공학박사 학위청구논문

A Machine Learning Approach for Freeway Speed Prediction

고속도로 속도예측을 위한 기계학습 접근법

2018년 2월

서울대학교 대학원

공과대학 건설환경공학부

최 윤 영

A Machine Learning Approach for Freeway Speed Prediction

지도교수 고 승 영

이 논문을 공학박사 학위논문으로 제출함.
2017년 11월

서울대학교 대학원
공과대학 건설환경공학부
최 윤 영

최윤영의 박사 학위논문을 인준함.
2017년 12월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Abstract

A Machine Learning Approach for Freeway Speed Prediction

Choi, Yoon-Young
Department of Civil and Environmental Engineering
The Graduate School
Seoul National University

Prediction of freeway traffic speed can be used for predictive traffic management to improve the quality of the intelligent transportation system. The data-driven prediction is widely used due to its predictive capability. Recently, the non - parametric method using machine learning shows excellent predictive capability. In these methods, the feature extraction or selection is used to mitigate the overfitting and reflect the congestion mechanism. Although this nonparametric approach can be used as advanced traveler information system due to its excellent capability, it cannot provide any information on the congestion mechanism. Lack of information makes it difficult to establish a strategy for use in operational management. This study proposes a highway speed prediction model based on machine learning approach with a feature selection that provides both high predictive performance and interpretation of traffic flow characteristics. To do this, a supervised feature selection is applied using principal component analysis (PCA) based variable grouping and ordering and support vector machine (SVM) based variable selection. Varimax rotation is also applied to obtain the simple structure. In the variable ranking, the variables in the PC are ranked by using the nonlinear correlation coefficient

which implies the predictive capability in the machine learning model. The cross-correlation coefficients were used in this study. With this grouped and ranked variables, the variables are selected by the forward selection method. The machine learning regression model in this study is SVM regression which has excellent generalization performance and low computational cost.

Empirical data evaluation was implemented based on the several month's data of Kyungbu freeway in Korea and the interstate (I-880) freeway in the United States. Comparing other approaches, the proposed feature selection approach well captured the characteristics of traffic flow among spatiotemporal variables. In particular, the feature selection performance is somewhat better than that of the artificial neural network feature extraction model, stacked auto-encoder, and the ensemble learning model, random forest. The vector space of the PCA is transformed into the traffic phase diagram between two spatiotemporal variables to obtain the implication of proposed approach in traffic engineering area. Based on the traffic phase interpretation, principal components with some loading of dependent variable can explain the propagation of traffic state. The proposed approach captures the propagation of traffic state well according to prediction step. The proposed approach would be used to establish strategies for avoiding congestion or preventing rear-end accidents because it has advantages in the multi-step prediction on congested areas and in identifying the congestion mechanism.

Keywords: Freeway Speed Prediction, Support Vector Machine, Principal Component Analysis, Feature Selection, Propagation of traffic state

Student Number: 2013-30982

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

Intelligent transportation systems (ITS) are widely used around the world and contributing to traffic management and comprehensive traveler information service. The traffic speed prediction of the freeway can be used to manage congestion or crash in the congested area. The key to ITS services is to provide accurate, timely and useful information to travelers and transportation professionals. To this end, the capability of traffic prediction under dynamic and complex traffic condition is most important and it enables proactive management or changes of traveler's decision based on the predictive information. It can be implemented by using advanced traveler information system (ATIS) or by performing an advanced traffic management system (AMTS) strategy.

Traffic speed can be predicted based on the theoretical model such as simulation. The simulation-based approach has the advantage of being able to explain the causal relationship, but it requires a lot of effort to make accurate predictions, and there are limitations to the accuracy due to the limitations of the simplified input parameters. The parametric or nonparametric data-driven approach has the advantage of accuracy. Especially the nonparametric based prediction technique shows high prediction performance due to the rapid advancement of traffic detection and machine learning technology.

The nonparametric based speed prediction method requires mitigating the over-fitting and capturing the appropriate traffic flow characteristics to achieve high prediction performance. A problem of the curse of dimensionality may arise where the dimension of the input data is high in the machine learning model. This indicates that the volume of space increases very quickly due to the increase in dimension, which makes the available data thinner. This adversely affects the algorithm, induces overfitting, and dramatically increases memory storage requirements and computational expense. Dimensionality reduction is required to solve this problem, and it is classified into feature extraction and feature selection. Feature extraction projects the original higher dimensional feature space into a new lower feature space. Feature selection, on the other hand, directly selects a subset of the relevant features for use in model construction. In feature extraction, additional analysis is difficult because the physical meaning of the original feature cannot be obtained in the transformed space. In contrast, feature selection preserves the physical meaning and provides better readability and interpretability, making it more useful for practical analysis. (Li et al., 2016). Because of the limitation of feature extraction, nonparametric based traffic speed prediction can be utilized only providing information and is difficult to use it in transportation management strategy. Therefore, it is necessary to capture the feature and interpret it as traffic flow characteristics. Traffic characteristics information from captured feature can contribute to establishing a traffic management strategy.

The variable selection can be implemented by the various theoretical methods, but there is a limit to providing integrated output due to individual evaluation of variables. For example, similar variables are evaluated identically and it causes the redundancy. On the other hand, there may be a significant variable when used with other information. A ranking method that takes into account the mutual relationships between variables is needed (Guyon et al., 2003).

1.2 Objectives and Scope

The objectives of this study are to propose a supervised feature selection approach that can provide both high predictive performance and interpretations of traffic flow characteristics and to derive the implication in the highway speed prediction by using it. The proposed approach consists of two steps, the first step makes a ranking of variables including mutual relationship among them and the second step selects the variables with support vector regression (SVR) based forward selection method.

The first step is principal component analysis (PCA) based variables ranking. The PCA is performed including dependent variable, the prediction target. Varimax rotation is also applied to obtain the simple structure that can easily interpret the variables space. The variables are assigned to the principal components (PCs) by using the maximum loading of each variable. The variables are divided into groups that are related to statistically independent PCs. The loading is the linear correlation between PC and a variable. Then, the ranking is determined in two ways, i.e., ranking between PC groups and ranking between variables in the same PC. The loading of the dependent variables is used to rank the PCs. This ranking criterion reflects the strength of the link between dependent variable and PC and can indicate the contribution of PCs in prediction. Next, the ranking of variables in PC is determined by using the auxiliary criterion of non-linear correlation. This auxiliary criterion is employed for overcoming the linear manner of PCA because the machine

learning based regression finds the non-linear structure. This study employed cross-correlation coefficients as a criterion of ranking of variables in each PCs.

The second step is Support vector machine (SVM) based variable selection and the forward selection is employed. The forward selection is a kind of wrapper method in supervised feature selection. The variables are progressively incorporated into larger and larger subsets. The selected feature subset is evaluated by predictive performance. Then, a new selected variable included or not according to the evaluation result. The wrapper method works iteratively until some stopping criteria are satisfied or the desired learning performance is obtained. This study employed performance measure as the root mean squared error (RMSE) and cross-validation evaluation which splits the data set into the training set and validation set. The SVR is employed for this forward selection because it has excellent generalization performance and low computational cost among machine learning based regression model.

The artificial neural network model is the most popular nonlinear model, but it has three major drawbacks: overfitting, local minima, and heavy computational cost (Vapink, 1995; Cristianini and Shawe-Taylor, 2000). An SVM is a machine learning method that can overcome the shortcomings of artificial neural networks. Compared to traditional artificial neural networks, SVM has the advantage of generalizing ability and having the global optimal solution (Vapnik, 1995; Vapnik et al., 1999). This is because SVM uses the principle of minimizing structural risk, unlike the neural network which adopts

the empirical risk minimization principle. Unlike the minimization of empirical risk which minimizes the training error and finds the local optimum, the principle of minimizing the structural risk is to minimize the upper bound of the generalization error. SVM is also suitable for complex systems and has robust performance when processing corrupted data (Müller et al., 1997; Gunn, 1998; Müller et al., 1999). Besides, the SVM model is suitable for real-time use and periodic model re-estimation due to its low computational cost and is easily deployed by users.

To find out the physical meaning of selected feature in terms of traffic flow, the vector space of PCA and result of selected variables were explored. The vector space of the PCA is transformed into the traffic diagram between two spatiotemporal variables. The traffic phase diagram by the proposed approach was discussed in terms of traffic flow characteristics. The result of variables selection was compared with other selection methods to find out the consistency of the discussed implication.

The remaining part of this study consists of literature review, methodology, empirical data evaluation, implication for traffic analysis, and conclusion. In the literature review, a simulation-based model and a data-driven model for traffic speed prediction were reviewed. Additionally, feature extraction and selection techniques in predicting traffic information were extensively reviewed. The methodology section introduced proposed approach including the theoretical description of PCA, Varimax rotation, SVM model

and variable selection technique. Also, random forest (RF), artificial neural network (ANN), and stacked auto-encoder (SAE) were briefly described for comparison and evaluation. In Empirical data evaluation, case studies of Korea highway and US highway were described. The proposed approach evaluated in terms of predictive performance, and computation cost by comparing with other approaches. In the implication for traffic analysis, traffic phase analysis was carried out to interpret the traffic flow characteristics from the PC used, and the consistency was examined using selected variables. Lastly, conclusions and limitations of the study were presented.

Chapter 2. Literature Review

2.1 Traffic Flow Based Model

The simulation-based model uses some traffic flow theory models and focuses on reproducing the traffic situation in the future time interval and measuring the traffic speed from the predicted state. It has the advantage of representing and applying important components in traffic modeling including geometric features and traffic signal. Although the traffic conditions and the traffic flow should be very similar, the simulation model uses a limited number of variables to simplify the actual situation. As a result, it is difficult to reproduce the actual traffic flow phenomenon.

Macroscopic models use equations from fluid flow theory to simulate the flow, density and mean speed for the future times. The simulation results come from the mathematical relations between traffic variables including traffic speed. METANET (Papageorgiou et al., 2010) is a representative example. However, it is not useful to predict urban area.

In the microscopic model, the behavior of an individual vehicle is simulated considering factors such as inter-vehicle interaction, driver behavior, and lane change. It uses car-following models and cellular automaton models. The Car-following model is time continuous ordinary differential equations that indicate the position and trajectory of the following vehicle according to the preceding vehicle. The cellular automaton is a simple model that separates

roads into small cells and moves the vehicles by cell according to predefined rules such as lane changing and acceleration. There is additional work to predict the OD matrix or intersection turn traffic. Examples are CORSIM (Halati et al., 1997), PARAMICS (Cameron and Duncan, 1996), INTEGRATION (van Aerde and Rakha, 2010).

Mesoscopic models combine the functions of macroscopic and microscopic models. It simulates individual vehicles but explains behavior and interactions based on general macro relationships. It is mostly used in large networks where it is infeasible to model microscopic models. DynaMIT (Ben-Akiva et al., 1998), the vehicle is shown individually, DynaSMART (Jayakrishnan et al., 1994) is determined the link speed by the speed-density relationship. Mesoscopic models are primarily used for assessment and traffic management, but may also be useful for ATIS.

Table 2.1 Literature review of traffic flow based model

Category	**Pros and cons**	**Simulation**
Macroscopic	Good for large network, cannot directly calculate, ignore detailed behavior	METANET
Microscopic	Details available, OD prediction is necessary, calculation cost is large	CORSIM PARAMICS INTEGRATION
Mesoscopic	More detailed than Macro, faster than Micro, disadvantages of both Macro and Micro	DynaMIT DynaSMART

2.2 Data Based Model

The data-driven model does not use the traffic flow theory, but rather it is determined from the data using statistical and machine learning techniques. The disadvantages are that large amounts of data are needed and the results are strongly linked to specific learning areas. As a result, there is a limit to transferability.

2.2.1 Parametric model

In the parametric method, the relationship between the parameters is set in advance in the model and the coefficient of the parameter is determined from the data. Overall, the parametric model is easy to understand but has a limit of low prediction performance.

Time series models including ARMA, ARIMA, and SARIMA have been studied from earlier days (Ahmed and Cook, 1979; Davis et al., 1990; Hamed et al., 1995). There is a limitation in that it is vulnerable to a state change due to mainly using single detection information. When linearity and normality conditions are met or assumed, they can be solved using a Kalman filter. (Okutani and Stephanedes, 1984; Whittaker et al., 1997; Stathopoulos and Karlaftis, 2003). There is an advantage to reflect the multivariate nature, but the normality assumption can be a disadvantage.

2.2.2 Non-parametric model

The nonparametric model is a technique for finding empirically suitable parameters and structures from infinite types without a predefined model structure. There is a disadvantage in that it has excellent predictive power but requires more data and is difficult to interpret. The nonparametric prediction of traffic speed attracted little attention although most of non-parametric traffic flow prediction focused on the travel time or traffic volume prediction.

Dougherty and Cobbett (1997) performed prediction of flow, speed and occupancy in the region of the Netherlands using back-propagation neural networks. Flow and occupancy predictions were effective, but speed prediction was not as good as performance and it was claimed to be due to low-speed vehicles in low traffic volume. It is also claimed to be impractical for implementation due to the vastness of the input data. Ishak and Alecsandru (2004) proposed an approach to optimize predictive performance using multiple artificial neural network topologies in different networks and traffic conditions. They evaluated the short-term speed prediction performance of the mix of neural network topologies according to different input settings and various prediction horizons (from 5 to 20 minutes). Optimal settings were determined regarding observed traffic conditions at upstream and downstream locations. The proposed approach showed better prediction performance than naive and heuristic approaches. Long-term predictions have also more dependent on the configuration of the long-term memory.

Ma et al. (2015) used the long short-term neural network to estimate the speed of the Beijing area. Long short-term neural network can overcome the problem of back-propagating error decay through memory blocks, and thus exhibits superior temporal prediction with long temporal dependency. The proposed model has better performance and stability than other dynamic neural networks.

Sun et al. (2003) proposed the local linear regression using the 32-day traffic-speed data in Houston, at 5-min intervals. The proposed model showed better performance than the k-nearest neighbor and kernel smoothing methods.

Yildirim and Çataltepe (2008) predicted the speed of remote traffic microwave sensors at different locations in Istanbul. From 5 minutes to 1 hour, the SVM model and k-nearest neighbor method were compared. Variable selection was used to improve the model. In particular, it was suggested that it would be advantageous to use sensors with high correlation. In this study, SVM produced better results than the k- nearest neighbor. Yao et al. (2017) proposed SVM model composed of spatial and temporal parameters. The SVM model using spatiotemporal parameters showed better performance than the historical model, the k-nn model, and the ANN model by using GPS data of Foshan city taxi in China. It was also advantageous for relatively multi-step prediction.

Table 2.2 Literature review of data-driven model

Category	**Method**	**Authors and year**
Parametric	ARMA, ARIMA, SARIMA	Ahmed and Cook (1979), Davis et al. (1990), Hamed et al. (1995)
	Kalman filter	Okutani and Stephanedes (1984), Whittaker et al. (1997), Stathopoulos and Karlaftis (2003)
Nonparametric	ANN	Dougherty and Cobbett (1997), Ishak and Alecsandru (2004), Ma et al. (2015)
	Local regression	Sun et al. (2003)
	SVM	Yildirim and Çataltepe (2008), Yao et al. (2017)

2.2.3 Feature selection or extraction in non-parametric

In the nonparametric model, review of feature selection or extraction was extended to the general traffic prediction due to the scarcity of related literature on traffic speed prediction.

Dia (2001) used PCA for unsupervised linear feature extraction and performed dimensional reduction by performing supervised non-linear classification on features. A time-lag recurrent network model was used to improve the prediction performance of traffic speed in Australia. Yildirimoglu and Geroliminis (2013) were used PCA to remove noise and to cluster the similar traffic patterns. They used Gaussian Mixture Model to predict the travel time in Californian freeway.

Abdulhai et al. (1999) used the time delay neural network to predict the flow and occupancy of California roads. The genetic algorithm (GA) was used to optimize the spatial and temporal input space, prediction horizon and data resolution. Zhong et al. (2005) applied GA to select the input variable. They designed both time delay neural network models as well as locally weighted regression models to predict short-term traffic for two rural roads in Alberta, Canada. They claimed that locally weighted regression models are faster and perform better. Vlahogianni et al. (2007) used GA for the temporal optimization of the input windows in each time delay neural network to overcome the problem of having a fixed line memory. They forecasted the short-term traffic volume in multiple locations of an urban signalized arterial

focusing on incorporating temporal and spatial volume characteristics.

Lv et al. (2015) performed unsupervised feature extraction using a stacked auto-encoder, and the importance of the variables was determined using the sparsity term. They concluded that the proposed model helped to capture generic traffic flow features and characterize spatial-temporal correlations in traffic flow prediction. However, as the hidden layer increases, the performance is not improved.

Li and Chen (2014) used the classification and regression tree to select the variables of travel time prediction and predict Taiwan's freeway travel time using ANN. The random forest is an ensemble learning method that uses a lot of decision trees using bagging (bootstrap aggregating) to extract multiple subsamples from a raw sample. The final result is to use the average for regression, or vote for classification. In variable selection, the variable importance is provided based on the reduction of the average accuracy of the variables through inclusion and omission processes during the bagging process. It was used to predict the real-time crash risk (Xu et al., 2013; Yu and Abdel-Aty, 2013; Shi and Abdel-Aty, 2015).

The grouping can remove variables that are less similar. Lee (2009) applied the ANN model for travel time prediction by reducing the dimension of input variables through k-mean clustering. Chen et al. (2010) used the fuzzy clustering technique, rough set, to reduce the dimension of the input variables and predict the urban travel time using the SVM model.

Table 2.3 Literature review of dimensionality reduction model

Concept	**Method**	**Authors and year**
Dimensionality reduction	PCA (Component)	Dia (2001), Yildirimoglu and Geroliminis (2013)
	Genetic Algorithm (GA)	Abdulhai et al. (1999), Zhong et al. (2005), Vlahogianni et al. (2007)
	Sparsity and auto-encoder	Lv et al. (2015)
	Decision tree	Classification and regression tree (Li and Chen, 2014), Random forest (Xu et al., 2013; Yu and Abdel-Aty, 2013; Shi and Abdel-Aty, 2015)
	Clustering	K-mean (Lee, 2009), Rough set (Chen et al., 2010)

2.3 Review Result and Study Direction

Most studies claimed that their machine learning model and efforts related to the improvement of the model was effective, but the strict comparison is difficult. This is because they contain a limitation of a specific concept in a specific area and not compare all of the different methods, and does do not use the same comparison criterion (Vlahogianni et al., 2004). Additionally, the model optimization was only implemented on their model, but the comparison model was a general model.

The ANN based models are the most popular nonlinear models, but they have three major drawbacks: overfitting, local minima, and heavy computational cost. An SVM method is a machine learning method that overcomes the shortcomings of artificial neural networks. Compared to traditional artificial neural networks, SVM has the advantage of excellent generalization ability and global optimal solution. This is because SVM uses the principle of minimizing structural risk, unlike neural network which adopts the principle of minimizing empirical risk. Unlike minimizing training errors and minimizing the empirical risk of finding local optima, the principle of minimizing the structural risk minimizes the generalization error. The results of SVM guarantee the global minima. SVM is also suitable for complex systems and has robust performance when processing corrupted data. Also, the SVM model is suitable for real-time use and capability test for prediction due to its

low computational cost and is easily deployable for users.

There are some limitations of feature selection or extraction approaches. The use of principal component has limitations in that it is difficult to interpret the features used because the component is the linear combinations of variables. The GA based research has a characteristic of probabilistic convergence and there are limitations in expecting consistent results and it is difficult to link them to interpretation. Sparsity, which uses a penalty term for the weight of a variable, is mainly used for the deep neural network technique of traffic volume prediction, but there are limitations in interpretation due to stochastic convergence and weight of variable. In the case of random forests that provide an importance for variable selection, variables with similar roles are likely to be similarly selected even if the importance is high since the relationship between variables is not considered. The clustering has the limitation in considering only similarities while the influence of variables contained in other clusters is also significant. In this study, the interpretations of selected PC or variables is the key for contribution and it is achieved by describing the process of approach, evaluating based on empirical data and finding the implications for traffic analysis. Also, the proposed approach is deterministic and considering the relationship between the variables.

Chapter 3. Method

3.1 Principal Component Analysis (PCA)

3.1.1 Basics of PCA

(1) Introduction

The PCA is a method of factor analysis. Factor analysis is a multivariate analysis technique that explains how various variables are connected to each other and explains it using common factors. The PCA is a technique that uses mathematical orthogonal transformation by the spectral decomposition of covariance matrices or correlation matrices using total variance (systematic variance). It is aimed to simplify the dimension and summarize multidimensional variables and to analyze the complex structure among the variables. The PCA derive independent, artificial variables called PC that is a linear transformation of variables. The importance of a PC is determined by the magnitude of its total variance. If a large part of the total variance is included by using some principal components, it is possible to reduce the size to minimize the loss of information.

In the 1930s, Hotelling (1933, 1936) named low-level independent factors as components and named PCA because it is sequentially maximizing the contribution of each component to the variance of the original variables. The PCA is a multivariate technique that transforms correlated variables that is

hard to interpret the structure into a few independent and conceptually meaningful principal components. It can be used for exploratory studies of multivariate data, structural simplification or summarization of data by dimensional reduction, sequential construction of independent components, identification of variables of dependent relation, evaluation or comparison of distribution types, a search of singular values or clusters, data fitting and model construction. The systematic variance is divided into common factor variance and specific variance, and on the other hand, common-factor methods only use the common variance. It is only possible if the column or row is a population.

When the overall average of the data set is 0, the first principal component w_1 of the data set X is defined as follows.

$$w_1 = \arg\max_{\|w\|=1} E\{(w^T X)^2\}$$

To find the k-th principal component, use $\widehat{X_k}$ that eliminated $k-1$ principal components in X.

$$\widehat{X_k} = X - \sum_{i=1}^{k-1} w_i w_i^T X$$

$$w_k = \arg\max_{\|w\|=1} E\{(w^T \widehat{X_k})^2\}$$

In the PCA, samples of higher dimensional space transform into samples of lower dimensional space (principal component) without linear

correlation by using orthogonal transformation. The PCA linearly transforms the data into a new coordinate system so that the data is mapped onto one axis. This transformation is defined by the fact that the first principal component has the greatest variance and the subsequent principal components have the greatest variance under the constraint that they are orthogonal to the previous principal components. The axis with the largest variance is the first principal component and the axis with the second largest is the second principal component. Thus, various applications are possible by decomposing into the components that best represent the sample differences. Important components are orthogonal because they are eigenvectors of the symmetric (covariance or correlation) matrix.

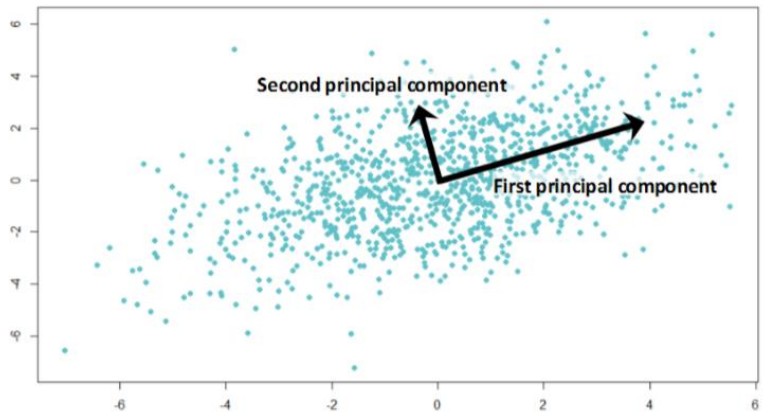


Figure 3.1 Concept of PCA

(Source: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/)

(2) Mathematical approach

The vector h_i projecting the original data ((z_i)) for the unit vector w with size 1 in two dimensions is given by $h_i = (z_i \cdot w)w = ({z_i}^T w)w$. The direction of the axis to maximize the variance of projection h_i is equivalent to the problem of maximization of $(z_i \cdot w)$ which is a variance of h_i. If the values are centered, the variance of the projected value, σ_w^2, is as follow.

$$\sigma_w^2 = \frac{1}{n}\sum_i (z_i \cdot w)^2 - (\frac{1}{n}\sum_i (z_i \cdot w))^2$$

This is developed as follows using the covariance matrix C.

$$\frac{1}{n}\sum_i (z_i \cdot w)^2 = \frac{1}{n}(Zw)^T Zw = \frac{1}{n} w^T Z^T Z w = w^T \frac{Z^T Z}{n} w$$

$$= w^T C\, w$$

This is a maximization problem in the unit vector w with a constraint of $w^T w = 1$. When the problem is solved by the Lagrangian multiplier method, the objective function (Q) is as follow.

$$Q = w^T C\, w - \lambda(w^T w - 1)$$

A point with $\frac{\partial \mathrm{Q}}{\partial \mathrm{w}} = 0$ is as follow.

$$\frac{\partial \mathrm{Q}}{\partial \mathrm{w}} = 2Cw - 2\lambda w = 0,\ Cw = \lambda w$$

This is equivalent to finding eigenvalues λ and eigenvectors w for the covariance matrix C. Since $\sigma_w^2 = w^T C\, w = w^T \lambda\, w = \lambda$, it can be seen that the eigenvalue represents the magnitude of variance.

(3) Eigenvalue decomposition

The PCA is based on the eigenvalue decomposition. For the matrix A, the scalar λ and the column vector v that satisfy the following equation are called eigenvalues and eigenvectors, respectively.

$$Av = \lambda v$$

This equation is developed using the identity matrix, E.

$$(A - \lambda E)v = 0$$

An eigenvector exists if there is no inverse matrix of $A - \lambda E$. If there is an inverse matrix, $v = 0$. The eigenvalue decomposition is a relationship using both the matrix P, which uses eigenvectors as column vectors and matrix Λ that uses eigenvalues as diagonal components. The values are sorted in descending order of magnitude, v_1 is the eigenvector of the first principal

component.

$$AP = \Lambda P$$

$$A[v_1, v_2, , \dots, v_n] = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix} [v_1, v_2, \dots, v_n]$$

If the matrix is orthogonal matrix with $A = P\Lambda P^{-1}$, $A = P\Lambda P^{\mathrm{T}}$. The symmetric matrix can always be applied the eigenvalue decomposition and diagonalized to an orthogonal matrix, orthogonally diagonalizable matrix. In general, orthogonal diagonalization is possible when a correlation matrix or a covariance matrix is used.

The determinant, k-squared, and inverse of a matrix can be easily expressed using eigenvalue decomposition.

$$\det(A) = \det(P\Lambda P^{-1}) = \det(\Lambda)$$

$$A^k = (P\Lambda P^{-1})^k = P\Lambda^k P^{-1}$$

$$A^{-1} = (P\Lambda P^{-1})^{-1} = P\Lambda^{-1} P^{-1}$$

Eigenvalue decomposition enables when matrix A has a linearly independent eigenvector. Linearly independent means that it cannot be represented by a linear combination of other vectors. The eigenvalues are unique for a matrix, but since the eigenvector matrix is not unique, a change of basis is possible to facilitate the interpretation.

(4) Application

The PCA can be applied using a covariance matrix, a scatterplot matrix, a variance-covariance matrix or a correlation matrix. Since the PCA is not scaled invariant, the scale can be unified when using the correlation matrix. The PCA can be defined as having a descending order of eigenvalues diagonal elements and non-negative eigenvalue since the covariance matrix $X^T X$ is a non-negative definite matrix. As a consequence, the first principal component obtained by the eigenvalue decomposition of the covariance matrix $X^T X$ is the axis that maximizes the variance of the projected data. If k principal components are selected from the p-dimensional eigenvalue decomposition, the data is projected into the k-dimensional subspace. The ratio of the sum of the eigenvalue of k principal components and total eigenvalue represents the ratio of variance. The PCs can be selected to the point where the eigenvalue changes suddenly by using scree plot, or PCs with eigenvalues greater than 1, which is the variance of one variable, can be used.

$$R^2 = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=j}^{p} \lambda_j$$

Factor loading shows a simple correlation between variables and factors in the un-rotated factor matrix. The variables that explained a component is

identified. The variables that have the greatest relationship with a component is determined by the factor loading. The squared sum of loadings of variables in a component indicates the variance of that factor explained by variables. The squared sum of loadings of variables in a rotated factor matrix cannot be interpreted in the same sense.

3.1.2 Varimax Rotation

The factor loading which indicates the variables contribute to the PC is used to interpret the PCA result. The sum of the squares of the loading of each variable in a PC is the eigenvalue of the PC. If the symmetric matrix, covariance matrix or the correlation matrix, is used for PCA, the variables can be grouped independently because the PCs are independent each other. However, since one variable is not loaded on one PC, the simple structure for easy handling can be obtained using the appropriate rotation. A simple structure means that each variable has a high loading on one factor and very low loadings on the other factor. The simple structure is an ideal state in which one variable can exert influence on only one factor. This is eligible because the eigenvectors of eigenvalue decomposition are not unique. There are rotation methods that maintain orthogonality such as Varimax or that does not maintain orthogonality such as Oblimin. The choice of the rotation method by their research purpose is recommended.

The Varimax is used an orthogonal basis to simplify the subspace without changing the actual coordinate system. It is the rotation used when the PCA is hard to analyze due to the dense result. By using Varimax rotation, the basis is transformed and the variables are assigned to a PC with the highest loading. It is eligible for grouping of variables based on independence. The Varimax uses the objective function that allows a given variable to be loaded heavily on a single PC and to maximize the overall variance. The objective function is as follows.

$$P\ \text{Varimax} = \arg\max_{P}\{\frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{p}(\Lambda P)_{ij}^{4} - \sum_{j=1}^{k}\frac{1}{p}\sum_{i=1}^{p}(\Lambda P)_{ij}^{2}\}^{2}$$

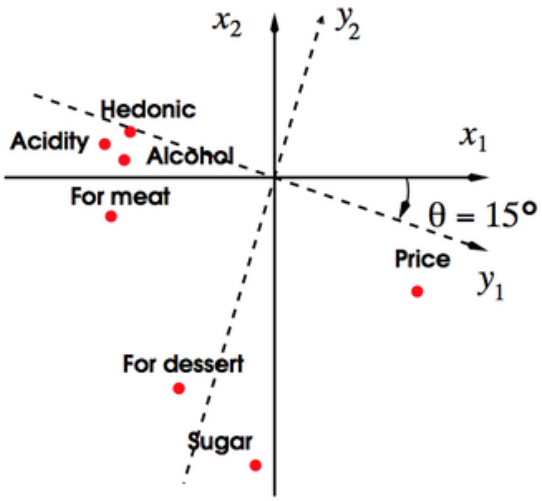


Figure 3.2 Example of Varimax rotation

(Source: https://www.mailman.columbia.edu/research/population-health-methods/exploratory-factor-analysis)

3.2 PCA Based Supervised Feature Selection

3.2.1 Variable Selection Procedure

In mechanical learning, over-fitting is generated by over-learning the training data and the model captures the unnecessary noise. In mechanical learning, over-fitting is generated by over-learning the training data, and the model captures the unnecessary noise. When the training has more evolved, the error for the training data is more decrease, but the error for the test data for validation increases.

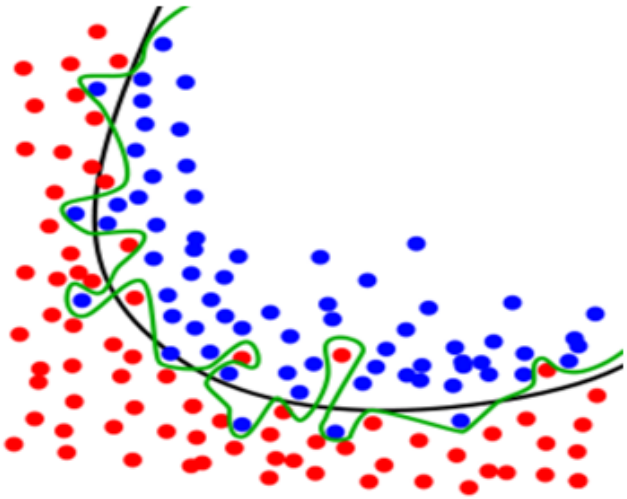


Figure3.3 Concept of overfitting
(Source: https://en.wikipedia.org/wiki/Overfitting)

In the machine learning, when the dimension of the input data is high, the curse of dimensionality arises, which is more important in the big data environment. When a curse of a dimension occurs, the amount of space increases very quickly due to the increase in dimensions, and it makes the available data thinner. This adversely affects the algorithm, induces overfitting, and greatly increases memory storage requirements and computational cost.

Dimensionality reduction is required to solve this problem, and it consists of feature extraction and feature selection. In the case of feature extraction, additional analysis is difficult because the physical meaning of the original feature cannot be obtained in the transformed space. In contrast, feature selection preserves physical meaning, providing better readability and interpretability, making it more useful for practical analysis.

In this study, two pieces of information are utilized for variable selection. First, variables including the dependent variables are grouped. The variable is assigned to the only one PC that has the most loading of the variable. The priorities are assigned in descending order of the loading of dependent variables in the PCs. Second, the variables in the same PC are prioritized using a cross-correlation coefficient (CCC), which indicates non-linear correlation with dependent variables. The use of non-linear correlation for priorities of variables in the same PC is to overcome the limitations of the linearity of the PCA. The loading is the linear correlation between the variables and PCs.

The CCC is a tool for measuring the nonlinear similarity between two

series from one point to another and used for signal processing, pattern cognition and so on. The coefficient is in the range -1 to 1 such as Pearson correlation. The larger the absolute value, the stronger the correlation strength. The CCC representing the nonlinear correlation implies the predictive strength in the nonlinear learning model. Therefore, in the group of the variables grouped by PC, Priority selection is possible. In a continuous function, if two series are f and g, the cross-correlation function $\otimes$ is defined as follows. f^* is the conjugate prior, and τ is the time lag.

$$(f \otimes g)(\tau) = \int_{-\infty}^{\infty} f^*(t) g(t+\tau) dt$$

Using ranking of the PC and the variable ranking within the PC, the key variables are selected enhancing the generalization performance. The most widely used method for the variable selection in machine learning is the wrapper method which selects the variables that improve the regression performance. There are forward selection methods that add variables, backward elimination methods that remove variables and step-wide methods. This study used the forward selection and the search order according to ranking is performed first among the groups and within the group. It is aimed that secures independent PCs as much as possible, then insufficient explanatory power can be recovered by adding the variables in the selected PCs.

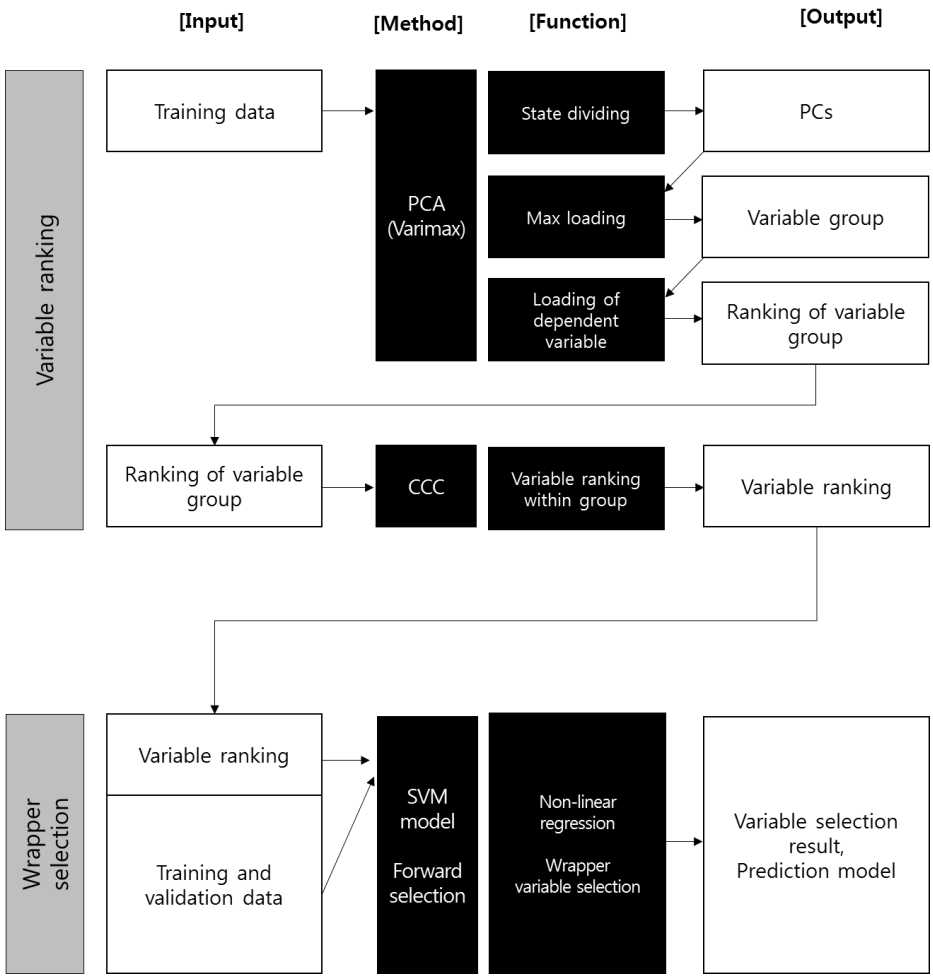


Figure3.4 Variable selection procedure

The generalization performance and low cost of computation are required due to the use of the wrapper that needs to develop the model lots of times. The SVM model has these properties and employed for the predictive model. The stopping criterion of the forward selection is the case where there

are k-th improvement failures or no searchable variables. Use the parameter k_1 between the groups and the parameter k_2 in the group. First, variables are selected by adding the PC using the representative variables for each PC, based on the dependent variable loading criteria identified by the PCA analysis. Then, for the selected PC only, test additional variables. In this way, it is possible to add spatiotemporal variables as much as possible and to improve prediction performance.

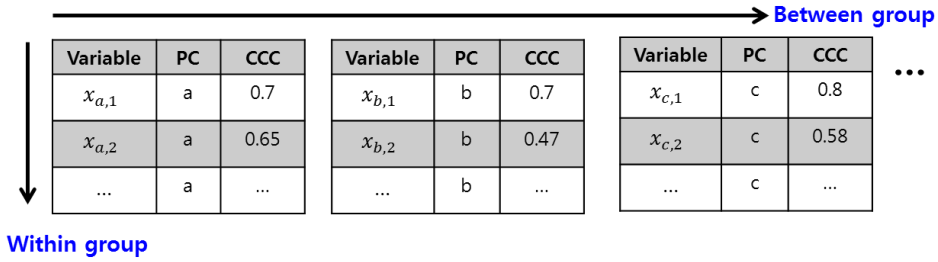


Figure 3.5 Forward selection of proposed approach

3.2.2 Support Vector Machine

The SVM is a classifier that maximizes the margin, the distance between support vectors. A support vector is a data element closest to the hyperplane in multidimensional space. The SVM makes up a set of hyperplanes or hyperplanes in high density or infinite dimensional space that can be used for classification, regression, or other operations. The good separation is achieved by the hyperplane that is the longest distance from the closest training data point

of a class, since the larger the margin, the lower the generalization error of the classifier.

Let w be a normal vector, y_i is a variable that indicates the group to which x belongs by 1 and -1, and b is the distance from the point where x is projected to w. In this case, the hyperplane satisfying $\cdot x - b = 1$ and $w \cdot x - b = -1$ is the line passing through the support vector. Since there is no other data between these distances (hard margin), the following constraints and object function are expressed.

$$arg \min_{(w,b)} \|w\|$$

$$\text{subject to } y_i(w \cdot x - b) \geq 1$$

The objective function is a problem of finding the norm of the normal vector, which is a problem of the square root, which is difficult to calculate. If it is two-dimensional, it is solved as follows.

$$\|w\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

By substituting $\frac{1}{2}\|w\|^2$ for $\|w\|$ in the objective function, the problem is easily solved with quadratic programming. The solution of decision variable, w and b, is not changed. The Lagrange multiplier can be used to

find the saddle point as follows.

$$arg \min_{(w,b)} \max_{\alpha \geq 0} \{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i (w \cdot x - b) - 1]\}$$

This is solved by partial differential equations as follows.

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i$$

Since the partial derivative value becomes 0 at the optimum point, it is as follows.

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

Substituting the partial differential results into the original problem changes the following.

$$L(w, b, \alpha_i) = \frac{1}{2}\left(\sum_i \alpha_i y_i x_i\right)^T \left(\sum_j \alpha_j y_j x_j\right) - \sum_i \alpha_i y_i \left(\sum_j \alpha_j y_j x_j\right)^T x_j + \sum_i \alpha_i$$

$$= -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j {x_i}^T x_j + \sum_i \alpha_i$$

If the second term of the negative sign in the original problem is the minimum, then L becomes the maximization of quadratic problem.

$$\max -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j {x_i}^T x_j + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0, \alpha_i \geq 0$$

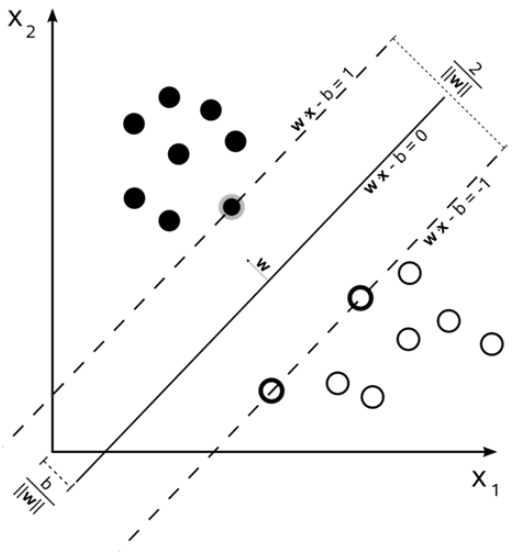


Figure 3.6 Concept of maximum margin
(Source: https://en.wikipedia.org/wiki/Support_vector_machine)

The SVM permits soft margins that allow some classification error to avoid overfitting. A misclassification tolerance term ξ_i is added to the objective function. The C is a regularization parameter. If it is small, the constraint is neglected. If it is large, the influence becomes large and a hard margin is obtained in infinite.

$$\min_{(w,b)} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{subject to } y_i(w \cdot x - b) \geq 1 - \xi_i$$

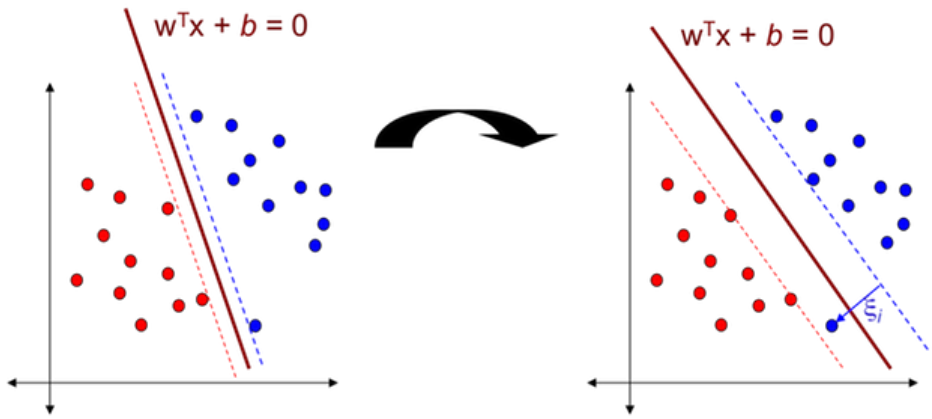


Figure 3.7 Soft margin concept in SVM

(Source: https://mubaris.com/2017/10/14/svm-python/)

The SVM uses a kernel trick that replaces the calculation of the inner product of the vector $(x_i{}^T x_j)$ with the kernel function to reflect the nonlinearity. Linear classification is performed at the modified high dimension, but nonlinear classification is performed in the original space. There are Linear, polynomial, and radial basis kernel function, and this study used radial basis function.

Table 3.1 Kernel function in SVM

Kernel	**Function**
Linear	$x * y$
Polynomial	$[(x * x_i) + 1]^d$
Radial basis function	$\exp(-\Upsilon\|x - x_i\|^2)$

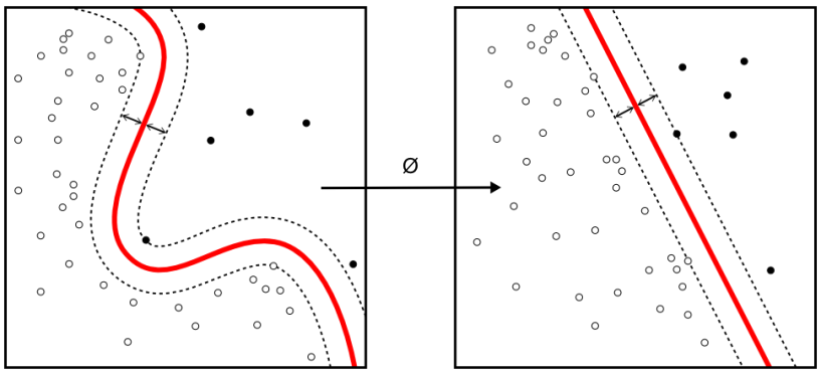


Figure 3.8 Kernel trick in SVM

(Source: https://en.wikipedia.org/wiki/Support_vector_machine)

The support vector regression (SVR), which is a regression model using SVM, can be expressed as below using the soft margin. Γ is the cost function.

$$f(x) = C\sum_{i=0}^{l}\Gamma(f(x_i - y_i) + 0.5\|w\|^2$$

The partial differential is:

$$w = \sum_{i=0}^{l}(\alpha_i - \alpha_i{}^*)\,\Phi(x_i)$$

As a consequence, the f(x) is:

$$f(x) = \sum_{i=0}^{l}(\alpha_i - \alpha_i{}^*)\left(\Phi(x_i)\cdot\Phi(x)\right) + b$$

The dot product of $\left(\Phi(x_i)\cdot\Phi(x)\right)$ can be replaced by the kernel function, and it is as follow.

$$f(x) = \sum_{i=0}^{l}(\alpha_i - \alpha_i{}^*)\,k(x_i, x_i) + b$$

3.3 Comparison Models

3.3.1 Random Forest

Decision trees analyze data and represent patterns existing between data as a combination of predictable rules. The model is called a decision tree because it is similar to a tree. The distinction of variables is learned so that the increase of homogeneity, impurity or uncertainty is reduced as much as possible. Entropy, Gini coefficient, etc. are used as indicators.

The random forest is an ensemble machine learning method that builds a lot of decision trees, and the final result is averaged in regression or voted in classification. The various sub-sampling with replacement is employed to build a lot of decision trees, called bagging (bootstrap aggregating). The random forest provides variable importance information based on the mean decrease accuracy in the process of bagging. The average accuracy is traced by inclusion or exclusion of a variable. This can be used as a source for selecting meaningful variables.

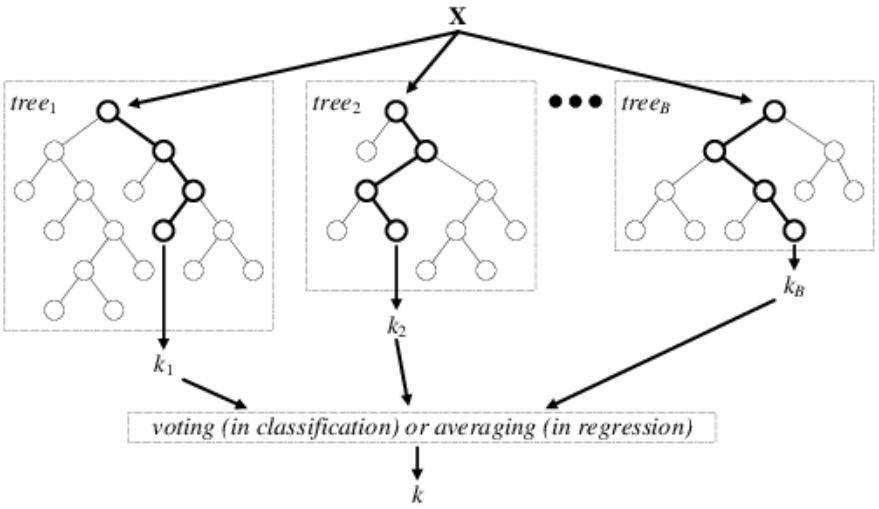


Figure 3.9 Concept of random forest
(Source: Verikas et al. 2016)

3.3.2 Artificial Neural Network

The ANN method is the most widely used learning method, and it is a method which is inspired by neural network of the brain. The ANN can model the complex non-linear structure using the weights of individual nodes that can form multiple layers. It was used in cases such as image recognition and speech recognition, which are difficult to solve based on rules, and for speed prediction. The ANN can be defined by a connection pattern between neurons in different layers, a learning process for updating the weights of connections, and an activation function for changing the weight input of neurons to activation.

In the feed forward neural network, which is the simplest neural network, information of the neural network is transmitted unidirectionally from

the input node to the output node through the hidden node, and there is no circulating path. It is learned by using backpropagation technique. The cost function is minimized and the activation function can be logistic, softmax, sigmoid, tanh, RELU, and so on. In ANN model, there is a limitation that it finds the local optimal solution. Also, the learning effect is not improved even if the layer is deepened, and the optimal performance is achieved in the first and second layers when the backpropagation technique is applied. The overfitting is more likely to occur than SVM for the purpose of minimizing the error.

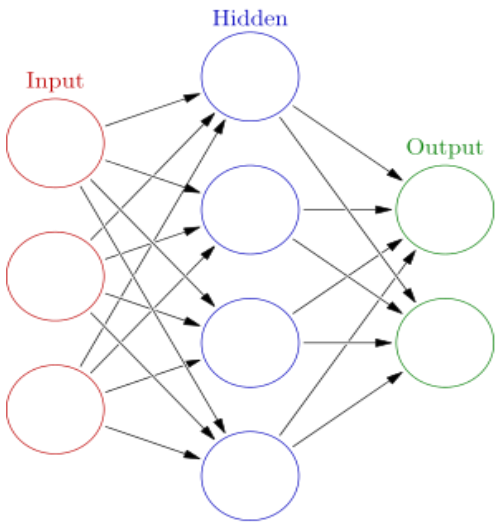


Figure 3.10 Concept of ANN
(Source: https://en.wikipedia.org/wiki/Artificial_neural_network)

3.3.3 Stacked auto-encoder

The auto-encoder is one of the deep learning methods to improve the ineffective limitation of ANN in multilayer. In contrast to ANN, the auto-encoder makes a structure that can be restored itself using encoding and decoding. To this end, the input variables are connected to the specific nodes and the results are obtained. This is an unsupervised learning method because the outputs are the same as the inputs.

The SAE is a technique that stacks multiple layers in a model structure and continuously encodes and decodes them. The local learning is performed using the auto-encoder to find out the features, and then the deep learning ANN structure can be completed using fine-tuning that links the original input and output data again.

The sparsity can be used to increase the efficiency of learning and mitigate the overfitting in SAE. This technique uses the weight matrix of the variable into the penalty term with the parameter in the objective function. The lasso ($l_1\ norm$) uses the absolute value while the ridge ($l_2\ norm$) uses the squared value as a penalty.

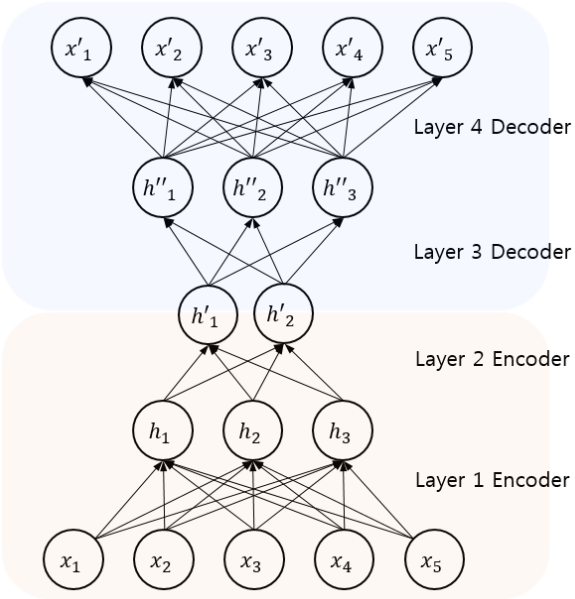


Figure 3.11 Concept of stacked auto-encoder

(Source: https://wikidocs.net/3413)

(a) Lasso (b) Ridge

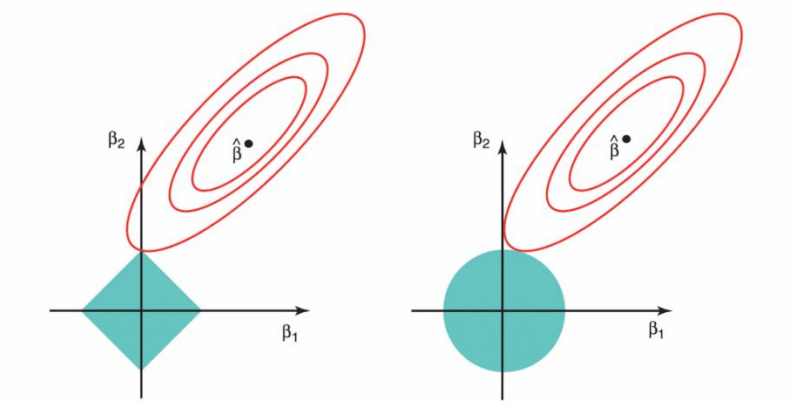


Figure 3.12 Concept of sparsity

(Source: https://onlinecourses.science.psu.edu/stat857/node/158)

Chapter 4. Empirical Data Evaluation

4.1 Evaluation Strategy

This chapter described the evaluation of the proposed machine learning approach. It was evaluated whether the approach improves the prediction performance by selecting significant variables for prediction horizon in real data. To this end, the proposed machine learning approach was used to predict 5-, 15-, and 30-minute of a Kyungbu freeway in Korea. Also, it was performed a 5-minute and 15-minute prediction of U.S. interstate freeway in California. The details of the process by which the proposed approach was applied to real data were also described. The performance measure used mean absolute error (MAE), root mean squared error (RMSE), and RMSE was used in the variable selection procedure.

$$MAPE(\%) = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$$MAE = \frac{\sum_{t=1}^{n}|A_t - F_t|}{n}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$$

Where, A_t : t-th observed value, F_t : t-th predicted value

The comparison models for performance evaluation were divided into variable selection group and the all variable group. The variable selection group consists of the case of using the single index of CCC, which is a nonlinear correlation coefficient, and the case of using the variable importance from the RF model as a single index, all using forward selection based on the SVM model. The number of the test for variable selection in comparison models was the same as the test number of the proposed model. In the all variable group, SVM model and RF model were developed using all variable. The overfitting can be identified by the predictive performance comparison with the all variable group. The overall improvement level of the ensemble learning can be compared with the variable selection. In the U.S. interstate freeway case, the ANN based models were further compared because the dimension of input space was relatively low. The performances were compared with basic ANN model and SAE model with the sparsity term.

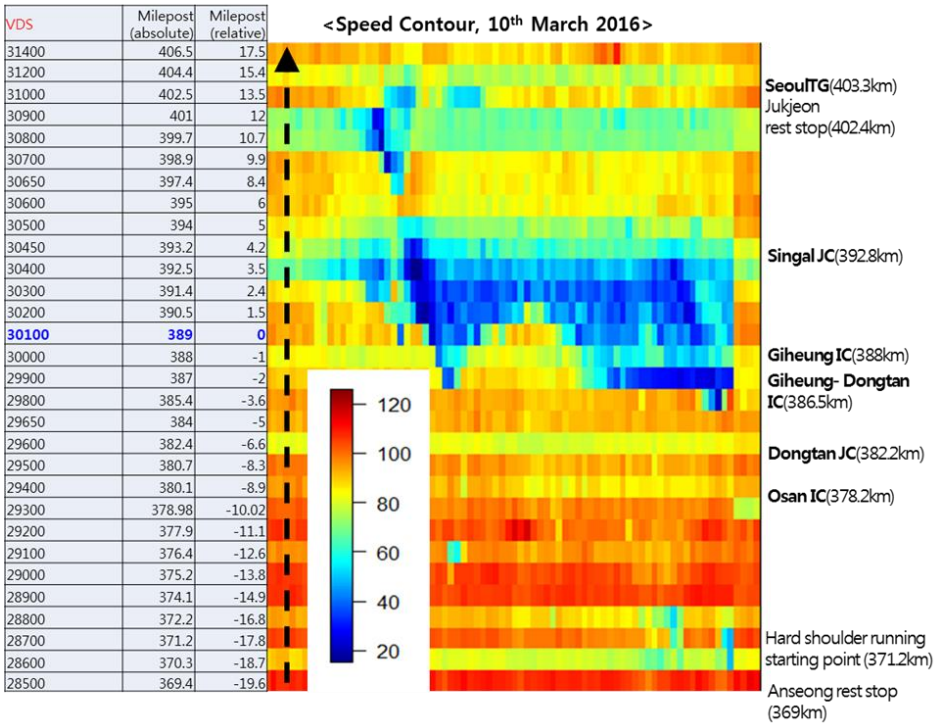
4.2 Case 1: Korean Freeway

4.2.1 Evaluation Site

The study area has 30 loop detectors of vehicle detection system (VDS) and is in 37.1km section (Milepost (MP) 369.4 km - 406.5 km) from Anseong rest stop to the Seoul tollgate in northbound. Data were collected for six months from March to August 2016, and the 5-minute aggregation data of speed (μ), occupancy (o) and traffic volume (q) were obtained. The cases of continuous observation for 2 hours were selected and performed the prediction for 5-minute, 15-minute and 30-minute using the 1-hour data as independent variables. The outliers were eliminated when the speed was 0 or more than 160 km/h. In the total 13,244 data sets, 9271 cases (70%, from March 19 19:15 to June 26 7:05) was used for training, and 3973 cases (30%, from June 26 7:10 to August 31 19:10) was used for validation. The time information is based on past 5 minutes.

The total variable space consists of 1080 spatiotemporal variables because of the three variables of μ, o, and q from 30 detectors with 12 time-steps. The time variable was added to reflect the effect of historical data and the variables space was finally 1081. The number of VDS and absolute and relative location and geometrical feature are shown in the following figure with the speed contour plot of March 10. The prediction targets were the 5-minute, 15-minute and 30-minute future of the VDS 30100 (389 km) located midway in the queue on March 10. The evaluation site operates hard shoulder running. The notations are μ for speed, o for occupancy, q for traffic volume. The

superscript i is relative MP, the negative is upstream, and the positive is downstream. The subscripts are in the order of time steps, -1t is 5 minutes ago, and -2t is 10 minutes ago and so on. For example, $\mu_{t=-1t}^{i=-1km}$ indicates the speed 5 minutes ago in the 1 km upstream.



VDS	Milepost (absolute)	Milepost (relative)
31400	406.5	17.5
31200	404.4	15.4
31000	402.5	13.5
30900	401	12
30800	399.7	10.7
30700	398.9	9.9
30650	397.4	8.4
30600	395	6
30500	394	5
30450	393.2	4.2
30400	392.5	3.5
30300	391.4	2.4
30200	390.5	1.5
30100	389	0
30000	388	-1
29900	387	-2
29800	385.4	-3.6
29650	384	-5
29600	382.4	-6.6
29500	380.7	-8.3
29400	380.1	-8.9
29300	378.98	-10.02
29200	377.9	-11.1
29100	376.4	-12.6
29000	375.2	-13.8
28900	374.1	-14.9
28800	372.2	-16.8
28700	371.2	-17.8
28600	370.3	-18.7
28500	369.4	-19.6

Figure 4.1 Basic information of evaluation site (Korean freeway)

4.2.2 Analysis

(1) 15-minute prediction

The prediction for 15-minute future at VDS 30100 (absolute MP 389 km, relative MP 0 km) was performed using the proposed approach. The one-hour data of 5 minute-aggregated μ, o and q from 30 VDS (the upstream section -19.6 km to the downstream section 17.5 km) and time information were used as the independent variable.

First, PCA analysis was performed. Since the time variable is not a variable representing the current state, it is excluded from the PCA analysis, and the variable selection is applied at first, and this indicates adding the reflection of the current state in the empirical prediction. In the first trial of PCA with Varimax rotation, the dependent variable was loaded on a PC with some independent variables which have maximum loading on the same PC. In this case, the PCA was re-performed using the dependent variable and the surrounding variables until the dependent variables were loaded alone in a PC. A total of four times PCAs were performed to find a PC loaded only dependent variable. The following table lists the PCs in the order of loading dependent variables. The cumulative 95% of the dependent variables were loaded in 18 PCs.

Table 4.1 PCA result for 15-minute prediction of Korean freeway

Trial number of PCA	**PC number**	**Loading of dependent variable**	**Cumulative loading of dependent variable**	**Number of maximum loaded variables**
4	2	0.735	24.7%	1
3	7	0.352	36.5%	-
2	26	0.314	47.1%	-
1	211	0.283	56.6%	-
2	2	0.253	65.1%	29
3	1	0.201	71.9%	10
4	3	0.135	76.4%	2
1	1	0.129	80.7%	367
4	5	0.126	85.0%	2
1	15	0.075	87.5%	3
3	2	0.069	89.8%	14
2	3	0.031	90.9%	23
2	4	0.030	91.9%	13
2	1	0.027	92.8%	36
…				

In the order of dependent variable loading, PCs are ranked. The independent variables are assigned to the PCs that have the maximum loading of the variables for grouping. In the group, the variables were ranked using the nonlinear correlation coefficient CCC. The result of ranking is shown in Figure 4.2.

In the ranking result, the first PC consists of the speed and the occupancy of 5 to 30 minutes ago in the upstream of -1 km, -2 km. The second PC includes variables of the speed and occupancy of 5 to 15 minutes ago in the 1.5 km downstream and 2.4 km downstream. The third PC includes the speed and the occupancy of the target point (VDS 30100) at five minutes ago. The fourth PC includes the speed and occupancy of 5 to 45 minutes ago in the bottleneck point, 4.2 km downstream and 3.6 km downstream. In addition, traffic volume between -17.8 km and -11.1 km, which is a section of hard shoulder running, and traffic volume information, about -5 km section, which is near Dongtan JC, are included. Then, the PCs were selected spatiotemporal variables in such a way from the nearby target point to far away. In the selected time range, upstream of 5 to 30 minutes ago is in a PC, but downstream of 5 to 15 minutes ago is in a PC because traffic flow from the downstream effects more rapidly to a predictive area than it from the upstream.

The variables were selected using the forward selection method with SVM and variable ranking. The stopping criteria were $k_1 = 3$ between PCs

and $k_2 = 1$ in the PC. The 18 tests were performed and took 169.6 seconds, and the test results are shown in Table 4.2. The RMSE of the optimal model is 12.788 in the validation data. The MAE was 7.521 and the average speed is 84.3 km/h.

The selected variables are shown in Figure 4.3. The bottleneck head of the queue, 4.2 km and the tail of the queue, -2 km were included. In the downstream, the speed of 4.2 km and the speed of 1.5 km were included. Target points included the speed of 5, 10, and 15 minutes ago. In the case of the upstream, the speed of 5 minutes ago and the occupancy of 10 minutes ago in the -2 km point were included, and the speeds of 5 minutes and 10 minutes ago in the -1 km point were included.

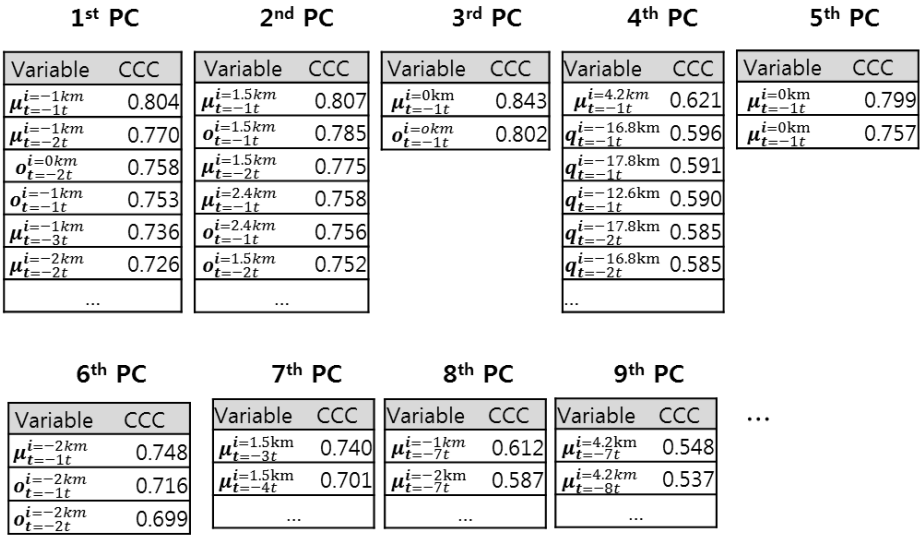


Figure 4.2 Variable ranking for 15-minute prediction of Korean freeway

Table 4.2 Result of forward selection for 15-minute prediction of Korean freeway

Number of trial	**Tested variable**	**RMSE of validation data**	**Calculation time (sec)**	**Selection result**
1	Time	24.333	11.5	O
2	$\mu_{t=-1t}^{i=-1km}$	16.614	9.6	O
3	$\mu_{t=-1t}^{i=1.5km}$	14.510	8.8	O
4	$\mu_{t=-1t}^{i=0km}$	14.002	8.5	O
5	$\mu_{t=-1t}^{i=4.2km}$	13.811	8.5	O
6	$\mu_{t=-1t}^{i=0km}$	13.555	8.7	O
7	$\mu_{t=-1t}^{i=-2km}$	12.984	8.9	O
8	$\mu_{t=-3t}^{i=1.5km}$	13.084	9.2	X
9	$\mu_{t=-7t}^{i=-1km}$	13.078	9.3	X
10	$\mu_{t=-7t}^{i=4.2km}$	13.094	9.3	X
11	$\mu_{t=-2t}^{i=-1km}$	12.942	9.5	O
12	$o_{t=-1t}^{i=1.5km}$	12.952	9.5	X
13	$o_{t=-1t}^{i=0km}$	12.968	9.4	X
14	$q_{t=-1t}^{i=-16.8km}$	12.993	9.6	X
15	$\mu_{t=-1t}^{i=0km}$	12.900	9.6	O
16	$o_{t=-1t}^{i=-2km}$	12.827	10.0	O
17	$o_{t=-2t}^{i=0km}$	12.871	9.9	X
18	$o_{t=-2t}^{i=-2km}$	12.778	10.0	O

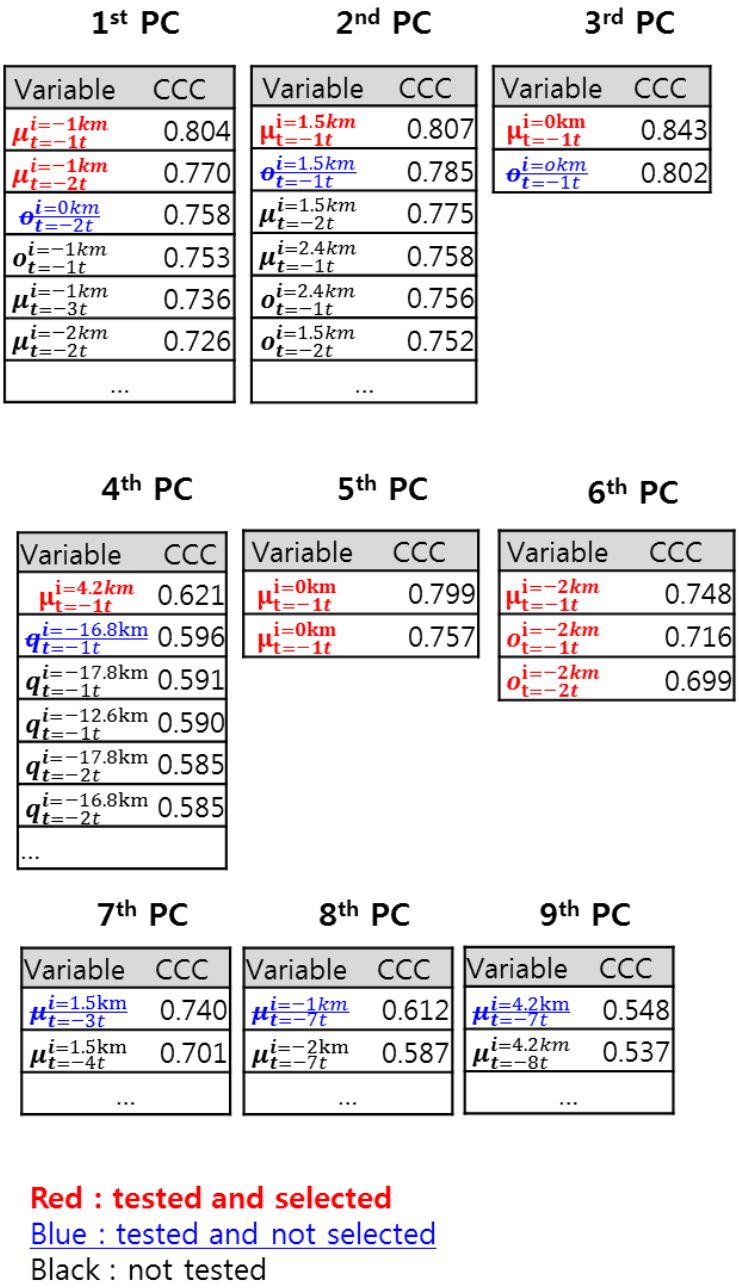


Figure 4.3 Result of variable selection for 15-minute prediction of Korean freeway

(2) 30-minute prediction

The prediction for 30-minute future at VDS 30100 (absolute MP 389 km, relative MP 0 km) was performed using the proposed approach. The independent variable space is the same as the 15-minute prediction. The same procedure of 15-minute prediction was applied for 30-minute prediction. PCA for 30-minute prediction was performed once and PC with only dependent variable was obtained. In the order of dependent variable loading, PCs that grouped the independent variables are ranked. In the group (PC), the variables were ranked using the nonlinear correlation coefficient, CCC. Variable ranking results were similar to the 15-minute prediction but somewhat changed. In the variable selection by SVM, the stopping criteria were $k_1 = 3$ between PCs and $k_2 = 1$ in the PC. The total RMSE of the optimal prediction model is 17.282 and the MAE is 10.74. The average speed was 84.3 km/h.

The result of variable selection is shown in Figure 4.4. In contrast to the 15-minute prediction, it can be seen that the spatiotemporal range is expanded by adding the speed and occupancy of -17.8 km point which is the starting point of hard shoulder running and immediately downstream of Dongtan JC, -6 km and -5 km point. The target information was reduced to only five minutes ago, and the head and tail information of the queue is used similar to the 15-minute prediction.

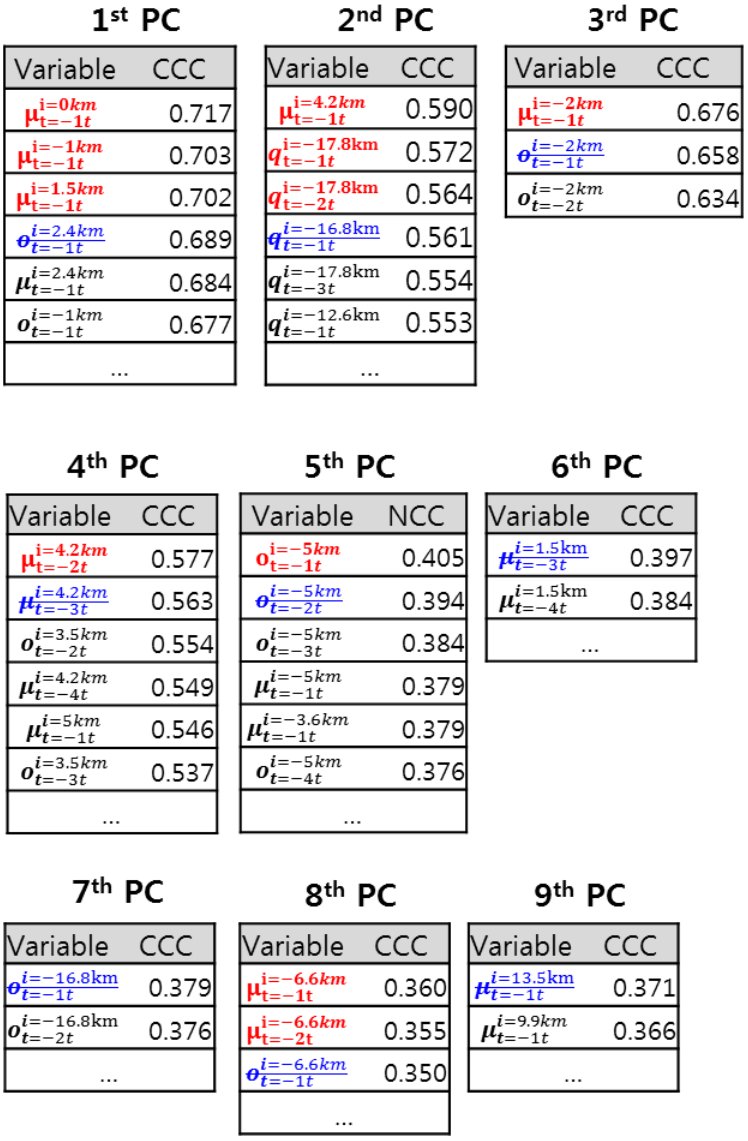


Figure 4.4 Result of variable selection for 30-minute prediction of Korean freeway

(3) 5-minute prediction

The prediction for 5-minute future at VDS 30100 (absolute MP 389 km, relative MP 0 km) was performed using the proposed approach. The independent variable space is the same as the 15-minute prediction. The same procedure of 15-minute prediction was applied for 30-minute prediction. A total of five times PCAs were performed to find a PC loaded only dependent variable. In the order of dependent variable loading, PCs that grouped the independent variables are ranked. In the group (PC), the variables were ranked using the nonlinear correlation coefficient, CCC. In the variable selection by SVM, the stopping criteria were $k_1 = 3$ between PCs and $k_2 = 1$ in the PC. The total RMSE of the optimal prediction model is 8.121 and the MAE is 4.628. The average speed was 84.3 km/h.

The result of variable selection is shown in Figure 4.5. In contrast to the 15-minute prediction, it can be seen that the spatiotemporal range is shortened. The information of -2 km and 4.2 km was eliminated while the information of -1 km and 1.5 km was added.

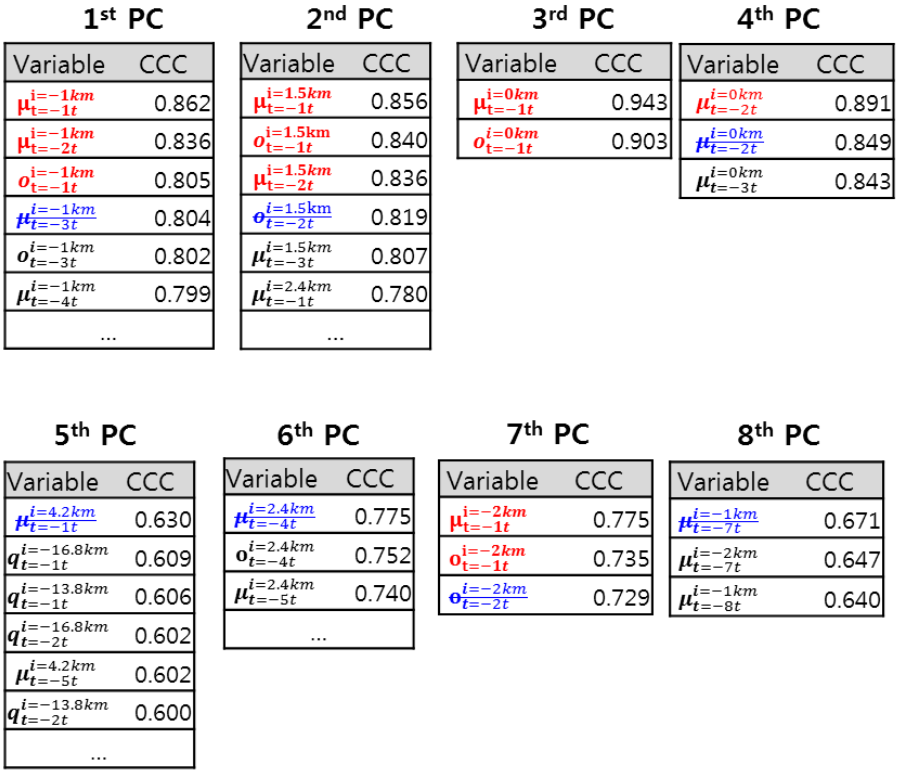


Figure 4.5 Result of variable selection for 5-minute prediction of Korean freeway

4.3 Case 2: Interstate Freeway

4.3.1 Evaluation Site

For the second case study, the data of interstate freeway I-880 in California was selected. Eight consecutive detectors, VDS 401611 - 400030, in 5.81 km (3.64 miles) section in the area near the San Jose area on the California Highway I-880 road were selected. The average length of detectors is 0.83km. The data of μ, o and q were collected, and the 30 seconds aggregated raw data was re-aggregated for 5-minutes. Data were collected for three months, February, March, and April 2016.

A total of 9713 samples are collected eliminating the outliers, and prediction is possible up to 15-minutes based on data up to 20 minutes ago. The training data and the validation data were divided into 70% and 30%. Since there are three pieces of information (μ, o, and q) in 8 detectors at four-time steps, there are 97 spatiotemporal variables as independent variables. Except for the weekends and night time and missing data, the data was fully observed over 40 minutes in all detectors.

The filtered samples were relatively similar in time to frequency distribution, and the influence on the specific time was leveled. The geometrical property, number of VDS in order from upstream to downstream and the speed contour plot of 1st February is shown in Table 4.3 and Figure 4.6. The congestion occurred in the VDS 1, and it is mitigated at VDS 3. The congestion cleared at VDS 5 but it occurs again at VDS 6. The VDS 3 was selected as

target variable for prediction. The average speed of target variable is 65.795 miles/h, and the standard deviation of speed was 13.22 miles/h.

Table 4.3 Evaluation site of interstate freeway

VDS number	**VDS ID**	**Postmile (absolute)**	**Postmile (relative, km)**	**Interval (km)**
8	401611	0.03	4.38	
7	408911	0.54	3.56	0.82
6	400514	1.28	2.37	1.19
5	400709	1.94	1.30	1.06
4	400508	2.55	0.2	0.98
3	403225	2.75	0.0	0.32
2	401440	3.44	-1.11	1.11
1	400030	3.64	-1.43	0.32

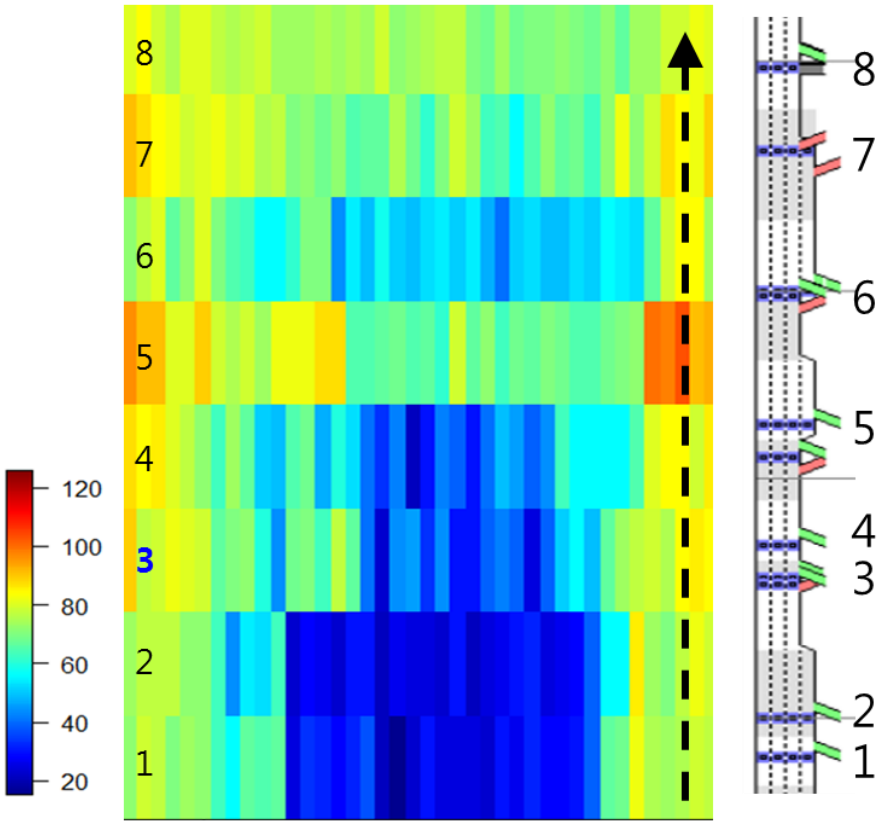


Figure 4.6 Characteristics of study area in interstate freeway

4.3.2 Analysis

(1) 5-minute prediction

The PCA was conducted using training data with dependent variables, and the Varimax rotation was performed. The PCA was performed using the correlation matrix because the scale of the information was different such as time, speed, traffic volume, and occupancy. The data decomposed into 97 PCs with the same number of variables. Dependent variables were loaded at the 57^{th} PC, and the loading of dependent variables loaded in 14 PCs was 98%. Variables were assigned to PCs that were the maximum loading of each variable, grouped by PC, and prioritized by dependent variable loading ranking.

The variables were grouped by assigning it to the maximum loaded PC. The variable ranking was determined, the between PCs by the loading of the dependent variable, and the within PC by the CCC. The variables were selected in the forward selection manner using the variables rankings by proposed approach and SVM model. In the variable selection by SVM, the stopping criteria were $k_1 = 3$ between PCs and $k_2 = 1$ in the PC. As a result of the variable selection, the total execution time is 95.7 seconds and the RMSE of validation data of final model is 6.961, MAE is 5.003.

In addition to the time variable, seven variables of the current state were selected and the result of variable selection is shown in Figure 4.7. The speed and the occupancy of 5 minutes ago at the 1.3 km downstream, and the speed of 5 minutes ago at the point of 0.2 km were selected. In the upstream,

speed and occupancy of 5 minutes, ago and the occupancy of 10 minutes ago at the -1.11 km point were selected. The target point was selected only at the speed of 5 minutes ago.

Table 4.4 PCA result for 5-minute prediction of interstate freeway

Number of PC	**Loading of dependent variable**	**Cumulative loading of dependent variable**
57	0.30047	30.0%
60	0.28346	58.4%
3	0.20044	78.4%
2	0.14542	93.0%
4	0.01520	94.5%
1	0.00821	95.3%
5	0.00738	96.1%
41	0.00504	96.6%
47	0.00298	96.9%
13	0.00266	97.1%
14	0.00201	97.3%
26	0.00190	97.5%
11	0.00185	97.7%
38	0.00180	97.9%
15	0.00166	98.0%

Table 4.5 Variable selection of 5-minute prediction of interstate freeway

Trial	**Variable**	**RMSE Test**	**Calculation time(sec)**	**Selection result**
1	$\boldsymbol{T}$	10.388	6.75	O
2	$\boldsymbol{\mu_{t=-1t}^{i=0km}}$	7.637	6.46	O
3	$\boldsymbol{o_{t=-1t}^{i=-1.11km}}$	7.499	6.34	O
4	$\boldsymbol{\mu_{t=-1t}^{i=1.30km}}$	7.173	6.42	O
5	$\boldsymbol{q_{t=-1t}^{i=4.38km}}$	7.240	6.40	X
6	$\boldsymbol{o_{t=-1t}^{i=2.37km}}$	7.217	6.47	X
7	$\boldsymbol{\mu_{t=-1t}^{i=0.2km}}$	7.094	6.67	O
8	$\boldsymbol{o_{t=-1t}^{i=1.30km}}$	6.968	6.88	O
9	$\boldsymbol{\mu_{t=-2t}^{i=3.56km}}$	7.031	7.24	X
10	$\boldsymbol{\mu_{t=-2t}^{i=0km}}$	6.999	7.20	X
11	$\boldsymbol{o_{t=-2t}^{i=-1.11km}}$	6.967	7.18	O
12	$\boldsymbol{\mu_{t=-2t}^{i=1.30km}}$	6.978	7.14	X
13	$\boldsymbol{\mu_{t=-1t}^{i=-1.11km}}$	6.961	7.09	O
14	$\boldsymbol{o_{t=-1t}^{i=-1.43km}}$	6.990	7.47	X

PC3

Variable	CCC
$\mu_{t=-1t}^{i=0km}$	0.806
$\mu_{t=-2t}^{i=0km}$	0.763
$o_{t=-1t}^{i=0km}$	0.754
$\mu_{t=-3t}^{i=0km}$	0.730
$o_{t=-2t}^{i=0km}$	0.717
...	

PC2

Variable	CCC
$o_{t=-1t}^{i=-1.11km}$	0.583
$o_{t=-2t}^{i=-1.11km}$	0.569
$\mu_{t=-1t}^{i=-1.11km}$	0.566
$o_{t=-1t}^{i=-1.43km}$	0.563
$\mu_{t=-1t}^{i=-1.43km}$	0.557
...	

PC4

Variable	CCC
$\mu_{t=-1t}^{i=1.30km}$	0.347
$\mu_{t=-2t}^{i=1.30km}$	0.337
$\mu_{t=-3t}^{i=1.30km}$	0.319
$\mu_{t=-4t}^{i=1.30km}$	0.303

PC1

Variable	CCC
$q_{t=-1t}^{i=4.38km}$	0.398
$q_{t=-1t}^{i=3.56km}$	0.387
$q_{t=-1t}^{i=2.37km}$	0.290
$q_{t=-1t}^{i=1.30km}$	0.182
$q_{t=-2t}^{i=4.38km}$	0.395
...	

PC5

Variable	CCC
$o_{t=-1t}^{i=2.37km}$	0.473
$o_{t=-2t}^{i=2.37km}$	0.477
$o_{t=-3t}^{i=2.37km}$	0.465
$o_{t=-4t}^{i=2.37km}$	0.451

PC 41, 47, 13, 14

Variable	PC	CCC
$\mu_{t=-1t}^{i=0.2km}$	41	0.673
$o_{t=-1t}^{i=1.30km}$	47	0.628
$\mu_{t=-2t}^{i=3.56km}$	13	0.257
$\mu_{t=-3t}^{i=3.56km}$	14	0.241

Red : tested and selected
Blue : tested and not selected
Black : not tested

Figure 4.7 Variable selection for 5-minute prediction in interstate freeway

(2) 15-minute prediction

The PCA was conducted for 15-minute prediction using training data with dependent variables, and the Varimax rotation was performed. The data decomposed into 97 PCs with the same number of variables. Dependent variables were loaded at the 50th PC. Variables were assigned to PCs that were the maximum loading of each variable, grouped by PC, and prioritized by dependent variable loading ranking.

The variables were grouped by assigning it to the maximum loaded PC. The variable ranking was determined, the between PCs by the loading of the dependent variable, and the within PC by the CCC. The variables were selected in forward selection manner using the variables rankings by proposed approach and SVM model. In the variable selection by SVM, the stopping criteria were $k_1 = 3$ between PCs and $k_2 = 1$ in the PC. As a result of the variable selection, the total execution time is 150.0 seconds for 21 of trials. The RMSE of validation data of final model is 8.046, MAE is 5.665.

In addition to the time variable, 12 variables of the current state were selected and the result of variable selection is shown in Figure 4.8. A spatiotemporal wider range of variables was used compared to the 5-minute prediction. It was used from the traffic volume of 20 minutes ago at the 8th detector to the speed of 5 minutes ago at the target point. In particular, various information at the downstream was selected.

PC1

Variable	CCC
$\mu_{t=-1t}^{i=0km}$	0.737
$\mu_{t=-2t}^{i=0km}$	0.691
$o_{t=-1t}^{i=0km}$	0.685
$\mu_{t=-3t}^{i=0km}$	0.647
$o_{t=-2t}^{i=0km}$	0.643
...	

PC3

Variable	CCC
$o_{t=-1t}^{i=-1.11km}$	0.548
$o_{t=-2t}^{i=-1.11km}$	0.527
$\mu_{t=-1t}^{i=-1.11km}$	0.526
$o_{t=-1t}^{i=-1.43km}$	0.532
$\mu_{t=-1t}^{i=-1.43km}$	0.521
...	

PC2

Variable	CCC
$q_{t=-4t}^{i=4.38km}$	0.400
$q_{t=-4t}^{i=3.56km}$	0.399
$q_{t=-1t}^{i=4.38km}$	0.398
$q_{t=-2t}^{i=4.38km}$	0.398
$q_{t=-3t}^{i=4.38km}$	0.395
...	

PC4

Variable	CCC
$\mu_{t=-1t}^{i=1.30km}$	0.322
$\mu_{t=-2t}^{i=1.30km}$	0.304
$\mu_{t=-3t}^{i=1.30km}$	0.295
$\mu_{t=-4t}^{i=1.30km}$	0.273

PC7

Variable	CCC
$o_{t=-2t}^{i=2.37km}$	0.470
$o_{t=-1t}^{i=2.37km}$	0.456
$o_{t=-3t}^{i=2.37km}$	0.443
$o_{t=-4t}^{i=2.37km}$	0.428

PC 41, 47, 13, 14

Variable	PC	CCC
$\mu_{t=-1t}^{i=0.2km}$	53	0.615
$\mu_{t=-1t}^{i=3.56km}$	21	0.263
$\mu_{t=-1t}^{i=2.37km}$	60	0.632
$o_{t=-1t}^{i=4.38km}$	9	0.357
$\mu_{t=-2t}^{i=3.56km}$	23	0.257

Red : tested and selected
Blue : tested and not selected
Black : not tested

Figure 4.8 Variable selection for 15-minute prediction in interstate freeway

4.4 Comparison of Predictive Capability

4.4.1 Case 1: Korean freeway

The predictive performance 15-minute prediction model was compared with the models of the variable selection group. The proposed model compared with the CCC based SVM model and RF based SVM model. The average RMSE of the proposed model was the best at 12.778, and the RF based variable selection model was second-ranked at 12.822. Based on the comparison result, the proposed model showed the best prediction capability and better efficiency of computational cost. In the model using all variables, the RMSE of the SVM model is 15.725 and it is confirmed that the overfitting phenomenon arises significantly in the case study. In the case of RF model with all variable, the RMSE is 13.778, which indicates that mitigation of overfitting by ensemble learning is lower than the effect of the proposed variable selection model.

In the case of the 30-minute prediction, the RMSE of the proposed model was the best as 17.282 and the RF based SVM model was the second best as 17.623. The RMSE difference between the two models was larger than the 15-minute prediction. The ability to mitigate the overfitting of the proposed model is better than RF based model when the prediction step is increased. Similar to 15-minute prediction, the computational cost of the proposed model is one-third of that of the RF based model. The overfitting phenomenon was confirmed compared to the SVM model that using all variable. It was also

confirmed that ensemble learning, RF, did not improve the predictive performance as much as the proposed model.

In the case of the 5-minute prediction, the RMSE of the proposed model was the best as 8.121 and the CCC based SVM model was the second best as 8.144. The RMSE difference between the two models was smaller than the 15-minute prediction. Similar to 15-minute prediction, the computational cost of the proposed model is one-third of that of the RF based model. The overfitting phenomenon was confirmed compared to the SVM model that using all variable. It was confirmed the difference in predictive performance between the RF model and the proposed model was smaller than the 15- and 30-minutes prediction.

Table 4.6 Performance comparison for 15-minute prediction of Korean freeway

Category	**Model**	**RMSE of validation data**	**Computational cost (sec)**
Variable selection	Proposed model	12.778	1040.3
	CCC based model	12.941	194.0
	RF based model	12.822	3343.0
All variable	SVM	15.725	710.2
	RF	13.778	3167.2

Table 4.7 Performance comparison for 30-minute prediction of Korean freeway

Category	**Model**	**RMSE of validation data**	**Computational cost (sec)**
Variable selection	Proposed model	17.282	949.5
	CCC based model	17.788	246.0
	RF based model	17.623	3478.2
All variable	SVM	19.237	686.3
	RF	18.581	3252.2

Table 4.8 Performance comparison for 5-minute prediction of Korean freeway

Category	**Model**	**RMSE of validation data**	**Computational cost (sec)**
Variable selection	Proposed model	8.121	1056.5
	CCC based model	8.144	185.3
	RF based model	8.177	3628.5
All variable	SVM	11.824	626.9
	RF	8.413	3474.5

4.4.2 Case 2: Interstate freeway

In the case of the 5-minute prediction, the RMSE of the proposed model was the best as 6.961 and the RF based SVM model was the second-ranked as 6.986. Other results of performance comparison were similar to the comparison of Korean freeway. In the case 5-minute prediction of interstate freeway, three comparison models were added, i.e., unsupervised PCA based variable selection model, ANN, and SAE, for performance comparison because of this case lower dimension of variable space than the case of Korean freeway.

The unsupervised PCA base variable selection model showed the RMSE of 8.377 and it was higher than the RMSE of the proposed model. In the case of ANN and SAE models, parameter optimization should be performed, but this was not possible for a limited time and only a few cases were performed. The node and layer of ANN performed only 4 cases of (3,2), (8,2), (5,5), (5,3,2), and (3,2) showed the best predictive performance. In the case of the SAE model, five nodes and layers of (3,2), (5,5), (15,5), (25,5), and (15,10,5) were performed with 5 cases of sparsity parameter. The best predictive performance was shown in (5,5) with sparsity parameters as lambda 0.002, beta 0.1, rho 0.005, epsilon 0.001. The best result of the ANN model is 9.809, which is the worst performance. The SAE model with sparsity term showed the RMSE of 7.162 and it was slightly higher than the RMSE of the proposed model. For the calculation cost, the proposed model took 1.6 minutes, but the RF took 2.9 minutes. The SAE model had an average of 273.5 minutes and the ANN had an

average of 437.7 minutes.

Table 4.9 Performance comparison for 5-minute prediction of interstate freeway

Category	**Model**	**RMSE of validation data**	**Computational cost (sec)**
Variable selection	Proposed model	6.691	97.1
	CCC based model	7.414	98.8
	RF based model	6.986	266.2
	Unsupervised PCA base model	8.377	95.9
All variable	SVM	7.458	26.6
	RF	7.002	175.3
	ANN	9.809	26,259.5 (average)
	SAE	7.162	16,790.6 (average)

In the case of the 15-minute prediction, the performance comparison was conducted as same as the case of Korean freeway. The RMSE of the proposed model was the best as 8.406 and the RF based SVM model was the second-ranked as 8.566. Other results of performance comparison were similar to the comparison of Korean freeway.

Table 4.10 Performance comparison for 15-minute prediction of interstate freeway

Category	**Model**	**RMSE of validation data**	**Computational cost (sec)**
Variable selection	Proposed model	8.406	153.0
	CCC based model	9.021	143.6
	RF based model	8.566	331.7
All variable	SVM	8.573	24.9
	RF	8.590	182.8

Chapter 5. Implication for Traffic Analysis

5.1 Traffic Phase of Principal Components

This chapter describes the implication of the proposed machine learning approach for traffic analysis. First of all, the terms were defined and explained. The traffic state is a homogeneous condition that is distinguished from other states by a traffic condition at a single point. The propagation of a traffic state indicates a phenomenon in which the traffic state of a single point evolves to another time-space. Kerner (2009) defined the traffic phase as a state of time and space that possess some unique feature. In this study, the definition of traffic phase by Kerner is used, but it is specifically defined as some unique state defined by the relationship of two spatiotemporal variables.

Some shock waves were used as the example of the propagation of traffic states in this study. The shock wave can be defined as the boundary condition of the time-space domain which shows discontinuity of traffic volume-density state (May, 1990). Typical congestion on the freeway is backward forming and forward recovery as shown in Figure 5.1. Backward forming is a phenomenon in which congestion expands in the upstream direction as it exists when congestion occurs in the bottleneck section. Forward recovery is the second most commonly occurring, with demand falling below capacity and the length of congestion decreasing to the downstream direction.

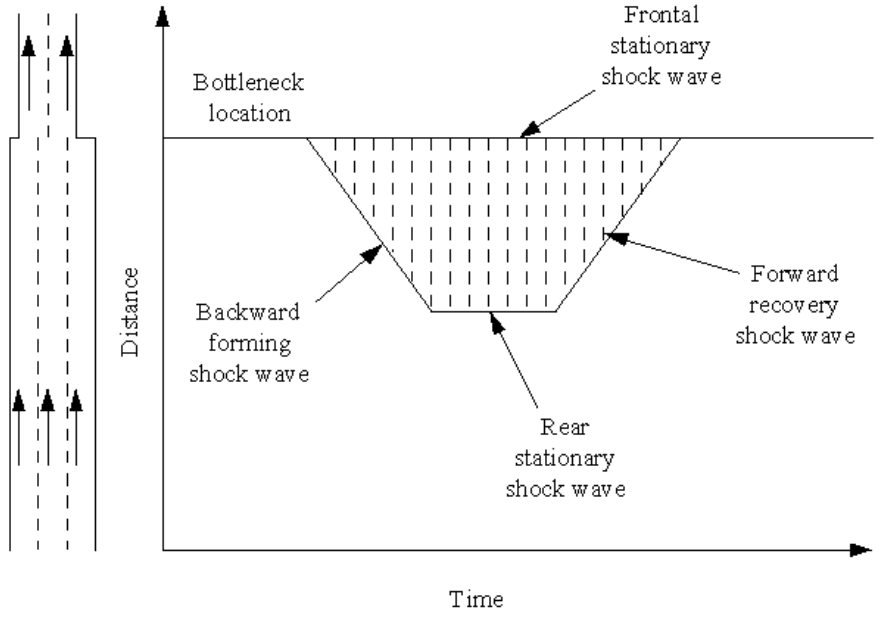


Figure 5.1 Shock wave phenomena at a freeway bottleneck

(Source: Traffic flow fundamentals (May, 1990, p.209))

(1) Numerical example setting

Numerical examples are introduced to analyze shock wave data. The example is simplified from the empirical analysis result that clearly shows the shock wave. The empirical analysis in Zheng et al. (2011) were used. The analysis area is the point where noise is low and shock waves of backward forming and forward recovery are observed at relatively uniform intervals.

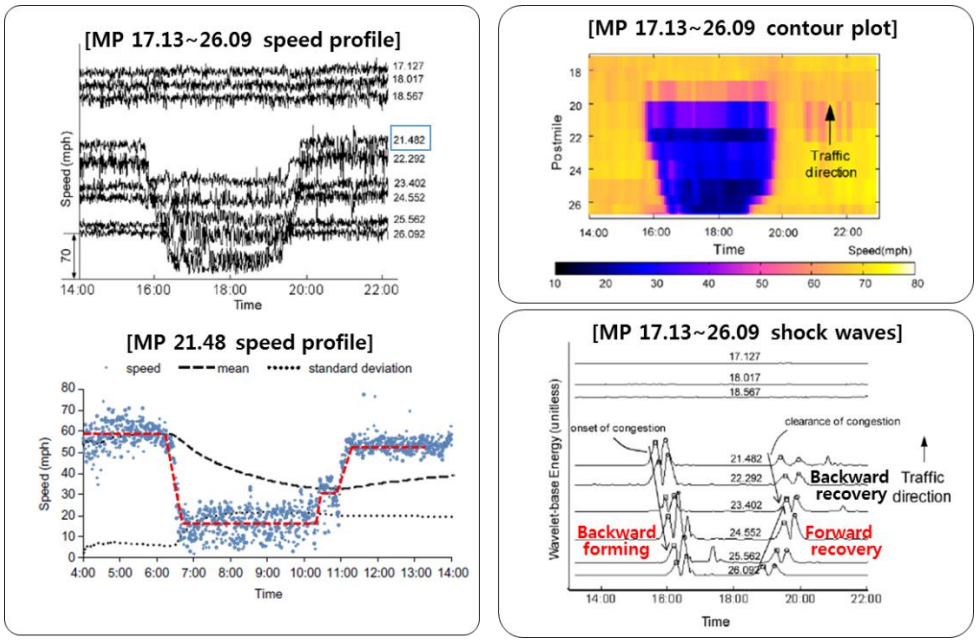


Figure 5.2 Basic source for the numerical example

(Source: Zheng et al., 2011)

To simplify the numerical example, the detector interval is made equal to 1.67km. All have the same shape of speed profile. The wave speed of the backward forming is 20km/h that moves the detector interval at every 5 minutes and the wave speed of the forward recovery is 10 km/h that moves the detector interval at every 10 minutes. The generated daily data that shown in Figure 5.2 was repeated for 30 days, and the noise of the standard normal distribution with the standard deviation of three was combined.

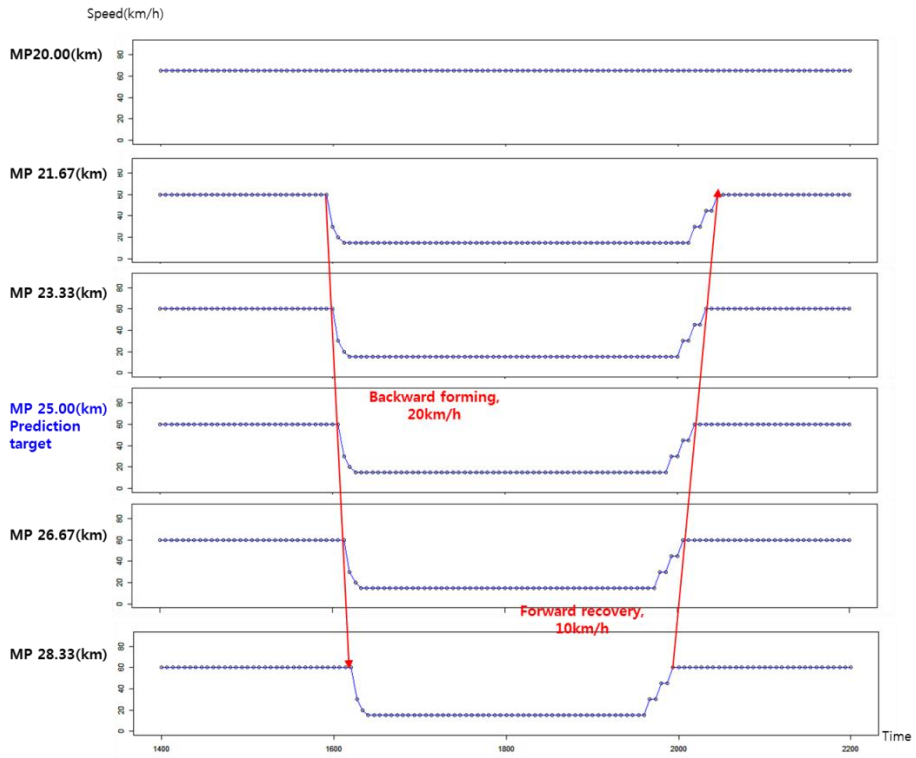


Figure 5.3 Basic speed profile of numerical example

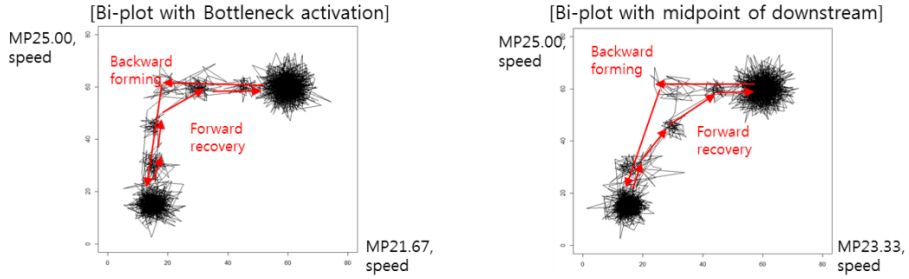
(2) Traffic phase diagram of numerical example

The two of speed variables of the numerical example were bi-plotted in the vector space, which is the concept of the space in which the PCA is performed. The bi-plot of the downstream congestion area and predicted target speed as shown in Figure 5.4 (a). The lower left corner of the graph is the congested traffic phase of both variables and the upper right corner is the phase where both variables are in free flow. The data point of shock waves appear at the upper left of both backward forming and forward recovery, and the bottleneck point shows a clearer difference than the midpoint.

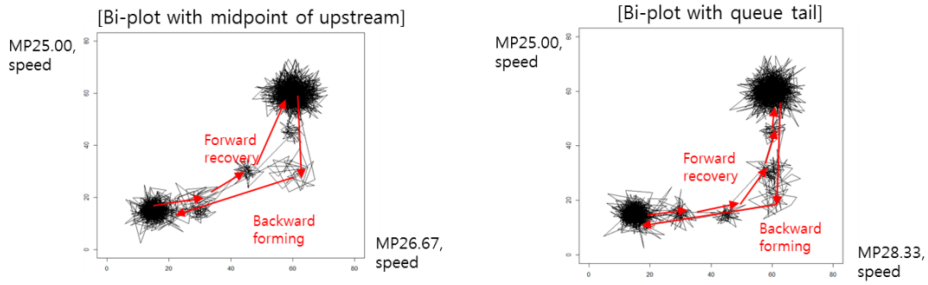
In the bi-plot of the upstream, Figure 5.4 (b), it is different from the case of downstream that the data point of the shock wave appears in the lower right corner. Similar to the downstream, the tail region of the queue where shock waves occur is plotted more clearly than the midpoint.

On the other hand, if it is not related to the shock wave, the bi-plot appears with only the state change of the target variable, Figure 5.4 (c).

(a) Downstream



(b) Upstream



(c) Not in the queue

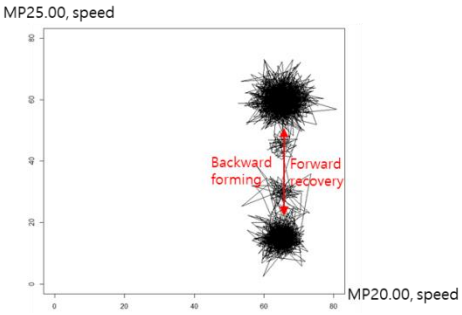


Figure 5.4 Speed bi-plot of numerical example (a) downstream, (b) upstream, (c) not in the queue

The results of the PCA on the bi-plot of the speed of prediction target and the speed of upstream or downstream are shown in Figure 5.5. In the case of downstream, since the condition of both variables are free flow or congested were the longest of the time, the PC connecting the two conditions is created first and this explains maximum of the variance of data in one axis. Next, a second PC is created, pointing to the data point generated by the upper left shock wave. In the upstream case, the first PC is the same as the downstream but the direction of the second PC, which is associated with the shock wave, is opposite to the case of downstream.

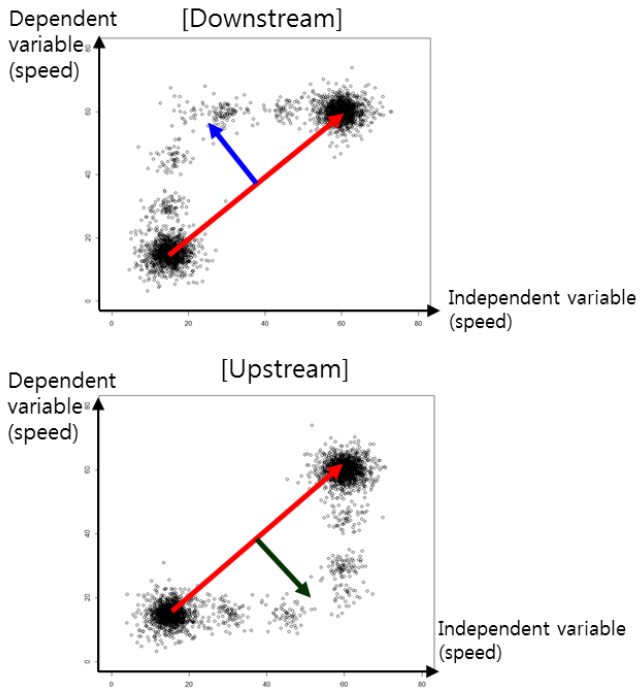
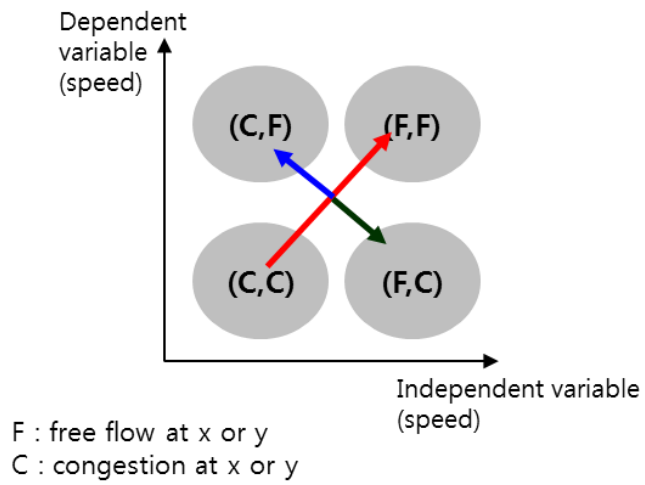


Figure 5.5 PCA result of numerical example

Based on the vector space of bi-plot, the traffic state of each of two speed variables is divided into the free flow (F) and the congested traffic (C), as shown in Figure 5.6. The traffic state of the two variables is composed of both free flow states (F, F), both the congested traffic (C, C), congested traffic only on the y-axis (F, C), and congested traffic on the only x-axis (C, F). The PCA results in the case of not in the queue showed the PCs are pointed their own variance, no effect each other. As shown in Figure 5.7, the overall vector space is represented by the traffic phase diagram, and the PCA results can be interpreted regarding traffic analysis. If there is a shock wave, both variables can be classified into both free flow phase (BF), both congestion traffic phase (BC), propagation of traffic state associated with the downstream (PD), and propagation traffic state associated with the upstream.

(a) In the queue



(b) Not in the queue

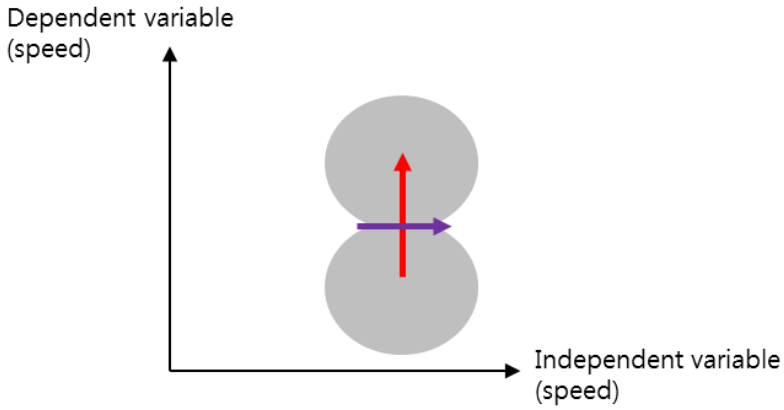


Figure 5.6 Overall vector space and PCs (a) in the queue, (b) not in the queue

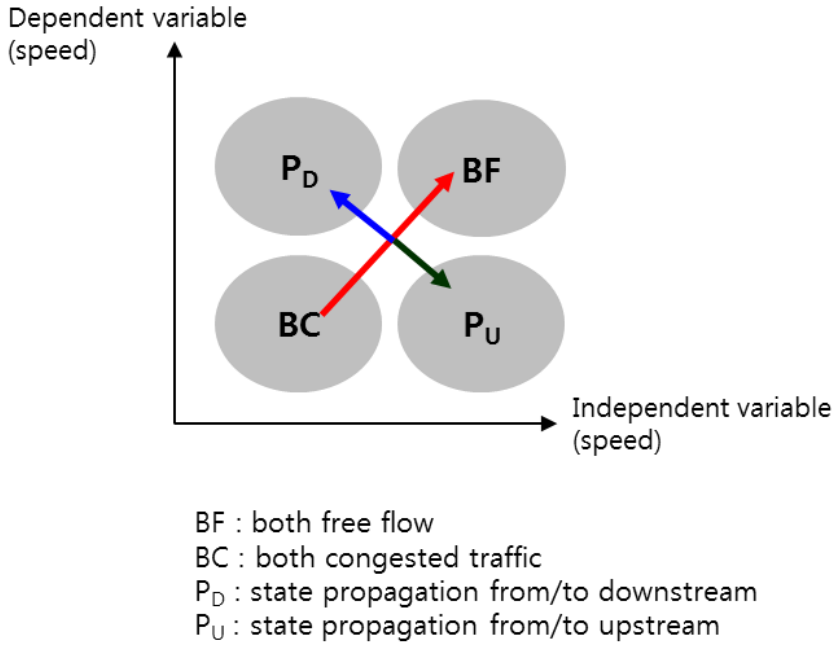


Figure 5.7 Spatiotemporal traffic phase diagram

(3) Identifying the PCs and its related variables

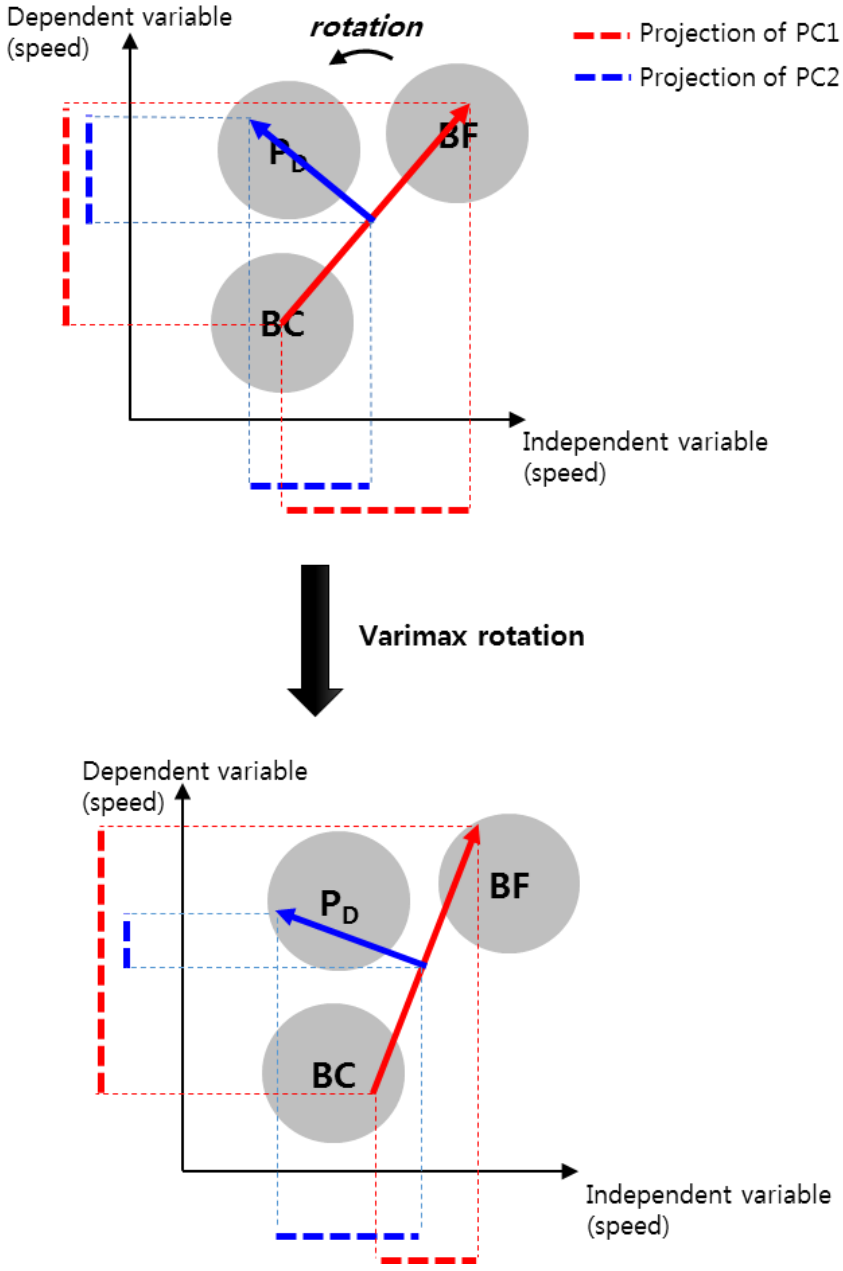
This sub-sections discussed two points. One is to identify the PCs representing the propagation of traffic state from the PCA results, and the other is to identify the spatiotemporal variables that describe the identified PC. The PCA generally used in machine learning as a feature extraction method in which the specific PCs are selected based on the PC's eigenvalue value and the entire data is projected to the corresponding space. However, this method is not easy to interpret in terms of variables because the original variables are transformed. In the proposed approach, the PCs which represents the propagation of traffic state are identified, and then the original variables identified that explains the identified PC for interpretation for traffic analysis.

To identify the PCs associated with the propagation of traffic state, the loading factor, which is the linear correlation coefficient between the original variables and the PC is used. However, since the result of initial PCA is not a simple structure, both propagations of traffic state related variables and dependent variables have maximum loading on the same largest PC. The result of PCA is changed when using the Varimax rotation that maintains orthogonality to make a simple structure, Figure 5.8. If there is the propagation of traffic state, the PC connecting the BF and the BC show the maximum loading of the dependent variable while the PC representing the propagation of traffic state show the maximum loading of the related independent variables. In this process, the dependent variable leaves some loading on the PC describing

the propagation of traffic state.

On the other hand, PCA result with variables that not in the queue are parallel to each variable, and the dependent variable has the loading of almost 0 in PC representing state change of other variables. Thus, a PC that is maximum loaded with dependent variable is PC that connects BF and BC, PCs with some loading of the dependent variable are PCs that describe the propagation of traffic state, and PCs with loading of dependent variables close to 0 can be identified as PCs that are not related to the propagation of traffic state. In addition, since independent variables that can explain each PC have max loading on each PC, spatiotemporal variables with max loading can be matched and identified and grouped. The change of loadings is represented in Table 5.1.

(a) In the queue (downstream)



(b) Not in the queue

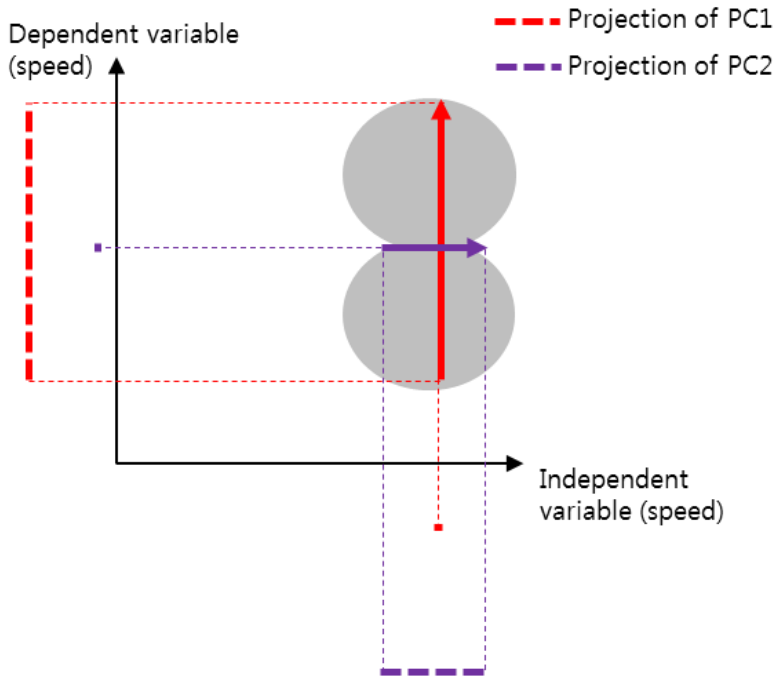


Figure 5.8 Change of loadings after Varimax rotation (a) in the queue (downstream), (b) not in the queue

Table 5.1 Loadings after Varimax rotation

(a) In the queue (before Varimax)

PC	**Dependent variable**	**Independent variable**
PC1	maximum	maximum
PC2	some	some

(b) In the queue (after Varimax)

PC	**Dependent variable**	**Independent variable**
PC1	maximum	some
PC2	some	maximum

(c) Not in the queue (after Varimax)

PC	**Dependent variable**	**Independent variable**
PC1	maximum	≒0
PC2	≒0	maximum

(4) Identifying PCs and variables in numerical example

In the two dimension, PCs are identified as described in subsection (3). It is necessary to examine whether the identification of the PCs and the variables related to the propagation of traffic state in two dimensions are the same manner in higher dimensions that is combined of upstream, downstream and not in the queue. Because it is difficult to explain in the high-dimensional case by using figure, it is explained by confirming PCA result of the numerical example.

Table 5.2 (a) shows the results of PCA using the 5-minute ago speed of all six detectors in the numerical example and the 15-minute future speed of MP 25.00 as a dependent variable. As a result, only the speed at MP20.00, which is not present in the queue, is loaded to PC 2 before the Varimax rotation, but all variables including dependent variable have maximum loading in PC 1 as described in subsection (3).

Table 5.2 (b) shows PCA loading values after the Varimax rotation. First, the maximum loading of the dependent variable is 0.440 in PC 4 and no other maximum loaded variables in the PC 4. The variables related to the propagation of the traffic state were loaded in PC 1 and PC 3 and the loadings of the dependent variable were 0.261 for PC 1 and the 0.289 for PC 3. Dependent variable loading on PC 2 is almost 0 and the speed at the MP 20.00 not in the queue was maximum loaded at PC 2. Taken together, it can be seen that interpretation of PCs for traffic analysis works at higher levels.

Table 5.2 Loadings in high dimensional PCA

(a) before Varimax

Variables	**PC1**	**PC2**	**PC3**	**…**
MP 20.00, -5min		1.000		
MP 21.67, -5min	0.905		0.079	
MP 23.33, -5min	0.951			
MP 25.00, -5min	0.968			
MP 26.67, -5min	0.954			
MP 28.33, -5min	0.907		0.072	
MP 25.00, +15min	0.915	0.000	0.003	

Note: The loadings of the independent variable is shown only over than 0.05

(b) after Varimax

Variables	**PC1**	**PC3**	**PC2**	**PC4**
MP 20.00, -5min			1.000	
MP 21.67, -5min	0.696			
MP 23.33, -5min	0.590			
MP 25.00, -5min	0.411			
MP 26.67, -5min		0.550		
MP 28.33, -5min		0.669		
MP 25.00, +15min	0.261	0.289	0.000	0.440

Note: The loadings of the independent variable is shown only maximum value

The PCA has conducted again with higher dimension by extending the independent variable up to 10 minutes ago, to make sure that the proposed approach can identify the PCs and variables related to the propagation of the traffic state. Table 5.3 shows that the approach described in the subsection (3) works consistently, even if the dimensions are enlarged.

Table 5.3 Loadings in higher dimensional PCA

(a) before Varimax

Variables	**PC1**	**PC2**	**PC5**	**...**
MP 20.00, -5min		0.509		
MP 20.00, -10min		0.502		
MP 21.67, -5min	0.905			
MP 21.67, -10min	0.890			
MP 23.33, -5min	0.949			
MP 23.33, -10min	0.936			
MP 25.00, -5min	0.965			
MP 25.00, -10min	0.954			
MP 26.67, -5min	0.950			
MP 26.67, -10min	0.942			
MP 28.33, -5min	0.905			
MP 28.33, -10min	0.902			
MP 25.00, +15min	0.863		0.109	

Note: The loadings of the independent variable is shown only over than 0.05

(b) after Varimax

Variables	**PC1**	**PC4**	**PC5**	**PC2**	**PC3**
MP 20.00, -5min				1.000	
MP 20.00, -10min					1.000
MP 21.67, -5min	0.596				
MP 21.67, -10min	0.689				
MP 23.33, -5min	0.494				
MP 23.33, -10min	0.595				
MP 25.00, -5min	0.368				
MP 25.00, -10min	0.449				
MP 26.67, -5min		0.458			
MP 26.67, -10min		0.547			
MP 28.33, -5min		0.547			
MP 28.33, -10min		0.640			
MP 25.00, +15min	0.201	0.245	0.552	0.000	0.000

Note: The loadings of the independent variable is shown only maximum value

(5) Occupancy and traffic volume data

Occupancy information is opposite in direction to speed information, but characteristics are quite similar. Plotting the speed of the dependent variable and the occupancy of the independent variable yields the same traffic phase diagram as reversing left to right of the speed-speed plot. Plotting the speed of the dependent variable and the occupancy of the independent variable yields the same traffic phase diagram as reversing left to right of the speed-speed plot. The traffic phase diagram looks different in Figure 5.9, but the PC that occurs in the space is the same as the speed case. Therefore, the speed and the occupancy of the same spatiotemporal data are likely to be loaded on the same PC.

In the case of traffic volume, there is a limit to explain the phenomenon of the propagation of the traffic state considering speed flow relation, because their relationship has the dual state. However, if there is a sudden change in traffic volume related to the propagation of the traffic state, it may be separated into some related PCs of the propagation of the traffic state.

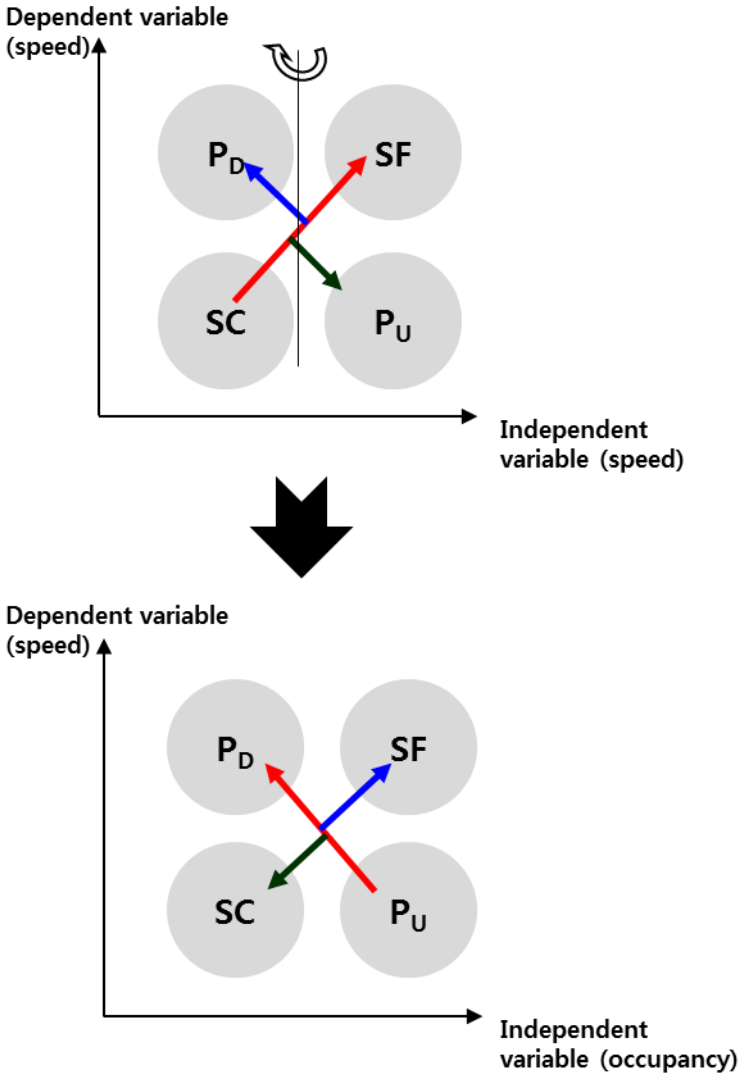


Figure 5.9 Spatiotemporal traffic phase diagram of speed-occupancy

5.2 Comparison of Selected Variables

Table 5.4 shows the results of the variable selection by the proposed approach and the other variable selection model. In the case of 15-minute prediction, Table 5.4. (a), the proposed model contains the information of the target point, the immediately upstream, the immediate downstream, and the information of the tale and the head of the queue. On the other hand, the RF based model also included both information of the tale and the head of the queue. The RF based model included more information about the near point and especially occupancy information of -17.8 km. The CCC based model did not include the information at the head of the queue and included the more information of the near points.

The comparison was conducted in the same manner for the 30-minute prediction and it is shown in Table 5.4 (b). The proposed model contains the information of the target point, the immediately upstream, the immediate downstream, and the information of the tale and the head of the queue. Compared to the 15-minute prediction, the upstream information was more added, including the traffic information at the point near Dongtan JC and especially at the starting point of hard shoulder running, -17.8 km. This can be regarded as a reflection of the fact that the propagation of the traffic state of the upstream evolves more slowly than the downstream.

The case of 5-minute prediction is shown in Table 5.4 (c). The result of the variable selection among models is not easy due to its similarity but the

proposed approach is the union of other models. The proposed model can capture the relevant feature variable than comparison models.

Table 5.4 Comparison of selected variable in Korean freeway (a) 15-minute (b) 30-minute, and (c) 5-minute prediction

(a) 15-minute prediction

Relative mile post	**Proposed approach**			**RF based SVM**			**CCC based SVM**		
	$\boldsymbol{\mu}$	$\boldsymbol{o}$	$\boldsymbol{q}$	$\boldsymbol{\mu}$	$\boldsymbol{o}$	$\boldsymbol{q}$	$\boldsymbol{\mu}$	$\boldsymbol{o}$	$\boldsymbol{q}$
4.2 km	$-1t$			$-1t$					
2.4 km				$-1t$	$-1t$		$-1t$	$-1t$	
1.5 km	$-1t$			$-1t$ $-2t$	$-1t$		$-1t$ $-2t$ $-3t$	$-1t$ $-2t$	
0 km	$-1t$ $-2t$ $-3t$			$-1t$ $-2t$ $-3t$	$-1t$		$-1t$ $-2t$ $-3t$	$-1t$	
-1 km	$-1t$ $-2t$			$-1t$ $-2t$	$-1t$		$-1t$ $-2t$ $-3t$	$-1t$	
-2 km	$-1t$	$-1t$ $-2t$		$-1t$	$-1t$		$-1t$ $-2t$		
-5 km									
-6.6 km									
-8.9 km					$-1t$				
-17.8 km					$-1t$				

(b) 30-minute prediction

Relative mile post	**Proposed approach**			**RF based SVM**			**CCC based SVM**		
	$\boldsymbol{\mu}$	$\boldsymbol{o}$	$\boldsymbol{q}$	$\boldsymbol{\mu}$	$\boldsymbol{o}$	$\boldsymbol{q}$	$\boldsymbol{\mu}$	$\boldsymbol{o}$	$\boldsymbol{q}$
4.2 km	$-1t$ $-2t$								
2.4 km					$-1t$				
1.5 km	$-1t$			$-1t$	$-1t$		$-1t$	$-2t$	
0 km	$-1t$			$-1t$			$-1t$ $-2t$		
-1 km	$-1t$			$-1t$			$-1t$ $-3t$	$-2t$	
-2 km	$-1t$			$-1t$	$-1t$ $-2t$		$-1t$ $-2t$	$-1t$	
-5 km		$-1t$							
-6.6 km		$-1t$ $-2t$							
-8.9 km					$-1t$				
-17.8 km			$-1t$ $-2t$						

(c) 5-minute prediction

Relative mile post	**Proposed approach**			**RF based SVM**			**CCC based SVM**		
	μ	***o***	***q***	***μ***	***o***	***q***	***μ***	***o***	***q***
4.2 km									
2.4 km									
1.5 km	-1t -2t	-1t		-1t	-1t		-1t -2t	-1t	
0 km	-1t -2t	-1t		-1t -2t	-1t		-1t -2t	-1t	
-1 km	-1t -2t	-1t		-1t -2t			-1t -2t	-1t	
-2 km	-1t	-1t		-1t	-1t		-1t		
-5 km									
-6.6 km									
-8.9 km									
-17.8 km									

The -17.8 km point is the starting point of the hard shoulder running. The traffic volume of this point was selected in the 30-minute prediction of the proposed model. From now on, it is discussed whether the operation of the hard shoulder running can affect the prediction of the target point and whether the traffic volume change at this point is significant in the 30-minute prediction.

The bi-plots between the 30-minute future of target point and the traffic volume at -17.8 km of 5 minute ago or at -13.8 km of 5 minute ago, or at -19.6 km of 5 minute ago are shown in Figure 5.9. The -19.6 km upstream from the -17.8 km point was not operating the hard shoulder running, there is no data point on the upper right-hand side of the bi-plot in Figure 5.10 because the traffic flow during the 5 minutes did not exceed 400. However, -17.8 km and its downstream point of -13.8 km, they were operating the hard shoulder running, they had a data point at the upper right corner in bi-plot as seen in Figure 5.10. The first PCs of each case were connected the BF and BC phase as same as in the speed-occupancy bi-plot. Then, the second PC directed the data point made when the hard shoulder running was operated. The second PC is at the right side of the first PC and it is the same as P_D in the case of speed or occupancy. This interpretation supports that the traffic volume related to the hard shoulder running was included in the PC such as the speed of 4.2 km downstream that showed the propagation of the traffic state.

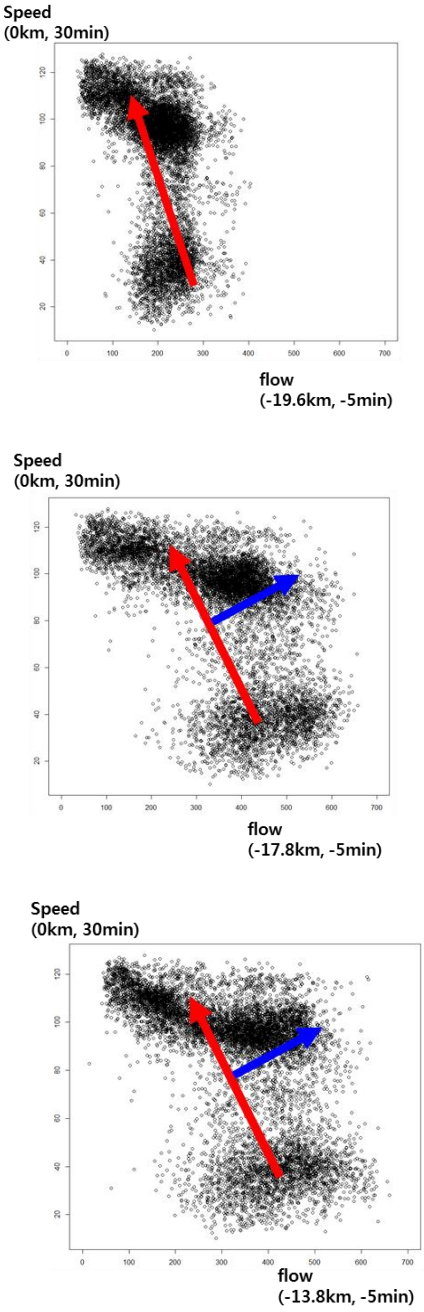


Figure 5.10 Bi-plot between $\boldsymbol{\mu}_{\boldsymbol{t=-6t}}^{\boldsymbol{i=0km}}$ and $\boldsymbol{q}_{\boldsymbol{t=-1t}}^{\boldsymbol{i=-13.8km}}$, $\boldsymbol{q}_{\boldsymbol{t=-1t}}^{\boldsymbol{i=-17.8km}}$, and $\boldsymbol{q}_{\boldsymbol{t=-1t}}^{\boldsymbol{i=-19.6km}}$

In the case of the Interstate freeway, the proposed model showed the best predictive performance simply using -1.1 km upstream and 1.3 km downstream and the target point, but the other model predicted performance was lower even they selected the variables from the wide range. In the 5-minute prediction, the feature selection of the proposed approach can be interpreted as mitigating overfitting and capturing traffic flow characteristics well.

In the case of 15-minute prediction, the proposed model expanded the spatiotemporal range as the prediction step became larger. On the other hand, the RF-based model and the CCC-based model are interpreted as the fact that the spatiotemporal range was rather reduced and the traffic flow characteristics are not properly reflected. In particular, the CCC-based model did not use upstream information at all.

Table 5.5 Comparison of selected variable in interstate freeway (a) 5-minute (b) 15-minute prediction

(a) 5-minute prediction

Relative mile post (mile)	**Proposed approach**			**RF based SVM**			**CCC based SVM**		
	μ	***o***	***q***	***μ***	***o***	***q***	***μ***	***o***	***q***
4.38									
3.56				-1t			-1t -2t		
2.37				-1t				-3t	-4t
1.30	-1t	-2t			-1t		-2t		
0.2				-1t					-1t
0.0	-1t -4t				-2t			-2t	
-1.11	-1t	-1t -2t							
-1.43								-2t	

(b) 15-minute prediction

Relative mile post (mile)	**Proposed approach**			**RF based SVM**			**CCC based SVM**		
	μ	***o***	***q***	***μ***	***o***	***q***	***μ***	***o***	***q***
4.38			-1t -4t						
3.56			-4t						
2.37	-1t	-2t		-1t -2t	-1t		-1t -2t -3t		
1.30	-1t			-1t	-1t			-1t -2t	
0.2	-1t			-2t	-2t			-2t	
0.0	-1t			-1t -3t	-1t		-1t	-1t	
-1.11	-1t	-1t -2t		-1t	-1t				
-1.43		-1t							

Chapter 6. Conclusions

This study proposed a machine learning approach for the freeway traffic speed prediction. The proposed machine learning approach is a prediction model that includes a variable selection that mitigates overfitting and captures traffic characteristics well. The objective of the proposed approach is to provide interpretation of what traffic phenomena are important for prediction from variable selection process and result.

In the proposed machine learning approach, variables are grouped using supervised PCA, and variable ranking of within the group is determined using CCC, which is a nonlinear correlation coefficient. Then the variables are selected using variable ranking and SVM model in the forward selection manner which adds the variables that improve the predictive performance. The empirical data evaluation for Korean freeway and U.S.A. interstate freeway were conducted. The proposed approach showed better performance than other variable selection models using variable ranking from the ensemble learning and the nonlinear correlation coefficient alone. Also, the proposed approach showed better predictive performance than the models using all variables, the RF model and the SAE model which is an unsupervised feature extraction model. Although the SAE model is not an optimal model, the proposed research technique is a dimensional reduction technique that very well describes the traffic characteristics. The computational cost of the proposed approach was

also more efficient than the ensemble learning and unsupervised feature extraction methods.

The implications of the proposed method were explored for the traffic analysis. The PC loading using dependent variables can reflect the phenomenon of propagation of the traffic state and it can be interpreted that it is important for speed prediction to reflect statistically different propagation properly. This indicates that predictive capability can be improved if the rapid changes in the traffic state, switch between the free flow and the congested traffic, can be predicted by reflecting the propagation of the traffic state from other time-space. The proposed method selected the appropriate variables according to the horizon step of the prediction and showed great predictive capability. Therefore, it is concluded that reflecting the wide range of propagation of the traffic state is an important factor to improve the predictive ability of the learning-based multi-step prediction model.

Unlike other machine learning models, the proposed model can provide information about the spatiotemporal congestion mechanism. The proposed approach can be used not only to provide information but also to establish an operation strategy for managing congestion and congestion-related crashes.

The proposed approach has a limitation of the data-driven approach. It can greatly be influenced by the data quality, Missing values, outliers, noise, and so on. In addition, the proposed method provides an interpretation of the

features that were limited in the machine learning approach, but it has a limitation that it cannot identify the microscopic mechanism. The information of the congestion mechanism obtained by the proposed approach can be used to determine the analysis focus in advance when the microscopic analysis is used to interpret the details of the congestion mechanism.

For the future research, the detailed mechanism can be identified through microscopic analysis using the congestion mechanism analyzed by the proposed approach. It is possible to investigate the common characteristics of the traffic flow from the spatiotemporal variables in the same PC. The impact of the data quality must be evaluated. Statistical simulation analysis can be used to assess the impact of noise on the proposed approach. The PCA and the CCC in the proposed approach can be substituted by similar methods. The PCA can be substituted for nonlinear PCA, and Granger causality can be applied instead of the CCC. Further research is needed to confirm the applicability to urban roads with signalized intersections or to use for non-recurrent congestion by the incident.

Reference

Abdulhai, B., Porwal, H., and Recker, W. (1999). Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks.

Ahmed, M. S., and Cook, A. R. (1979). *Analysis of freeway traffic time-series data by using Box-Jenkins techniques* (No. 722).

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., and Mishalani, R. (1998, February). DynaMIT: a simulation-based system for traffic prediction. In *DACCORD Short Term Forecasting Workshop* (pp. 1-12).

Cameron, G. D., and Duncan, G. I. (1996). PARAMICS—Parallel microscopic simulation of road traffic. *The Journal of Supercomputing*, *10*(1), 25-53.

Chen, Y., Van Zuylen, H. J., and Qipeng, Y. A. N. (2010, January). Travel time prediction on urban networks based on combining rough set with support vector machine. In *Logistics Systems and Intelligent Management, 2010 International Conference on* (Vol. 1, pp. 586-589). IEEE.

Cristianini, N., and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Davis, G. A., Nihan, N. L., Hamed, M. M., and Jacobson, L. N. (1990). Adaptive forecasting of freeway traffic congestion. *Transportation Research Record*, (1287).

Dia, H. (2001). An object-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research*, *131*(2), 253-261.

Dougherty, M. S., and Cobbett, M. R. (1997). Short-term inter-urban traffic forecasts using neural networks. *International journal of forecasting*, *13*(1), 21-31.

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*(1), 5-16.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

Halati, A., Lieu, H., and Walker, S. (1997). CORSIM-corridor traffic simulation model. In *Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and OpportunitiesUrban Transportation Division, ASCE; Highway Division, ASCE; Federal Highway Administration, USDOT; and National Highway Traffic Safety Administration, USDOT.*

Hamed, M. M., Al-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.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, *1*(1), 27-35.

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.

Jayakrishnan, R., Mahmassani, H. S., and Hu, T. Y. (1994). An evaluation tool for advanced traffic information and management systems in urban networks. *Transportation Research Part C: Emerging Technologies*, *2*(3), 129-147.

Kerner, B. S. (2009). Earlier Theoretical Basis of Transportation Engineering: Fundamental Diagram Approach. In *Introduction to Modern Traffic Flow Theory and Control* (pp. 173-219). Springer, Berlin, Heidelberg.

Lee, Y. (2009, July). Freeway travel time forecast using artifical neural networks with cluster method. In *Information Fusion, 2009. FUSION'09. 12th International Conference on* (pp. 1331-1338). IEEE.

Li, C. S., and Chen, M. C. (2014). A data mining based approach for travel time prediction in freeway with non-recurrent congestion. *Neurocomputing*, *133*, 74-83.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 94.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F. Y. (2015). Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 865-873.

Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, *54*, 187-197.

May, A. D. (1990). *Traffic flow fundamentals*.

Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V. (1997, October). Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks* (pp. 999-1004). Springer, Berlin, Heidelberg.

Müller, K. R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V. (1999). Using support vector machines for time series prediction. *Advances in kernel methods—support vector learning*, 243-254.

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

Papageorgiou, M., Papamichail, I., Messmer, A., and Wang, Y. (2010). Traffic simulation with metanet. In *Fundamentals of traffic simulation* (pp. 399-430). Springer, New York, NY.

Shi, Q., and Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, *58*, 380-394.

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

Sun, H., Liu, H., Xiao, H., He, R., and Ran, B. (2003). Use of local linear regression model for short-term traffic forecasting. *Transportation Research Record: Journal of the Transportation Research Board*, (1836), 143-150.

Van Aerde, M., and Rakha, H. (2010). *INTEGRATION Rel. 2.30 for Windows-User's Guide, Volume I*. Volume II. Technical report, M. Van Aerde & Associates, Ltd., Blacksburg, Virginia.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), 988-999.

Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J., and Olsson, M. C. (2016). Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness. *Sensors*, *16*(4), 592.

Vlahogianni, E. I., Golias, J. C., & Karlaftis, M. G. (2004). Short-term traffic forecasting: Overview of objectives and methods. *Transport reviews*, *24*(5), 533-557.

Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2007). Spatio-Temporal Short-Term Urban Traffic Volume Forecasting Using Genetically

Optimized Modular Networks. *Computer-Aided Civil and Infrastructure Engineering*, *22*(5), 317-325.

Whittaker, J., Garside, S., and Lindveld, K. (1997). Tracking and predicting a network traffic process. *International Journal of Forecasting*, *13*(1), 51-61.

Xu, C., Tarko, A. P., Wang, W., and Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis and Prevention*, *57*, 30-39.

Yao, B., Chen, C., Cao, Q., Jin, L., Zhang, M., Zhu, H., and Yu, B. (2017). Short-Term Traffic Speed Prediction for an Urban Corridor. *Computer-Aided Civil and Infrastructure Engineering*, *32*(2), 154-169.

Yildirim, U., and Çataltepe, Z. (2008, October). Short time traffic speed prediction using data from a number of different sensor locations. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on* (pp. 1-6). IEEE.

Yildirimoglu, M., and Geroliminis, N. (2013). Experienced travel time prediction for congested freeways. *Transportation Research Part B: Methodological*, *53*, 45-63.

Yu, R., and Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis and Prevention*, *51*, 252-259.

Zheng, Z., Ahn, S., Chen, D., and Laval, J. (2011). Applications of wavelet transform for analysis of freeway traffic: Bottlenecks, transient traffic, and traffic oscillations. *Transportation Research Part B: Methodological*, *45*(2), 372-384.

Zhong, M., Sharma, S., and Lingras, P. (2005). Refining genetically designed models for improved traffic prediction on rural roads. *Transportation planning and technology*, *28*(3), 213-236.

국문초록

통행 속도 예측은 예측 기반 교통운영관리에 활용되어 지능형 교통체계의 서비스를 개선할 수 있다. 시뮬레이션 기반 속도 예측은 인과관계를 설명할 수 있으나 정확성에 한계를 갖고 있는 반면, 자료기반 기법은 인과관계를 설명할 수 없으나 우수한 예측성능을 갖고 있다. 특히 기계학습 기법을 이용하는 비모수적 기법의 성능이 우수한데, 과적합을 완화하고 교통혼잡 특성을 반영하기 위한 특징 추출이나 선택을 이용하여 예측성능을 높이고 있다. 이러한 비모수 기반 속도 예측은 높은 성능으로 인해 정보시스템으로는 활용이 가능하지만, 혼잡 매커니즘에 대한 정보를 제공하지 못해 교통운영관리에 사용하기 위한 전략 수립에 어려움이 있다.

본 연구는 혼잡 매커니즘에 대한 정보를 제공하면서 예측성능이 우수한 특징 선택기법을 이용한 기계학습기반의 고속도로 속도 예측 모델을 제안하고자 한다. 이를 위해 주성분 분석과 지지벡터머신을 이용한 지도(supervised) 특징선택을 이용한다. 예측 목표인 종속변수를 포함한 주성분 분석을 수행하고, 단순구조를 얻기 위한 Varimax 회전을 사용한다. 최대 적재(loading)를 이용하여 각 변수를 주성분에 할당하여 그룹화 한 뒤, 종속변수의 분산인 적재값을 이용하여 순위화하면, 종속변수의 변량을 설명할 가능성이 높은 순서대로 주성분 및 주성분에 포함된 변수를 식별할 수 있게 된다. 이어서, 주성분 내 변수들 간에는 기계학습 모형에서의 예측성능을 순위화 하기 위해 비선형 성관계수를 사용하여 순위화한다. 본 연구에서는 교차상관(cross-correlation)계수를 사용하였다. 변수를 그룹화하고, 그룹 간/그룹 내 우선순위를 이용하여 변수를 순차적으로 투입해 나가는 전향선택(forward selection) 방식의 변수선택 기법을 사용한다. 기계학습 회귀모델로는 일반화 성능이 우수하며, 계산비용도 낮은 지지벡터머신의 회귀모델을 이용한다.

제안된 특징선택 기법은 한국의 경부고속도로 일부 구간과 미국의

interstate (I-880)고속도로에서 경험자료를 통해 평가하였다. 그 결과 특징을 잘 선택해 내었으며, 특히 경험적 연구에서 앙상블학습 기법인 random forest, 비지도 특징추출기법인 stacked auto-encoder 기반의 인공신경망 기법보다 다소 우수한 차원감소 성능을 보이고, 계산 비용에서도 상당히 효율적이었다.
제안된 특징선택 기법의 교통공학적 함의를 도출하기 위해 주성분 분석의 벡터공간(vector space)을 두 시공간 변수간의 traffic phase diagram으로 변환하여 해석해 보았다. 그 결과 종속변수 적재가 있는 주성분들은 한 시공간의 교통상태가 다른 시공간으로 전이(propagation)되는 현상을 설명할 수 있으며, 통계적으로 서로 다른 상태전이를 구별할 수 있다. 또한 제안된 기법을 사용한 모델의 선택변수와 다른 기법(random forest, cross-correlation 계수)을 이용한 선택 변수와 비교 결과, 제안된 모델이 가장 시공간적으로 폭넓게 변수를 선택하였으며 혼잡 대기행렬의 head, tail, 가변차로 운영 시점부 등을 prediction step에 따라 적절하게 반영하는 것으로 나타났다. 제안된 특징선택 기법은 속도예측을 위한 교통상태 변화를 적절히 선정해내어, 혼잡 매커니즘에 대한 정보를 제공할 수 있는 고속도로 속도 예측을 구현하였다. 또한 다른 차원축소 기법에 비해서도 평가된 경험적 분석에서 가장 우수한 예측 성능을 보이는 것으로 나타났다. 특히, prediction step에 따라 적절한 특징을 선택하여 multi-step prediction에 우수한 성능을 보였다. 따라서 본 연구가 제안한 기계학습 접근법은 고속도로 혼잡관리 및 혼잡에 의한 추돌사고 관리를 위한 교통운영 전략수립에 활용될 수 있을 것으로 기대된다.

주요어: 고속도로 속도 예측, 지지벡터머신, 주성분 분석, 특징 선택, 교통상태 전이
학번: 2013-30982