in the areas of medical diagnosis; syntactic, textile, speech, and face recognition; signal processing; etc.

2.1. Pattern recognition algorithm

To develop a pattern recognition algorithm for traffic flow forecasting, consider a discrete time-series $q = \{q_1, q_2, ..., q_n\}$ (e.g., 5 min. traffic volumes) where n is the total number of points in the series. The first goal of the prediction algorithm is pattern matching to find the "nearest" value or group of near values, also called the nearest neighbors, of the current state q_n in the past data. Then, we predict q_{n+1} on the basis of those nearest values; e.g., if the size of the "neighborhood" were $k = 1$, and the nearest value

were q_j, then we would predict q_{n+1} on the basis of q_{j+1}. The definition of the current state of the time-series can be extended to include several consecutive values $\{q_{n-l}, q_{n-l+1}, \ldots, q_{n-1}, q_n\}$ where l is a pattern size such that $1 \le l \le n-1$. A segment in the series is defined as a difference vector $s = (s_1, s_2, \ldots, s_{n-1})$ where $s_i = q_{i+1} - q_i, 1 \le i \le n-1$. We map these differences onto trinary variable d_i, by encoded each of them as a 0, 1 or 2 where $d_i = 0$ if $q_i > q_{i+1}$, 1 if $q_i < q_{i+1}$ and 2 if $q_i = q_{i+1}$ to define the direction in any pattern. Hence, a pattern in the time-series can be represented as $p_d = (d_{n-l}, \ldots, d_{n-1})$, a vector of 0's, 1's, and 2's. Smaller size patterns may refer to readily identifiable shapes, whereas larger patterns may have more complex shapes, as shown in Figure 1. The size of the pattern used for matching has an important impact on minimizing the error and correctly predicting the direction of series change. Therefore, the size of pattern must be optimized to obtain the best results.

The success in predicting the future states directly depends on the pattern matching algorithm. The overall procedure is shown as the following algorithm. This algorithm is identical to a *k*-nearest neighbor search except that a longer pattern is matched.

- Step 1: Start with a minimal neighborhood size, *k*.
- Step 2: Start with a minimal pattern size, *l*.
- Step 3: Form the pattern of size *l* describing the current state, i.e.,
  $p_d = (d_{n-l}, \ldots, d_{n-1})$
- Step 4: Search the time-series $(d_1, \ldots, d_{n-l-1})$ to find the nearest match(es) for $p_d = (d_{n-l}, \ldots, d_{n-1})$. Each nearest match corresponds to an index j, for which the matching pattern is $p'_d = (d_{j-l}, \ldots, d_{j-1})$. The difference vector associated with this trinary-coded vector is $s'_d = (s_{j-l}, \ldots, s_{j-1})$, and the final difference associated with match number h is s_j^h.
- Step 5: Estimate the value y_{n+1} on the basis of the final differences for all of the nearest neighbors:
  $$q_{n+1} = q_n + s_m \text{ where } s_m = \sum_{h=1}^{k} \frac{s_j^h}{k}$$
- Step 6: Calculate the root mean squared error (RMSE) between the actual and predicted values, for these choices of neighborhood size and pattern size, for the entire estimation set.
- Step 7: Repeat step 3 to 6 for patterns of size l+1, l+2,…, l_{max}.
- Step 8: Repeat step 2 to 7 for neighborhood sizes of k+1, k+2,…, k_{max}.
- Step 9: Choose the optimal pattern recognition model which yields minimal RMSE by optimizing the neighborhood and the pattern sizes.

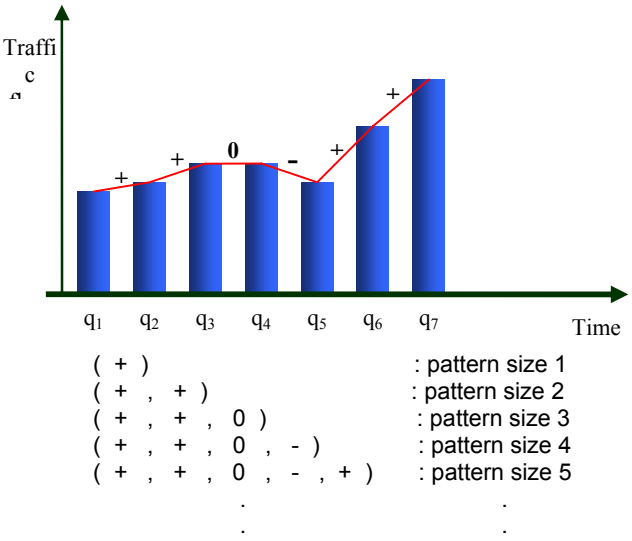

Figure 1. Example of pattern size and direction

3. Model performance

3.1. Traffic flow data set

Traffic flow data were collected at a site monitored by the traffic management system in the Center for Advanced Transportation Technology (CATT) at the University of Maryland. The site chosen within CATT is located on the southbound direction of I-95, near its interchange with MD-32. This is a four-lane directional highway segment, as shown in Figure 2. The Annual Average Daily Traffic (AADT) at the study site was reported as 193,550 vehicles in the year 2004. From this site, a database of 2 months of aggregate 5-minute traffic volumes was assembled from March 1 to April 30, 2004, resulting in 17,020 observations. The data set contains rare periods of missing observations, where data is not available for up to 5 minutes.

In order to use the pattern recognition model for forecasting the future state of traffic flow based on the past sequences of traffic patterns, the time series data was divided into two parts: an estimation data set and a test data set. We used the data from March 1 to April 29 as an estimation data set (16,753 observations) and the data of April 30 as a test data set (267 observations).

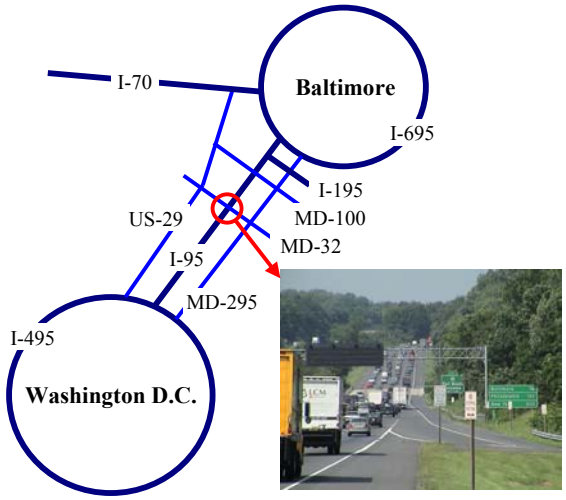

Figure 2. Spatial location of study site.

3.2. Model evaluation

We compared the performance of the proposed pattern recognition model with the well-known nearest neighbor non-parametric regression by investigating the RMSE between the actual and the predicted traffic flow (for an example of the nearest neighbor formulation of non-parametric regression, see [5]). We also measured the number of times the changes in the actual and predicted values go in the same direction, and we report the ratio of this count to the total number of predictions, which we call the *direction success percentage*. The computer program used to perform the pattern matching and nearest neighbor non-parametric regression was written in C language.

For our design of experiment, we used neighborhood sizes k from 1 to 10 for both the pattern recognition model and nearest neighbor non-parametric regression. The pattern recognition algorithm used pattern sizes l from 1 to 10. Figures 3 and 4 show the comparison of RMSE and the rate of direction success between the two models, as a function of the size of the neighborhood. We observe from Figure 3 and 4 that the RMSE of the proposed pattern recognition model, with the best-performing pattern length, is less than that of the nearest neighbor non-parametric regression for all neighborhood sizes. The rate of direction success of the pattern recognition model (about 62%) is much better than that of the nearest neighbor non-parametric regression (about 50%), which is quite encouraging. Notice that a value of 50% does not necessarily mean that the nearest neighbor method is purely an unbiased random walk, because a number of incorrect local fluctuations could be offset, in the predicted profile generated, by a single correct direction change of large magnitude. The increased performance of the pattern recognition model suggests that there are some repeatable patterns occurring within the data stream, and their recognition occasionally produces better predictions. Of course, the improvement is only from about 50% to about 62%, which means that random fluctuations within the data stream are quite noticeable as well.

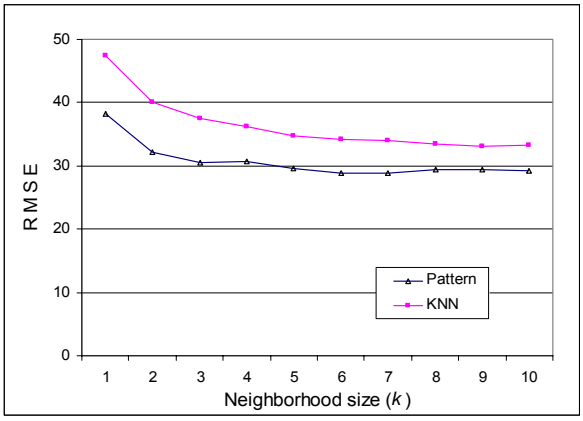

Figure 3. RMSE of pattern vs. k-NN

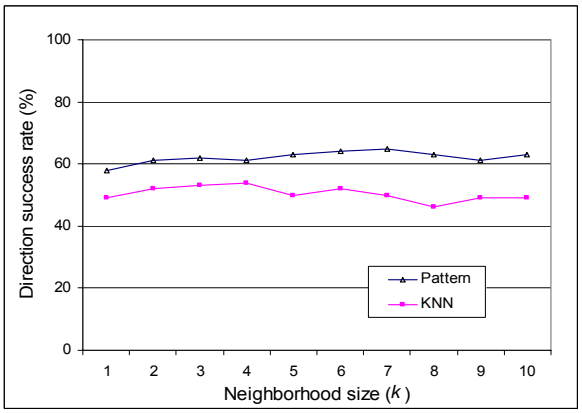

Figure 4. Direction success rate of pattern vs. k-NN

Figure 5 shows the RMSE performance of the different pattern sizes, for the different neighborhood sizes. Several things are clear from this picture. First, marked improvements can be seen in RMSE performance as the neighborhood size grows from 2 to 6, but the benefit seems to fall off at this point. Since larger neighborhoods require more computation, it makes sense only to increase while the benefits are clear. Second, the optimal pattern size is around 5 for most cases. There is a general improvement in RMSE performance as the pattern size increases up to 5, which means that a larger state size does indeed provide a better match for current vs. historical data. However, the improvement does not continue past pattern sizes of 5, suggesting that 25-minute patterns can be found in the data and are important for prediction, not much evidence exists of pattern durations greater than this.

Figure 6 shows the actual and the predicted traffic flow profiles for some time intervals of the test data set (April 30) for both methods, using the neighborhood size of 7. It should be observed that the proposed pattern recognition model is more successful in predicting the future state of traffic flow than the nearest neighbor nonparametric regression.

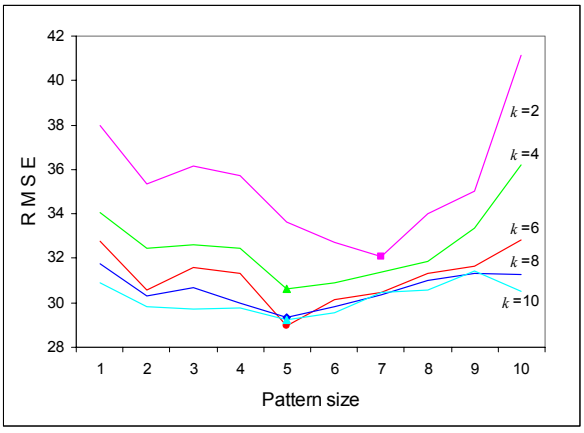

Figure 5. Optimal pattern size as a function of neighborhood size

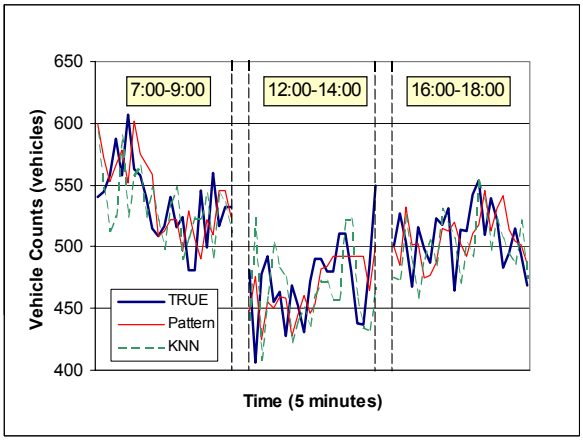

Figure 6. Traffic flow profiles of two models with actual data

4. Conclusions and future studies

Short term traffic flow forecasting has played a key role in proactive and dynamic traffic control systems. Hence, many attempts have been made to predict traffic flow along a roadway. Current *k*-nearest neighbor non-parametric regression models are "memoryless" in the sense that the current state entirely determines the future state without considering the past sequence of traffic flow patterns that produced that state.

In this paper, we have proposed a pattern recognition technique, which enables us to consider the past sequences of traffic flow patterns to predict the future state and have described that how a pattern recognition model can help to overcome this problem. It was found that the pattern recognition model is capable of predicting the future state of traffic flow better than the *k*-nearest neighbor non-parametric regression model with a smaller definition of state. The RMSE and the rate of direction success of the pattern recognition model between the actual and predicted traffic flow are superior to those of the *k*-nearest neighbor non-parametric regression model. The optimal pattern size suggests that long-term traffic "events" of duration 25 minutes or less seem to recur in the traffic stream and are important sources of information for prediction purposes.

Future studies should be pursued to further improve the performance of the proposed pattern recognition model. Particularly, the pattern recognition algorithm needs to be applied to a variety of traffic conditions. It could be also extended to forecast multiple intervals

into the future, which will allow for the development of more advanced and sophisticated traffic control strategies under intelligent transportation systems (ITS). While the current method establishes patterns based only on the signs of changes in traffic volumes, one might also hope to exploit the information contained in the magnitudes of these changes. There is a trade-off, however, since the added complexity in pattern specification makes the matching process more difficult. We hope that this paper is a good platform for the development of more effective nearest neighbor non-parametric regression models.

References

[1] Cheslow, M., Hatcher, S.G., Patel, V.M.: An Initial Evaluation of Alternative Intelligent Vehicle Highway Systems Architectures. MITRE Report 92w0000063, 1992

[2] Kennedy, R.L., Lee, Y., Roy, B.V., Reed, C.D., Lippmann, R.P.: Solving Data Mining Problems Through Pattern Recognition. Prentice Hall, New Jersey, 1998

[3] Mendenhall, W., Sincich, T.: A Second Course in Statistics: Regression Analysis. Prentice Hall, New Jersey, 1996

[4] Davis, G.A., Nihan, N.L.: Nonparametric Regression and Short-term Freeway Traffic Forecasting. ASCE Journal of Transportation Engineering, Vol. 117, No. 2, 1991, pp. 178-188

[5] Smith, B.L., Demetsky, M.J.: Traffic Flow Forecasting: Comparison of Modeling Approaches. ASCE Journal of Transportation Engineering, Vol. 123, No. 4, 1997, pp. 261-266

[6] Smith, B.L., Williams, B.M., Oswald, R.K.: Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting. Transportation Research C, Vol. 10, No. 4, 2002, pp. 303-321

[7] Smith, B.L., Oswald, R.K.: Meeting Real Time Traffic Flow Forecasting Requirements with Imprecise Computations. Computer-Aided Civil and Infrastructure Engineering, Vol. 18, No. 3, 2003, pp. 201-213

[8] Clark, S.: Traffic Prediction Using Multivariate Nonparametric Regression. ASCE Journal of Transportation Engineering, Vol. 129, No. 2, 2003, pp. 161-167

[9] You, J., Kim, T.J.: Development and Evaluation of a Hybrid Travel Time Forecasting Model. Transportation Research C, Vol. 8, No. 1, 2000, pp. 231-256

[10] Cassidy, M., Windover, J.: Driver Memory: Motorist Selection and Retention of Individualized Headways in Highway Traffic. Transportation Research A, Vol. 32, 1998, pp. 129-137

[11] Transportation Research Board: Highway Capacity Manual. National Research Council, Washington D.C., 2000

[12] Duda, R.O., Hart, P. E., Stork, D. G.: Pattern Classification. 2nd edn. John Wiley & Sons, New York, 2001

[13] Singh, S.: Pattern Modeling in Time-Series Forecasting. Cybernetics and Systems, Vol. 31, Iss. 1, 2000, pp. 49-66

[14] Singh, S., Stuart, E.: A Pattern Matching Tool for Time-Series Forecasting. presented at the 14th International Conference on Pattern Recognition, Vol. 1. Brisbane, Australia, 1998, pp. 103-105