IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 2, Ser. III (Mar. –Apr. 2022), PP 38-57
www.iosrjournals.org

The Potential Use of Large-Scale Data for Multiple Time-Step Forecasting for Motorway Traffic Flow in Advanced Data Management Systems

Hyun-ju Choi[1], Jin-soo Lee[2]
[1](NETTREK CO., LTD, Republic of korea)
[2](Urban science institute, College of urban science/ Incheon national university, Republic of korea)

Abstract:
The dynamic short-term forecasting of traffic variables such as the traffic flow is one of the essential factors in Intelligent Transportation Systems (ITS). Many novel forecasting models have been proposed in the literature and applied in practice. Despite the remarkable accomplishment of forecast modeling in ITS, many challenges remain, such as simplifying the processes, conducting feasibility studies, and extending forecasting horizons with acceptable levels of prediction error from the perspective of field engineers. Regarding contemporary ITS, one of the most crucial elements of the successful fulfillment of proactive ITS strategies is to estimate multiple time-period traffic demand levels with as much accuracy as a single-point forecasting approach. Recently, the wide-ranging introduction of the Advanced Data Management System (ADMS) provides data-driven Non-Parametric Regression (NPR) again, offering considerable practical opportunities in the area of traffic forecast modeling. In this vein, a multiple time-period forecasting model, based on k-Nearest Neighbor Non-Parametric Regression (KNN-NPR) is presented in this paper to address the aforementioned obstacles. The model is tested with large quantities of 5-minute freeway traffic data, and the forecasting horizon of the model is extended 12 steps ahead into the future. The results show for one-step-ahead cases that the KNN-NPR model is clearly superior to two other models which are compared models here, i.e. Kalman filtering and ARIMA, in terms of forecasting accuracy. Moreover, for multiple-steps-ahead cases, the performance of the model is comparable, at the very least, to the one-step-ahead results of the benchmark models. In addition, it is shown that the forecasting capability of KNN-NPR should be re-examined, at the very least, under the condition of data accessibility such as that offered in the ADMS.

Key Word*: Big data; Advanced data management system; Motorway traffic volume; Multiple time-period forecasting; Non-parametric regression*

Date of Submission: 03-04-2022 Date of Acceptance: 16-04-2022

I. Introduction

Since the early 1980s, dynamic short-term predictions of traffic variables such as volume, speed and occupancy levels have been key research topics in Intelligent Transportation Systems (ITS). In order to estimate future traffic variables, various forecasting models ranging from simple to sophisticated have been proposed. These mainly use current time-series data. Despite these efforts, there is a consensus among traffic experts that additional benefits of ITS implementation, especially in the ITS sub-systems such as Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management systems (ATMS), can be realized with multi-interval predictions rather than through the use of future information estimated by a family of single-interval prediction models. To accomplish this, several sophisticated methodologies based on mathematical or algorithm approaches have been reported in an effort to discover the nature of current states of traffic variables instantly and, in turn, to estimate future multi-interval states of traffic variables using the revealed knowledge of the current states. From a traffic engineer's perspective, advanced multi-interval prediction models, although they generate useful estimations extending to several future time steps, are too complicated to formulate mathematically, and in many cases, it is not easy for many field staff members to (re-)calibrate the parameters and/or to modify the structure of a model that is installed and operated in an ITS system. System developers and operators, therefore, now deeply acknowledge the need for a user-friendly and simplified forecasting technique that can estimate future multi-interval traffic variables without requiring a full understanding of the complexity of advance models. Additionally, this is one of the challenges that should be resolved before the easy applicability of the developed model can be realized by forecasting modelers.

In recent years, the wide implementation of ITS has made vast quantities of historical data more available than ever before given its leading-edge technologies. Advanced Data Management Systems (ADMS),

which assemble analyze and manage vast amounts of historical and current data systematically and which supply other systems with ad-hoc information, have been rapidly and widely introduced. Under the current situation of easy accessibility to massive amounts of historical data, data-mining techniques such as nonparametric regression, expert systems, and case-based reasoning are promising approaches for estimating future multiple states more easily and simply than more sophisticated and complex models. Despite this bright opportunity, a few studies of multi-interval predictions based on a data-mining approach have been reported (Smith et al. 1996; Chang et al. 2010, 2011; Yoon and Chang 2014), although some investigations have been made into the possibilities of single-interval predictions (Davis and Nihan 1991; Smith et al. 2002; Smith and Oswald 2003; Clark 2003; Qi and Smith 2004; Turochy 2006).

Clearly, research on dynamic multi-interval predictions remains a crucial issue in the ITS area for those who seek to execute ITS strategies more proactively and successfully. Several outstanding studies (Okutani and Stephanedes 1984; Smith et al. 1996; Kirby et al. 1997; Lan and Miaou 1999; Innamaa 2000; Vlahogianni et al. 2005) have been done on the important issue of multi-interval traffic flow forecasting. Despite the fact that the models generate multi-interval predictions effectively, a full understanding of the complexity of the models does not come easily to field experts who do not have sufficient experience in the area of predictions. Additionally, it was indicated by Yoon and Chang (2014) that future research on multi-period forecasting for the motorway traffic flow should be done to a level of acceptable accuracy.

The main objective of this article is to rediscover the hidden potentialities of k-Nearest Neighbor Non-Parametric Regression (KNN-NPR) in multi-step-ahead forecasting of motorway traffic volumes under the conditions of large-scale historical data available. Excluding the intricacy of other methodologies, the model introduced here is designed with important factors of real applications, such as simplicity, user-friendliness, convenience, and expansibility in building algorithms and operating the built model. In order to demonstrate the potential in real and effective practical applications of ITS, the model is experimentally tested with real-world data. This research also focuses on the following two points of NPR forecasting which significantly contribute to both pattern selection and forecasting reliability: the optimal determination and characteristics of two critical parameters, which are the optimal number of nearest neighbors in the neighborhood and the optimal embedding size of the state vector for the m-th future time step in KNN-NPR; and various forecasting functions to estimate the future state more accurately. In this way, the potentialities of NPR in multi-step-ahead forecasting for motorway traffic is diagnosed and recovered under the condition of large-scale data available in the 21st century.

II. Literature review and problem statement

Literature review

Various fine models to predict traffic variables such as volume, speed, occupancy, and travel time have been proposed and widely applied in the literature. Models ranging from naïve to hybrid were developed to solve specific forecasting problems, and all have strengths and weaknesses. Studies on short-term prediction can be divided into single-interval and multi-interval approaches according to the prediction horizon (h) at the forecast point (t). Most of the models are mainly utilized to estimate Single-Interval (SI) predictions of the Time Interval TI (t+1) at TI (t); Multi-Interval (MI) models generate predictions of TI (t+h), h =1, 2, 3, … at TI (t). The SI prediction horizon can easily be extended to the MI prediction horizon by modifying the structure of the SI model, but the problem of prediction accuracy arises, because there is a concurrent increase in the uncertainties associated with future states when the length of the prediction horizon is extended (Chang et al. 2010). In other words, the prediction accuracy of most short-term prediction models based on the (linear or non-linear) directionality and variation of the current state dramatically decreases as the number of time steps ahead increases. Only a few studies, therefore, have attempted to solve the import issue pertaining to dynamic MI traffic flows. Additionally, there are numerous reviews of short-term forecasting efforts in transportation research (See Vlahogianni et al. 2004; Karlaftis et al. 2011; Vlahogianni et al. 2014).

The first aim of this study is to develop a MI-prediction methodology to generate traffic volumes. The studies cited in the present paper, therefore, concentrate on MI-predictions to estimate the traffic variables of travel time, speed, and traffic flow, thus sparing an iterative literature review of SI predictions by many articles. MI-prediction approaches fall into the following categories: linear regression (Lan and Miaou 1999; Sun et al. 2003, Kim et al. 2009), nonlinear time-series analysis (Okutani and Stephanedes 1984, Artificial Neural-Network (ANN) family model (Kirby et al. 1997; Park and Rilett 1998; Innamaa 2000; Ishak and Alecsandru 2004; Vlahogianni et al. 2005; Hamad et al. 2009), and KNN-NPR studies (Smith and Demetsky 1996; Sun et al. 2003; Chang et al. 2010, 2012; Yoon and Chang 2014).

Regression models are employed to forecast dependent variables, such as the traffic flow or speed, using a mathematical function. Lan and Miaou (1999) proposed a generalized linear model based on a Bayesian switching rule to predict traffic flows. Sun et al. (2003) proposed a Local Linear Regression (LRR) model to estimate multi-step traffic speeds and compared the performance of the model to those of nonparametric

approaches (KNN and Kernel methods) and historical profiles; they also indicated that LRR is the best from among the models and that nonparametric approaches are second best. However, the two critical parameters of the embedding dimensions, the number of lagged observations, of the state vector and the k-values, the optimal number of neighbors, of KNN for all future time steps were not optimized in their study. However, the traffic variables related to traffic flow systems are highly correlated (Chien et al. 2002), and it is not easy to solve very complicated nonlinear relationships using regression models (Chang et al. 2012; Yoon and Chang 2014). Additionally, Bayesian linear regression was applied to estimate the departure-time-based link travel time when it is longer than the length of a time interval (Kim et al. 2009).

The nonlinear time-series analysis approach to explain the dynamic behavior of current traffic conditions is based on mathematical modeling. Okutani and Stephanedes (1984) employed Kalman Filtering (KF) to forecast signalized traffic flow with smoothed traffic data, finding that the prediction error concurrently increases with more than one time step ahead. Due to the high level of complexity, in mathematical terms, of a model such as a sophisticated regression model, nonlinear time-series models have had few prediction applications (Smith and Demetsky 1995).

The Artificial Neural Network (ANN) model is a promising approach for solving nonlinear prediction problems efficiently. ANNs are also promising means of forecasting traffic states with multiple input/output schemes (Adeli 2001). Due to these advantages, numerous studies based on various ANNs, from the traditional Back Propagation (BP) algorithm to sophisticated hybrid ANNs with other advanced models have been proposed. Kirby et al. (1997) employed BP-based conventional ANNs to estimate the short-term traffic volume. Park and Rilett (1998) proposed a modular ANN which outperformed other methods (conventional ANNs and historical profile, real-time profile, and exponential smoothing methods) in terms of prediction accuracy. In ANN-based MI forecasting, a family of Multi-Layer Perceptron (MLP) ANNs with BP algorithms has shown promise. In research on ANNs, Innamaa (2000) employed MLP-ANN to estimate multiple time-period traffic flows, and Ishak and Alecsandru (2004) utilized MLP-ANN for MI speed predictions. ANNs sophisticatedly combined with other techniques have increasingly been proposed to solve learning-optimization problems more efficiently or to consider the nature of nonlinear-or-nonstationary time series as well. A hybrid MLP-ANN with a genetic algorithm was utilized for estimating (intensive) signalized traffic flows (Vlahogianni et al. 2005) in a MI-prediction scheme, and a combined MLP-ANN with Empirical Mode Decomposition (EMD) based on the Hibert-Huang transform (Huang et al. 1998) was used to predict MI link speeds (Hamad et al. 2009).

KNN-NPR given accessibility to vast quantities of historical data supported by ADMS is a viable candidate for solving MI-prediction problems more easily and efficiently than other approaches. Smith and Demetsky (1996) reported a MI prediction model based on NPR that was used to generate motorway traffic volumes. Chang et al. (2010) employed a KNN-NPR strategy to estimate MI path travel time for a bus transit surmounting multiple time lags, which are unavoidable during surveying current path travel time information Later, Yoon and Chang (2014) utilized KNN-NPR for urban signalized traffic volume forecasting and showed that the NPR approach can, at the very least, perform effectively and stably in terms of its forecasting accuracy and hit rate in spite of the MI prediction horizon and the intensive evolution of temporal traffic state. Additionally, Chang et al. (2012) used a NPR approach to estimate multivariate missing traffic variables in multiple time periods, showing that the NPR can outperform Seasonal Auto-Regressive Integrated Moving Average (SARIMA), one of most widely used parametric approaches, without distorting the macroscopic relationships between traffic variables and with an acceptable level of estimation error.

In terms of the model structure, the elements of MI forecasting in these studies, except for a few of the KNN-NPR approaches, are the length of the time interval = [1~15 min], the number of multiple time steps ahead (h) = [2~6], and the total prediction horizon = [3~30 min]. Therefore, the total horizon times in MI forecasting are at most 30 min or less. The prediction accuracy in many cases (steeply) decreases to an unacceptable level when there is an increment in the number of multiple time steps, as any extension of the length of the prediction horizon or increase in the number of multiple time steps ahead usually brings about a concurrent increment of future uncertainties, which detrimentally affects the stability of the estimated future state and thus degrades the prediction accuracy (Chang et al. 2010). On the other hand the total horizon times of MI prediction based on KNN-NPR with a 15-min interval length reach 60~240 minute, i.e., h = 4~16, with an acceptable degree of prediction error [5~10%] (Smith and Demetsky, 1996; Chang et al. 2010). This shows that KNN-NPR approaches are feasible for MI predictions, as the NPR approach assumes that the bulk of knowledge about a relationship without understanding the nature of the system being modeled lies in past information rather than in the artificial relationship discovered by a person-developing model (Eubank 1988).

Problem statements

Several studies have investigated MI predictions based on various methods, from traditional to refined in ITS. Despite these efforts, from the perspective of a traffic engineer, there are several practical obstacles associated with simple-and-wide applications of MI predictions. Due to the high level of complexity, in terms of

mathematical modeling, of parametric approaches such as LRR and KF, parametric methods have few traffic prediction applications (Smith 1995). Sophisticated ANN models such as a combination of ANN and advanced methods such as fuzzy-neural, genetic-neural, or wavelet-neural techniques are conceptually complex. In these cases, it is not easy to (re-)calibrate and (re-)determine the optimal structure and parameters of the ANN by field personnel when prevailing conditions change. Additionally, the models are likely to be misunderstood or misapplied by field personnel, especially if they do not possess the expertise to recalibrate the models or conduct production-basis studies within a limited budget and/or time (Smith and Oswald 2003). It appears that data-driven approaches such as nearest neighbor regression, expert systems, and case-based reasoning under the data-access conditions supported by data management systems are more user-friendly for many field experts, such as ITS system builders, as compared to complex (mathematical) approaches, because the experts, in many cases, are likely not to have sufficient knowledge of traffic flow behavior and complex mathematical modeling, instead having only field experience in the area of searching algorithm and the database structure of the system. Therefore, there is an ongoing need for a MI forecasting model that does not require a full understanding of the model and traffic flow behaviors and that can generate robust estimations while remaining convenient and highly applicable.

In terms of traffic flow behavior, a time-series traffic flow is a complex system which changes very dynamically. The characteristics of traffic flows are chaotic (Disbro and Frame 1989) and the patterns of traffic flows vary dynamically depending on the prevailing traffic conditions (Smith et al. 2002). Vlahogianni et al. (2006) showed the properties (nonstationarity, nonlinearity, deterministic structure, chaos, and transitional movements) of short-term time-series traffic flows. These characteristics of a traffic flow are closely related to both unknown parameters and the uncertainties of future states, which in turn affect the prediction accuracy to an unacceptable level in many cases, especially MI forecasting (Chang et al. 2010; Yoon and Chang 2014). To comprehend these issues related to uncertainty and unknown parameters, the complex information contained in the vast and various historical data must be obtained and analyzed continuously. Additionally, this analysis process to determine the necessary data source and the information that is required is a challenge. On the other hand, this represents a data-driven approach that may be able to solve this problem without a full understanding of the complex characteristics of the traffic flow system; the past cases most similar to the current state can be selected and then used to estimate future system behavior (Smith and Oswald 2003).

In the past, ITS real-time database systems only stored unused or even deleted past data periodically. For this reason, many short-term prediction approaches, i.e., artificial person-developing models, to capture system dynamics were developed under the condition of access to small amounts of data, such as current data, without any data management system to support the vast quantities of historical data, including key information on the future state. Therefore, few studies of MI forecasting based on data-driven approaches such as case-based reasoning or nonparametric regression have been conducted compared to short-term forecasting methods based on real-time data. Recently, ADMS was introduced widely due to its state-of-the-art information and searching technologies, offering good opportunities to access 'big' data, including historical and current data, to various users such as traffic experts and other (sub-)systems of the ITS. Additionally, this condition of real-time access to big data has presented a real promising and practical application of data-driven approaches such as NPR in ITS forecasting area.

III. Methodology

The methodology based on KNN-NPR to forecast multiple time-period traffic volumes presented in this paper is presented in the four subsections of Section 3. The theoretical background of NPR is briefly described in Section 3.1. The following three elements of KNN-NPR are described in Sections 3.2 to 3.4, respectively. These are (1) the state space and prediction horizon, (2) the distance metric, and (3) the forecasting function. Lastly, the KNN-NPR forecasting algorithm is discussed with its pseudo-code in Section 3.5.

Theoretical background

To overcome the challenges of artificial parametric modeling, NPR has been continuously developed over the last 30 years. The nearest neighbor is referred to as the k-nearest neighbor in NPR. The NPR approach presumes that most knowledge about complex relationships among variables is inherent in the bulk data rather than the synthetic information generated by a human-made model (Eubank 1988). To put it another way, this approach is a sort of tactical and practical approach based on a decision-making process using past similar experience which is included in past experiences i.e., vast quantities of historical data without an understanding of the nature of the target system. NPR has a strong theoretical background. Estimations in NPR are generated by independent variables, as potential neighbors (n)$\rightarrow\infty$, nearest neighbors (k)$\rightarrow\infty$ with k/n$\rightarrow$0, after which the KNN method yields asymptotically minimum risk decisions (Devijver 1982). The method was extended to time-series data, showing that the nearest-neighbor estimation by the straight average converges to the minimum mean-square error forecast, while the convergence rate of nonparametric density estimations is also clearly

optimal among nonparametric estimators under a mixed condition (Yakowitz 1987). The nature of NPR theory implies that the KNN method for a state space of m size should produce results comparable, at least, to any m-th-order parametric method (Smith et al. 2002). It was found that, with MI forecasting and missing-data imputation for traffic variables, a nearest approach can at least outperform ARIMA, a widely used parametric time-series approach, in terms of estimation accuracy as a vast amount of high-quality data is available, i.e., n→∞ (Chang et al. 2012; Yoon and Chang 2014). This arises because the root of the NPR approach is in pattern recognition (Karlsson and Yakowitz 1987; Davis and Nihan 1991).

State space and prediction horizon

The system dynamic is consecutive, but the consecutive state is divided and aggregated by the length of the time interval in discrete dynamic systems. Therefore, the status of the system is time-series in nature. Most approaches, therefore, to solve time-series problems seek to define the state space as a series of values recorded during the past d time intervals. In our case, system values are traffic flow measurements. In other words, the state vector at the time interval (t) consists of each record with a measurement during each time interval [t, t-1, t-2, …, t-d], where d is the embedding size, the suitable number of lags, of the state space. In KNN, the system dynamic is mined by the attractor, i.e., a state vector. For a D-dimension attractor, the embedding dimension d is at least equal to or greater than 2D+1; i.e., $d \geq 2D + 1$ (Takens 1981). For example, a state vector with D=1 and an embedding size d at the time interval (t) for the traffic flow records measured every 5 minutes can be written as follows:

$$x(t) = [q(t), q(t-1), q(t-2), \dots, q(t-d)] \quad (1)$$

Here, $q(t)$ is the traffic flow during the current time interval (t), $q(t-1)$ is the traffic flow during the previous 5-minute time interval (t-1), and so on.

Once a state vector has been defined, a prediction problem can be formulated with a prediction horizon. The formulation in this study for the multi-interval forecasting problem with a one-dimensional state vector is defined as follows:

Given $x_m(T)$ with d_m
Predict $\hat{q}(T+m)$

Here, $x_m(T) = [q(T), q(T-1), \dots, q(T-d_m)]$ is the (current) state vector with d_m for the m-th future time step at the current prediction point (T); d_m is the suitable embedding size for $x_m(T)$; and $\hat{q}(T+m)$ is the estimated traffic volume during the future time interval $(T+m)$.

The independent and dependent variables are defined by parametric approaches. The input and output state vectors, in contrast, are defined by means of nonparametric regression. To search for potential neighbor nominees in a historical database and to record past future-state nominees onto the output space, both an input state vector for a potential neighbor nominee and an output vector for a future-state nominee related to the input state vector are elements in the process of the KNN-NPR algorithm.

Let us define the n-day historical dataset made up of n input-state-vector candidates, $x_m^j(t) = [q_j(t), q_j(t-1), \dots, q_j(t-d_m)]$ for the m-th future time step where j = 1, 2,…, n and t<T, which are connected as $x_m(T)$ at a current prediction point (T). In addition, the d_m value of the input state space at t is time-dependent on the d_m-size section of the time sequence of a day, which is related to the d_m value of current state space at T. Note that more past knowledge can be utilized with non-time dependency than with time dependency. This restriction has a major advantages as regards as the prediction accuracy and execution time of a family of data mining-based time-series approaches such as KNN-NPR in the case of multi-interval traffic forecasting (Chang et al. 2014): (1) the time dependency can beneficially effect the estimate of the directionality and variation of the future state and then the prediction accuracy, as traffic volumes recurrently and/or steeply vary on a weekly-daily-hourly basis; (2) it reduces the quantity of past data to about 1/[the number of time sequences (per day)], which in turn is closely related to the search time during the process of building the KNN, although a long search time is no longer a challenge in KNN-NPR due to leading-edge information technologies (Smith and Oswald 2003; Chang et al. 2010; Yoon and Chang 2014).

The output vector, o_m^j of $x_m^j(t)$ in this study consists of two elements and is defined as Eq. (2). The first element is the historical traffic volume $q_j(t+m)$ at time interval t+m. The second is the state distance u_m^j,

stated in Section 3.3, between $x_m(T)$ and $x_m^j(t)$. With the definition of these state vectors, the [input]→[output] structure for $x_m^j(t)$ and o_m^j with $x_m(T)$ is $[q_j(t), q_j(t-1), \dots, q_j(t-d_m)] \rightarrow [q_j(t+m), u_m^j]$.

$$o_m^j = [q_j(t+m), u_m^j] \tag{2}$$

Distance metric

KNN-NPR approaches to estimate future condition commonly use the past experience, i.e., cases similar to the current case, included in the bulk knowledge. In order to determine suitable past-state cases that are similar to a current-state case on the basis of "closeness" in NPR, a distance metric such as the L_M distance is commonly used to mathematically measure the state distance, i.e., the closeness, in the independent variable space, where the L_M distance is referred to as M={1,2, …, ∞} in the Manhattan, Euclidean, and max distance metrics. The L_M distance considers each value of a state vector equally. Note that a weighed distance metric of a higher dimension may be more feasible in an instinctive sense, whereas it is obviously heuristic in nature and requires careful consideration by the modeler (Smith et al. 2002). Additionally, it is not easy for field staff to (re-)calibrate the value of the weight when prevailing conditions change even slightly. In many cases, the field staff may not have the special knowledge necessary to calibrate the weight values without a full understanding of an operating model.

The Euclidean distance (ED, L_2) u_m^j for m-th future time step is used in this study to measure the nearness between $x_m(T)$ and $x_m^j(t)$, where j = 1, 2,…, n. It is defined and can be rewritten with $x_m(T)$ and $x_m^j(t)$ as Eqs. (3) and (4), respectively. Note that there are several approaches to estimate the similarity in the NPR approach. The ED is sensitive to noise, which may be a momentous signal for the future state, especially at a turning point of the state. The traffic flow state shows some fluctuation and varies rapidly in nature. In such a case, the ED is a promising technique as it can immediately capture the directionality of the current state, especially when the directionality steeply varies or is extensively disturbed (Yoon and Chang 2014).

$$u_m^j = \left[\sum |x_m(T) - x_m^j(t)|^2\right]^{1/2} \tag{3}$$

$$u_m^j = \begin{bmatrix} |q(T) - q_j(t)|^2 + |q(T+1) - q_j(t+1)|^2 \\ +, \dots, +|q(T+d_m) - q_j(t+d_m)|^2 \end{bmatrix}^{1/2} \tag{4}$$

Forecasting function

Before the description of the forecasting function, let us assume that the k-nearest-neighbor data set is built by the neighbor-searching-and-updating procedure of the forecasting algorithm presented in Section 3.5. The data set for a given $x_m(T)$ and k value at the forecasting point (T) consists of both the selected input state vectors $x_m^i(t)$ and the selected output vectors o_m^i corresponding to $x_m^i(t)$, respectively, where i = 1, 2,…, k and k/n→0. In order to build both the neighborhood consisting of $x_m^i(t)$ and the output composed of o_m^i for the future multiple time step (m) at the forecasting point (T), the following database structure is used in this study:

<Neighborhood>		<Output>	
[i]	[k-nearest neighbors, $x_m^i(t)$]		[k-output vectors, o_m^i]
1	$[q_1(t), q_1(t-1), \dots, q_1(t-d_m)]$	→	$[q_1(t+m), u_m^1]$
2	$[q_2(t), q_2(t-1), \dots, q_2(t-d_m)]$	→	$[q_2(t+m), u_m^2]$
…	…	…	
k	$[q_k(t), q_k(t-1), \dots, q_k(t-d_m)]$	→	$[q_k(t+m), u_m^k]$

Once the above data set has been built, the forecast is estimated with the various forecasting methods (FMs) in NPR. Note that the components of the nearest input state vectors are the independent variables and that the elements of the nearest output vectors provide the basis with which to estimate the dependent variables, i.e., the predictions, by a FM. The following seven methods defined as Eq. (5) to Eq. (10) in this study are used to generate the future state with the selected output vectors (and the selected neighbors). The first method is mathematical straight average that ignores all available information about the future state provided by the distance metric or the correlation of each selected neighbor to the current state. The others, in contrast, are efforts to improve on the performance of the straight average in terms of prediction accuracy by considering the

easily applicable information obtained during the process of KNN-NPR building, as follows (despite the fact that these are heuristic in nature): the nearness of each neighbor to the current state (weighting by the inverse of the state distance) as Eq. (6), the relationship between the key elements of each neighbor to the those of the current state (adjusted by the ratio of $q(T)$ to the $q_i(t)$ of each neighbor) as Eq. (7), the overall correlation between the elements of each neighbor to those of the current state(adjusted by the ratio of the average state of each neighbor to that of the current state) as Eq.(8), and a combination of two (or three) of these techniques as Eq. (9)-(11).

FM 1 expressed as Eq. (5) is the straight average of the selected output elements. This considers all of the dependent variables evenly and applies an equal weight to each selected output. FM 2 expressed as Eq. (6) is based on the notion that past states more similar to the current state have more prior information about the future state and therefore should have more of an impact on the determination of the future state. Instead of simple averaging, method 2 weighs the selected output elements by the ratio of the inverse of the corresponding ED to the sum of the inverse of the state-distance elements. FM 3 expressed as Eq. (7) assumes that the output elements, adjusted by the ratio of average of the elements of the current state to that of the elements of each selected neighbor prior to averaging, provide more inferred information about the future state, especially when the time horizon is extended in the scheme of multiple time-period forecasting. FM 4 expressed as Eq. (8) assumes that the output elements can provide more deduced information on the future state if they are modified by the ratio of $q(T)$ of the current state to $q_i(t)$ of each neighbor prior to averaging. FM 5 expressed as Eq. (9) combines methods 2 and 3, assuming that the prediction can be improved more by applying ED weighing instead of simple averaging to the adapted output elements using the ratio of the average of the elements of the current state to that of the elements of each selected neighbor in the case of multi-interval forecasting as compared to adjustments alone. FM 6 expressed as Eq. (10) integrates methods 3 and 4 into a straight average by averaging the ratios of the two prior to simple averaging. This method assumes that a composite modification of the output elements by averaging the ratio of the two will generate more accurate predictions than adjustments alone prior to straight averaging. FM 7 expressed as Eq. (11) combines methods 3 and 4 into method 2. This method applies the composite adjustment of methods 3 and 4 and then applies a weighing technique according to the inverse of the Euclidean distance. This assumes that the output elements adjusted by averaging the ratios of methods 3 and 4 will provide more inferred information about the future state by applying the ED as compared to that by straight averaging.

$$\hat{q}(T+m) = \sum_{i=1}^{k} q_i\,(t+m)/k \tag{5}$$

$$\hat{q}(T+m) = \sum_{i=1}^{k} \frac{q_i(t+m)}{u_m^i} \Big/ \sum_{i=1}^{k} \frac{1}{u_m^i} \quad , \quad u_m^i > 0 \tag{6}$$

$$\hat{q}(T+m) = \left[\sum_{i=1}^{k} q_i(t+m) \left\{\sum_{j=0}^{d_m} \frac{q(T-j)}{d_m+1} \Big/ \sum_{j=0}^{d_m} \frac{q_i(t-j)}{d_m+1}\right\}\right]/k \tag{7}$$

$$\hat{q}(T+m) = \sum_{i=1}^{k} q_i\,(t+m)\{\mathrm{q(T)}/\mathrm{q_i\,(t)}\}/k \tag{8}$$

$$\hat{q}(T+m) = \left[\sum_{i=1}^{k} \left\{q_i(t+m) \left(\sum_{j=0}^{d_m} \frac{q(T-j)}{d_m+1} \Big/ \sum_{j=0}^{d_m} \frac{q_i(t-j)}{d_m+1}\right)\right\}/u_m^i\right] \Big/ \left[\sum_{i=1}^{k} 1/u_m^i\right] \tag{9}$$

$$\hat{q}(T+m) = \left[\sum_{i=1}^{k} q_i(t+m) \left\{\left(\sum_{j=0}^{d_m} \frac{q(T-j)}{d_m+1} \Big/ \sum_{j=0}^{d_m} \frac{q_i(t-j)}{d_m+1}\right) + (q(T)/q_i\,(t))\right\}/2\right]/k \tag{10}$$

$$\hat{q}(T+m) = \left[\sum_{i=1}^{k} q_i(t+m) \left\{\left(\sum_{j=0}^{d_m} \frac{q(T-j)}{d_m+1} \Big/ \sum_{j=0}^{d_m} \frac{q_i(t-j)}{d_m+1}\right) + (q(T)/q_i\,(t))\right\}/2u_m^i\right] \Big/ \left[\sum_{i=1}^{k} 1/u_m^i\right] \tag{11}$$

KNN-forecasting algorithm

The k-nearest neighbor classification algorithm finds a group of k objects in the (training) data set. The three key components (the state vectors, the distance metric, and the forecasting functions) of KNN-NPR approach described earlier are integrated into the KNN-NPR multiple-time-period forecasting algorithm presented in this study. The forecasting algorithm searches for and attracts the neighbor-and-output candidates

from the historical data and updates these in the neighborhood and output set using determinant, i.e., the Euclidean distance, through an iterative process. It then generates the future states of multiple time periods using the FMs. The forecasting algorithm consists of three steps: (1) initialization, (2) building the neighborhood and output data set, and (3) generating the forecast. The pseudo-code for the KNN-NPR multiple-time-period forecasting algorithm is as follows:

Given the multiple time step (m), the state vector $x_m(T)$, and k value at TI(T):

1) Initialize the list of the neighbors and the elements of the outputs for all future m steps ahead to include cases 1, 2,…, k of the aforementioned database.

2) For each neighbor candidate $x_m^j(t)$ and output candidate $q_j(t+m)$, j = 1, 2,…, n.

   2-1) Calculate u_m^j between $x_m(T)$ and $x_m^j(t)$ by Eq. (4)

   2-2) If $u_m^j < u_m^{max}$ then
   (where $u_m^{max} = \max [u_m^1, u_m^2, \dots, u_m^k]$)

      2-2-1) Withdraw $x_m^i(t)$, $q_i(t+m)$ and u_m^{max} from the database
      (where $x_m^i(t)$ and $q_i(t+m)$ are associated to u_m^{max}, $1 \le i \le k$)

      2-2-2) Update $x_m^j(t)$, $q_j(t+m)$ and u_m^j onto the database

      2-2-3) Find new u_m^{max} in the updated database

3) Estimate $\hat{q}(T+m)$ by Eqs. (5) - (11)

IV. Application and findings

Study design

The KNN-NPR forecasting methodology described in Section 3 is tested with real freeway data collected by a loop detector and is compared to benchmark techniques. It is analyzed in an effort to demonstrate the efficiency of the method. In Section 4.1, the test traffic flow data and the characteristics of the traffic data are briefly explained. The performance measures are then defined to analyze and determine the values of the key parameters, d_m and k, of the developed model, and to evaluate the efficiency of the method through a comparative study with the benchmark models. Finally, the benchmark models are outlined. In Section 4.2, the optimal values of the two key parameters of the presented model for each multiple time step are analyzed and determined with the key performance measure and the features of the two parameters on the basis of the analysis results in the scheme of multi-interval forecasting are then discussed. At the beginning of Section 4.3, the seven FMs defined earlier are analyzed with the results of the rank test and the performance measures, after which the best approach is selected for a deeper analysis. Finally, the presented methodology with the selected FM is compared with the benchmark models to evaluate the quality of the predictions, and some of the findings from the analysis results of the comparative study are discussed.

As stated earlier, the basic approach of KNN-NPR has its origin in pattern recognition, and the quality and quantity of available data critically contribute to the effectiveness of KNN-NPR. A vast quantity of historical traffic flow data, therefore, was collected for the experimental study. The traffic flow data is measured by a paired-loop detector and managed by OASIS (The Center for Operations Analysis and Supportive Information), a nation-wide ADMS at the Korea Express Corporation. The test bed is located on Expressway #50, one of six main lines, as shown in Fig. 1. The data consists of traffic flow measurements in 5 min intervals during a 22-week period between July and November of 2010. In order to satisfy the data quality specifications, the target detector, rebuilt before the period of the data collection, was carefully selected. The size of the historical data was 44,352 measurements ([288*7*22] = 288 5-min sequences per day for 7 days a week during the 22 weeks). The target day for the experimental study was the last Friday of the last week. Additionally, the data was not adjusted by a technique, for example smoothing, so as to retain the dynamic variation of the time-series traffic flow, despite the fact that the performance of the forecasting model may decrease slightly in such a case in terms of prediction accuracy.

Figure 1: Test site

Before determining the performance measure, it makes sense briefly to analyze the features of the target system state, as shown in Fig. 5. The traffic state is separated into two levels, low and high, and the period of transition from one to the other is short. The traffic flow states at the high level during the morning peak period change more 7 times compared to those at the low level during an off-peak period, and several local peak periods after the morning peak exist which do not exhibit the typical two peaks, morning and afternoon. In addition, the observations at the high level show coarse conditions, including sharp noise.

These features of the dynamically mixed traffic flow state with a wide degree of variation are closely related to the uncertainties of future states in the forecasting problem. The Mean of the Absolute Difference, AD=$|q(t+1)-q(t)|$, and the Absolute Percentage Difference, APD = $[|q(t+1)-q(t)|/q(t)*100]$ of the target traffic flow are 14.63 and 7.10, respectively, and the Mean and Standard Difference of the Relative Percentage Difference, RPD = $[(q(t+1)-q(t))/q(t)*100]$, are respectively -0.60 and 9.53, as shown in Fig. 2. These statistical results indicate that the traffic flow state varies by about 7.6% from the absolute average and that the percentage of the variation ranges from -30% to +40%. In addition, it is almost inevitable that a prediction failure occurs in the case of single-interval forecasting, when the analysis results of the performance of the forecasting approach do not match the above results. Therefore, it is clear that there are wide variations in time-series traffic flows and that the time-series traffic flow is a very dynamic complex system in nature.

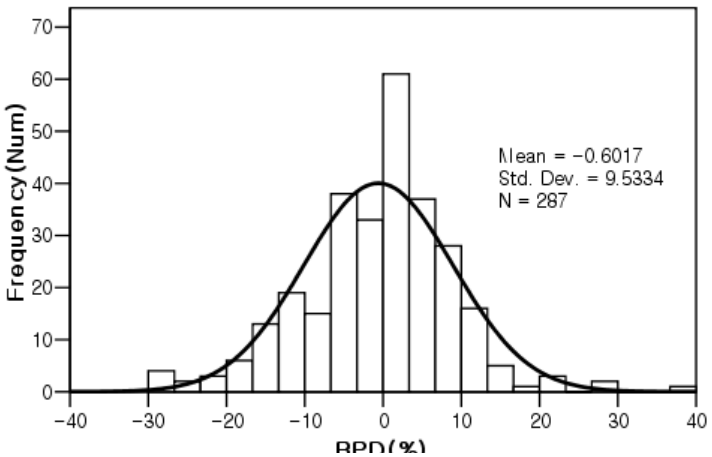


Figure 2: Distribution of the RPD

Through the brief analysis above, the following performance measures were selected. Mean Absolute Percentage Error (MAPE, %), the average of $[|\hat{q}_i - q_i|/q_i*100]$ (where q_i is the actual traffic flow of sample i and $\hat{q}_i$ is the forecasted traffic flow of sample i) provides the most useful basis for comparison when a state system exhibits wide variations (Smith et al. 2003; Yoon and Chang 2014). In addition, Mean Absolute Error (MAE), the average of $|\hat{q}_i - q_i|$, was used with MAPE. MAPE was also employed as a performance measure when finding the optimal value of the key parameters d_m and k of the presented model. Moreover, the Mean and Standard Deviation of the Relative Percent Errors [RPE = $(\hat{q}_i - q_i)/q_i*100$], MRPE and SDRPE, was used for a more detailed analysis.

As mentioned in the end of Section 2.2, this study makes an effort to improve the performance of NPR. The presented KNN-NPR model must outperform, as refined models do, historical average approaches while also showing a level of performance comparable, at the very least, to that of parametric approaches. In the context of the above two cases, to evaluate the performance of the presented methodology in this article, a comparative study was conducted with three traditional models: a simple naïve model was used as a worst-case approach, whereas the best case relied on the two well-known approaches of the seasonal ARIMA (p, d, q) (P, D, Q) and the Kalman filter, both of which are widely applied in the forecasting area of ITS.

Naïve models usually employ the historical average, i.e., the historical pattern, of the variable. The historical average is modified by the ratio of the current value to the historical average corresponding to the current value. The naïve model used in this study is defined as Eq. (12). Average historical traffic flows were calculated for each time-of-day and same-day-of-the-week points during the previous 8 weeks starting from the target day. Note that the 8-week historical average traffic flow rates minimized the forecasting error by the historical straight average model.

$$\hat{q}(T+1) = q_{hm}(t+1) \times q(T)/q_{hm}(t) \quad (12)$$

Here, $\hat{q}(T+1)$ is the traffic volume during the time interval (T+1) at forecasting point (T), $q(T)$ is the traffic volume during the current time interval (T), and $q_{hm}(t)$ and $q_{hm}(t+1)$ are the historical average traffic volumes during time interval (t) and $(t+1)$ corresponding to (T) and (T+1), respectively.

Before the brief statement pertaining to the ARIMA model used in this study, let us skip the Kalman filter [See Kalman 1960; Kalman and Bucy 1961]. ARIMA(1,0,1)(0,1,1)s was reported as the best selection from among ARIMA models in traffic estimation problems (Smith et al. 2002; Williams and Hoel 2003). This ARIMA form was also successively applied for traffic volume forecasting (Smith et al. 2002; Williams and Hoel 2003) and the imputation of the traffic variables (Chang et al. 2012), showing results comparable to those of NPR approaches. A seasonal (1,0,1)(0,1,1)2016, therefore, was used to forecast the 2,016 consecutive traffic volumes (12 time intervals per hour, 24 hours per day during the 7 days of the week).

Analysis and determination of key parameters

The performance of a KNN-NPR-based forecasting model is closely related to two key parameters: the embedding size, i.e. the number of lags, of the state space and the number of nearest neighbors of the neighborhood. The two parameters strongly contribute to the selection of the similar patterns, which in turn mostly determines the information about the future state, as NPR has its genesis in pattern selection. Therefore, it is crucial to analyze and determine the best or optimal values of the two parameters at the same time, instead of one of the two.

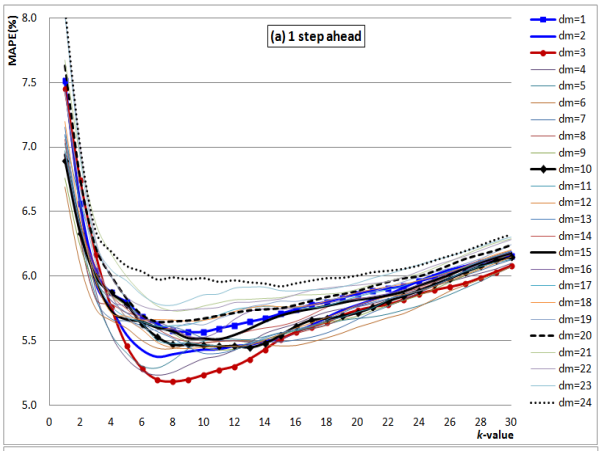
Given a one-dimensional (D=1) state space, the embedding size (d) is equal to or greater than 3 according to Taken's definition of $d \geq 2D+1$. This presents the possibility of reconstructing the state space of any dynamical system with $d < 2D+1$ (Packard et al., 1980). Thus far, there is no universal technique for the (re-) construction of the state space. In many cases, therefore, an analyst uses experimental-case-based approaches to find the applicable d value. On the other hand, the condition n→∞, k→∞ with k/n→0 of NPR is limited in the real world because the historical information available is finite. The experimentalist, hence, should determine the suitable k value for the test data. In this case, a useful approach to find the best or optimal d and k values is the enumeration method, which enumerates possible combinations of parameter values, i.e., d_m and k in this study, after which the best or second-best value is analyzed and predetermined with the performance measure. In order to find the optimal d_m and k values for each time step ($m \leq 12$), an experimental test in this research was conducted for the combined cases of $d_m \leq 25$ and $k \leq 30$ in increments of 1 for all multiple time steps, respectively. With the test data, the KNN-NPR multiple-time-period forecasting algorithm in Section 3.5 was applied with 2,592,000 cases (d_m values [1~25] * k values [1~30] * time step ahead [1~12] * 288 intervals per day = [25*30*12*288]), and the optimal d_m and k values for each time step ahead were then simultaneously analyzed and respectively identified using MAPE.

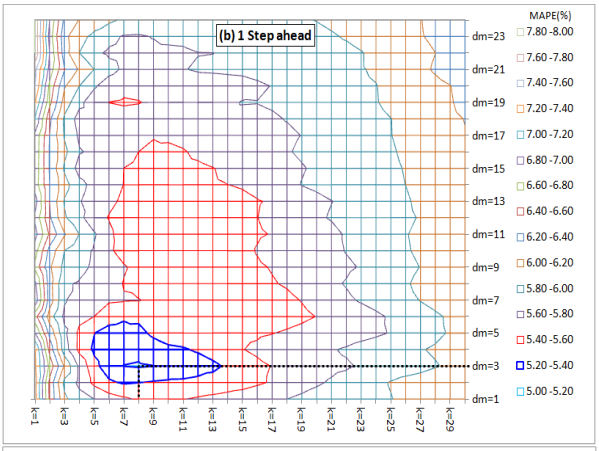
The effects of the d_m and k values on the prediction accuracy for the cases of future time steps [1, 2, 3, 7, 12] with Forecasting Method (FM) 1 are shown in Fig.3. For one step ahead, as shown Fig. 3(a), the MAPE for each d_m value steeply decreases to the minimum errors and then progressively increases with little variation when the k value increases. It does this as well for each k value when the d_m value increases. Therefore, the forecasting difference explained with the two parameters is geometrically concave. This fact clearly indicates the following: (1) in the case of D=1, Taken's definition, i.e., $d \geq 2D+1$, is valid, the best or optimal d value exists, and the definition may not valid when d→∞, (2) the temporal development of signalized traffic volume state is closer to chaotic, at least in this study, than it is to stochastic, and (3) the best or optimal d value is effective with the best or optimal k value and vice versa. It should be carefully noted, therefore, that the two

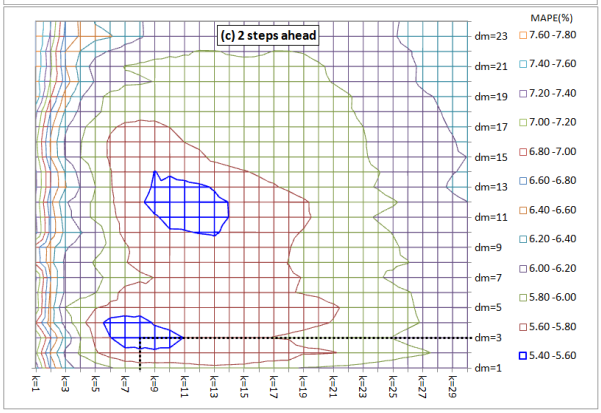
parameters should be simultaneously analyzed with a suitable performance measure; otherwise, the KNN-NPR approaches may fail to generate desirable results, as some studies (using NPR as one of the comparative approaches) inadvertently do. Additionally, the quantities of historical data must be large enough to meet the condition $k/n \approx 0$.

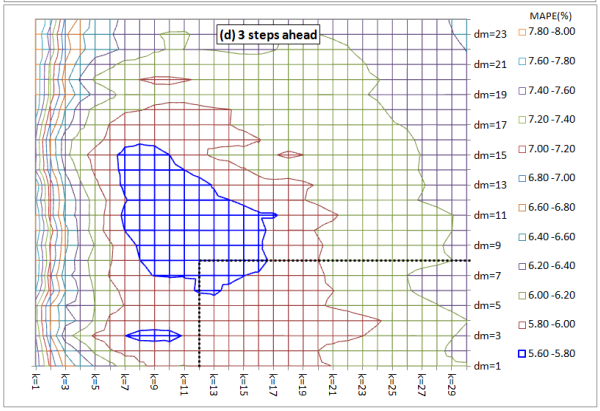
Figs. 3(b) - (f) show the evolution of the Optimal Error Space (OES), in which the forecasting error is less than the minimum error +0.2% when utilizing the best d_m and k value. With a time step ≤ 3, OES expands and the center of the OES moves toward the increment of the d_m and k value, then becoming virtually stationary with little variation, despite the fact that time step ahead increases. These characteristics of OES indicate (1) the x and y coordinates of OES can be used for the optimal d_m and k value for each time step ahead with consideration of the acceptable error instead of the best values, and (2) the optimal d_m and k values for all multiple time steps can be, in advance, easily analyzed and updated on a weekly or monthly basis with both the KNN-NPR multiple-time-period forecasting algorithm presented in this article and test data. Additionally, once the best or second-best d_m value for each time step ahead has been determined, an iterative analysis of it is no longer necessary.

The optimal d_m and k values of all multiple time periods determined by the analysis above are shown as Table 3. The optimal d_m values were divided between time steps (T+2) and (T+3), with the exception of FM4, after which they did not increase when the future time steps ($T \geq 3$) increased. FM4 showed similar results on the boundary line between the time steps (T+3) and (T+4). These results indicate that the uncertainties of the future state are not reduced even if the number of lags is scaled up. Thus, the state of the prevailing traffic flow rate during the previous 30~40 minutes is sufficiently optimal, in the case of a 5-min time interval length, to image the future state when the time step $T \geq 3$. Regarding the optimal k values, the values shows similar analysis results to those of the d_m values, barring FM2. FM2 showed a gradual increment of the k value to time step (T+7), after which the k values were stationary. The analysis results of k value reveal that the increment of the nearest neighbors according to the increment of the future time step does not consecutively contribute to the diminution of the uncertainties of the future state, specifically given as x_m with d_m, k/n→α (α≈0) with a finite n. This is true because past experiences can be divided into (finite) sub-state patterns, on which NPR is deeply dependent in nature to identify a future state.









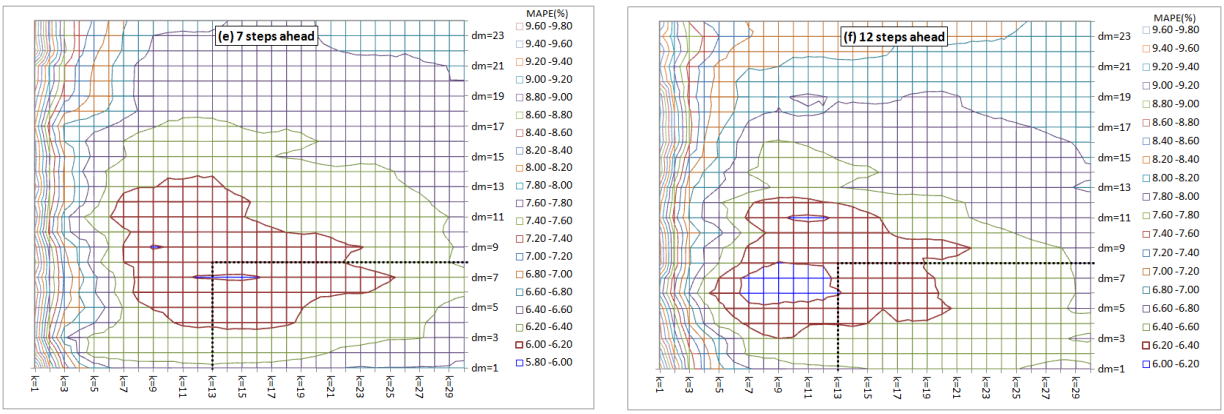


Figure 3: Effects of the d_m and k values on the forecasting error

The prediction error, according to the increment of future time step, increased log-likely with little or some variation, shown as Fig. 4. The error behaviors can be divided into three groups: the worst, the second-best, and best performer groups are FM [4], FM [1, 2, 6, 7], and FM [3, 5] respectively. Considering the prediction accuracy and stability, FM4 adjusted by $q(T)/q_i(t)$ was the worst case, showing a roller-coast trend in the MI scheme, although FM4 for one step ahead was more accurate than the two conventional models, i.e., FM1 and 2. This indicates that there is, therefore, careful consideration required when any type of FM "adjusted by $q(T)/q_i(t)$" is applied for MI forecasting. The conventional FMs did not improve the prediction accuracy, but they were more stable than the others, apart from FM3 and FM5. On the other hand, FM 3 and 5 were more stable with little variation and were more accurate than the others. Moreover, they have "adjusted by the rate of the average of $x_m(T)$ to that of $x_m^i(t)$" in common on the frame of FM1 (or FM2), excluding "adjusted by $q(T)/q_i(t)$." It was found that the nonlinear directionality and variance of the future state may be, at the least in our case, effectively adjusted by the ratio of the current state average to the neighboring state average, as the best or optimal d_m value and the degree of similarity contribute to the basis of the future state.

Table 1: Analysis results of the optimal d_m and k value

Forecasting method		Time step ahead												
		T+1	T+2	T+3	T+4	T+5	T+6	T+7	T+8	T+9	T+10	T+11	T+12	Ave.
d_m	1	3	3	8	8	8	8	8	8	8	8	8	8	7
	2	3	4	8	8	8	8	8	8	8	8	8	8	7
	3	3	3	8	8	8	8	8	8	8	8	8	8	7
	4	3	3	3	6	6	6	6	6	6	6	6	6	5
	5	3	3	8	8	8	8	8	8	8	8	8	8	7
	6	3	3	8	8	8	8	8	8	8	8	8	8	7
	7	3	3	8	8	8	8	8	8	8	8	8	8	7
k	1	8	8	12	13	13	13	13	13	13	13	13	13	12
	2	9	12	13	13	13	14	15	15	15	15	15	15	14
	3	11	12	14	15	15	15	15	15	15	15	15	15	14
	4	8	8	11	11	11	11	11	11	11	11	11	11	11
	5	12	12	14	15	15	15	15	15	15	15	15	15	14
	6	8	8	11	11	11	11	11	11	11	11	11	11	11
	7	9	9	14	14	14	14	14	14	14	14	14	14	13

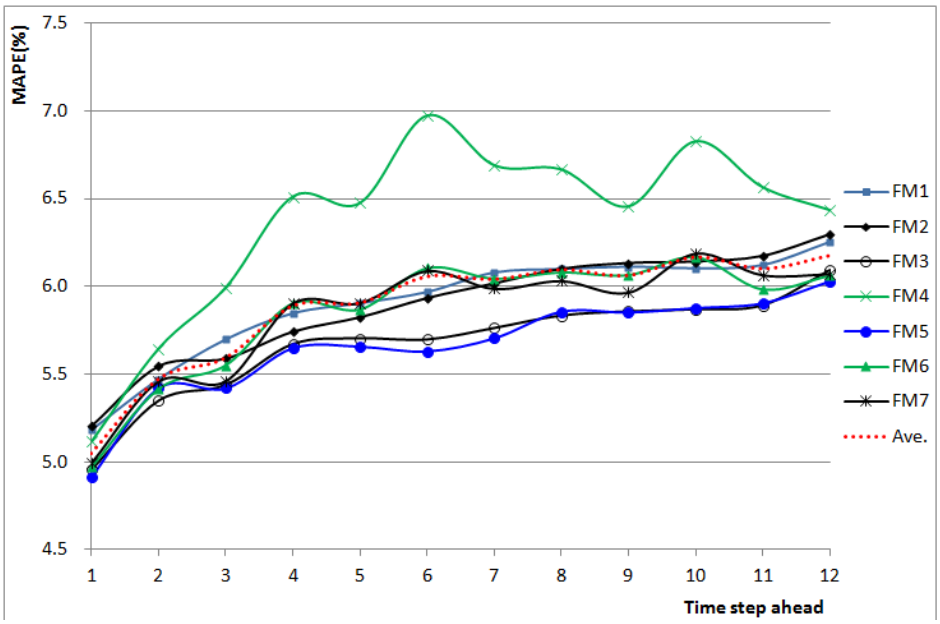


Figure 4: Error distribution of the forecasting methods

Analysis of results and findings

Before a detailed analysis of the results, let us define the analysis period of the target data as 5~24 hours for following reasons: (1) the light traffic volumes during 0-5 hours usually do not show a wider variation than the target time period, (2) the future traffic volumes generated by the forecasting models are mainly used during the target time period, and (3) a 24-hour analysis of the results, therefore, can bring about biased results unintentionally. For ranking the models in terms of forecasting accuracy, a Friedman and Wilcoxon signed-rank test with a significance level of α=0.05 was performed to compare the accuracy of the presented seven FMs and three benchmark models with the Absolute Percentage Errors (APEs). The test converts the lowest measure, i.e., APE, at each prediction point to rank 1, the second lowest to rank 2, and so on, and ranks the forecasting methods with mean rank scores. Note that the test results are also the same, using Absolute Errors (AEs), as the rank test operates on the ranks of the related values instead of the observed values.

The analysis results of the rank test and the performance measure for the single and multiple forecasting cases are shown in Tables 2 and 3, respectively. As shown in Table 2, the best performing method of the seven presented Forecasting Methods (FMs) in terms of APE converted to ranks were FM 5, "adjusted by the rate of the average of $x_m(T)$ to that of $x_m^i(t)$" on the frame of "weighed by the inverse of the distance" in the single- and multi-interval forecasting cases. The second-best methods were FM 3, 6, and 7, which have "adjusted by the rate of the average of $x_m(T)$ to that of $x_m^i(t)$" in common. The "straight" and "weighed by the inverse of the distance" models were comparable, showing no differentiation, and were not included in the upper group. The poorest performing method, excluding FM 1 and 2, was FM 4, which combine "adjusted by $q(T)/q_i(t)$" only on the frame of FM1; this result is different from those in a comparative study (Smith et al., 2002). In contrast, for two members (FM 6, 7) of second-best performing group, the attribute "adjusted by $q(T)/q_i(t)$" is integrated into FM 5 or 3, which are the best and second-best FMs in terms of the MAPE. Therefore, it is clear that "adjusted by the rate of the average of $x_m(T)$ to that of $x_m^i(t)$" (with "weighed by the inverse of the distance") contribute more to the future state than "adjusted by $q(T)/q_i(t)$." For the benchmark models, none of them were included in the upper or middle performing group. The ARIMA and KF models, however, outperformed the naïve model as the worst case; they are comparable to each other, although ARIMA outperformed KF somewhat in terms of the MAPE. These findings suggest, therefore, that the KNN-NPR with the seven presented FMs shows, at the very least with FM 5 and 3, more improved performance than the conventional parametric approaches of the ARIMA and KF models.

Table 2: Statistical results of the significance test

Single interval			Multiple intervals		
Method	Mean rank	MAPE	Method	Mean rank	MAPE
FM5	5.08	4.92	FM5	3.75	5.66

FM6	5.17	4.95	FM3	3.81	5.69
FM3	5.21	4.94	FM7	3.94	5.83
FM7	5.24	4.99	FM6	3.98	5.87
FM1	5.38	5.16	FM2	4.06	5.89
FM2	5.47	5.20	FM1	4.09	5.90
FM4	5.55	5.09	FM4	4.38	6.36
ARIMA	5.71	5.61	-	-	-
Kalman filter	5.74	5.75	-	-	-
Naïve	6.45	6.43	-	-	-

Table 3: Summary of the analysis results of performance measures

MOE	Method		Time step ahead											
			T+1	T+2	T+3	T+4	T+5	T+6	T+7	T+8	T+9	T+10	T+11	T+12	Ave.
MAPE (%)	Naïve		6.43	-											6.43
	ARIMA		5.61	-											5.61
	KF		5.75	-											5.75
	KNN NPR	FM1	5.16	5.49	5.67	5.89	5.89	6.00	6.12	6.10	6.10	6.08	6.11	6.21	5.90
		FM2	5.20	5.55	5.58	5.77	5.81	5.93	6.02	6.09	6.13	6.13	6.15	6.29	5.89
		FM3	4.94	5.46	5.40	5.69	5.74	5.71	5.79	5.83	5.87	5.88	5.90	6.11	5.69
		FM4	5.09	5.64	5.92	6.51	6.50	6.97	6.69	6.69	6.44	6.86	6.61	6.44	6.36
		FM5	4.92	5.42	5.39	5.66	5.68	5.63	5.72	5.85	5.86	5.88	5.91	6.04	5.66
		FM6	4.95	5.41	5.59	5.92	5.84	6.07	6.02	6.11	6.40	6.15	5.94	6.03	5.87
		FM7	4.99	5.43	5.45	5.87	5.90	6.06	5.95	6.04	5.97	6.19	6.06	6.08	5.83
MAE	Naïve		17.58	-											17.58
	ARIMA		14.91	-											14.91
	KF		15.32	-											15.32
	KNN NPR	FM1	13.94	14.77	15.59	16.14	16.17	16.53	16.78	16.75	16.70	16.60	16.68	16.99	16.14
		FM2	13.93	14.99	15.32	15.83	15.91	16.28	16.48	16.65	16.68	16.68	16.74	17.15	16.05
		FM3	13.33	14.72	14.80	15.61	15.80	15.82	16.06	16.11	16.09	16.12	16.15	16.69	15.61
		FM4	13.87	14.97	16.01	17.75	17.40	18.58	18.40	18.54	17.79	18.54	18.19	17.79	17.32
		FM5	13.27	14.63	14.77	15.52	15.63	15.58	15.83	16.10	16.04	16.09	16.12	16.49	15.51
		FM6	13.44	14.48	15.13	16.22	15.90	16.45	16.64	16.98	17.42	16.70	16.28	16.66	16.03
		FM7	13.49	14.53	14.78	16.10	16.04	16.51	16.47	16.82	16.36	16.80	16.54	16.74	15.93
MRPE (%)	Naïve		-0.22	-											-0.22
	ARIMA		0.19	-											0.19
	KF		0.64	-											0.64

	KNN NPR	FM1	-0.26	-0.23	0.25	0.37	0.46	0.51	0.64	0.80	0.89	0.90	0.95	0.88	0.51
		FM2	-0.38	-0.35	-0.02	0.06	0.14	0.27	0.44	0.65	0.75	0.87	0.90	0.89	0.35
		FM3	-0.14	-0.16	0.03	0.18	0.36	0.48	0.60	0.87	0.98	1.06	1.07	1.05	0.53
		FM4	-0.19	-0.16	-0.06	-0.53	-0.29	-0.12	-0.09	-0.15	-0.34	0.59	0.66	0.51	0.07
		FM5	-0.22	-0.27	-0.15	-0.01	0.15	0.28	0.40	0.66	0.80	0.94	0.99	0.99	0.38
		FM6	-0.20	-0.17	-0.26	-0.25	-0.03	0.13	0.24	0.38	0.80	0.62	0.73	0.79	0.23
		FM7	-0.31	-0.32	-0.32	-0.25	-0.07	0.01	0.15	0.39	0.50	0.65	0.70	0.65	0.15
SDRPE (%)	Naïve		8.56	-											8.56
	ARIMA		7.95	-											7.95
	KF		7.62	-											7.62
	KNN NPR	FM1	7.08	7.43	7.52	7.88	7.83	7.96	8.07	7.97	7.94	8.06	8.11	8.12	7.83
		FM2	7.13	7.56	7.57	7.87	7.87	7.99	8.02	8.05	8.10	8.15	8.25	8.32	7.91
		FM3	6.92	7.62	7.49	7.81	7.79	7.82	7.85	7.76	7.79	7.76	7.89	8.05	7.72
		FM4	6.93	8.00	7.99	8.59	8.92	9.70	9.27	8.81	9.05	9.25	8.89	8.49	8.68
		FM5	6.89	7.70	7.54	7.81	7.81	7.84	7.87	7.87	7.88	7.86	7.96	8.06	7.76
		FM6	6.75	7.66	7.59	8.02	7.99	8.29	8.34	8.14	8.58	8.24	8.06	7.91	7.97
		FM7	6.82	7.77	7.53	7.95	8.05	8.32	8.23	8.10	8.26	8.33	8.16	8.07	7.97

The analysis results of performance measures are summarized in Table 3. Note that the MAPE, MAE and SDRPE with MRPE results using the reactive approach during the target time period are 6.09, 16.81 and 8.21 with -0.07, respectively; these threshold values provide the criteria, at least, for the decision of forecasting failure in the single-interval forecasting case. For the naïve model, the forecasting failure occurred with performance measures of 6.43>6.09, 17.58>16.81 and 8.56>8.21 with -0.22, respectively, whereas the results of the other nine models satisfied the threshold values of the performance measures. The naïve model, therefore, was eliminated from the more detailed analysis.

Regarding single-interval forecasting, the analysis results [4.92~5.20, 13.27~13.94, 6.82~7.08 with -0.38~-0.14] of the seven FMs were clearly more accurate than those [5.61~5.75, 15.32~15.32, 7.62~7.5 with 0.19~0.64] of the benchmark models. The best performer among them was FM5, as in the results of the rank test, with regard to MAPE and MAE [4.92, 13.27], although it was third best with SDRPE [6.89 with -0.22]. As regards multi-interval forecasting, the average performance results, [5.66~5.90, 15.51~16.14, 7.76~7.97 with 0.07~0.53] of the FMs with the exception of FM4 were comparable, despite multiple time steps ahead, to those of the benchmark models applied to generate single-interval predictions. The best case of the multiple-forecasting scope was also FM5 with the average performance measure [5.66, 15.51 and 7.76 with 0.38 respectively]. Consequently, the best performer in the case of two prediction horizons was clearly FM5 with the analysis results of the rank test and performance measures. The more detailed analysis, therefore, included FM5.

A time-series comparison between the actual and predicted traffic flow rates was done with three cases, as shown in the dotted boxes in Fig. 5 (a)-(c). Despite the use of multiple time steps ahead, the predictions estimated by the method presented here concurred with the observations, showing a remarkable increase in the early morning, intensive fluctuations between 12 and 15 PM, oscillation with little variation from early to late at night, and then a steep decrease very late at night. In the case of the rapid and instable increase and decrease denoted with the dotted boxes (a), It is unavoidable, when using forecasting approaches that only rely on the current state, to fail to proactively capture the directionality of the future state at the turning point, as one subsequent time-lag after the turning point is at least necessary to reconstruct the current state anew. The proposed method, in contrast, instantaneously or in advance identified the directions and variances of future states for multiple time steps, despite the fact that these behaviors recur on a time-of-day and day-of-week basis with little variation. This is true because KNN-NPR is highly dependent on the recognition of past experience.

As for the wide variation and harsh noise shown in boxes (b) and (c), respectively in Fig. 5, the time-series analysis techniques nearly lose the ability to detect the directionality despite the use of single-interval forecasting in the presence of fluctuating or harsh noise, subsequently generating crossed oscillating estimations.

As a result, the forecasting failure occurs when the technique cannot detect the directionality of the state (Vlahogianni et al. 2005, Yoon and Chang 2014). On the other hand, the present model immediately or in advance reflects the trends of the time-series state with some acceptable differences from three steps ahead, after which it minimizes, more than three steps ahead, the differences either by capturing any fluctuation immediately or at least one or two steps later or by smoothing in the middle of the oscillations. Thus, the proposed model, not showing, at least, zigzag-like estimations, does show somewhat better performance although it partially loses the ability to capture the directions as compared to the time-series approaches. This directly indicates that traffic condition data is a mixed dynamical system that is closer to chaotic rather than stochastic and that the developed methodology is capable of strong performance without the requirement of a full understanding of the characteristics of traffic flows when seeking to recognize the intrinsic complex patterns of the bulk knowledge included in historical data.

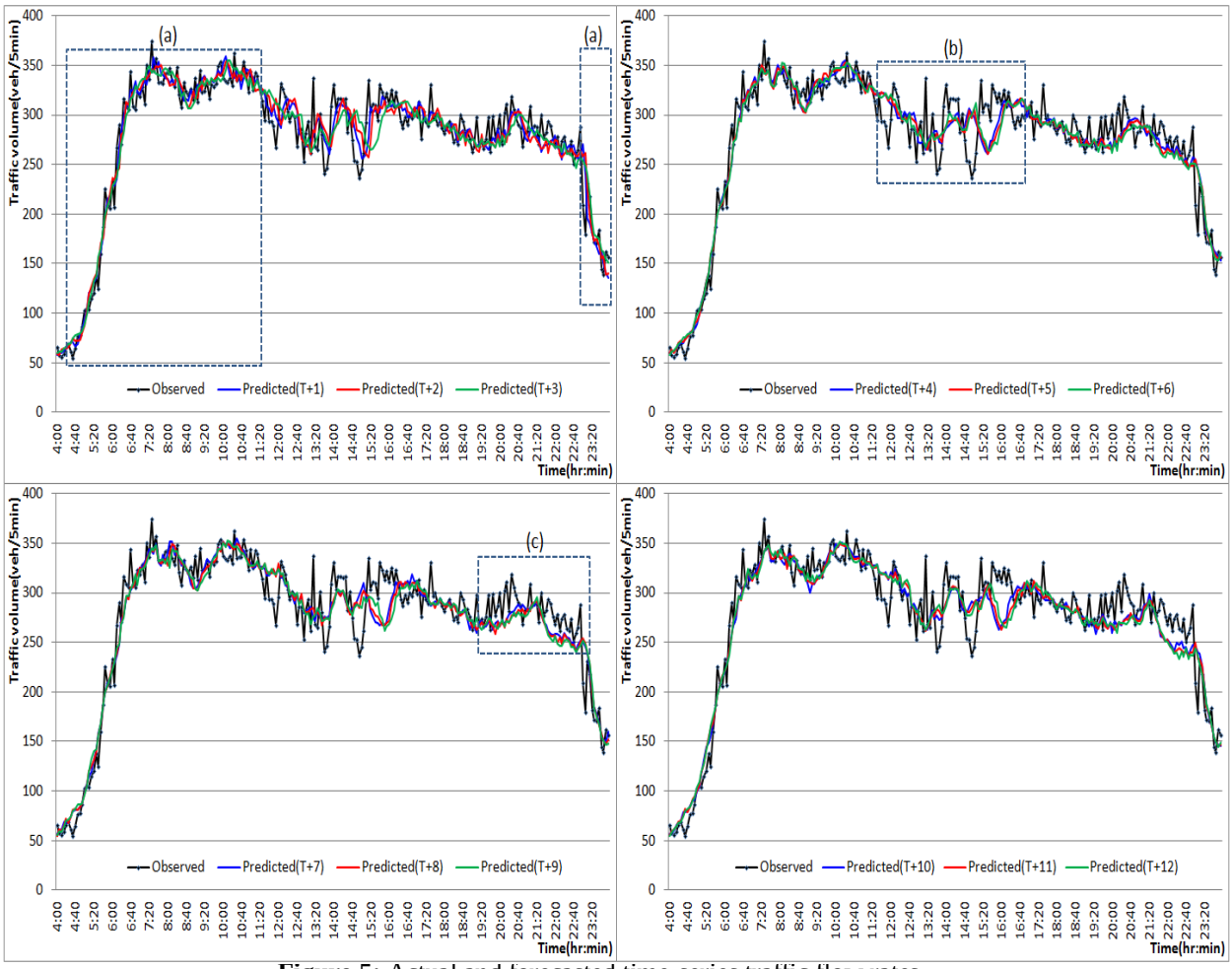


Figure 5: Actual and forecasted time-series traffic flow rates

Fig. 6 shows the predictions by ARIMA (T+1), KF (T+1), the developed model (T+1, and T+12) against the observations. The correlation coefficients, which directly indicate the level of accuracy, were 0.984, 0.984, 0.987, and 0.982, respectively. This shows that the one-step-ahead estimations predicted by the developed model are closer to the actual observations than those of the two comparative models. Moreover, the 12-step-ahead predictions do not great differences either. The statistical results, i.e., the t-value, from a paired t-test with a significance level of α=0.05 were 0.437, 1.723, 0.064, and 1.326, respectively, all of which are less than 1.96. Thus, the three models for one step ahead are significant at the 5% level. In particular, the developed model, despite the use of 12 steps ahead, passed the t-test. The results of statistical analysis therefore demonstrate that the developed method is a promising approach for MI traffic flow forecasting.

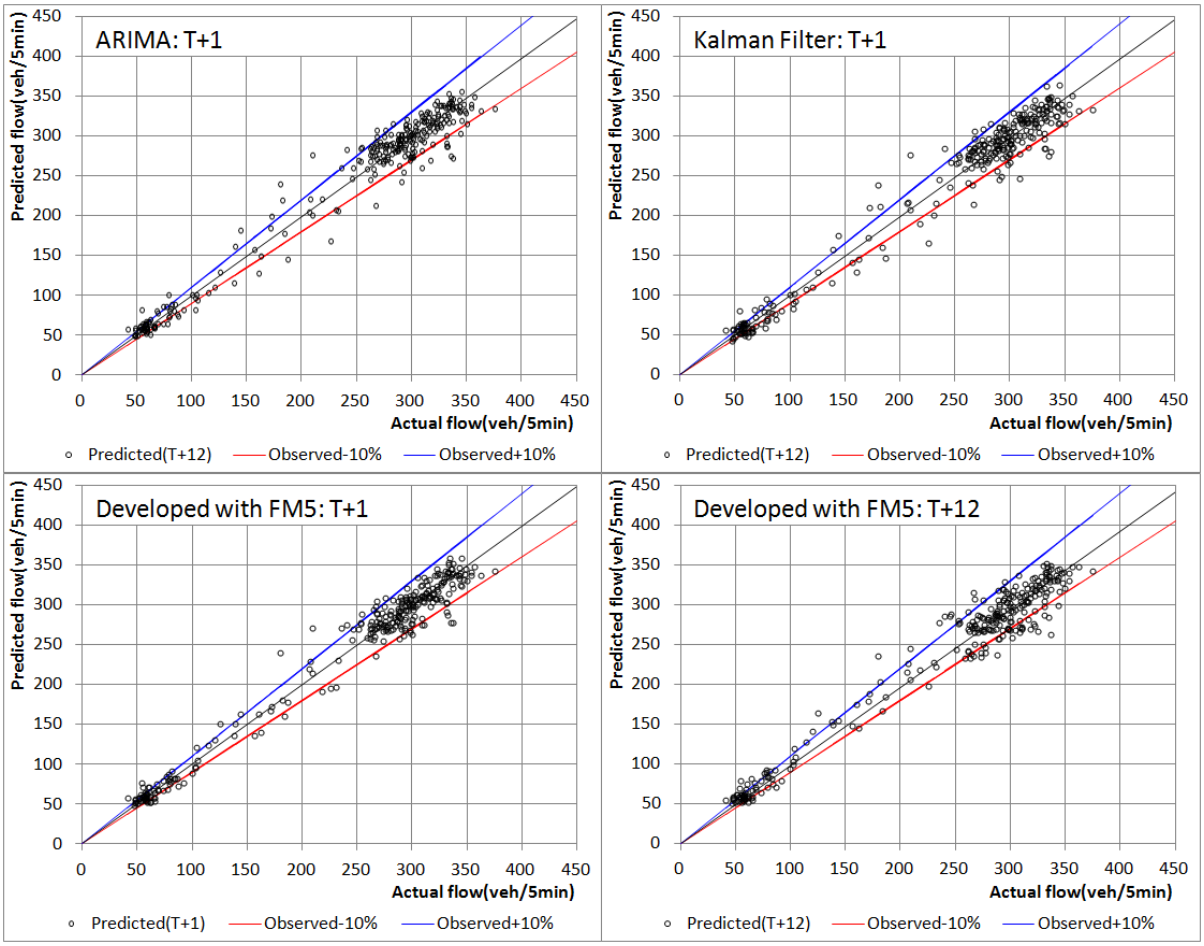


Figure 6: Predictability and stability of the compared three models of the traffic flow rate

Table 4: The results of performance measures for the three states

Performance measure	State	ARIMA	Kalman filter	Developed model	
		T+1	T+1	T+1	T+12
MAPE	Low	10.93	11.60	9.38	11.16
	Transition	11.72	12.44	9.99	10.33
	High	4.60	4.67	4.13	5.33
MAE	Low	6.80	7.30	5.74	6.90
	Transition	20.64	21.99	17.71	18.20
	High	13.87	14.15	12.50	16.10

The traffic flow state in Fig. 6 is divided into three levels: low ($q \leq 100$), transition ($100 < q \leq 250$), and high ($250 < q$), and the results of key performance measures are shown in Table 4. Note that the averages and standard deviations of the traffic volumes for the low, transition, and high states are 63.3 and 11.9, 173.7 and 48.5, and 303.4 and 26.5, respectively. Therefore, MAPE of low state is higher than those of the other states, despite the fact MAE is lower than those of the others. This indirectly indicates that the traffic volume state appears stable late at night, ostensibly, but it appears much closer to an unstable state than those of the other time periods in terms of the forecasting problem. For the one-step-ahead case, the developed model clearly outperformed the compared models in all states in terms of MAPE and MAE, whereas in the case of multiple time steps ahead, the performance of the proposed model was at least comparable to those of the other two models. Especially in a transition state, the developed model showed better performance for all time steps ahead than the two benchmark models. These analysis results therefore suggest that the KNN-NPR methodology, with the proposed forecasting functions, is capable of stronger performance than single-interval parametric approaches and that the methodology, furthermore, is a feasible promising alternative for multiple-time-period forecasting under unstable, meta-stable and stable in terms of the relative amounts of variation.

Consequently, the developed model, in the case of single-interval predictions, outstandingly outperformed the benchmark models, the two parametric approaches (ARIMA and Kalman filter) and the naïve

model according to the analysis results of both a rank test and performance measures. Additionally, the worst case of the forecasting method, i.e., forecasting method 4, was followed by the three comparative models. Note that when the time step ahead is extended, which in turn concurrently increases the uncertainties of the future state, the prediction error dramatically and unavoidably increases. As regards multi-interval forecasting, the performance results of the developed model were either comparable to or partially better than those of single-interval predictions by the benchmark models, which estimate the state one time step ahead. This fact indicates that the uncertainties of the future state can easily and efficiently be diminished by the known information from the intrinsic and complex patterns without a full understanding of the characteristics of the traffic flow. In addition, the developed model, with the exception of forecasting method 4, showed stable predictability under various states without a complicated (re-)calibration process of the parameters, which is one of challenges associated with the operating of the model.

V. Conclusions

This study was conducted to improve the performance of Non-Parametric Regression (NPR) with (1) adjusted forecasting functions, (2) an optimization strategy for the determination of the key parameter values, and (3) the time dependency of the input state to reduce the quantity of researched historical data with consideration of the execution time of the developed model and to employ more credible information for the reduction of uncertainties of future states.

The developed KNN-NPR methodology with the modified forecasting functions showed more effective performance than those of two conventional approaches, i.e., the straight average approach and the weighted average approach using inverse of the state distance in terms of forecasting accuracy. In a comparative study, the proposed model clearly outperformed the benchmark models, ARIMA, the Kalman filter model, and a naïve model according to the analysis results of a rank test and performance measures. The results of the comparative test showed that NPR can provide better performance than, or is at the very least be comparable to, the parametric approaches tested here, due to the fact that the optimized parameter values, the d_m and k values, and the time dependency of the input state highly contribute to build more inferred nearest neighbors, which in turn is closely related to the reduction of the uncertainties of future states by the key element, i.e., "the rate of the average of $x_m(T)$ to that of $x_m^i(t)$," of the forecasting functions. In addition, the fact that the performance of NPR as presented in this study exceeded those of ARIMA and KF not only strongly suggests that the (temporal) evolution of the traffic volume is characteristically chaotic rather than stochastic but also obviously validates arguments pertaining to the evolutionary behavior of short-term traffic flow state.

The key parameters, the d_m and k values, of the proposed model were simultaneously optimized and the relationship of d_m and k values to prediction error was demonstrated, indicating very carefully that some studies in which only the k values were optimized (or just analyzed and not to reaching a suitable condition) showed undesirable results unintentionally. The analysis results of the d_m and k values showed that an optimal error space with a concave combination shape geometrically exists in the relationship between the d_m and k values and the forecasting differences. On the other hand, this fact indicates that the existence of an optimal error space is strong evidence that intrinsic and complex patterns exist regardless of whether the associated interspace is clear or obscure. Future research into a local optimization strategy to find a suitable k value at each forecasting point rather than the strategy used in this study for a global optimal k value for all forecasting points therefore should be conducted for a more accurate estimation of the future state without a consecutive optimal k value adjustment on a weekly or monthly basis, despite the fact that the two values are mostly fixed in practice, until prevailing condition changes.

Finally, the results of this article demonstrate the high potentiality of NPR in future research. There are other opportunities to improve on the performance of NPR with different state definitions and/or distance metrics and/or forecasting functions, which may improve the performance of NPR. Additionally, both the time dependency of the input state with a time extender and much more historical data theoretically may lead to better results. Therefore, an empirical investigation into suitable quantities of databases to support NPR forecasting should be done from the perspective of data management. Moreover, further study for multi-step-ahead forecasting should be conducted for urban-signalized-arterial traffic flows, which exhibit wide-intensive fluctuations in nature.

References

[1]. Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46, 175-185.

[2]. Adeli, H. (2001). Neural networks in civil engineering. Computer-Aided Civil and Infrastructure Engineering: 1989-2000, 16(2), 126-142.

[3]. Chang, H., Baek, S., Shah, A.A., Lee, J.D., and Mahalik, N.P. (2007). Development of distributed real-time decision support system for traffic management centers using microscopic CA model. Iranian Journal of Science & Technology, Transaction B, Engineering, 31(B2), 155-166.

[4]. Chang, H., Park, D., Lee, S., Lee, H., and Baek, S. (2010). Dynamic multi-interval bus travel time prediction using bus transit data. Transportmetrica, 6(1), 19-38.
[5]. Chang, H., Seong, J.N., Lee, Y., Yoon, B. (2011). Dynamic freeway path travel time prediction based on nonparametric regression approach using dedicated short-range communications data. Proceeding in 90th Annual Meeting of Transportation Research Board, Washington, DC.
[6]. Chang, H., Park, D., Lee, Y., and Yoon, B. (2012). Multiple time period imputation technique for multiple missing traffic variables: nonparametric regression approach. Canadian Journal of Civil Engineering, 39, 448-459.
[7]. Chien, S., Ding, Y., and Wei, C. (2002). Dynamic bus arrival time prediction with artificial neural network. ASCE Journal of Transportation Engineering, 128, 429-438.
[8]. Clark, S. (2003). Traffic predicting using multivariate nonparametric regression. Journal of Transportation Engineering, 129(2), 161-168.
[9]. Davis, G. and Nihan, N., (1991) Nonparametric regression and short-term freeway traffic forecasting. Journal of Transportation Engineering, 117, 178-188.
[10]. Devijver, P. (1982). Statistical pattern recognition. Applications of pattern recognition, K. S. Fu, ed., CRC Press, Boca Raton, Fla., 15-36.
[11]. Disbro, J.E. and Frame, M. (1989). Traffic flow theory and chaotic behavior. New York State Department of Transportation Report FHWA/NY/SR-98/91, New York.
[12]. Eubank, J.D. (1988). Spline smoothing and nonparametric regression. Marcel Dekker Inc., NY.
[13]. Farmer, J.D., and Sidorowich, J.J. (1987). Predicting chaotic time series. Physical Review Letter, 59, 845-848.
[14]. Guegan, D., and Leroux J. (2009). Forecasting chaotic systems: The role of local Lyapunov exponents. Chaos, Solitons & Fractals, 41, 2401-2404.
[15]. Hamad, K., Lee, E., Shourijeh, M.T. and Faghri, A. (2009). Near-term travel speed prediction utilizing Hibert-Huang transform. Computer-Aided Civil and Infrastructure Engineering, 24, 551-576.
[16]. Huang, N., Zheng, S., Long, S., Wu, M., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., and Liu, H. (1998). The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London: Series A, Mathematical and Physical Sciences, 454, 93-95.
[17]. Innamaa S. (2000). Short-term prediction of traffic situation using MLP-neural networks. Proceeding in 7th World Congress of Intelligent Transportation System, Turin, Italy.
[18]. Ishak, S., and Alecsandru, C. (2004). Optimizing traffic prediction performance of neural networks under various topological, input, and traffic condition settings. Journal of Transportation Engineering, 130 (4), 452-465.
[19]. Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, Transactions of the ASME 82D, 35-45.
[20]. Kalman, R.E., and Bucy, R.S. (1961). New results in linear filtering and prediction theory. Journal of Basic Engineering, 60, 95-108.
[21]. Karlaftis, M.G., and Vlahogianni, E.I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C, 19, 387-399.
[22]. Karlsson, M., and Yakowitz, S. (1987). Rainfall-runoff forecasting methods, old and new. Stochastic Hydrology and Hydraulics, 1, 303-318.
[23]. Kim, J., Rho, J., and Park, D. (2009). On-line estimation of departure time-based link travel times from spatial detection system. International Journal of Urban Sciences, 13(1), 63-80.
[24]. Kim, D., Park, D., Rho, J., Baek, S., and Namkoong, S. (2007). A study on the construction of past travel time pattern for freeway travel time forecasting: focused on loop detectors. International Journal of Urban Sciences, 11(1), 14-29.
[25]. Kirby, H.R., Watson, S.M., and Dougherty, M.S. (1997). Should we use neural networks or statistical models for short-term motorway traffic forecasting? International Journal of Forecasting, 13, 43-50.
[26]. Lam, W.H.K., Tang, Y.F., and Tam, M.L. (2006). Comparison of two non-parametric models for daily traffic forecasting in Hong Kong. International Journal of Forecasting, 25, 173-192.
[27]. Lan, C.J., and Miaou, S.P. (1999). Real-time prediction of traffic flows using dynamic generalized linear models. Transportation Research Record, 1678, 168-178.
[28]. Liu, Z., Sharma, S., and Datla, S. (2008). Imputation of missing traffic data during holiday periods. Transportation Planning and Technology, 31(5), 525-544.
[29]. Mulhern, F.J., and Caprara, R.J. (1994). A nearest neighbor model for forecasting market response. International Journal of Forecasting, 10, 191-207.
[30]. Okutani, I., and Stephanedes, Y.J. (1984). Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B, 18B(1), 1-11.
[31]. Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S. (1980). Geometry from a time series. Physical Review Letters, 45, 712-716
[32]. Park, D., and Rilett, L. (1998). Forecasting multiple-period freeway link travel times using modular neural networks. Transportation Research Record, 1617, 163-170.
[33]. Park, D., Kim, N., Park, H., and Kim, K. (2012). Estimation trade-off among logistics cost, CO2 and time: a case study of container transportation systems in Korea. International Journal of Urban Sciences, 16(1), 85-98.
[34]. Qi, Y., and Smith, B.L. (1987). Identifying nearest-neighbors in large-scale incident data archive. Transportation Research Record, 1987, 89-98.
[35]. Robinson, P. (1983). Nonparametric estimators for time series. Journal of Time Series Analysis, 4, 185-207.
[36]. Shah, A.A., Kim, H., Baek, S. Chang, H., and Ahn, B. (2008). System architecture of a decision support system for freeway incident management in Republic of Korea. Transportation Research Part A, 42(5), 799-810.
[37]. Smith, B.L. (1995). Forecasting freeway traffic flow for intelligent transportation system applications. Doctoral dissertation, Department of Civil Engineering, University of Virginia, Charlottesville, VA.
[38]. Smith, B.L., and Demetsky, M.J. (1995). Short-term traffic flow prediction: neural network approaches. Transportation Research Record, 1453, 98-104.
[39]. Smith, B.L., and Demetsky, M.J. (1996). Multiple-interval freeway traffic flow forecasting. Transportation Research Record, 1554, 136-141.
[40]. Smith, B.L., and Demetsky, M.J. (1997). Traffic flow forecasting: comparison of modeling approaches. Journal of Transportation Engineering, 123(4), 261-266.

[41]. Smith, B.L., Williams, B.M., and Oswald, R.K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C, 10, 303-321.
[42]. Smith, B.L., and Oswald, R.K. (2003). Meeting real time traffic flow forecasting requirements with imprecise computations. Computer-Aided Civil and Infrastructure Engineering, 18(13), 201-213.
[43]. Sun, H., Liu, X., Xiao, H., He, R.R., and Ran, B. (2003). Use of local linear regression model for short-term traffic forecasting. Transportation Research Record, 1936, 143-150.
[44]. Takens, F. (1981). Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, 898, 366-381.
[45]. Turochy, R.E. (2006). Enhancing short-term traffic forecasting with traffic condition information. Journal of Transportation Engineering, 132 (6), 469-474.
[46]. Tong, H. (1993). Nonlinear time series: a dynamic approach. Oxford University Press, New York.
[47]. Vlahogianni, E.I., Golias, J.C., and Karlaftis, M.G. (2004). Short-term traffic forecasting: overview of objectives and methods. Transport Reviews, 24(5), 533-557.
[48]. Vlahogianni, E.I., Karlaftis, M.G., and Golias, J.C. (2005). Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. Transportation Research Part C, 13(3), 211-234.
[49]. Vlahogianni, E.I., Karlaftis, M.G., and Golias, J.C. (2006). Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. Transportation Research Part C, 14(5), 351-367.
[50]. Vlahogianni, E.I., Karlaftis, M.G., and Golias, J.C. (2014). Short-term traffic forecasting: Where we are where we're going. Transportation Research Part C, 43(1), 3-19.
[51]. William, B.M., and Hoel, L.A. (2003). Modelling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. Journal of Transportation Engineering, 129(6), 664-672.
[52]. Yakowitz S. (1987). Nearest-neighbor methods for time-series analysis. Journal of Time Series Analysis, 8(2), 235-247.
[53]. Yoon, B. and Chang, H. (2014). Potentialities of data-driven non-parametric regression in urban signalized traffic flow forecasting. Journal of Transportation Engineering, 140(7).

Hyun-ju Choi, et. al. "The Potential Use of Large-Scale Data for Multiple Time-Step Forecasting for Motorway Traffic Flow in Advanced Data Management Systems." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(2), 2022, pp. 38-57.