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 22-37
www.iosrjournals.org

Multiple Time-step Forecasting of Urban Traffic Flow Based on Data-Driven Non-Parametric Regression

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:
Future traffic volume is one of crucial elements in the area of advanced traffic signal control for intelligent transportation systems (ITS). Unfortunately, the temporal development of the urban traffic volume, i.e. the urban traffic volume (UTV), shows rapidly intensive fluctuations in nature. This is inevitably or frequently related to the well-known forecasting failure in the ITS forecast modeling. Despite the several remarkable achievements of single-interval UTV forecasting in the literature, it remains one of the major challenges when predicting reliable multi-interval UTVs for more proactive and tactical traffic signal operations. On the other hand, data-driven, knowledge discovery approaches to ITS forecast modeling are becoming known for their applicability given their use of tremendous quantities of historical data supported by advanced data management systems. In order to address the chronic uncertainty associated with the future UTV state efficaciously given the golden opportunities in the era of data technology, a system-oriented pattern-extraction methodology based on k-nearest neighbor non-parametric regression to predict multiple time-period UTVs is proposed in this paper. The model was tested against a large volume of real-world UTV data and then its performance was compared to those of two comparative models, the simple rolling average and the ARIMA, which are widely used for one-step-ahead forecasting. The analysis results revealed that the proposed model, despite its use of multi-step-ahead forecasting, clearly outperformed the other two models. This suggests that similar pattern-extraction methods, with vast quantities of historical data available, can serve as robust approaches for the direct recognition of the temporal evolution of the UTV state without a time-delayed response or the complexities of advanced models.

Key Word*: Big data; Urban traffic flow; Advanced data management system; K-nearest neighbor non-parametric regression; Multi-interval forecasting*

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

I. Introduction

Conventional adaptive signal operation in urban areas has become more advanced to reduce the social costs associated with traffic congestion. Future traffic volume is used as one of essential input parameters of advanced traffic signal operation, which will play key roles in advanced urban traffic management system as part of modern intelligent transport systems (ITS). In order to implement more proactive signal-control strategies extending over several signal cycles in practice, a robust forecasting model to generate reliable future multi-interval signalized traffic volume is crucial. Despite the numerous accomplishments of forecasting traffic variables in the literature and the efforts made to realize the short-term forecasting of urban signalized traffic flow, it seems, in many cases, that current forecasting models from an engineering perspective do not meet the prediction accuracy levels required in practice.

Regarding the temporal development of traffic flow state, signalized traffic flow reveals naturally rapid and intensive fluctuations that are closely related to the uncertainty problem in ITS forecast modeling, which in turn is linked to the well-known forecasting failure in many cases. One promising alternative to cope with this chronic problem could be data-driven nonparametric regression (NPR), which is theoretically founded on the initial deterministic conditions, i.e. chaos state theory. NPR completely relies on cases similar to itself which intrinsically exist in a vast volume of historical data, sparing the mathematical complexities of advanced models. Fortunately, current data technologies make tremendous amounts of current and historical traffic data available in real time. In addition, prerequisites such as big data accessibility and high-speed searching capability for NPR-based forecast modeling in actual ITS applications are currently possible given the state of leading-edge information technologies (Smith and Oswald, 2003; Chang et al., 2010). In spite of the necessity for advanced signal control and the real-time availability of vast volumes of historical data for knowledge-discovery-based prediction modeling, research on NPR for the multiple-time-period forecasting of urban signalized traffic flow has not been available in the literature.

Reliable multi-interval forecasting of urban signalized traffic flow clearly remains as one of the ongoing challenges in the academic research on real-world applications of modern ITS forecast modeling. In order to address the need for practical and academic progress, this study proposes a data-driven forecasting method based on k-nearest neighbor non-parametric regression (KNN-NPR) for the reliable multi-step forecasting of signalized traffic flow. The model is designed to meet the requirements associated with system-oriented components for its direct applicability in advanced data management systems (ADMS), excluding aspects such as the high mathematical complexities and complicated parameters of sophisticated models. The performance of the model is illustrated in an experiment with real-world signalized traffic flow data which exhibits intensive fluctuations in nature and meets the proper data requirements, i.e. the pattern variety level and availability of sufficient historical data in KNN-NPR. Additionally, some findings pertaining to the state evolution and predictability of signalized traffic flow are tentatively discussed considering the potential of data-driven NPR forecasting approaches.

II. Backgrounds

Future information about traffic flow variables (such as volume, speed, occupancy rate and travel time) is an essential element in ITS. Therefore, it is natural that a number of academic and empirical investigations into ITS forecasting have been made (and are still ongoing) as part of the effort to accomplish the ultimate goal of forecast modeling to overcome the uncertainty problem related to future traffic flow state in the literature. Despite the remarkable achievements thus far, the literature review in this paper, given its aims as discussed earlier, is focused on forecasting research for traffic volume, i.e. the urban traffic volume (UTV), which characteristically reveals intensive and rapid variations. Additionally, it is important to note that several studies include broad reviews of ITS forecasting research (see Vlahogianni et al., 2004, 2014; Karlaftis and Vlahogianni, 2011; Chang et al., 2012b; Mori et al. 2015).

Modeling approaches to UTV prediction can be categorized as parametric and nonparametric approaches which have their respective origin in different state theories. The parametric model is based on stochastic theory, whereas the nonparametric model has its origin in chaos theory. Accordingly, the two theoretical foundations of the modeling approaches are irreconcilable with regard to the state of the development of traffic flow systems (Smith et al., 2002; Vlahogianni et al., 2005; Chang et al., 2012b; Yoon and Chang, 2014). For this reason, it is crucial that the temporal state evolution of UTV is closer to which of the two state theories, as the reliability of a forecasting model (in terms of its prediction accuracy) relies strongly on the underlying assumption of the state-development behaviors of the target system.

In the approach which relies on probability theory, a few mathematical models, such as the time-series analysis (TSA) and state-space (SS) models have been used statistically to estimate the future state of UTV. The representative TSA model is the family of autoregressive integrated moving average (ARIMA), i.e., ARIMA(p,q,d). ARIMA(0,1,1), simple exponential smoothing with growth, was selected as the optimal means of effectively estimating the future state of 1-min aggregated UTV (Hamed et al., 1995), whereas ARIMA(3,0,3) was chosen as the suitable combination for the prediction of 3-min UTV data (Stathopoulos and Karlaftis, 2003). Despite the fact that they offer acceptable performance level with single-interval prediction of motorway traffic volume, i.e. the continuous traffic volume (CTV), it was pointed out that the ARIMA-family models remain seriously deficient with regard to their ability to explain the extreme conditions of UTV (Vlahogianni et al., 2005; Yoon and Chang, 2014). SS, with the same theoretical underpinnings as Kalman filter (KF), is one of promising alternatives to overcome the weaknesses of ARIMA models (Stathopoulos and Karlaftis, 2003). KF was employed to predict 15-min aggregated UTV data that does not reveal intensive oscillations (Okutani and Stephanedes, 1984). A multivariate SS model was proposed and tested against three-minute UTV data (Stathopoulos and Karlaftis, 2003), outperforming ARIMA in terms of prediction accuracy. It was also indicated by Stathopoulos and Karlaftis (2003) that other challenges related to the SS-based forecasting method remain to be addressed before the prediction accuracy of UTV can be improved to match that of CTV. Despite these efforts, it remains a challenge for parametric models to avoid concurrent and repetitious forecasting failures when the state of UTV varies abruptly and intensively (Vlahogianni et al., 2005; Yoon and Chang, 2014).

With reference to the models based on chaos theory, a few remarkable studies based on nonlinear or nonparametric techniques have been reported. Note that there is the predominant notion that the temporal development of the UTV state is much closer to a chaotic or mixed one than a stochastic one (Stathopoulos and Karlaftis, 2001; Vlahogianni et al., 2006; Yoon and Chang, 2014). Therefore, machine learning in the form of an artificial neural network (ANN) and knowledge discovery in data in KNN-NPR may be particularly useful in the area of UTV prediction. Combined ANN models with advanced techniques such as fuzzy and genetic algorithms have been utilized tactically in an effort to explain the phase transition of UTV states or for the dynamic optimization of model parameters. Fuzzy-ANN models patently improved the performance of conventional ANN (and KF) methods in terms of the prediction accuracy, instantaneously reacting to a phase

transition of the UTV state in less than five minutes (Yin et al. 2002; Stathopoulos et al., 2008). Nevertheless, it can be seen that fuzzy-ANN predictors are deficient with regard to decision making when attempting to capture the phase transition of UTV states consecutively under the condition of repetitive and rapid variations (Yoon and Chang, 2014). It is notable that a static ANN model coupled with a genetic algorithm for the (meta-) optimization of ANN elements (the step size, momentum and the number of hidden units) was shown to be very capable of proactively recognizing the recurrent and rapid phase transition of the 3-min aggregated UTV state, naturally showing very acceptable prediction accuracy (Vlahogianni et al., 2005). In addition to advanced ANN approaches, it was demonstrated that a data-driven methodology based on KNN-NPR (tested using 90-second aggregated data), with a very large volume of historical data available, shows robust potential for the reliable forecasting of UTV, efficaciously capturing the directionality and variance of the intensive and rapid state of UTV without a time delay, even when the model was intentionally simplified for actual applications in ADMS (Yoon and Chang, 2014).

Based on the literature (on the two approaches to forecasting UTV), it appears that the parametric models, in spite of being one step ahead, do not yet have the ability to estimate the recurrent extreme cases reliably in UTV forecasting despite the fact that they show acceptable levels of accuracy for CTV (Vlahogianni et al., 2005; Yoon and Chang, 2014). Regarding nonparametric models, their performances, which with regard to single-interval UTV forecasting are obviously better than those of the parametric models in terms of prediction accuracy (Yin et al. 2002; Vlahogianni et al., 2005; Stathopoulos et al., 2008), have reached an acceptable degree of accuracy for one-step prediction, although their performances for UTV were not as accurate as those for CTV. In addition, it was also indicated that no model which uses either a parametric or a nonparametric approach is clearly superior to the others in terms of accuracy when dealing with CTV with a single-interval forecasting horizon (Yoon and Chang, 2014).

Despite the painstaking efforts made thus far, crucial challenges that should sufficiently and effectively be addressed from the perspectives of academic research and real applications remain. One of these is to extend the forecasting horizon to multiple time steps ahead with an acceptable degree of predictability, which will inevitably be required for more proactive and strategic traffic control strategies. From the standpoints of model structure and performance, the forecasting horizon can easily be extended to more than one step ahead structurally, and a concurrent problem related to prediction accuracy then inescapably arises, especially when the temporal state of a system such as UTV evolves rapidly and intensively. This is because the number of future uncertainties increases concurrently and dramatically, when the time length of the prediction is extended (Chang et al., 2010). Accordingly, it is obvious that the multi-interval forecasting of UTV is vital in modern ITS and thus should be realized in order to ensure reliable or improved accuracy levels which exceed the current capabilities. Moreover, it was noted in our previous study that there still are promising potentialities for the multiple forecasting of UTV using data-driven NPR (Yoon and Chang, 2014).

III. Data-driven methodology for multiple forecasting

Theoretical background

The model presented in this study for the multiple-step-ahead forecasting of UTV is based on NPR. NPR has its theoretical foundation in chaotic systems, in which the successive states of a dynamic system strongly and deterministically rely on the initial conditions. In a data-driven forecasting problem based on the NPR approach, if k-past cases, i.e. k-nearest neighbors (KNN), similar to a current case, i.e. the initial condition, can be extracted from neighbor candidates (n) included in a historical dataset, the k-past future cases associated with the KNN set can be efficaciously employed to estimate the future state. As such, KNN-NPR is a promising approach for pattern-based analysis when large amounts of high-quality historical data are available. In addition, KNN-NPR is called NPR, distinguishing it from parametric regression (PR), which has its theoretical foundation in statistical stochastic process.

It is crucial to acknowledge that NPR has a strong theoretical basis in the nonlinear analysis of time-series data. The well-known KNN strategy for nonlinear classification was extended to a time-series analysis for short-term prediction (Karlsson and Yakowitz, 1987). If the data condition for a variety of patterns is satisfied as $n\rightarrow\infty$, $k\rightarrow\infty$ with $k/n\rightarrow 0$, the estimations generated by a NPR predictor are at least asymptotically optimal, making minimal risk decisions (Karlsson and Yakowitz, 1987; Davis and Nihan, 1991). Specially, NPR has its origin in pattern recognition without any assumption of statistical distribution; thus, it has strong advantages over PR approaches when used for the reconstruction of chaotic or mixed states (Altman, 1992; Mulhern and Caprara, 1994; Guegan and Leroux, 2009). Due to the dynamic local pattern recognition of NPR, it can also offer reliable performance capabilities in the forecasting of the temporal evolution of a chaotic state, regardless of whether or not the boundary condition of the patterns is clear (Yoon and Chang, 2014).

The data-oriented approach of NPR presents promising opportunities in forecasting modeling and practice in the case that a sufficient high-quality historical database is available. With regard to modeling approaches, NPR is a type of simplified but robust decision-making process based on a diversity of natural local

patterns which intrinsically exist in the bulk of past knowledge rather than on the artificial understanding interpreted by man-made models usually based on the complexity of mathematical formulas (Smith and Oswald, 2003). Hence, NPR has practical and tactical advantages over sophisticated models which have hindered field staff who in many cases do not have sufficient expertise related to the essential components of advanced forecast models for actual applications, such as knowledge about traffic flow behaviors, an understanding of and knowhow pertaining to modifications of the structures and algorithms of advanced models, and knowledge about the determination and (re)calibration of model parameters. It was also indicated that the bottlenecks have not yet received sufficient attention for real-world practice in the academic research on advanced ITS forecast modeling (Yoon and Chang, 2014). From the perspective of an ITS practitioner, NPR could be a feasible alternative to surmount the obstacles of advanced models in practice, as such ITS practitioners have deep knowhow about the structure and conditions of the data collected and managed in their systems. Furthermore, the run-time issue of data-driven NPR models, which inevitably arises in the data-searching process, already has been addressed by the development of cutting-edge information technologies (Smith and Oswald, 2003; Chang et al., 2010).

Of course, academic efforts regarding short-term prediction of traffic volume using NPR have been sufficient (Davis and Nihan, 1991; Smith and Demetsky, 1996-7; Smith et al., 2002; Clark, 2003; Sun et al., 2003; Chang et al., 2012b; Yoon and Chang, 2014; Dernas et al., 2015). By virtue of these accomplishments, NPR shows at the very least performance capabilities similar to those of PR when the evolutionary behavior of the state is closer to a stochastic state (Smith et al., 2002), whereas NPR approaches obviously outperform PR approaches in terms of prediction accuracy when the temporal development of a state is more analogous to a mixed or chaotic behavior (Smith and Demetsky, 1997; Chang et al., 2012b; Yoon and Chang, 2014). Despite the fact that NPR is likely a bright approach to UTV time-series analysis, only one investigation to predict intensive and rapid UTV with the time horizon of one step ahead has been made by Yoon and Chang (2014) in the transportation area. It was also indicated in their study that there is the potential to use NPR for multi-interval forecasting of UTV, representing an area which remains to be addressed thoroughly. In addition to traffic volume forecasting, NPR has been efficiently employed for various ITS estimation purposes, such as for the multiple imputation of missing traffic variables (Chang et al., 2012a), multiple time-period prediction of link travel time (Cai et al., 2016) and for path travel time forecasting under the conditions of multiple time-lagged observations (Chang et al., 2010; Chang et al., 2011), showing improved or acceptable performances in terms of estimation accuracy. Accordingly, it is clear that NPR is one of the most promising and suitable approaches toward ITS forecasting in the age of big data technology.

Components for forecasting algorithm

The data-driven methodology presented in this study for multi-interval UTV prediction is implemented through a KNN-NPR forecasting algorithm which consists of the two steps, similar pattern building and forecast generation. Pattern building is carried out through a search process that involves three state vectors (current, input, output) and a similarity measure, and forecast generation is then conducted by a forecasting function using the built patterns, i.e. similar cases. The three key elements, i.e. the three state vectors, the similarity measure, and the forecasting function, are defined in the following three subsections, after which they are integrated into the forecasting algorithm that is described using pseudocode in the last subsection.

Definition of state vectors

In discrete system analysis, (continuous) time is divided with a fixed length of time into time intervals, i.e. (discrete) time series, which are used to reconstruct the state development of a dynamic system, i.e. the state vector, in chronological order. In other words, the dimension and embedding size of the time series should be defined before the state vector is defined. One-dimensional time series for the forecasting of multiple time periods in our case, hence, is defined as $[(t), (t-1), \dots, (t-d_m+1)]$ for m-step ahead at the point of forecasting $(\mathrm{t}+\Delta\mathrm{t})$, $\Delta\mathrm{t} \to 0.0$, where m $(\geq 1,$ integer) denotes the forecasting horizon, i.e. the time period in the future, for which traffic flow states are estimated, and d_m $(\geq 1, \text{integer})$ is the embedding size, i.e. a suitable number of lags, of the time series for m-step ahead. Accordingly, the current state vector $x_c^m(t)$ used geometrically to reconstruct a time-series of a current traffic flow at the defined time intervals $[(t), (t-1, \ \dots, (t-dm+1)]$ can be defined as Eq. (1), where qt is the measured traffic volume (vehicles/ a time length) during the current time interval (t), $q(t-1)$ is the measured traffic volume during the previous time interval (t-1), and so on. In addition, there exist a number of state-vector types which are ultimately useful (Smith et al., 2002), indicating that no formal type of state space has been reported in the academic literature. Furthermore, for the widely used definition of the state space, Takens' theorem (1981) can be referenced.

$$x_c^m(t) = [q(t), q(t-1), \dots, q(t-d_m+1) \qquad (1)$$

The current state vector is employed (as a sort of initial condition state) to extract cases similar to it from n potential cases which intrinsically exist in the historical database. Each similar case consists of an input state vector and an output state vector, and the output vector is associated to the input vector. The j-th input state vector $x_j^m(\tau), j \in n$, at the past time interval (τ) for $x_c^m(t)$ is defined by Eq. (2), and τ is the (past) running time index, $\tau \leq t - m$.

$$x_j^m(\tau) = \left[q_j(\tau), q_j(\tau - 1), \dots, q_j(\tau - d_m + 1)\right] \quad (2)$$

The output state vector o_j^m related to $x_j^m(\tau)$ is also defined with Eq. (3), where $q_j(\tau + m)$ is the past traffic volume at time interval (τ+m), u_j^m is the geometric distance between $x_c^m(t)$ and $x_j^m(\tau)$, and $q_{h,j}^m$ is the average of the elements of $x_j^m(\tau)$.

$$o_j^m = [q_j(\tau + m), u_j^m, q_{h,j}^m] \quad (3)$$

Definition of similarity function

In time-series data-driven NPR, a suitable technique for quantifying the degree of "nearness" between a past and a current case should be determined. Nearness is principally used as a critical criterion to extract past similar cases from the (historical) database in the data search process, which in turn is closely related to the reliability of the NPR predictor in terms of the estimation accuracy. Therefore, selecting a suitable one from the possible similarity functions necessitates careful consideration, especially in the case of UTV, the temporal state development of which varies widely and intensively in nature, as stated before.

The Minkowski distance metric, i.e., the L_p distance, is a widely used metric among various similarity measures ranging from the (weighed) geometric distance to the statistical goodness of fit test, and the Euclidean distance (ED), the L_p distance when $p = 2$, is also widely utilized in the independent variable space of NPR. Nearness measured by the ED metric is sensitive to noise in the signal. Namely, the geometric distance (between the input and output state vectors) in the case of UTV sharply increases due to the extremes at the steep turning points of temporal evolution of the two state vectors. It was also demonstrated that this sensitivity can be efficaciously used to quantify the degree of nearness between the two state vectors in UTV forecasting, resulting in acceptable estimation error (Yoon and Chang, 2014). Additionally, it was pointed out that a weighed distance metric is heuristic in nature, implying that careful consideration be made by ITS practitioners (Smith et al., 2002). Moreover, such a weighed approach can generate biased nearness in the case of UTV, weighing extreme cases heavily or lightly. In this context, the ED, i.e., the $L_{p=2}$ distance, is also employed to measure the similarity in this study. Here, the ED, u_j^m, between $x_c^m(t)$ and $x_j^m(\tau)$ for m-step ahead can be defined by Eq. (4).

$$u_j^m = \left[\sum \left| x_c^m(t) - x_j^m(\tau) \right|^2\right]^{1/2} \quad (4)$$

Thus, the nearness between the two vectors with the d_m-embeding size is quantified with Eq. (5) using the corresponding elements according to the time period.

$$u_j^m = \left[\sum_{i=0}^{d_m - 1} \left| q_c(t - i) - q_j(\tau - i) \right|^2\right]^{1/2} \quad (5)$$

All (optimal) values of the d_m parameter for each m-step ahead are not equal in many cases; thus the nearness calculated by Eq. (5) is repeatedly computed equal to the number of forecasting horizons (m) in the similar-pattern search process. In practice, measuring all sub-nearness values of the elements of the two vectors for all steps ahead can be time-consuming, as overlapping sub-nearness is redundant from the standpoint of data access and computing issues. This inefficiency is related to the performance of data-driven NPR predictors in terms of the execution time required in practice, as well. In order to deal with this time-consuming problem effectively, a nearness state vector, u_j, introduced in this study for all possible embedding sizes with the maximum embedding size (d_{max}), is defined as Eq. (6), where d_{max}= max$\{d_m\}$. u_j consists of u_j^d, which represents the nearness for each possible d, d=[1, 2, … ,d_{max}]. Each element of u_j is also generated and built sequentially while the nearness between the two state vectors with d_{max} is quantified using Eq. (5). In this way, using the nearness state vector reduces the execution time (to measure the overlapping sub-nearness and to browse historical data) to the degree of [$\sum d_m$-d_{max}].

$$u_j = \left[u_j^1, u_j^2, \dots, u_j^{d_{max}}\right] \quad (6)$$

Definition of forecasting function

Forecast generation for the future multiple time step (m) is carried out with a forecasting function using the k_m-output vectors which are built on a data structure in the process of the KNN-NPR multiple forecasting algorithm presented in this study, where k_m is a suitable k-value for m-step ahead. The data structure to record KNN information for m, therefore, is defined as follows, and the forecasting function is then given with the KNN database. The KNN dataset has an input-output structure, as $x_i^m(\tau) \rightarrow o_i^m$, $i \in k_m$ and $k_m/n \rightarrow 0.0$. The input sector is made up with the k_m-nearest neighbors, i.e. k_m-input state vectors similar to $x_c^m(t)$ in the nearness measure, which serve as the NPR independent variables. In contrast, the k_m-output vectors of the output sector that are associated with the k_m-input vectors serve as the baseline for the estimation of the dependent variable produced by the forecasting function stated in the next paragraph.

i	Input ($x_i^m(\tau)$)	→ Output (o_i^m)
1	$[q_j(\tau), q_j(\tau-1), \dots, q_j(\tau-d_m+1)]$	$\rightarrow [q_1(\tau+m), u_1^m, q_{h,1}^m]$
2	$[q_j(\tau), q_j(\tau-1), \dots, q_j(\tau-d_m+1)]$	$\rightarrow [q_2(\tau+m), u_2^m, q_{h,2}^m]$
…	…	…
k_m	$[q_j(\tau), q_j(\tau-1), \dots, q_j(\tau-d_m+1)]$	$\rightarrow [q_k(\tau+m), u_k^m, q_{h,k}^m]$

Infinite techniques for a forecasting function(FF) are possible according to the modeler's initiatives and goals, and various FFs ranging from the straight average to the combination of the weighted average and available supplementary information have been reported in literature (Smith et al., 2002; Yoon and Chang, 2014). Two typical FFs that are usually employed are the straight average (SA) and the weighted average by the inverse of distance, i.e., nearness, (WAID) for the NPR dependent variable values of the k-output vectors. The SA technique, also termed the arithmetic mean, imposes identical weights on each dependent variable value, not considering the nearness information. On the other hand, the WAID applies different weights to each dependent variable in accordance with reversed nearness value. Therefore, the intrinsic information included in the nearness value can be used as a pivotal driver, from the standpoint of decision making, to diminish the uncertainties of the future state. It should also be noted that the utilization of the nearness shares the basic purpose of experience-based approaches, such as NPR, successfully to deal with the limitations of human knowledge, even though it is heuristic in nature.

With regard to UTV forecasting, the adjusted-by-ratio (AR) value, which is "adjusted by the ratio of the average of the elements of the current state vector to that of the elements of each selected input state vector" that is coupled with SA or WAID was proposed as a means of improving the performance of the two averaging methods (Yoon and Chang, 2014). It has also been reported that the AR-and-SA (ARSA) and AR-and-WAID (ARWAID) functions for one-step-ahead forecasting outperform the two averaging functions in terms of estimation reliability. Therefore, the ARSA and ARWAID methods are adjusted and employed for multiple forecast generation in this study, as defined by Eqs. (7) and (8), respectively, where $\hat{q}(t+m)$ denotes the predicted traffic volume for m-step ahead at forecasting point $(t + \Delta t)$, q_c^m is the average of the elements of $x_c^m(t)$ with d_m for m-step ahead, and ε is the minimal nearness value with 0.0001.

$$\hat{q}(t+m) = \sum_{i=1}^{k_m} \frac{q_i(\tau+m)}{k_m} \cdot \frac{q_c^m}{q_{h,i}^m} \quad (7)$$

$$\hat{q}(t+m) = \left[\sum_{i=1}^{k_m} \frac{q_i(\tau+m)}{(u_i^m+\varepsilon)} \cdot \frac{q_c^m}{q_{h,i}^m}\right] / \left[\sum_{i=1}^{k_m} \frac{1}{u_i^m+\varepsilon}\right] \quad (8)$$

KNN-NPR multiple forecasting algorithm

For the realization of data-driven models such as NPR for ITS forecasting, an advanced sorting and searching technology should be supported in order to build past cases similar to the current case from the large-scale historical database. Despite the necessity of advanced searching algorithms, it is beyond the scope of this article to address search technologies which collect similar cases, and we assume that a suitable sorting and searching algorithm facilitates the proposed methodology in this study. In addition, existing studies based on NPR for ITS forecasting have commonly used forecasting algorithms that integrate the key components of their models, excluding any sorting and searching methods (Smith and Demetsky, 1996; Smith et al., 2002; Chang et al., 2010; Chang et al., 2011; Chang et al., 2012 a,b; Yoon and Chang, 2014).

As noted above, the three components (the three state vectors, the similarity function, and the two forecasting functions) stated previously and the updating process of the selected similar cases are combined into the KNN-NPR multiple forecasting algorithm for the realization of the presented methodology in this research. The pseudocode of the algorithm shown in Fig. 1 consists of three steps: (1) initialization, (2) building k_m-objects, i.e., the k_m-output vector list, and (3) multiple-forecast generation. At the point of forecasting $(t + \Delta t)$, the d_m and k_m values are given, the d_{max}-size current state vector $x_c(t)$ that includes all possible d-size current

state vectors is constructed, and q_c (the set of average current states, $[q_c^1, q_c^2, \dots, q_c^{d_{max}}]$, for all possible d-size current state vectors) is calculated, where d=[1, 2, …, d_{max}]. After the initialization of the k_m-objects, the building of the k_m -objects, involving the preprocessing and candidate-output-vector updating steps is conducted. In the preprocessing step, the input state vector $x_j(\tau)$ with d_{max} and the past future states $q_j(\tau + m)$ for each m-step ahead case at the (past) running time period (τ) are attracted from the historical database, after which the nearness state vector u_j (between $x_c(t)$, i.e. the current state vector and $x_j(\tau)$ for all possible d-sizes) and the set of average historical states $q_{h,j}$ (for all possible input state vectors within d_{max}) are computed. In the candidate update step, o_j^m, i.e. $[q_j(\tau + m), u_j^m, q_{h,j}^m]$, of $x_j^m(\tau)$ is updated on the k_m-object list using each respective elemental value of u_j associated with d_m. Finally, future traffic volume states for multiple time steps are produced on the basis of the built k_m-output vector by Eqs. (7) and (8).

Given forecasting horizon (m) at forecasting point (t + Δt), d_m- and k_m-values,
the current state vector $x_c(t)$ with d_{max}, $q_c = [q_c^1, q_c^2, \dots, q_c^{d_{max}}]$, and $k_m/n \rightarrow 0.0$:
1) Initialize the list of the k_m-objects with $[q_i(\tau + m), u_i^m, q_{h,i}^m]$
(where $q_i(\tau + m)$ and $q_{h,i}^m$ have zero values, u_i^m is the maximum nearness value, and i=1, 2, …, k_m)
2) For each $x_j(\tau)$ with d_{max} and $q_j(\tau + m)$ at the (past) running time period (τ)
(where j=1, 2, …, n, and $\tau \le t - m$)
2-1) Calculate u_j and $q_{h,j}$ between $x_c(t)$ and $x_j(\tau)$ by Eq. (5)
(where, $u_j = [u_j^1, u_j^2, \dots, u_j^{d_{max}}]$ and $q_{h,j} = [q_{h,j}^1, q_{h,j}^2, \dots, q_{h,j}^{d_{max}}]$)
2-2) For each future m-step ahead
-Select u_j^m and $q_{h,j}^m$ from u_j and $q_{h,j}$ with d_m
If $u_j^m < u_{max}^m$ then
(where $u_{max}^m = \max\{u_1^m, u_2^m, \dots, u_k^m\}$)
2-2-1) Withdraw $[q_i(\tau + m), u_{max}^m, q_{h,i}^m]$ from the k_m-object list
(where $[q_i(\tau + m), q_{h,i}^m]$ are related to u_{max}^m, $1 \le i \le k_m$)
2-2-2) Update $[q_j(\tau + m), u_j^m, q_{h,j}^m]$ on the k_m-object list
2-2-3) Search for a new u_{max}^m on the updated k_m-object list
3) Generate $\hat{q}(t + m)$ by Eqs. (7) and (8)

Figure 1: Pseudocode for the KNN-NPR multiple forecasting algorithm

IV. Experiments and empirical findings

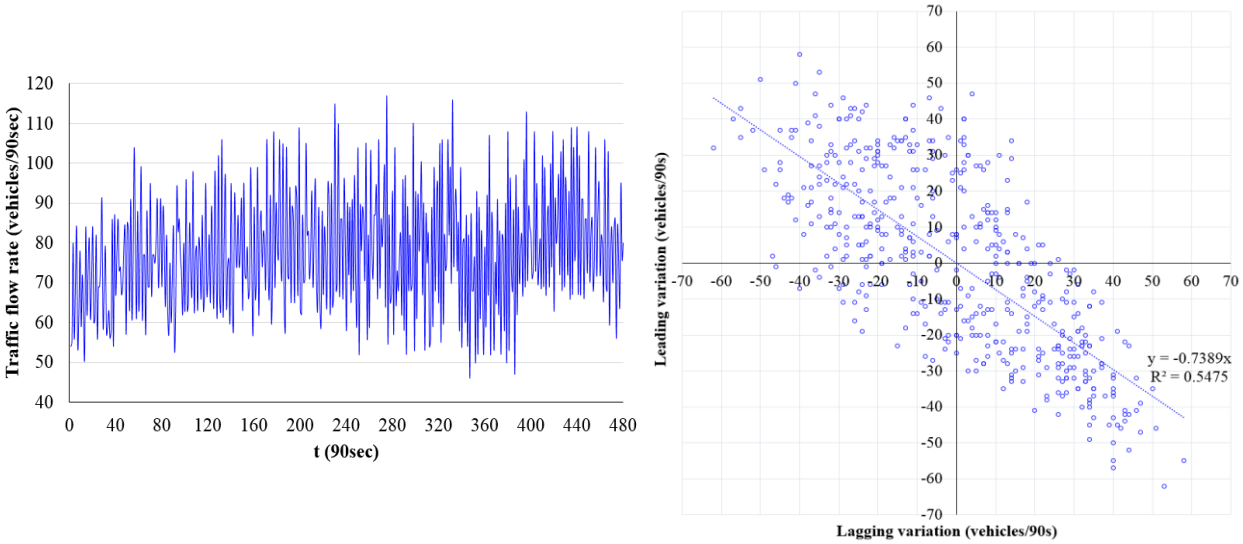
Study design

In this subsection, the test data and its features are examined with regard to predictability, and yardstick models and performance measures are then carefully determined based on the diagnosed characteristics in order to validate the capabilities of the KNN-NPR multiple forecasting model, presented in this article, for the wide and intensive state evolution of urban signalized traffic flow.

In the present study, we used data, identical to that employed in our previous research (Yoon and Chang, 2014) as collected from a major signalized arterial (in Seoul, South Korea) where an advanced signal control system known as COSMOS (the cycle, offset, split model of Seoul) is operated, permitting right turns on red (RTOR). The test site is usually congested during the half of the day from noon to midnight. The available data from 50 weeks starting with the first week of January to the third week of December in 2010 were made up with traffic volume measurements in 90-second intervals. The data consist of 336,000 observations (=960 90-second sequences per day × 7 days a week × 50 weeks); thus, the quantity of the data, i.e. the nearest neighbor candidates (n), is enough to meet the data condition, k/n→0.0, of NPR with a suitable k-value. For the evaluation of the presented methodology, we also used the data from the Friday of the final week as the target UTV system. Additionally, the data did not undergo any preprocessing technique such as smoothing and imputation so as to maintain the inherent dynamic of temporal state evolution that naturally includes negative noise and meaningful signal for future states.

From the standpoint of forecast modeling, it is logical to diagnose the temporal state behaviors of the target system to settle the minimal capability of a predictor. The temporal state of the target UTV shown in Fig. 2(a) fluctuates intensively, exhibiting extremes at repeated turning points, mainly due to consecutive and complex interruptions by traffic signals that are located upstream. Fig. 2(b) shows the directional instability of the temporal state evolution between leading and lagging variations, where the leading and lagging variations are defined as $[q(t)\text{-}q(t\text{-}1)]$ and $[q(t\text{+}1)\text{-}q(t)]$, respectively. This relationship is biased toward different

directions, largely dispersing on the second and fourth quadrants, which in turn indicates that the directionality gained with the current state cannot be utilized to capture that of a future state in many cases. Specifically, the leading variation from -30 to 30 is fairly close to the well-known random walk under a closed boundary condition within a lagging variation of ±50, whereas the directionality is reversed with absolute differences of two variations of 0~115 when the leading variation is greater than 30 or less than -30. This noise-like behavior indicates that only one chance exists to capture the directionality of the future state, which is closely related to an acceptable prediction without a delay time to recover or adjust the directionality in the forecasting process.



(a) (b)

Figure 2: Features of the target system: (a) temporal variation, (b) volatility of state evolution

In the area of ITS forecasting, there still is on-going academic debate with regard to parametric and non-parametric modeling approaches as to whether the temporal state behavior of a traffic flow system is closer to a stochastic or a chaotic case. The KNN-NPR forecasting model presented here has its theoretical basis in the initial deterministic system, whereas parametric approaches based on statistical stochastic process are grounded in stochastic state theory. Additionally, no multiple prediction model widely used for the aforementioned intensive UTV with acceptable levels of prediction accuracy in practice exists from the viewpoint of transportation practitioners, though a small number of refined studies are available in the literature. Accordingly, a comparative study to validate the robustness of the KNN-NPR method was carried out with two benchmark models, ARIMA and the simple rolling average (SRA) models, as the best and worst cases, respectively.

The ARIMA family of models, one of the most widely used time-series analysis techniques, is founded on discrete-time stochastic processes, a combination of random walk and Markov process. In practice, the ARIMA(p, d, q) model is one of the most efficaciously applied ARIMA models, where p, d, and q represent the order of the autoregressive, differencing, and moving-average components, respectively. With considerable amount of historical data and a long seasonal period length with 336,000 observations over a course of 350 days and, at the very least, 960 time sequences per day in our case, it is difficult for the seasonal ARIMA framework to meet the required time in an ITS system, requiring six days and seven nights as run time from the standpoint of the system time. Moreover, it was demonstrated that the temporal behavior of UTV does not contain the conventional conception of periodicity (hourly, daily, or even weekly periodicities) statistically, instead revealing a random walk that deviates from recurrent patterns (Stathopoulos and Karlaftis, 2001). The ARIMA (p, d, q) with non-seasonal components is, hence, employed as the best-case benchmark model with consideration of real-time execution in a real-world ITS system. With regard to SRA(q), also known as the simple moving average, as the worst case model, it could be also used in practice to simply analyze the state evolution trend line if there is no forecasting model available to capture the temporal dynamics of the system such as the target UTV addressed in this research. On the other hand, there exists the possibility that SRA could outperform, in terms of estimation accuracy, a forecasting model which fails to capture the correct directionality and variation of a future state at sequential turning points and inescapably experiences perfect forecasting failures.

Four performance measures were carefully chosen for macroscopic and microscopic analyses. Regarding the macroscopic analysis, two error functions and a statistical ranking test were employed, as follows. If the traffic flow rate varies greatly from -40.7% to 50.8% over the average traffic flow, as shown Fig. 2(a), the mean absolute percentage error (MAPE, %) defined by Eq. (9) provides the most instinctive and useful

basis for a comparison (Smith et al., 2002), where n is the number of test samples, and q_i and $\hat{q}_i$ are the measured and predicted traffic flow rates of simple i, respectively. The MAPE was also used to identify the optimal parameter values, the k_m and d_m values, of the proposed KNN-NPR methodology. A MAPE analysis was supplemented with the mean absolute error (MAE, vehicles) defined by Eq. (10).

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{q}_i - q_i|}{q_i} \times 100, \quad q_i > 0 \tag{9}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{q}_i - q_i| \tag{10}$$

Additionally, the Friedman test, a non-parametric statistical test, using absolute errors, $|\hat{q}_i - q_i|$, was utilized in order to quantify the performance of the four method, i.e., the two FFs of the proposed model and the two benchmark models, with the mean rank. Regarding the microscopic analysis, a state-variation scattergram composed of the actual variation, $q_i(t+1) - q_i(t)$, and the predicted variation, $\hat{q}_i(t+1) - q_i(t)$, as shown in Fig. 3 was used to clarify the degree of forecastability fully with perfect forecast line (y=1.0x). The dots on the dotted line represent perfect prediction, whereas the others in the first and third quadrants denote that under- or overestimation on the basis of the complete forecast line, despite the fact that the correct directionality of the future state is captured. In contrast, the dots in the second and fourth quadrants, except for acceptable errors, are the cases in which the direction of the predicted state is estimated as opposite that of the future state, exhibiting a seesaw-like condition, i.e. a total forecasting failure.

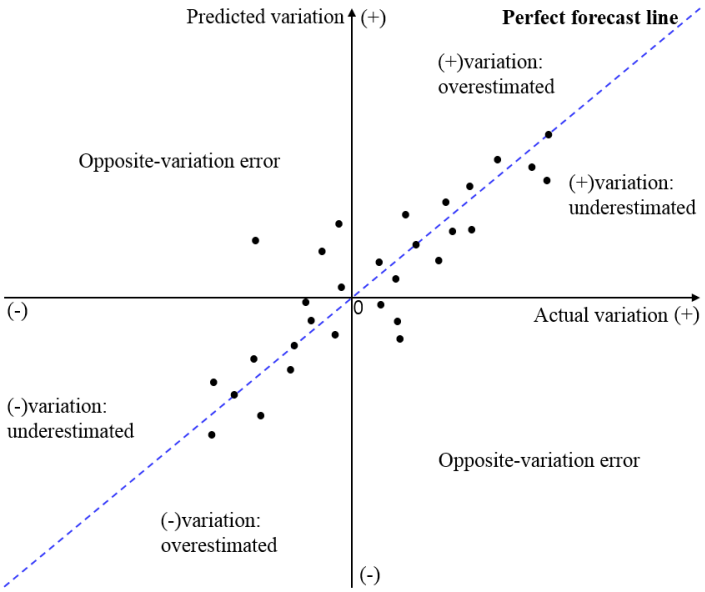


Figure 3: State-variation scattergram

Model-parameter analysis and findings

Once the condition of the variety of historical patterns is satisfied, the performance of a data-driven NPR approach in terms of the estimation accuracy is fundamentally contingent upon the recognition of historical patterns similar to the current state evolution of the target system.

The embedding size, d, used to reconstruct the current state evolution and the number, k, of neighbors similar to it have crucial effects on the degree of prediction accuracy, participating in the decision-making process for the future state.

Regrettably, it was indicated that a generally accepted technique to determine the optimal d and k values (at once) still does not exist in KNN-NPR (Yoon and Chang, 2014). Concerning the d value, the definition by Takens' embedding theorem (1981), $d \geq 2D + 1$ in D-dimensional Euclidean space, is widely referred to and employed as a minimal d-value condition. On the other hand, it was found that the state space with a d value of less than 2D+1 can also reconstruct the dynamic features of a system effectively depending on the circumstance (Packard et al., 1980). With regard to the suitable k value, potential candidates (n) included in the limited historical data available are finite in the real world, which implies that the finite k value, i.e. the number of nearest neighbors, should be optimally inferred according to the given d value, meeting the theoretical data condition of NPR, stated earlier as $n \to \infty$, $k \to \infty$ with $k/n \to 0.0$. It is, therefore, crucial to

determine the suitable values of the two parameters simultaneously, i.e. d_m and k_m, of the proposed KNN-NPR methodology for each m-step ahead in order to ensure an acceptable degree of predictability.

A forecasting simulation, as used in our previous study (Yoon and Chang, 2014), with the forecasting algorithm shown in Fig. 1 was also applied to analyze the effects of the two parameters of the proposed KNN-NPR methodology on the forecastability and to determine their optimal values for additional evaluation studies. A scenario of the simulation for each combination of future time step (m) ahead, m=[1,2,3,4], and FF consisted of all possible combinations of the two parameters in increments of 1 as 1,000 = d_m values [1-20] × k_m values [1-50]. The total number of prediction cases was, therefore, 384 × 10^4 (=two FFs × m values[1-4] × target time sequences[1-480] × parameter scenarios [1-1,000]). The prediction error for each prediction case was analyzed with MAPE, and the 80 prediction-error curves for the combinations of d_m values and m values (according to all k_m values) were then analyzed for each FF.

The effects of the two parameters on the forecastability of the proposed methodology with FF2, i.e. ARWAID as defined by Eq. (8), are shown in Fig. 4 for four steps ahead in each case. The behaviors of the MAPE curves for each step ahead, when the d_m-value increases in the order of 1→20, can be distinguished into the five following steps: (1) the curve steeply approaches the proximity of the boundary of the second-best minimal error space, (2) it goes through the error space gradually with little variation, (3) it transfers from the error space to the best minimal error space over the distinct interspace between the two spaces, (4) it reaches the minimal-error state gradually, and (5) it then gradually increases with little variation. In this way, the relationship between the prediction error and the increment of the d_m value shows a convex shape at the very least in our case.

The results of the d_m-value analysis reveal meaningful facts related to the roles of the d_m value and to the empirical findings. First, d_m^o, which represents the optimal d_m value for each m-step ahead, can be much larger than the minimal d value, i.e. 3, according to Takens' definition in our case of D=1, i.e. a one dimensional time-series state, but an increment of more than d_m^o+β of the d_m value cannot capture the directionality of the future state, although the d_m-value satisfies the conditions of Takens' definition. Second, the convex shape of the prediction error directly implies that the boundary condition of d_m^o obviously exists. Additionally, it can be seen that the d_m^o value is affected and is limited by the complex combinations of interruptions by traffic signals located upstream. Finally, d_m values that exist in the best error space show acceptable levels of error within the minimal error of +0.5% of the best case of the d_m value, which in turn implies that suitable d_m values between d_m^o-α to d_m^o+β, α ≤ β, can be analyzed and determined in advance within acceptable levels of prediction error from the standpoint of the (re) calibration of the model parameters. Moreover, it is indicated that the d_m^o value can be analyzed and updated through an experimental study on a daily, weekly or even monthly basis (Chang et al., 2012b) or can be fixed without recalibration in real-world applications (Yoon and Chang, 2014).

With regard to the k_m-value, the prediction error for all d_m-value cases of four steps ahead steeply and then gradually decreases to a minimal error condition with acceptable variation, after which it progressively increases as the k_m value increases. This significantly implies that a categorized pattern of future states closely related to the temporal development of the current state exists regardless of whether the interspace between the patterns is clear or not (Yoon and Chang, 2014), even when future time steps increase. Moreover, it was noted that a reliable decision-making step to diminish the uncertainties of the future state to the degree of an acceptable error level in NPR can be carried out with a given suitable value of k_m that can be analyzed and determined periodically on a daily or weekly basis (Smith and Oswald, 2003; Chang et al., 2010, 2012b; Yoon and Chang, 2014). For each m-step ahead, the optimal error space as explained within the minimal error, with the best k_m value, of +0.1% also exists as in the d_m-value case, despite the fact that the d_m-value case is not the best case. Additionally, the prediction error with the k_m value equal to 1 does not reach the acceptable level of error even when the d_m value is the best case. Accordingly, it can be seen, at the very least in our case, that the temporal evolution of the UTV state during the time period of d_m^o+m is closer to the initial deterministic (or mixed) case rather than a stochastic case from the perspective of forecast modeling.

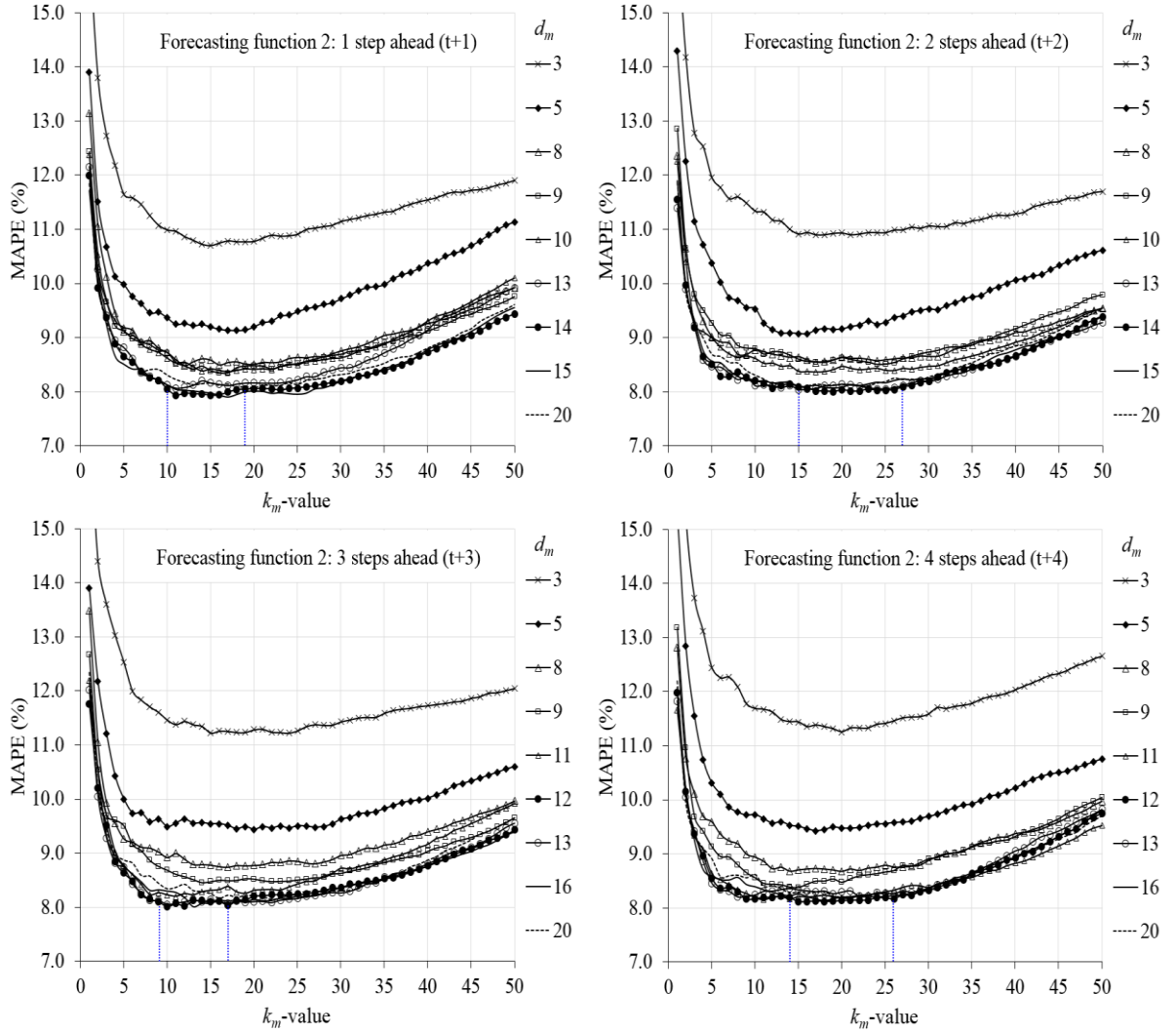


Figure 4: Effects of the model parameters on forecastability

Based on the analysis results of the two parameters in terms of reliability, the best d_m value and optimal k_m value, i.e. d_m^o and k_m^o, for each time step ahead are determined, as shown in Table 1, to conduct more analyses and to demonstrate the robustness of the proposed methodology through a comparative study. The best d_m values for the two FFs, i.e. FF1 and FF2 by ARSA and ARWAID, do not increase but rather decrease to satisfy the minimal or optimal prediction error levels when the number of time steps ahead is extended. Interestingly, this fact is different from the role and features of the d value as indicated in a forecasting study of uninterrupted traffic volume done by Chang et al. (2012b). It appears, at the least in our case, that the d_m^o value can be influenced by the complex interruptions of traffic signals and that the evolution of the UTV state during the time period of d_m^o+m can be classified by a NPR attractor with a combination of the d_m^o and k_m^o values. Regarding the range of k_m values associated with optimal error space, the two marginal values of the range for two and four steps ahead increase and the span of the two values is expanded with d_m^o values identical to those used with one and three steps. On the other hand, the d_m^o value varies from 14 to 12 in the case of three steps ahead, and the two marginal values of the range and the associated differences for each FF then decrease at once. These facts indicate that (1) the temporal evolution of UTV has an interface between two different states in the process of the phase transition, and then the span of the k_m values is enlarged to diminish the uncertainty in the decision-making process, and (2) a (future) state changes into different states after the phase transition, and the size of the d_m^o value and then the range of k_m values are subsequently shortened to capture one of the diverged states. Additionally, the d_m^o values and k_m^o values as shown in Table 1 are used as the optimal parameter values to analyze the results, and the k_m^o value is the median of the range of k_m values.

Table 1: Selection of optimal values of the model parameters

FF	Step ahead	d_m^o value	Optimal error space		k_m^o value
			Range of k_m	Average error	
FF1	t+1	14	11-16	7.96	14
	t+2	14	15-25	8.06	20
	t+3	12	9-12	8.14	10
	t+4	12	14-23	8.16	18
FF2	t+1	14	10-19	7.95	14
	t+2	14	15-27	8.04	21
	t+3	12	9-17	8.08	13
	t+4	12	14-23	8.15	20

Results and findings

The test results are summarized in Table 2, where the two FFs as the best-performer group are distinguished from the two benchmark models as the worst-performer group by means of the three performance measures. In the case of the worst-performer group, SRA and ARIMA are the best and the worst, respectively. Note that SRA(14) showed minimal prediction error in terms of MAPE, and ARIMA(3,0,8) was estimated and selected as the best ARIMA form. The finding of the worst performance, close to a forecasting failure, of ARIMA is consistent with the results of Stathopoulos and Karlaftis (2003), Vlahogianni et al. (2005), and Yoon and Chang (2014). More interestingly, the SRA outperforms ARIMA on all performance measures. This fact indicates that moving average techniques to forecast intensive traffic flow can serve to avoid the worst forecasting failure when there is no suitable model available.

With regard to for the best-performer group, the fact that FF2, the ARWAID, performs better than FF1, the ARSA, is in accordance with the results of Yoon and Chang (2014). Additionally, the WAID component found in FF2 improves upon the performance of FF1, which corresponds to the results of Smith et al. (2002) and Chang et al. (2012a,b). The proposed methodology for one step ahead remarkably outperforms the benchmark models based on the forecasting accuracy gains (%) [64.69, 53.22] in terms of MAPE and [51.72, 36.32] by means of MAE as opposed to ARIMA and SRA, respectively. Noting that the test bed consists of four lanes in each direction, the MAEs per lane for the four steps of the two FFs are less than 1.5 vehicles, a value which is very acceptable in terms of real-world applications of advanced traffic signal operations. In particular, the MAPEs for the two FFs do not increase dramatically but instead increase gradually within +0.17% according to the extension of the future time step, which is in good agreement with the result of Chang et al. (2012b) but not with the result of Vlahogianni et al. (2005). This fact indirectly indicates that the temporal development of an intensive UTV state has intrinsic patterns, clear or not, as NPR has its origin in strong pattern recognition. In addition, FF2 is selected as the best performer to conduct a more in-depth analysis through a comparative study of the two benchmark models.

Table 2: Summary of the results

Model	Prediction	MAPE (%)	MAE (veh/90sec)	Mean rank
KNN-NPR				
FF1	t+1	7.98	5.88	1.92
	t+2	8.01	5.90	1.51
	t+3	8.10	5.95	1.53
	t+4	8.14	5.97	1.51
FF2	t+1	7.96	5.84	1.90
	t+2	8.00	5.89	1.49
	t+3	8.03	5.89	1.47
	t+4	8.12	5.95	1.49
ARIMA	t+1	22.56	16.50	3.31
SRA	t+1	17.03	12.51	2.86

The observations are directly compared with the predictions from a time-series variation assessment in Fig. 5. Despite the fact that the observed values widely and steeply vary from 39 to 117 and exhibit no stable state temporarily, the proposed model with FF2 very successfully captures the direction and variation of the future state with no time-delay response in most cases; moreover, in the few cases that it fails to capture these

characteristics, it instantly recovers and reconstructs after only one time lag. Moreover, in several extreme cases, it under- or overestimates the future state, failing to estimate the future variation accurately despite the fact that the future direction is proactively and successively captured. On the other hand, the predictions for four steps ahead are in good agreement with those of one step ahead with acceptable differences. Therefore, the prediction error between the observations and predictions for four-step aggregation is reduced 4.14% in terms of the MAPE. In this context, it can be seen that the predictions, in spite of the over- and underestimations in several extreme cases, can be effectively utilized as input variables for more proactive signal control spanning a period of four steps in practice. It is also indicated that there still another opportunity to improve prediction accuracy through combinations of different similarity measure and forecasting function.

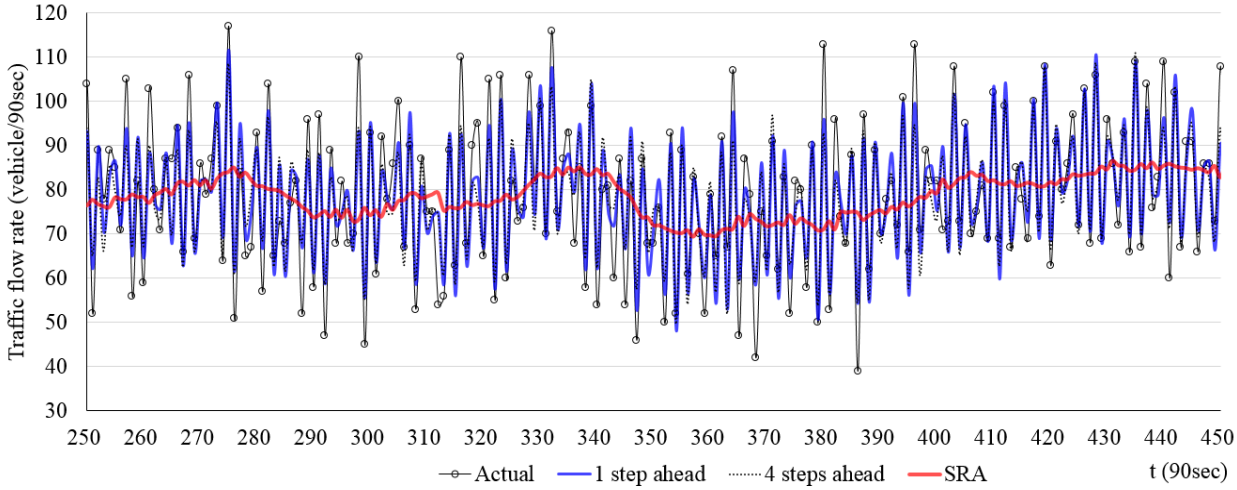


Figure 5: Time-series comparison of the observation and the prediction

The predictability of FF2 for four steps ahead and the benchmark models for one step ahead are shown as a state-variation scattergram in Fig. 6, which directly compares the actual variation against the predicted variation in each case, fully elucidating the predictability of the intensive state of the target UTV with a perfect forecast line (PFL), i.e. the actual variation = 1.0×predicted variation. In addition, the results of the perfect forecast analysis are summarized in Table 3. Concerning the benchmark models, the performances, shown in Table 3 and in Fig. 6, are not acceptable from the viewpoint of real-world practice and also appears to be closer to a forecasting failure from the standpoint of prediction modelling in spite of the one-step prediction method. The state variations of SRA are closer to the PFL than those of ARIMA, as the future state of UTV rapidly changes upwards and downwards of the estimations by SRA rather than SRA can encapsulate the direction of the future state proactively. In the case of ARIMA, it appears at the very least that the state variation of ARIMA mostly fails to explain PFL with an unacceptable linear relationship, a correlation (r) of 0.68, determination (r2) of 0.46, and a coefficient (a) of 0.32. This undesirable performance of ARIMA directly reveals that the time-series behavior of the UTV state is at least not closer to a stochastic case than a chaotic or mixed case, due to the academic fact that ARIMA has stochastic state theory as its theoretical foundation.

Regarding the proposed methodology, state variations between the observed and the predicted cases are distributed much closer to PFL in Fig. 6. The same-direction variation (+/+ and -/-) occupies up to 92.5%, whereas the opposite-direction variation (+/-) is limited to less than 7.50%, as shown in Table 3. Nevertheless, the opposite-direction variation is at most less than 3.12% if the cases that satisfy the hit rate are excluded from the opposite variation calculation. Note that a hit rate (%) within the actual variation of ±10 vehicles is used in our case. Namely, cases of perfect forecasting failure not capturing the directionality of the future state and not generating estimations within the acceptable hit rate at once are limited at most to 3.12%. For the nearness between the two state variations with regard to PFL, the cohesion of the state variations with the acceptable [correlation and determination] limit of the coefficient as of [0.96, 0.92] does not exhibit a distinguishing difference for any of the time steps. In this way, the linear relationship between the two variations explains the PFL by at least 87.68% or more based on the coefficient values. The hit rate of the proposed method reaches more than 81.04%, showing hit-rate gains (%) of more than 84.3 and 144.6 against SRA and ARIMA, respectively. This notable performance of the proposed methodology is strong academic evidence that the temporal evolution of the UTV state is closer to a chaotic or mixed case in nature, as the theoretical foundation of NPR is based on initial deterministic, i.e. chaotic, system theory. As such, the proposed model based on NPR theory has self-evident advantages in the forecast modeling of the intensive UTV state. Consequently, it is

obvious that the KNN-NPR methodology proposed in this article has strong potential for use in the multi-step forecasting of intensive signalized traffic flow with desirable performance capabilities.

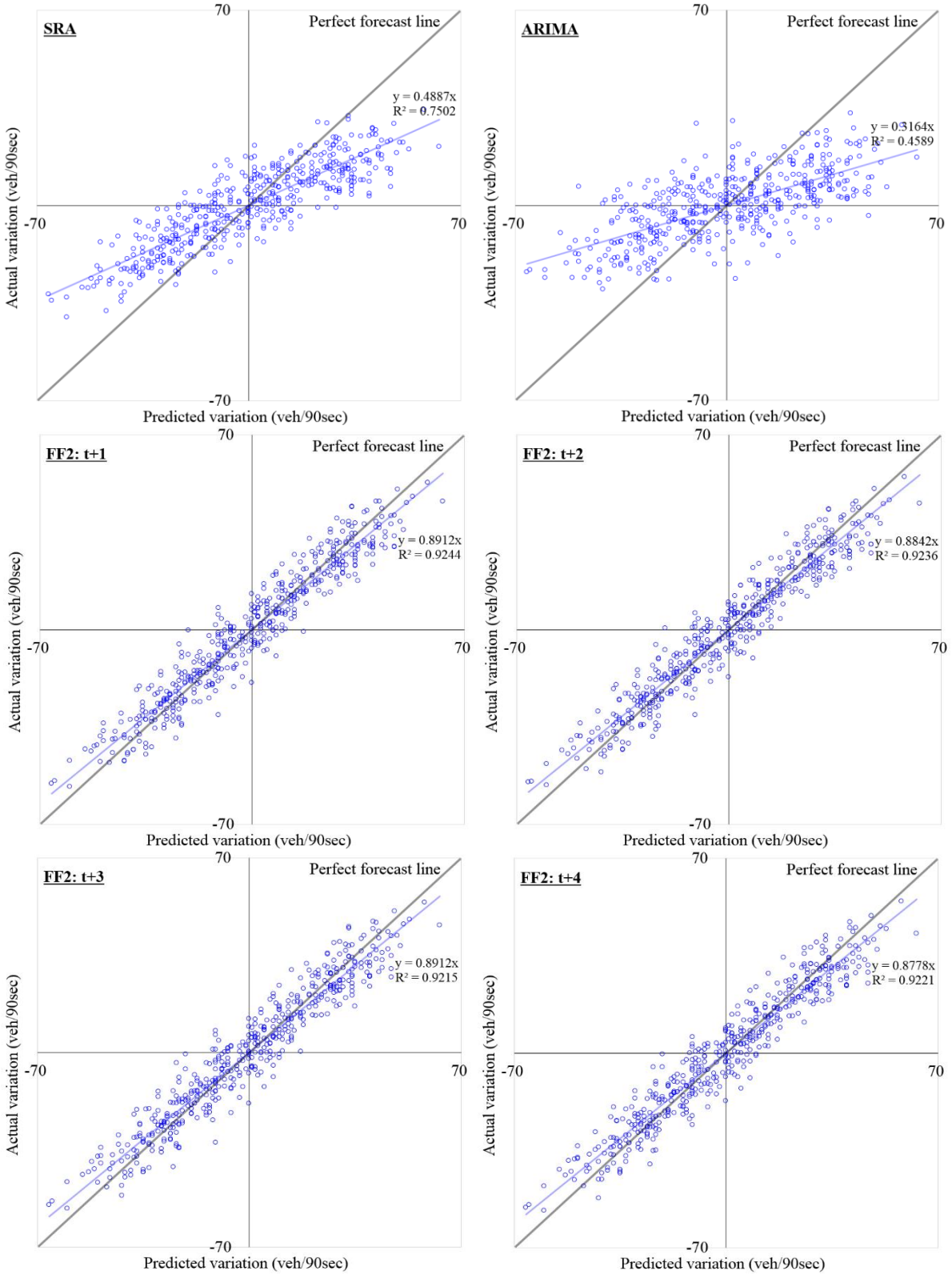


Figure 6: State-variation scattergrams for perfect predictability

Table 3: Results of the perfect forecasting analysis

Model	Step ahead	Linear relation			Directionality (%)			Hit rate (%)			
		r	r2	a	+/+	-/-	+/-	+/+	-/-	+/-	total
FF1	t+1	0.96	0.92	0.89	48.54	44.17	7.29	80.69	84.91	60.00	81.04
	t+2	0.96	0.92	0.88	48.54	44.17	7.29	81.12	85.38	57.14	81.25
	t+3	0.96	0.92	0.89	48.75	44.49	6.46	82.48	84.65	54.84	81.67
	t+4	0.96	0.92	0.88	48.75	43.95	7.29	83.33	83.41	60.00	81.67
FF2	t+1	0.96	0.92	0.89	48.54	45.21	6.25	84.12	85.71	63.33	83.54
	t+2	0.96	0.92	0.88	48.13	44.38	7.49	81.82	84.51	58.33	81.25
	t+3	0.96	0.92	0.89	48.33	44.17	7.50	83.62	85.38	66.67	83.13
	t+4	0.96	0.92	0.88	48.33	44.17	7.50	84.91	82.08	63.89	82.08
ARIMA	t+1	0.68	0.46	0.32	40.83	36.04	23.13	40.82	38.15	11.71	33.13
SRA	t+1	0.87	0.75	0.49	46.46	40.83	12.71	50.22	43.88	21.31	43.96

Note 1: r, r2, and a denote the correlation, determination, and coefficient, respectively.
Note 2: +/+, -/-, and +/- represent the same direction in the first quadrant, the same direction in the third quadrant, and the opposite direction in the second or fourth quadrant in Fig. 3, respectively.

V. Conclusions

Despite the considerable achievement in ITS forecasting modeling, multiple forecasting of intensive signalized traffic flow remains a major challenge to be addressed successfully, even though few notable studies of one-step prediction of signalized traffic volume based on sophisticated models have been reported in the literature. In addition, it is necessary to pioneer a new area of ITS forecasting in the era of big data, switching prediction approaches from artificial modeling by modelers to data-based knowledge discovery, which provides another opportunity to address chronic issues such as the uncertainty problem successfully. To speak strictly from the viewpoints of traffic practitioners, there is no available model with which to estimate the future multiple states of intensive signalized traffic with acceptable levels of prediction error in practice. In this context, a multi-period forecasting method for the prediction of intensive signalized traffic flow to handle the above problems effectively is proposed in this research.

The proposed methodology is developed based on a data-driven KNN-NPR approach. The methodology is designed with a simplified data structure suitable for (advanced) data management systems, which is one of the major advantages from the perspectives of field staff, who have sufficient experiences about the analysis and management of ITS data. In the validation study with real-world signalized traffic volume data, the proposed model wholly outperformed the two benchmark models, i.e. the simple rolling average and ARIMA, in terms of the prediction difference and perfect predictability. Specifically, the performances of the proposed model for four steps ahead were clearly superior to those of the comparative models for one step ahead. Despite the use of multi-step forecasting, the proposed methodology efficaciously captures the directionality and variation of the future state with no time delay. It instantly reconstructs the correct state after only one time lag when it infrequently fails in capturing the directionality in advance. Based on the results, it can be seen that the temporal evolution of the signalized traffic flow state can be constructed with simplified NPR approaches and can be predicted on a multi-step horizon to a degree of acceptable prediction accuracy in keeping with the motorway traffic flow. In addition, there remain other opportunities to improve on the performance of the proposed methodology by using a more sensitive nearness measure, a more sophisticated forecasting function, a combination of these measures, or a combination of NPR and advanced models in the future. Furthermore, future investigations should attempt to reduce the execution time of data-driven NPR approaches to the level of high execution-time models.

References

[1]. Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46, 175-185.
[2]. Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., and Sun, J. (2016). A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. Transportation Research Part C, 10.1016/j.trc.2015.11.002, 62, 21-34.

[3]. 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.

[4]. Chang, H., Seong, J.N., Lee, Y., and Yoon, B. (2011). Dynamic freeway path travel time prediction based on nonparametric regression approach using dedicated short-range communications data. Proc. 90th Annual Meeting of Transportation Research Board, Washington, DC, 2011.

[5]. Chang, H., Park, D., Lee, Y., and Yoon, B. (2012a). Multiple time period imputation technique for multiple missing traffic variables: nonparametric regression approach. Canadian Journal of Civil Engineering, 39, 448-459.

[6]. Chang, H., Lee, Y., Yoon, B., and Baek, S. (2012b). Dynamic near-term traffic flow prediction: system-oriented approach based on past experiences. IET Intelligent Transport Systems, 6(3), 292-305.

[7]. Clark, S. (2003). Traffic predicting using multivariate nonparametric regression. Journal of Transportation Engineering, 129(2), 161-168.

[8]. Davis, G.A. and Nihan, N.L., (1991). Nonparametric regression and short-term freeway traffic forecasting. Journal of Transportation Engineering, 117, 178-188.

[9]. Dernas, M., Placzek, B., Porwik, P. and Pamula, T. (2015). Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction. IET Intelligent Transport System, 10.1049/iet-its.2013.0164, 9(3), 264-274.

[10]. Guegan, D., and Leroux J. (2009). Forecasting chaotic systems: The role of local Lyapunov exponents. Chaos, Solitons & Fractals, 41, 2401-2404.

[11]. Hamad, M.M., AI-Masaeid, H.R., and Said, Z.M.B. (1995). Short-term prediction of traffic volume in urban arterials. Journal of Transportation Engineering, 121(3), 249-254.

[12]. 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.

[13]. Karlsson, M., and Yakowitz, S. (1987). Rainfall-runoff forecasting methods, old and new. Stochastic Hydrology and Hydraulics, 1, 303-318.

[14]. Mori, U., Mendiburu, A., Alvarez, M., and Lozano, J.A. (2015). A review of travel time estimation and forecasting for advanced traveler information system. Transportmetrica A: Transport Science, 11(2), 119-157.

[15]. Mulhern, F.J., and Caprara, R.J. (1994). A nearest neighbor model for forecasting market response. International Journal of Forecasting, 10, 191-207.

[16]. Okutani, I., and Stephanedes, Y.J. (1984). Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B, 18B(1), 1-11.

[17]. Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S. (1980). Geometry from a time series. Physical Review Letters, 45, 712-716

[18]. Smith, B.L., and Demetsky, M.J. (1996). Multiple-interval freeway traffic flow forecasting. Transportation Research Record, 1554, 136-141.

[19]. Smith, B.L., and Demetsky, M.J. (1997). Traffic flow forecasting: comparison of modeling approaches. Journal of Transportation Engineering, 123(4), 261-266.

[20]. 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.

[21]. 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.

[22]. Stathopoulos, A., Dimitriou, L., and Tsekeris, T. (2008). Fuzzy modeling approach for combined forecasting of urban traffic flow. Computer-Aided Civil and Infrastructure Engineering, 23, 521-535.

[23]. Stathopoulos, A., and Karlaftis, M.G. (2001). Temporal and spatial variations of real-time traffic data in urban areas. Transportation Research Record, 1768, 135-140.

[24]. Stathopoulos, A., and Karlaftis, M.G. (2003). A multivariate state-space approach for urban traffic flow modeling and prediction. Transportation Research Part C, 11(2), 121-135.

[25]. 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.

[26]. Takens, F. (1981). Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, 898, 366-381.

[27]. 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.

[28]. 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.

[29]. 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.

[30]. Vlahogianni, E.I., Karlaftis, M.G., and Golias, J.C. (2014). Short-term traffic forecasting: Where we are and where we're going. Transportation Research Part C, 43(1), 3-19.

[31]. Yin, H., Wong, S.C., Xu., J, and Wong, C.K. (2002). Urban traffic flow prediction using a fuzzy-neural approach. Transportation Research Part C, 10, 85-98.

[32]. Yoon, B. and Chang, H. (2014). Potentialities of data-driven nonparametric regression in urban signalized traffic flow forecasting. Journal of Transportation Engineering, 10.1061/(ASCE)TE.1943-5436.0000662, 140(7), 04014027.