Optimizing Traffic Prediction Performance of Neural Networks under Various Topological, Input, and Traffic Condition Settings

By
Sherif Ishak[1]
Department of Civil and Environmental Engineering
Louisiana State University
Baton Rouge, LA 70803
Phone: (225)-578-4846
Fax: (225)-578-8652
Email: SISHAK@LSU.EDU

And

Ciprian Alecsandru
Graduate Student
Department of Civil and Environmental Engineering
Louisiana State University
Baton Rouge, LA 70803

Paper submitted for presentation at the
Transportation Research Board
82nd Annual Meeting
Washington, D.C. 2003
and for publication in the Transportation
Research Record journal

Revised and Re-Submitted on October 25, 2002

[1] Corresponding Author

ABSTRACT

This paper presents an approach to optimize the short-term traffic prediction performance on freeways using multiple artificial neural network topologies under different network and traffic condition settings. The study was conducted to encourage multi-model techniques that are capable of improving the performance over single-model approaches. Using a mix of traditional and modern neural network topologies, the short-term speed prediction performance was extensively evaluated under different input settings and various prediction horizons (from 5 to 20 minutes). The input patterns were constructed from the most recent past data collected at multiple detector stations. To enable the network to learn from historical information, a long-term memory component was attached to the input patterns in the form of a time index. This allows the network to build internal representation of recurrent conditions that are likely to be observed at the same time of day, in addition to the short-term memory that is often encapsulated in the most recent information. Optimal settings were determined by maximizing the performance under different traffic conditions observed at the target location, as well as upstream and downstream locations. Comparative statistical analysis with naïve and heuristic approaches showed that the optimized neural network approach resulted in much better performance. The study shows that no single topology outperformed the others. However, for nearly 33% of the cases, the Co-Active Neuro-Fuzzy Inference System (CANFIS) topology was found optimal. The study also shows that the optimal settings were consistently more dependent on the long-term memory component as prediction horizon increases. Interestingly enough, the optimization of prediction performance under different traffic conditions allows for the identification of situations when prediction accuracy is unacceptable by traffic information dissemination systems.

KEYWORDS

Traffic prediction, speed prediction, freeway operation, artificial neural networks, and performance optimization.

INTRODUCTION

Real time and predictive information on traffic conditions is becoming increasingly critical to the successfulness of advanced traffic management systems (ATMS). With the remarkable advancement in data collection and dissemination technology, the traveling public has now the ability to access such information at both pre-trip planning stage and en-route. With accurate and reliable information, travelers can make appropriate decisions to bypass congested segments of the network, change departure times and/or destination, whenever appropriate. Such decisions are likely to affect the potential demand at various points on the network and provide opportunities for better utilization of the existing infrastructure capacity. On the other hand, traffic management centers need such information to support their primary management and control functions.

Urban freeways, a primary source of mobility, are heavily traveled by both commuters and non-commuters who continuously experience excessive delays and queuing conditions on a daily basis. Freeway travelers often seek information on traffic conditions in the form of travel times and delays along their trips. Such information provides an indirect measure of travel cost, which most travelers seek to minimize. Since travel decisions are often made in advance, travelers are often seeking predictive information on traffic conditions along their selected route and within the duration of the trip. Research in the past few years has addressed this need with a variety of short-term traffic prediction models that attempt to capture the dynamic nature of traffic conditions. A review of the literature on previously developed models and research activities in this area reveals that there is still a need to improve on the prediction performance and capability of existing models. This research study presents an approach to optimize the performance of neural networks in short-term traffic prediction by exploring different network topologies and settings under different traffic conditions.

BACKGROUND

The concept of intelligent transportation systems has introduced several functions and services that have the potential to improve the efficiency, safety and productivity of our surface transportation system. Among them is our ability to make better decisions on our travel choices when current and predicted information becomes available. This concept has motivated researchers to seek traffic prediction models that are capable of forecasting traffic flow, speed, and travel times in short terms (from 5 to 30 minutes). Several research efforts were thus conducted in the last few years to support ITS applications and provide the travelers with travel time information at the pre-trip planning stage and en-route. Kaysi et al. (*1*) and Ben-Akiva et al. (*2*) recommended that traffic routing strategies under recurring and non-recurring congestion be based on forecasting of future traffic conditions rather than historical and/or current traffic conditions. This is because travelers' decisions are affected by future traffic conditions rather than current traffic conditions. Several prediction methods have been implemented in research in the past two decades. Ben Akiva et al. (*3*) grouped those methods into three categories: (a) statistical models, (b) macroscopic models, and (c) route choice models based on dynamic traffic assignment. Time series models have been extensively used in traffic forecasting for their simplicity and strong potential for on-line implementation (see for example; *4-14*).

Recently, Chen and Chien (*15*) conducted a study using probe vehicle data to compare the prediction accuracy under direct measuring of path-based travel time versus link-based travel times. The study showed that under recurrent traffic conditions, path-based prediction is more accurate than link-based prediction. Chien and Kuchipudi (*16*) presented the results of using real-time and historical data for travel time prediction. Another study by Kwon et al. (*17*) used an approach to estimate travel time on freeways using flow and occupancy data from single loop detectors and historical travel time information. Forecasting ranged from a few minutes into the future up to an hour ahead. The study showed that current traffic conditions are good predictors for the near future, up to 20 minutes, while long-range predictions need the use of historical data. Another study by Smith et al. (*18*) compared parametric and nonparametric models for traffic flow forecasting. The study showed that prediction under seasonal autoregressive integrated moving average (ARIMA), a parametric modeling approach to time series, outperforms other non-parametric approaches, like regression based on heuristically improved forecast generation.

Recently, several studies have investigated the use of artificial neural networks to model short-term traffic prediction. For instance, Park and Rilett (*19*) proposed two modular Artificial Neural Networks (ANN) models for forecasting multiple-period freeway link travel times. One model used a Kohonen Self Organizing Feature Map (SOFM) while the other utilized a fuzzy c-means clustering technique for traffic patterns classification. Rilett and Park (*20*) proposed a one-step approach for freeway corridor travel time forecasting rather than link travel time forecasting. They examined the use of a spectral basis neural network with actual travel times from Houston, Texas. Another study by Abdulhai et al. (*21*) used an advanced time delay neural network (TDNN) model, optimized using a Genetic Algorithm, for traffic flow prediction. The results of the study indicated that prediction errors were affected by the variables pertinent to traffic flow prediction such as spatial contribution, the extent of the loop-back interval, resolution of data, and others. Lint et al. (*22*) presented an approach for freeway travel time prediction with state-space neural networks. Using data from simulation models, they showed that prediction accuracy was acceptable and favorable to traditional models. Several other studies applied neural networks for predicting speed, flows, or travel times (see for instance; *23-28*).

OBJECTIVES

The objective of this study is to optimize the performance of neural networks in short-term traffic prediction of speed on freeways. The study investigates the performance of four different neural network architectures under different settings and traffic conditions. In addition to the previously applied architectures (MLP and Modular networks), two new architectures are introduced: the PCA (Principal Component Analysis) network and the Co-Active Neuro-Fuzzy Inference System (CANFIS). In each network, we investigate the effect of introducing a long-term memory component, in addition to the short-term memory components, on the prediction performance in the range of 5 to 20 minutes. To ensure optimal performance under different settings, the prediction performance was tested under a wide spectrum of spatial and temporal traffic conditions.

DESCRIPTION OF THE NEURAL NETWORK MODELS

Four different architectures were applied in the study to compare the prediction performance of each under different configurations and traffic conditions. Each model is described briefly in this section.

Multi-Layer Perceptron (MLP) Network

The MLP has been extensively used in many transportation applications for its simplicity and ability to perform nonlinear pattern classification and function approximation. It is, therefore, considered the most widely implemented network topology by many researchers (*29*, *30*). Its mapping capabilities are believed to approximate any arbitrary function. An MLP consists of three types of layers: input, hidden, and output. It is normally trained with the backpropagation algorithm, which is based on minimizing the sum of squared errors between the desired and actual outputs. Since this topology is very common, it has been repeatedly explained in detail in the literature, and thus, will not be covered here to avoid redundancy and to allow other new topologies to be explained.

Modular Network

Modular networks are a special class of multiple parallel feed-forward MLPs. The input is processed with several MLPs and then the results are recombined. This type offers specialization of function in each sub-module and does not require full interconnectivity between the MLP's layers. Modular networks are often faster to train due to the smaller number of weights for the same size network. The topology used in this study is composed of two primary components: local expert networks and a gating network (*31, 32*). The basic idea is linked to the concept "divide-and-conquer", where a complex system is better attacked when divided into smaller problems, the solutions of which lead to the solution of the entire system. Using a modular network, a given task will be split up among some local expert networks, thus reducing the load on each in comparison with one single network that must learn to generalize from the entire input space. A gating network eventually combines the output from the local experts to produce an overall output. FIGURE 1(a) shows the topology of the modular network.

Hybrid Principal Component Analysis (PCA) Network

PCA is a technique that finds an orthogonal set of directions in the input space and provides a way to find the projections into these directions in an ordered fashion. The orthogonal directions are called eigenvectors of the correlation matrix of the input vector and the projections of the corresponding eigenvalues. PCA has the ability to reduce the dimensionality of the input vectors, and therefore, can be used for data compression. When used in conjunction with MLP, the PCA can reduce the number of inputs to the MLP and improve its performance. In this study a hybrid PCA/MLP network is used, combining both unsupervised and supervised learning paradigms in one topology. The PCA projects the input vector onto a smaller dimensional space, and thus, compressing the input for the MLP network. It should be emphasized that PCA is a well known statistical procedure that is used in feature extraction from high-dimensional space (*29, 30, 32*). The topology of the hybrid PCA network is illustrated in FIGURE 1(b).

Co-Active Neuro-Fuzzy Inference System (CANFIS)

CANFIS belongs to a more general class of adaptive neuro-fuzzy inference systems (ANFIS) (*31*). In the context of this study, CANFIS is used as a universal approximator of any nonlinear function. The characteristics of CANFIS are emphasized by the advantages of integrating neural networks with fuzzy inference systems (FIS) in the same topology. The powerful capability of CANFIS stems from pattern-dependent weights between the consequent layer and the fuzzy association layer. Like the radial-basis function network (RBFN), CANFIS is locally tuned. The architecture of CANFIS is illustrated in FIGURE 1(c). The fundamental component for CANFIS is a fuzzy axon that applies membership functions (MF) to the inputs. Two membership functions are commonly used: general bell and Gaussian. The network contains also a normalization axon to expand the output into the range of 0 to 1. The second major component in the type of CANFIS used in this study is a modular network that applies functional rules to the inputs. The number of modular networks matches the number of network outputs and the number of processing elements in each network corresponds to the number of MFs. CANFIS also has a combiner axon that applies the MF outputs to the modular network outputs. Finally, the combined outputs are channeled through a final output layer and the error is backpropagated to both the MFs and the modular networks. The CANFIS architecture used in this study is composed of four layers as shown in FIGURE 1(c). The function of each layer is described as follows. Each node in layer 1 is the membership grade of a fuzzy set (A, B, C, or D) and specifies the degree to which the given input belongs to one of the fuzzy sets, which are defined by three membership functions. Layer 2 receives input in the form of the product of all output pairs from the first layer. The third layer has two components. The upper component applies the membership functions to each of the inputs, while the lower component is a representation of the modular network that computes, for each output, the sum of all the firing strengths. The final layer calculates the weight normalization of the output of the two components from the third layer and produces the final predictions of speed at different prediction horizons.

STUDY AREA AND DATA COLLECTION

The study was conducted on a freeway segment of I-4 in Orlando, Florida, using 30-second loop detector speed data. The traffic surveillance system compiles data from a 40-mile six-lane corridor instrumented with 70 inductive dual loop detector stations spaced at nearly 0.5 miles in both directions. The real time and archived data is accessible on the web at the URL: trafficinfo.engr.ucf.edu. The loop detector data is collected in real time via a dedicated high-speed link between the I-4 Regional Traffic Management Center (RTMC) in Orlando and the intelligent transportation system lab at the University of Central Florida. Speed, volume counts and lane occupancies are downloaded and compiled into an SQL server that supports multiple publicly accessible web applications such as real time and short-term travel time predictions between user-selected on- and off-ramps. The web-based short-term traffic prediction system was implemented using a nonlinear time series model that was tested extensively in a previous study (*14*). In this study data was collected from three adjacent stations over 1-mile section. A total of 28 days were randomly selected in the year 2001 during the morning peak period from 6:00 AM to 10:00 AM in the westbound direction. To reduce the noise and random fluctuations in speed, the data resolution was reduced from 30-second averages to 5-minute averages using a 5-minute moving time window that is updated every minute. The recent past 5-minute speed averages were based on data observed as far back as 10 minutes from the current time. The 5-minute moving average and the 10-minute rolling horizon were arbitrarily selected for this study. Further testing may deem necessary to verify if the selected values have a significant impact on the predictive performance of the adopted approach. The 28 days were randomly split into 14 days for training, 4 for validation, and 10 for testing.

TRAINING

Training each of the four network topologies was conducted using NeuroSolutions (*33*). In order to achieve optimal performance, different settings were attempted by varying the number and type of inputs to each network. The input to each network was classified into two components: short-term memory (STM) and long-term memory (LTM). The STM component was represented by speed data observed in the past 10 minutes at the three adjacent stations. Input patterns were constructed over time and space to capture temporal and spatial variations of traffic conditions. However, to optimize the performance of the networks, input patterns were constructed from three spatial settings: current and upstream (*y*, *z*), current and downstream (*y*, *x*), or the three stations combined (*x*, *y*, *z*). The LTM component was introduced in addition to the STM component to test the network's ability to learn from similar historical traffic conditions observed at the same time on other days. The LTM component was represented by a time index attached to each constructed speed pattern. The time index was referenced to the beginning of the peak period at 6:00 AM and expressed in increments of 1 minute. The essence of using LTM component is to improve the performance at relatively longer prediction horizons by making the network time-cognizant during prediction. Each network was trained to predict the average 5-minute speeds at 5, 10, 15, and 20-minute horizons. During the training phase the performance of the networks was monitored via the validation set to avoid overtraining. Training is terminated when the mean square error (MSE) for the cross-validation set does not decrease for 50 consecutive training cycles, a common procedure to prevent overtraining.

Except for the hybrid PCA network, the input patterns representing STM were in the form:

Type 1 input: $\{x(t), x(t\text{-}5), y(t), y(t\text{-}5)\}$
Type 2 input: $\{y(t), y(t\text{-}5), z(t), z(t\text{-}5)\}$
Type 3 input: $\{ x(t), x(t\text{-}5), y(t), y(t\text{-}5), z(t), z(t\text{-}5)\}$
Where
$x(t)$ = average 5-minute speed at downstream station at time t
$x(t\text{-}5)$ = average 5-minute speed at downstream station at time t-5
$y(t)$ = average 5-minute speed at current station at time t
$y(t\text{-}5)$ = average 5-minute speed at current station at time t-5
$z(t)$ = average 5-minute speed at upstream station at time t
$z(t\text{-}5)$ = average 5-minute speed at upstream station at time t-5

The hybrid PCA network has the ability to reduce the dimensionality of the input space by locating the principal components. Therefore, the input patterns were presented in high dimensional vectors of 10 observations taken one minute apart from time *t*. In other words, the input patterns were in the form:

Type 1 input: $\{x(t), x(t\text{-}1),\ldots x(t\text{-}9), y(t), y(t\text{-}1),\ldots y(t\text{-}9)\}$
Type 2 input: $\{y(t), y(t\text{-}1),\ldots y(t\text{-}9), z(t), z(t\text{-}1),\ldots z(t\text{-}9)\}$
Type 3 input: $\{x(t), x(t\text{-}1),\ldots x(t\text{-}9), y(t), y(t\text{-}1),\ldots y(t\text{-}9), z(t), z(t\text{-}1),\ldots z(t\text{-}9)\}$

The output vector for each input pattern was constructed in the form $\{S(t+5), S(t+10), S(t+15), S(t+20)\}$, where $S(t+5)$, $S(t+10)$, $S(t+15)$, $S(t+20)$ are the average 5-minute speeds taken at 5, 10, 15, and 20 minute predictions, respectively.

TESTING

In order to test the performance of each network after training, a set of 10 peak periods collected from 10 different days was presented to the network. For each input pattern in the testing set, multiple predictions were made at 5, 10, 15, and 20 minute horizons. Each predicted value was compared against the actual observed value to calculate two measures of performance: average absolute relative error and root mean square error. Each measure is defined as follows:

$$AARE = \frac{\sum_{i=1}^{N} \left| \frac{S_i - V_i}{V_i} \right|}{N} \tag{1}$$

And

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (S_i - V_i)^2}{N}} \tag{2}$$

Where
AARE = Average absolute relative error in speed
RMSE = Root mean square error
S_i = Predicted speed (mph) for observation *i*
V_i = Actual speed (mph) for observation *i*
N = Number of observations

The two measures were used to compare the performance of the four networks under the following settings: desired prediction horizon (5, 10, 15, and 20 minutes), input type (*xy, yz,* and *xyz*), and inclusion of LTM component (Yes/No). In addition to the previous controlled variables, the performance was also checked against various combinations of traffic conditions at each of the three stations. At each station, traffic conditions were broken down into four levels of congestion: level 1 (speed >60 mph), level 2 (40-60 mph), level 3 (20-40 mph), and level 4 (<20 mph). For all possible combinations of the four levels of congestion at each station, the network performance was evaluated to determine the optimal settings. This procedure was primarily used to address two questions. First, which network and what settings are optimal for predictions at each of the four horizons tested? Second, what traffic conditions are associated with the largest errors or worst prediction performance so that we can identify the level of confidence in our predictions? Both questions are critical to the successful online implementation of a traffic prediction system. The answer to the first question will identify which network and what settings are best for each traffic condition in terms of minimization of the *AARE* and *RMSE*. This allows for a hybrid prediction system that is optimized to produce the best performance under different prediction horizons and traffic conditions. The answer to the second question distinctively identifies the traffic conditions during which prediction accuracy is unacceptable, and therefore, should not be disseminated to the public or disseminated with an appropriate measure of uncertainty.

OPTIMAL SETTINGS

The optimal settings were selected for each of the four levels of prediction horizon using *AARE* and *RMSE* independently. Despite the fact that both measures are often used to quantify the performance of prediction models, consistency between the two measures is not, by their mathematical definition, guaranteed. Therefore, optimal settings were selected and presented separately. While some optimal settings were different by measure, others were consistent, indicating that both measures are in agreement. For each prediction horizon, a total of 25 cases were presented as a result of different combinations of traffic conditions at each of the three stations. The 25 cases, however, do not include all possible combinations of traffic conditions, but only the ones covered in the testing data set. Theoretically, with four levels of congestion at each station, the total number of combinations would be 4x4x4 = 64. However, some combinations were not observed in the testing set due either to its limited size or to their infrequent occurrence in general. Therefore, the optimization was limited to the 25 cases. Additional testing with a larger data set can be used to optimize the performance for other cases.

The optimal settings for each of the four prediction horizons based on *AARE* are shown in TABLE 1 and TABLE 2. Both tables show the predictive performance of the adopted approach as compared to that of the naïve and heuristic approaches (*18*). The naïve approach assumes that predicted values can simply be derived from current observations, assuming no change in traffic conditions. The heuristic approach, however, relies on the use of historical information that is observed at the same time and location. Mathematically,

$$S(t+i) = \frac{S(t)}{H(t)} H(t+i) \tag{3}$$

Where,

$S(t+i)$ = Predicted speed at time $t + i$, where i is the prediction horizon

$H(t+i)$ = Historical speed value observed at time $t + i$ and derived from training data set

$S(t)$ = Observed speed at time t

$H(t)$ = Historical speed observed at time t

The tables show the predictive performance in terms of the *AARE* for each of the 25 cases based on the three approaches. Each record in the table shows the optimal settings for the neural network (NN) approach in terms of the network topology, the type of input, the relevance of LTM, and the *AARE* for each set of traffic condition combinations. Each table shows that the optimized NN approach outperforms both the naïve and heuristic approaches for all cases. This can be clearly seen when comparing the *AARE* of each approach for each case. However, to test if the differences in performance are statistically significant, both Friedman and Wilcoxon tests were conducted for each prediction horizon. The Friedman test showed that for all prediction horizons and all 25 cases, the prediction performance of the NN approach was significantly different from the naïve and the heuristic approaches. The Wilcoxon signed ranks test also showed that the two-sided probabilities using normal approximation were less than 0.001 when comparing the NN approach with the naïve and the heuristic approaches. This indicates that the performance of the NN approach was significantly different from the other two as well.

Considering the NN approach, the tables also show that no particular network topology seemed to have outperformed the others for all cases. The same applies to the type of inputs and the inclusion of LTM component. For 5-minute predictions of speed, the *AARE* did not exceed 7% for 22 out of 25 cases. Two cases produced errors as high as nearly 15% and the third 11.7%. A possible explanation for this is that the predictions associated with those conditions are independent of the information relayed by STM and LTM components, and therefore, the network topologies could not build sufficient internal representations of such cases. For 10-minute predictions, 21 cases exhibited errors less than 10%, 2 less than 15%, and 2 as high as 25% and 28%. The worst two cases in both 5-minute and 10-minute predictions were the same (case 9 and 13). The same explanation may apply. However, both cases exhibited higher errors with 10-minute predictions than with 5-minute predictions, confirming the intuition that prediction accuracy diminishes with longer horizons. Similar conclusions can made from 15-

and 20-minute predictions, which show almost consistently higher errors for the worst cases as the prediction horizon increases. The optimal settings were also selected based on the *RMSE* for each prediction horizon. Comparisons show some discrepancies between optimal settings for each case based on each measure. The cases with identical optimal settings based on *AARE* and *RMSE* are marked with an asterisk in both tables.

PERFORMANCE ENVELOPES

For each of the four levels of prediction horizon the performance envelopes were plotted for each case using the minimum and maximum values of *AARE* and *RMSE*. The performance envelopes, shown in FIGURE 2(a) and FIGURE 2(b), can be used to identify cases where prediction errors do not meet a maximum acceptable threshold value that is appropriate for online implementation. For instance, if we set the *AARE* threshold value to 10%, then FIGURE 2(a) can be used to identify all cases where optimal settings do not yield errors less than or equal to 10%. Consequently, when such cases are encountered in real world, the high levels of uncertainty in predictions must then be realized, if not eliminated entirely from traffic information dissemination systems. This is an essential concept for the traveling public to maintain high reliability in information disseminated by traffic management centers.

It should be emphasized here that the high errors associated with such cases may be attributed to under-representation of those cases in the training data. This often leads to the network's inability to generalize under such conditions. Additional training with data collected from conditions poorly represented could essentially lead to improvement in the overall prediction performance. Even with the potential improvement of performance as a result of additional training, it is reasonable to expect that there will be certain conditions where prediction accuracy is practically unacceptable. For any traffic prediction system to be successfully implemented, such conditions must be identifiable to recognize the limitations of the prediction model.

EFFECT OF TOPOLOGY AND LTM COMPONENT

The optimal settings defined in the study included the network topology and the relevance of LTM component as input to the network. In this section we identify the topology that is dominant in the optimal settings for each prediction horizon, as shown in FIGURE 3(a). The figure shows that no specific topology appears to consistently dominate the optimal settings of all prediction horizons. However, CANFIS is shown to outperform the other topologies for 10- and 20-minute predictions and to demonstrate comparable performance to the modular network for 5-minute predictions. It should be noted, however, that for 20-minute predictions, the best performance of CANFIS, as shown in TABLE 2(d), exhibits relatively high predictive errors that may not be practically acceptable by the traveling public. The figure also shows that when all cases are combined, CANFIS outperforms all other topologies and dominates nearly 33% of the overall optimal settings.

Another important factor that was introduced in this study is the relevance of LTM component and its impact on the prediction performance. As mentioned earlier, the LTM component assists the network in retaining some of the historical information in its weights during the training process. Such memory component is useful in predicting the onset of congestion and in making longer-horizon predictions when predicted conditions are less dependent on information relayed by the STM component. The inclusion of both components can essentially lead to a model capable of predicting recurring (LTM component) and non-recurring conditions (STM component). FIGURE 3(b) illustrates the role of the LTM component in each prediction horizon. The figure shows the percentage of cases whose optimal settings included the LTM component. The figure clearly shows that the relevance of LTM is more pronounced in longer-horizon predictions. For instance, nearly 65% of the cases included LTM component in their optimal settings for 20-minute predictions, as compared to 40% for 5-minute predictions. This trend suggests that the critical role of LTM in making predictions more accurate in longer horizons.

COMPARISON BETWEEN OPTIMAL AND NON-OPTIMAL PERFORMANCE

Prediction performance was optimized under different network settings and various traffic conditions. In order to quantify the performance improvements achieved by optimization with traffic conditions versus optimization with network settings only, we compare the optimal to non-optimal performance for each case. Non-optimal performance refers to optimization with network settings only and without consideration of traffic conditions. This results in selecting the best network topology and input settings for all traffic conditions. Optimal performance in this section, on the other hand, refers to optimization with network settings and traffic conditions. To facilitate the comparison, the reduction in errors of both scenarios was calculated. FIGURE 4(a) shows the percentage reduction in *AARE* for each prediction horizon scenario. The figure shows significant improvements in prediction performance as a result of optimization with traffic conditions. The average improvement in terms of percentage reduction of errors was 6%, 6.1%, 7.9%, and 8.2% for 5-, 10-, 15-, and 20-minute predictions, respectively. Some improvements were as high as 20 to 30% such as cases 4, 11, and 13. Similar results were attained in the comparative evaluation of performance based on *RMSE*. FIGURE 4(b) shows the improvements in terms of the reduction in *RMSE* (mph) for each case and each prediction horizon. The reduction of *RMSE* was as high as 14 mph for case 4 and 9 mph for case 17. A better illustration of the performance improvements in all cases combined can be seen in FIGURE 5(a) and FIGURE 5(b). Each figure shows the cumulative percentage of cases with a reduction in *AARE* and *RMSE*, respectively, that is less than or equal to a specific value. For instance, the figures show that 90% of the cases showed improvements of 16% or less in terms of *AARE* reduction and nearly 4 mph in terms of *RMSE* reduction. Such improvements were exclusively attributed to performance optimization with traffic conditions over optimization with network settings only.

CONCLUSIONS

This study presented an approach to optimize the performance of neural networks in short-term traffic prediction. Four neural network architectures (MLP, Modular, Hybrid PCA, and CANFIS) were trained and tested under various network settings. The input to the networks was divided into two main components: STM and LTM. The STM component was represented by spatiotemporal information observed in the past 10 minutes and expressed in terms of 5 minute speed averages. The LTM component was represented by the time stamp associated with each STM component in order to make the trained network time-cognizant. This technique was primarily introduced to allow the networks to learn from the historical information on traffic conditions during similar peak periods in the past. This was necessary to improve the prediction performance during recurring conditions when future predictions are less dependent on information encoded in the STM component. The performance was measured in terms of two types of errors: average absolute relative error and root mean square error.

The optimal settings were selected to minimize the prediction errors under different network settings, various traffic conditions, and multiple prediction horizons. The optimal settings were based on the testing results obtained from four network topologies trained with the same data set. The network settings were varied by changing the input type in the STM component and toggling the LTM component for each of the four network topologies considered. Traffic conditions were broken down into four levels at each of the three stations. This resulted in a total of 25 combinations of different traffic conditions. Each of the 25 cases was optimized independently to identify the optimal network topology and the optimal network settings. The prediction performance of the optimized NN approach was compared to that of both naïve and heuristic approaches. Statistical tests showed that the *AARE* produced by the NN approach was significantly less than that of the naïve and heuristic approaches for all prediction horizons.

For the NN approach, the study showed that no particular network topology has consistently outperformed the others for all prediction horizons and all cases. It was also found that the performance optimization under different traffic conditions has the advantage of identifying cases where none of the NN models were able to produce acceptable performance. While additional training with more data may improve the performance for some of those cases, it is still unequivocally critical to identify the major limitations of the prediction model and the cases where its performance falls below the minimum acceptable by traffic management centers. This is a critical issue to the dissemination of reliable information to the public and for the successful implementation of the prediction models. The study also showed that despite the variations in the optimal network topologies, CANFIS network has outperformed the other topologies for 10- and 20-minute predictions and demonstrated comparable performance to the modular network for 5-minute predictions. For nearly 33% of all cases combined, CANFIS network was the optimal topology.

Another important finding is the effect of the LTM component on the optimal performance. The results showed that the LTM component was more frequently seen in the optimal settings as the prediction horizon increases. Nearly 65% of the cases included the LTM component in their optimal settings for 20-minute predictions, as compared to 40% for 5-minute predictions. This trend emphasizes the critical role of LTM in making predictions more accurate in longer horizons. Finally, the study pointed out the comparative evaluation of prediction performance under optimal and non-optimal traffic condition settings. Using the reduction in *AARE* and *RMSE* the performance improvement in each case and for each prediction horizon was evaluated. An average improvement in *AARE* was shown in the range of 6% to 8.2% and as high as 20% to 30% for a few cases. The reduction in *RMSE* was also shown to be relatively large. The comparative evaluation clearly demonstrates the benefit of optimizing the performance under different traffic conditions at the same station, and both upstream and downstream stations.

The study presented an approach for the development of a more efficient traffic prediction system with multiple neural network topologies and multiple network and traffic condition settings. The conclusions presented in this paper are primarily based on the optimization results derived from the testing data set. Generalization can only be made by assuming that the testing data set is a truly representative and unbiased sample. To verify this assumption, further testing can be made using a larger testing data set. Also, the approach presented in this paper was extensively examined at one location. Based on the results, the approach can be applied to other locations as well. For locations that exhibit similar traffic conditions during the peak periods, the settings obtained in this study may be transferred directly without retraining. However, testing is recommended with data collected from the other locations first. If the testing results are not satisfactory, then the current settings may not be applicable without additional performance optimization at the new location by following the steps described in this study. Another potential improvement can be attained using a hybrid model-based and memory-based approach. A strictly model-based approach would be analogous to the one described in this paper, but can be augmented with a memory-based approach such as case-based reasoning to compensate for the lack of knowledge comprehensibility issues that neural networks suffer from.

REFERENCES

1. Kaysi, I., M. Ben-Akiva, and H. Koutsopoulos. "An Integrated Approach to Vehicle Routing and Congestion Prediction for Real-Time Driver Guidance." Paper presented at the 72nd TRB Annual Meeting, Washington, D.C., 1993
2. Ben-Akiva M., A. de Palma, and I. Kaysi. "Dynamic Network Models and Driver Information Systems". *Transportation Research*, Vol. 25A, No. 5, 1991.
3. Ben-Akiva, M., E. Kroes, J. de Ruiter, S. Smulders, J. Whittaker, and K. Lindveld. "Real-Time Prediction of Traffic Congestion." Vehicle navigation & information systems: conference record of papers. New York, N.Y. Institute of Electrical and Electronics Engineers, 1992. p. 557-562, 1992.
4. Stamatiadis, C. and W. Taylor. "Travel Time Predictions for Dynamic Route Guidance with a Recursive Adaptive Algorithm." Paper presented at the 73rd TRB Annual Meeting, Washington, D.C., 1994.
5. Ahmed, M., and A. Cook. "Analysis of Freeway Traffic Time Series Data by Using Box-Jenkins Techniques." *Transportation Research Record No. 722*, TRB, National Research Council, Washington, D.C., pp 1-9, 1982.
6. Gazis, D., and C. Knapp. "On Line Estimation of Traffic Densities from Time Series of Traffic and Speed Data." *Transportation Science*, Vol. 5, No. 3, 1971.
7. Lu Jian. "Prediction of Traffic Flow by an Adaptive Prediction System." *Transportation Research Record* 1287, TRB, National Research Council, Washington, D.C., pp 54-61, 1990.
8. Okutani, I. and Y. Stephanedes. "Dynamic Prediction of Traffic Volume through Kalman Filtering Theory." *Transportation Research B*, Vol. 18B, No. 1, pp 1-11, 1984.
9. Hamed, M. M., H.R. Al-Masaeid, and Z.M.B. Said. "Short-Term Prediction of Traffic Volume in Urban Arterials." *ASCE Journal of Transportation Engineering*, Vol. 121, No. 3, pp 249-254, 1995.
10. Lee, S. and D. Fambro. "Application of the Subset ARIMA Model for Short-Term Freeway Traffic Volume Forecasting." *Transportation Research Record. No. 1678* (1999) p. 179-188.
11. Chang, J. L. and S.P. Miaou. "Real-Time Prediction of Traffic Flows using Dynamic Generalized Linear Models." *Transportation Research Record, No. 1678,* pp 168-178, 1999.
12. D'Angelo, M., H. Al-Deek, and M. Wang. "Travel Time Prediction for Freeway Corridors**."** *Transportation Research Record. No. 1676* p. 184-191, 1999.
13. Williams, B. M. "Multivariate Vehicular Traffic Flow Prediction: An Evaluation of ARIMAX Modeling" Transportation Research Board Annual Meeting in Washington D.C., 2001.

14. Ishak, S. and H. Al-Deek. "Performance Evaluation of a Short-Term Time-Series Traffic Prediction Model" Accepted for publication in the *ASCE Journal of Transportation Engineering*, Vol. 128, No. 6, 2002.
15. Chen, M. and S. Chien. "Dynamic Freeway Travel Time Prediction Using Probe Vehicle Data: Link-based vs. Path-based." Transportation Research Board Annual Meeting in Washington D.C., 2001.
16. Chien S. and C. M. Kuchipudi. "Dynamic Travel Time Prediction with Real-Time and Historical Data". Transportation Research Board Annual Meeting in Washington D.C., 2002.
17. Kwon, J., B. Coifman, and P. Bickel. "Day-to-Day Travel Time Trends and Travel Time Prediction from Loop Detector Data." Transportation Research Board Annual Meeting in Washington D.C., 2000.
18. Smith, B. L., B. M. Williams and R. K. Oswald "Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting" *Transportation Research C*, 10(4), pp 303-321, 2002.
19. Park, D. and L. Rilett. "Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Networks." *Transportation Research Record. No. 1617* p. 163-170, 1998.
20. Rilett, L., D. Park. "Direct Forecasting of Freeway Corridor Travel Times Using Spectral Basis neural Networks." *Journal of Transportation Engineering*, ASCE, Vol. 125, no. 6 (Nov./Dec. 1999).
21. Abdulhai, B., H. Porwal, and W. Recker. "Short-Term Freeway Traffic Flow Prediction Using Genetically Optimized Time-Delay-Based Neural Networks." Transportation Research Board Annual Meeting in Washington D.C., 1999.
22. Lint, V., S. P. Hoogendoorn, and H.J. Zuylen. "Freeway Travel Time Prediction with State-Space Neural Networks." Transportation Research Board 81st Annual Meeting, 2002.
23. Fu, L. and L.R. Rilett. "Estimation of Time-Dependent, Stochastic Route Travel Times using Artificial Neural Networks." *Transportation Planning and Technology*, 24(1), pp. 25-36, 2000.
24. Park, D., L.R. Rilett, and G. Han. "Spectral Basis Neural Networks for Real-Time Travel Time Forecasting." *ASCE Journal of Transportation Engineering*, Vol. 125, No. 6, pp. 515-523, 1999.
25. Van Der Voort, M., M. Dougherty, and S. Watson. "Combining Kohonen Maps with ARIMA Time Series Models to Forecast Traffic Flow." *Transportation Research part C*, 4(5), pp. 307-318, 1996.
26. Yasdi, R. "Prediction of Road Traffic using a Neural Network Approach" *Neural Computing and Applications*, 8, pp. 135-142, 1999.
27. Zhang, H. M. "Recursive Prediction of Traffic Conditions with Neural Network Models" *ASCE Journal of Transportation Engineering*, Vol. 126, No. 6, pp. 472-481.
28. Zhang, H., S. G. Ritchie, and Z.P. Lo. "Macroscopic Modeling of Freeway Traffic using an Artificial Neural Network" *Transportation Research Record, No. 1588,* pp 110-119, 2000.
29. Duda, R., P. Hart, and D. Stork. Pattern Classification. Second Edition, John Wiley & Sons, Inc., 2001.

30. Ham, F. M. and I. Kostanic. *Principles of Neurocomputing for Science and Engineering*. McGraw Hill, NY, 2001.
31. Jang, J.S.R., C.T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing*. Prentice Hall, N.J., 1997.
32. Principe, J. C., N. R. Euliano, and W.C. Lefebvre. *Neural and Adaptive Systems*. John Wiley and Sons, Inc., NY, 2000.
33. Lefebvre, Curt, “Neuro Solutions, Version 4.10”, NeuroDimension, Inc., Gainesville, FL, 2001.

LIST OF TABLES

LIST OF FIGURES

TABLE 1 Optimal Settings for 5- and 10-Minute Predictions Based on *AARE*

(a) 5-Minute Prediction Horizon

Case	Congestion Indicator			Optimal Settings for NN Approach			Average Absolute Relative Error		
	Downstream Station (X)	Current Station (Y)	Upstream Station (Z)	Network	Inputs	LTM Component	NN Approach	Naïve Approach	Heuristic Approach
7	3	3	2	MODULAR	XY	No	0	1.448	1.449
15	2	3	3	MLP	ZY	Yes	0	0.428	0.85
22*	1	2	2	CANFIS	XY	No	0	0.025	0.087
16	2	3	2	MODULAR	ZY	Yes	0.001	0.624	0.582
20	2	1	2	CANFIS	ZY	Yes	0.001	0.041	0.093
23	1	2	1	CANFIS	XY	No	0.002	0.047	0.134
8*	3	3	1	MODULAR	XYZ	No	0.003	0.602	0.532
17*	2	3	1	PCA-hybrid	XY	Yes	0.006	0.445	0.453
11*	2	4	4	CANFIS	ZY	Yes	0.008	0.506	0.737
19	2	2	1	CANFIS	XY	No	0.008	1.046	0.873
24*	1	1	2	MODULAR	XYZ	No	0.009	0.03	0.127
3	3	4	2	PCA-hybrid	XYZ	No	0.012	0.607	1.436
5*	3	3	4	PCA-hybrid	XYZ	No	0.014	0.146	0.274
1	3	4	4	MODULAR	XY	Yes	0.017	0.784	1.618
21*	2	1	1	CANFIS	XY	No	0.023	0.648	0.598
25*	1	1	1	CANFIS	ZY	No	0.024	0.487	0.438
2	3	4	3	MODULAR	XY	Yes	0.032	0.6	1.304
18	2	2	2	CANFIS	ZY	No	0.033	1.417	1.444
6	3	3	3	PCA-hybrid	XYZ	No	0.047	0.717	1.215
12*	2	4	3	MODULAR	XY	No	0.061	0.566	0.767
4*	3	4	1	MLP	XYZ	No	0.064	0.486	0.304
14*	2	3	4	PCA-hybrid	ZY	Yes	0.069	0.217	0.5
10*	3	2	1	MODULAR	XYZ	Yes	0.117	1.263	0.845
9*	3	2	2	MLP	XYZ	No	0.147	1.536	1.337
13	2	4	2	MLP	XYZ	Yes	0.149	0.518	0.404

(b) 10-Minute Prediction Horizon

Case	Congestion Indicator			Optimal Settings for NN Approach			Average Absolute Relative Error		
	Downstream Station (X)	Current Station (Y)	Upstream Station (Z)	Network	Inputs	LTM Component	NN Approach	Naïve Approach	Heuristic Approach
7	3	3	2	MODULAR	XY	Yes	0.001	2.575	1.06
20	2	1	2	PCA-hybrid	XYZ	No	0.001	0.416	0.094
23*	1	2	1	MLP	XY	No	0.001	0.166	0.073
22*	1	2	2	PCA-hybrid	XYZ	No	0.002	0.096	0.034
25*	1	1	1	CANFIS	ZY	No	0.003	0.728	0.805
12	2	4	3	MLP	XY	No	0.004	1.942	1.022
11*	2	4	4	CANFIS	ZY	Yes	0.006	1.48	0.6
24*	1	1	2	CANFIS	XY	No	0.006	0.089	0.032
5	3	3	4	CANFIS	ZY	Yes	0.01	0.696	0.166
3	3	4	2	CANFIS	XY	Yes	0.013	2.199	0.661
15	2	3	3	MLP	ZY	Yes	0.016	1.74	0.645
4*	3	4	1	MLP	XYZ	No	0.018	0.354	0.638
17	2	3	1	CANFIS	XY	Yes	0.018	0.711	0.472
21*	2	1	1	MODULAR	XY	No	0.024	1.444	1.16
8*	3	3	1	MODULAR	ZY	No	0.029	0.702	0.632
6	3	3	3	PCA-hybrid	ZY	Yes	0.033	2.136	0.659
1	3	4	4	CANFIS	XYZ	Yes	0.035	2.899	0.89
16	2	3	2	CANFIS	XY	Yes	0.043	1.703	0.97
2*	3	4	3	MODULAR	XY	Yes	0.046	2.183	0.795
19	2	2	1	CANFIS	XY	No	0.09	1.942	1.833
10*	3	2	1	CANFIS	XY	Yes	0.098	1.25	0.973
14	2	3	4	PCA-hybrid	ZY	Yes	0.116	1.418	0.554
18*	2	2	2	CANFIS	ZY	No	0.148	3.788	2.452
13*	2	4	2	MODULAR	ZY	No	0.256	0.472	0.687
9	3	2	2	MLP	XYZ	No	0.285	2.407	2.112

Cases marked with * have the same optimal settings for both *AARE* and *RMSE*.

TABLE 2 Optimal Settings for 15- and 20-Minute Predictions Based on *AARE*

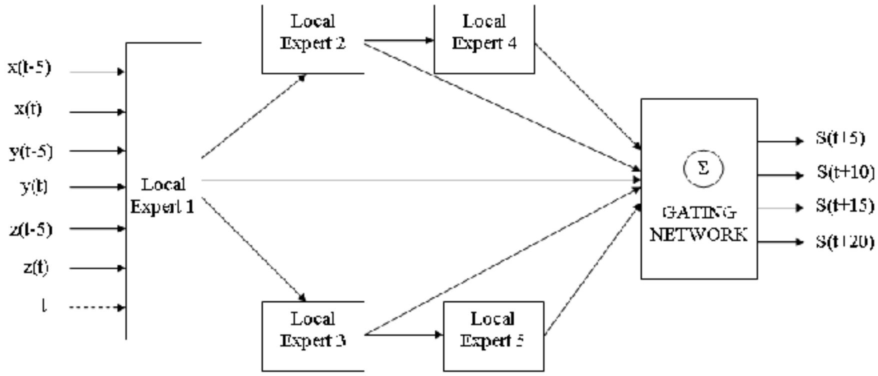
(c) 15-Minute Prediction Horizon

Case	Congestion Indicator			Optimal Settings for NN Approach			Average Absolute Relative Error		
	Downstream Station (X)	Current Station (Y)	Upstream Station (Z)	Network	Inputs	LTM Component	NN Approach	Naïve Approach	Heuristic Approach
8*	3	3	1	MLP	XY	Yes	0	1.2	0.669
7	3	3	2	Hybrid PCA	XYZ	Yes	0.001	2.111	1.243
3	3	4	2	Hybrid PCA	XYZ	Yes	0.002	1.965	0.716
25*	1	1	1	Hybrid PCA	XY	No	0.002	1.49	1.58
12	2	4	3	MLP	XY	Yes	0.008	1.601	0.78
22*	1	2	2	MLP	XY	Yes	0.009	0.083	0.01
24	1	1	2	CANFIS	XY	Yes	0.014	0.111	0.03
15	2	3	3	Hybrid PCA	XY	No	0.02	1.596	0.665
23	1	2	1	MLP	XY	No	0.021	0.166	0.077
5*	3	3	4	Hybrid PCA	XYZ	Yes	0.024	0.546	0.135
4*	3	4	1	CANFIS	YZ	Yes	0.027	0.462	0.677
20*	2	1	2	MLP	XY	Yes	0.028	0.221	0.152
1	3	4	4	MLP	XY	Yes	0.037	2.072	0.639
16	2	3	2	MLP	XYZ	No	0.039	1.875	1.333
17	2	3	1	MODULAR	XY	Yes	0.043	0.793	0.493
2*	3	4	3	CANFIS	YZ	No	0.044	2.326	0.703
6*	3	3	3	MODULAR	XY	Yes	0.05	2.01	0.712
11	2	4	4	MODULAR	XYZ	No	0.05	2.265	1.182
21*	2	1	1	MODULAR	YZ	No	0.081	2.069	1.523
9	3	2	2	CANFIS	YZ	No	0.132	2.015	1.447
19	2	2	1	CANFIS	YZ	Yes	0.175	2.845	2.278
13*	2	4	2	CANFIS	XY	Yes	0.181	0.583	0.729
10*	3	2	1	CANFIS	XY	Yes	0.201	2.138	1.762
18*	2	2	2	CANFIS	YZ	Yes	0.224	4.8	2.422
14	2	3	4	CANFIS	YZ	No	0.491	2.638	1.434

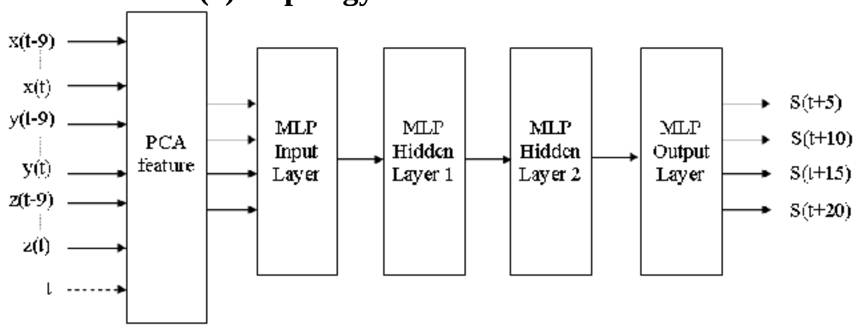
(d) 20-Minute Prediction Horizon

Case	Congestion Indicator			Optimal Settings for NN Approach			Average Absolute Relative Error		
	Downstream Station (X)	Current Station (Y)	Upstream Station (Z)	Network	Inputs	LTM Component	NN Approach	Naïve Approach	Heuristic Approach
23*	1	2	1	MLP	XY	Yes	0.003	2.13	0.964
15*	2	3	3	Hybrid PCA	XYZ	Yes	0.004	2.659	1.267
24	1	1	2	Hybrid PCA	XYZ	Yes	0.008	1.593	0.746
8	3	3	1	Hybrid PCA	XY	No	0.009	1.924	1.992
17	2	3	1	MLP	XY	Yes	0.009	2.565	0.742
5	3	3	4	MLP	XY	Yes	0.01	0.082	0.033
7	3	3	2	CANFIS	XY	Yes	0.016	0.18	0.024
20	2	1	2	Hybrid PCA	XY	No	0.018	1.804	0.661
3	3	4	2	MLP	XY	No	0.019	0.236	0.056
22	1	2	2	Hybrid PCA	XYZ	Yes	0.021	1.058	0.247
25	1	1	1	CANFIS	YZ	Yes	0.024	0.5	0.677
12	2	4	3	MLP	XY	Yes	0.026	1.58	1.278
16	2	3	2	MLP	XY	Yes	0.026	2.587	0.692
4*	3	4	1	MLP	XYZ	No	0.031	2.285	1.366
1*	3	4	4	MODULAR	XY	Yes	0.04	0.997	0.66
6	3	3	3	CANFIS	YZ	No	0.047	2.603	0.769
2	3	4	3	MODULAR	XY	Yes	0.049	1.886	0.662
11	2	4	4	MODULAR	XYZ	No	0.063	2.12	1.182
13*	2	4	2	MODULAR	YZ	No	0.065	3.341	2.55
9*	3	2	2	CANFIS	YZ	No	0.145	3.272	1.68
14*	2	3	4	CANFIS	YZ	Yes	0.148	3.383	2.429
19	2	2	1	CANFIS	XY	Yes	0.168	0.58	0.711
10*	3	2	1	CANFIS	XY	Yes	0.174	2.656	1.769
21*	2	1	1	CANFIS	YZ	Yes	0.191	5.27	2.57
18*	2	2	2	CANFIS	YZ	No	0.302	1.745	0.498

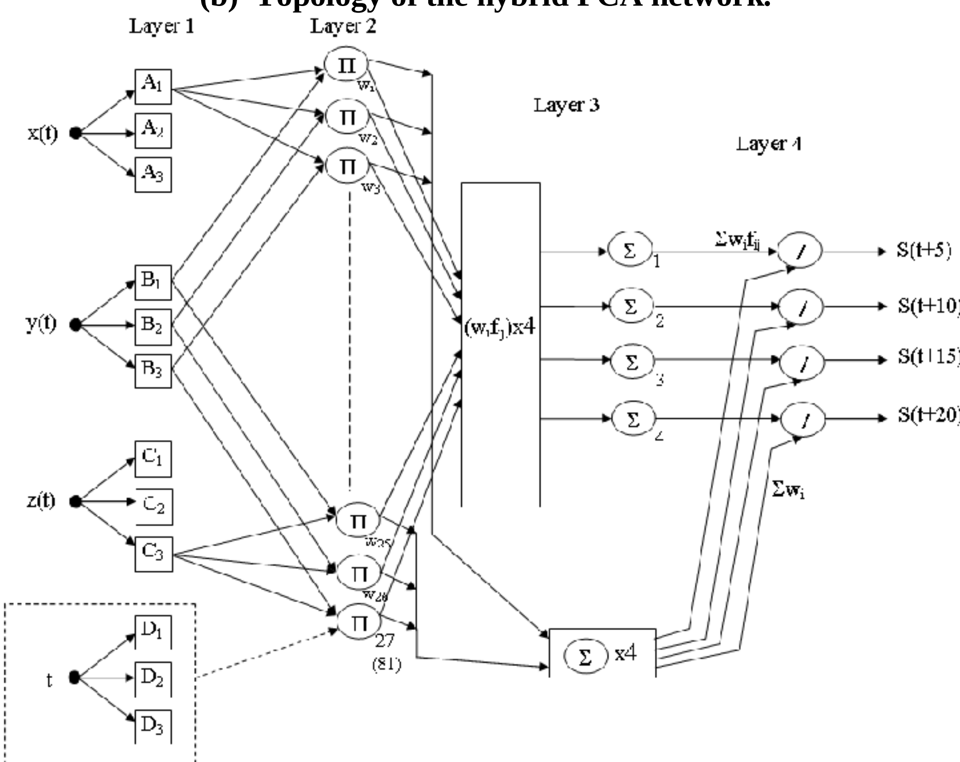
Cases marked with * have the same optimal settings for both *AARE* and *RMSE*.



(a) Topology of the modular network.

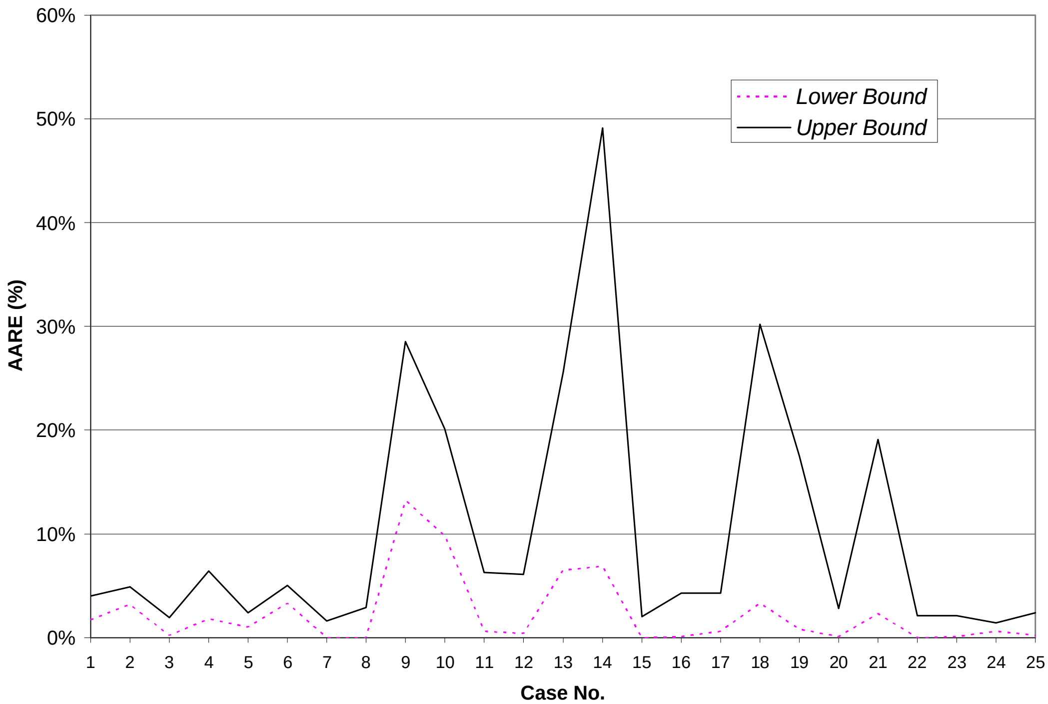


(b) Topology of the hybrid PCA network.

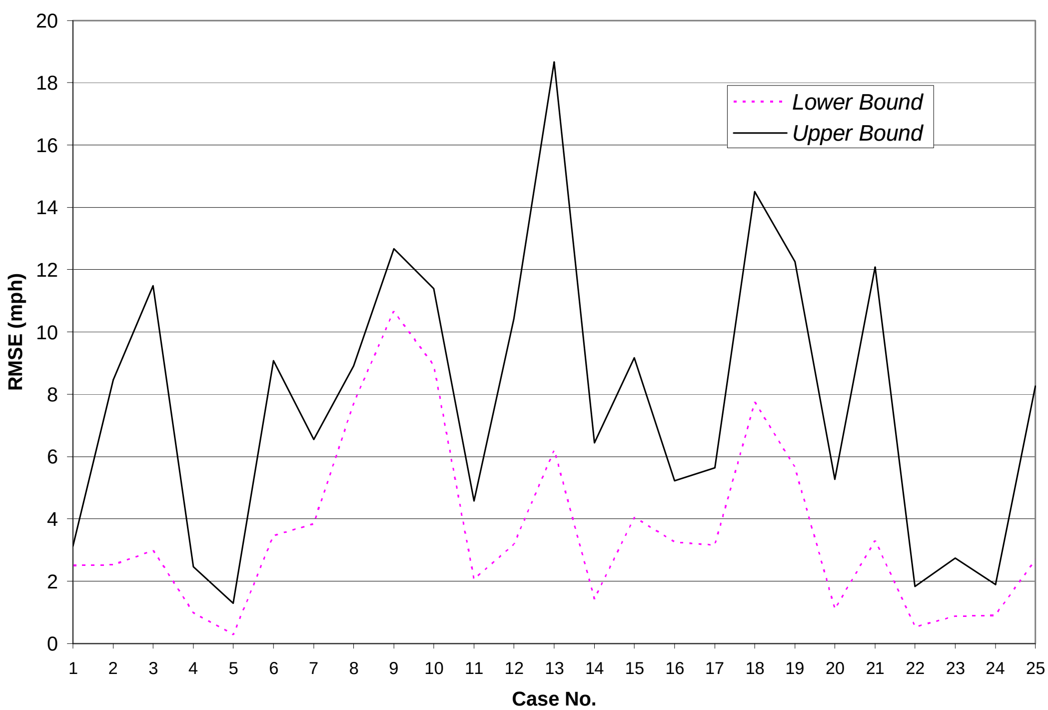


(c) Topology of CANFIS

FIGURE 1 Illustration of network topologies.

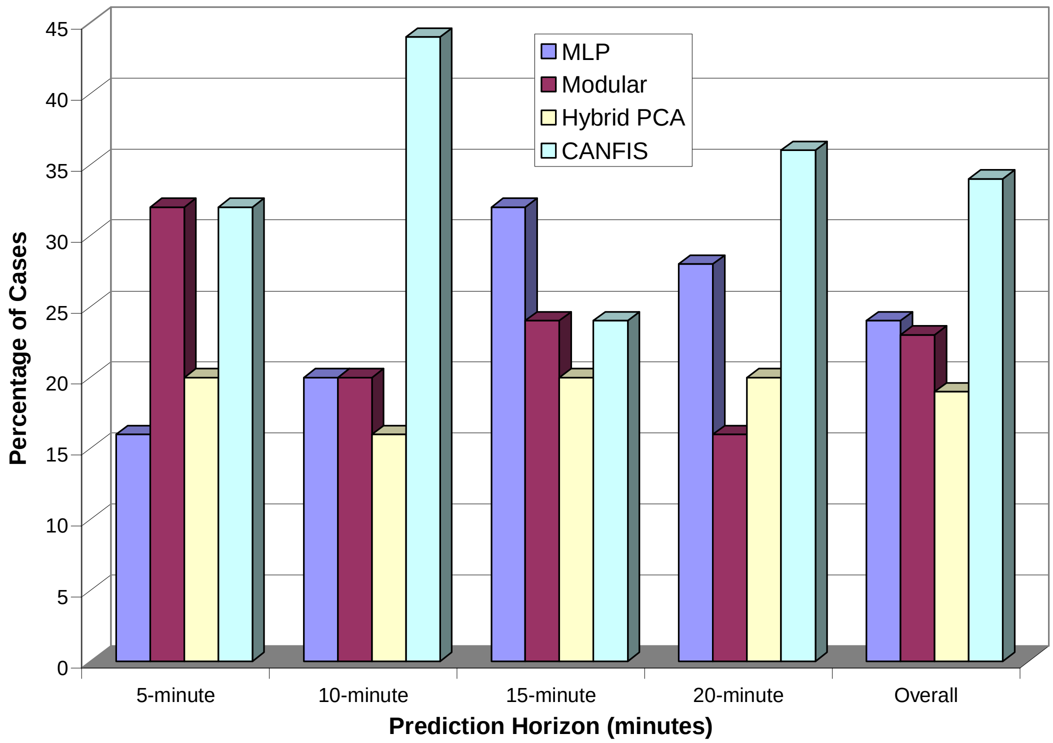


(a) Performance envelopes based on *AARE*.

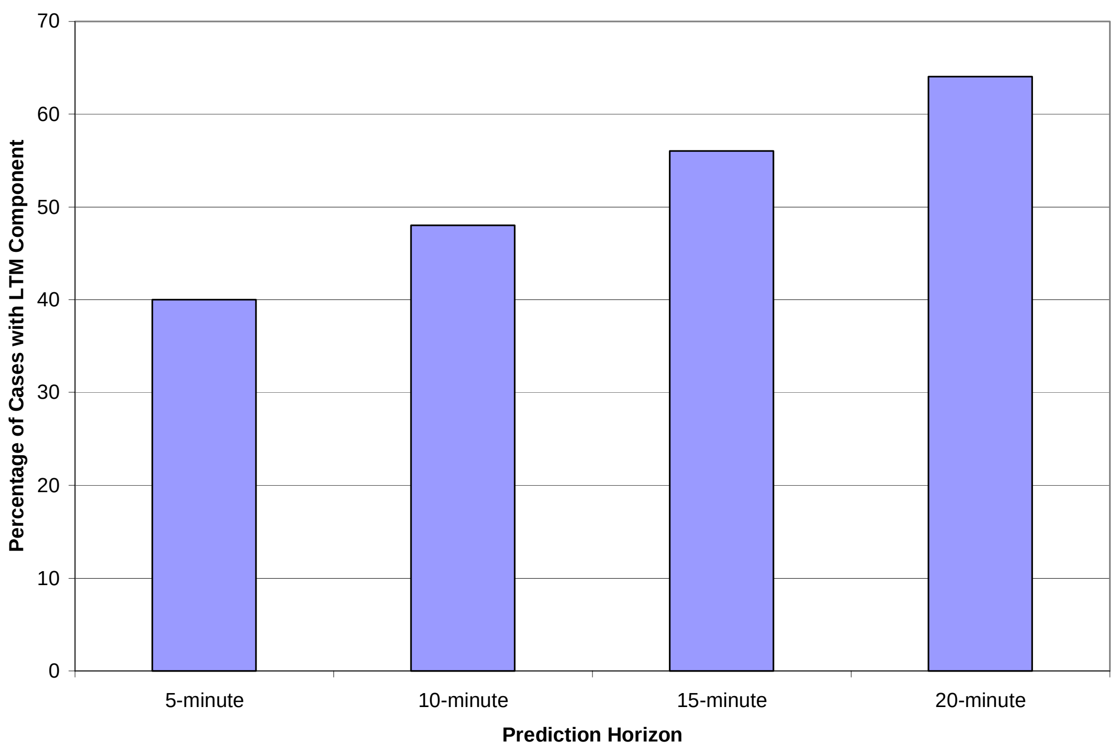


(b) Performance envelopes based on *RMSE*.

FIGURE 2 Performance envelopes associated with optimal settings for the NN approach.

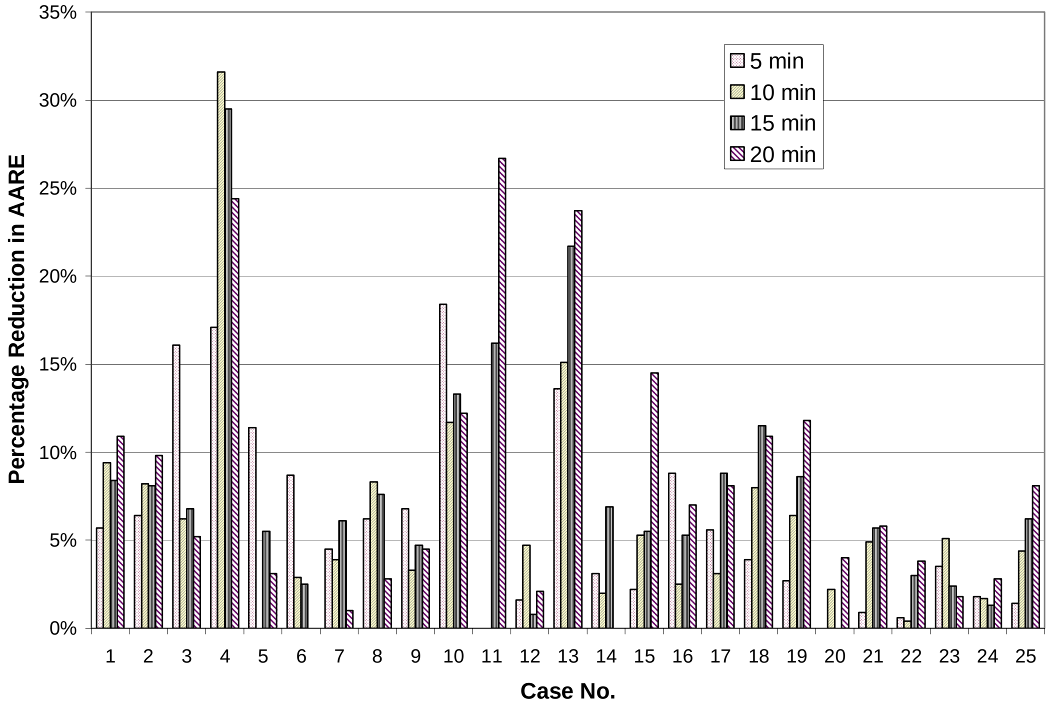


(a) Frequency of optimal networks with prediction horizons based on *AARE*.

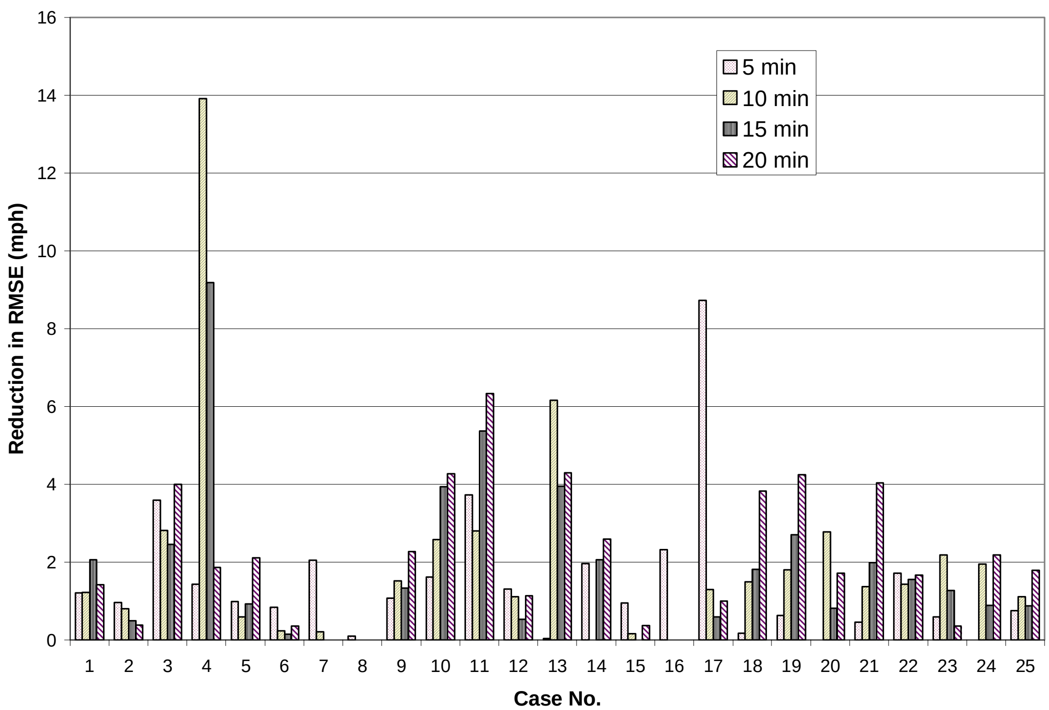


(b) Percentage of cases with LTM component in optimal settings.

FIGURE 3 Effect of network topology and LTM component on optimal settings.

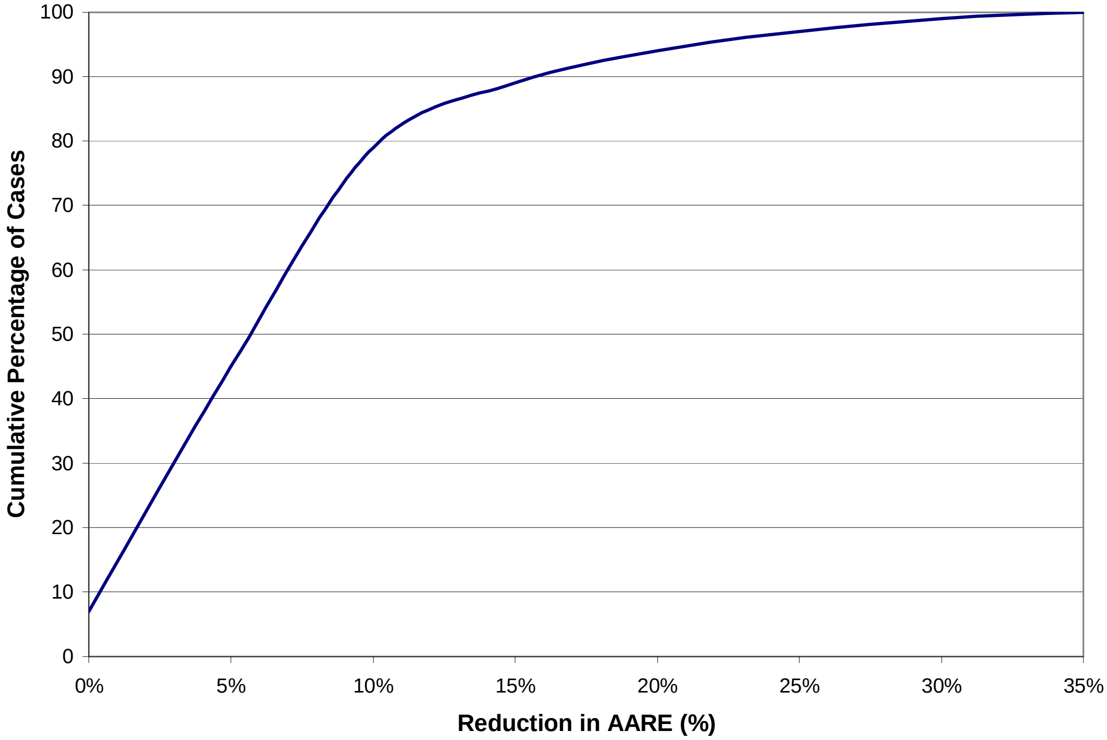


(a) Percentage reduction in *AARE*.

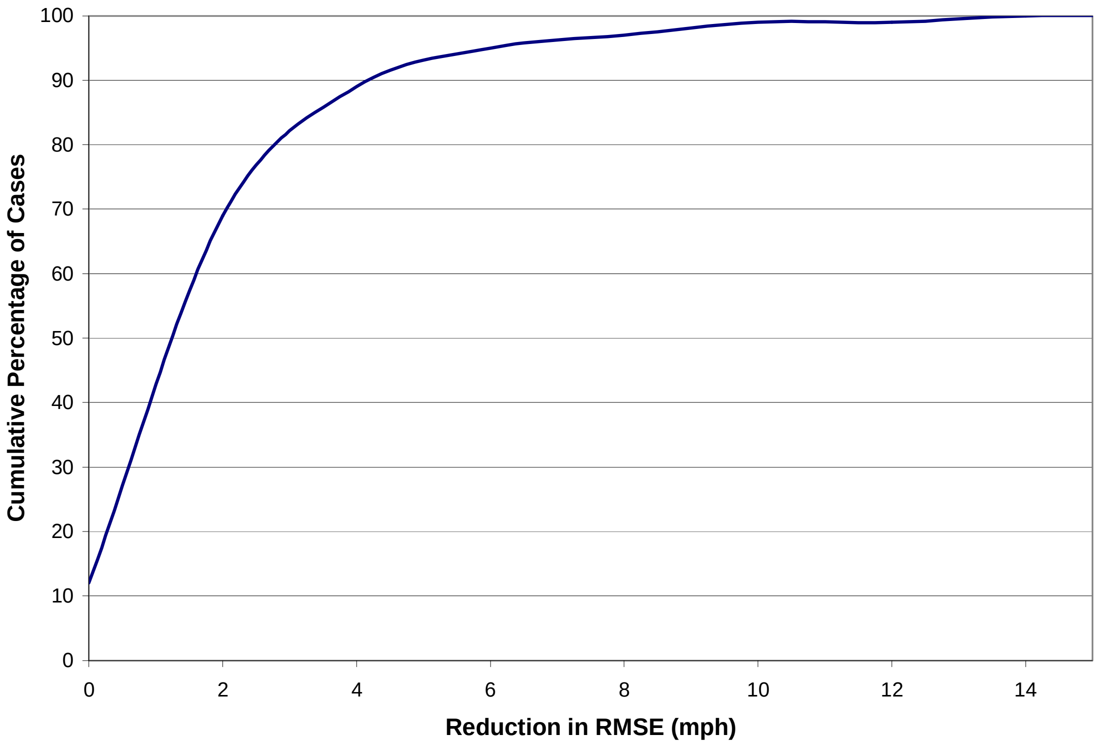


(b) Reduction in *RMSE*.

FIGURE 4 Reductions in *AARE* and *RMSE* achieved by optimization with traffic conditions.



(a) Percentage reductions in *AARE*.



(b) Reduction in *RMSE*.

FIGURE 5 Performance improvements achieved by optimization with traffic conditions.