Title

Short Term Freeway Traffic Flow Prediction Using Genetically-Optimized Time-Delay-Based Neural Networks

Authors

Abdulhai, Baher
Porwal, Himanshu
Recker, Will

Publication Date

1999

CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Short Term Freeway Traffic Flow Prediction Using Genetically-Optimized Time-Delay-Based Neural Networks

Baher Abdulhai
Himanshu Porwal
Will Recker

California PATH Working Paper
UCB-ITS-PWP-99-1

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation; and the United States Department Transportation, Federal Highway Administration.

Report for MOU 360

January 1999

ISSN 1055-1417

Short Term Freeway Traffic Flow Prediction Using Genetically-Optimized Time-Delay-Based Neural Networks

Baher Abdulhai
Assistant Professor and Manager
Intelligent Transportation Systems Research
Department of Civil Engineering
University of Toronto
Toronto, Ontario, Canada M5S 1A4
Tel: (416) 946-5036
Fax: (416) 978-5054
baher@ecf.utoronto.ca

Himanshu Porwal*
Graduate Student Researcher,
hporwal@ea.oac.uci.edu

Will Recker*
Professor and Director
recker@translab.its.uci.edu

***Institute of Transportation Studies**
Department of Civil and Environmental Engineering
University of California Irvine
Irvine, CA 92696-3600

PATH

Final Report, December 1998

TABLE OF CONTENTS

ABSTRACT 1

INTRODUCTION 2

EXISTING MODELS VS. THE PROPOSED METHODOLOGY 2

GENETICALLY OPTIMIZED NEURAL NETWORKS 4

THE TIME DELAY NEURAL NETWORK MODEL 5

FREEWAY SITE DATA DESCRIPTION 8

EFFECT OF EXTENT OF PREDICTION ON PREDICTION ACCURACY 10

EFFECT OF SPATIAL CONTRIBUTION ON PREDICTION ACCURACY 11

EFFECT OF DATA RESOLUTION ON PREDICTION ACCURACY 12

VALIDATION OF TRANSFEREABILITY TO REAL TRAFFIC DATA 13

COMPARISON TO THE MLF MODEL (ZHANG1997) 14

SUMMARY AND CONCLUSION 14

ACKNOWLEDGMENT 15

REFERENCES 15

APPENDIX A 26

Short Term Freeway Traffic Flow Prediction Using Genetically-Optimized Time-Delay-Based Neural Networks

Baher Abdulhai
Assistant Professor and Manager, Intelligent Transportation Systems Research
Department of Civil Engineering, University of Toronto,Toronto, Ontario, Canada M5S 1A4

Himanshu Porwal*
Graduate Student Researcher

Will Recker*
Professor and Director

***Institute of Transportation Studies**
Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA 92696-3600

ABSTRACT

Proper prediction of traffic flow parameters is an essential component of any proactive traffic control system and one of the pillars of advanced management of dynamic traffic networks. In this paper, we present a new short term traffic flow prediction system based on an advanced Time Delay Neural Network (TDNN) model, the structure of which is optimized using a Genetic Algorithm (GA). After presentation of the model's development, its performance is validated using both simulated and real traffic flow data obtained from the California Testbed in Orange County, California. The model predicts flow and occupancy values at a given freeway site based on contributions from their recent temporal profile as well the spatial contribution from neighboring sites. Both temporal and spatial effects were found essential for proper prediction. An in-depth investigation of the variables pertinent to traffic flow prediction was conducted examining the extent of the “look-back” interval, the extent of prediction in the future, the extent of spatial contribution, the resolution of the input data, and their effects on prediction accuracy. Results obtained indicate that the prediction errors vary inversely with the extent of the spatial contribution, and that the inclusion of three loop stations in both directions of the subject station is sufficient for practical purposes. Also, the longer the extent of prediction, the more the predicted values tend toward the mean of the actual, for which case the optimal look-back interval also shortens. Interestingly, it was found that coarser data resolution is better for *longer* extents of prediction. The implication is that the level of data aggregation/resolution should be comparable to the prediction horizon for best accuracy. The model performed acceptably using both simulated and real data. The model also showed potential to be superior to such other well-known neural network models as the Multi layer Feed-forward (MLF) when applied to the same problem.

Keywords: Traffic Flow Prediction, Neural Networks, Genetic Algorithms, Traffic Management.

INTRODUCTION

Advanced Traffic Management and Information System components typically rely directly on traffic monitoring data as inputs to their underlying decision logic. These systems utilize either historical, current, or projected traffic data. In this context, the problem of reactive versus anticipatory, or proactive, traffic control received considerable attention in the past few years. The prime question is whether to formulate control decisions to react to latest observed traffic conditions or rather attempt to forecast or anticipate short-term future conditions as the basis for control decisions. Reactive control, as the name implies, reacts to already-observed conditions of the traffic stream. Not only do such control systems await problems to arise before reacting but also, the conditions of the traffic system may have changed by the time the control decisions are formulated and implemented. In such cases, the system might operate under already-dated control strategies. Alternatively, proactive control reacts to near-term anticipated conditions. Consequently, the traffic network would always operate (theoretically at least) under control strategies that are more relevant to the prevailing conditions. Therefore, anticipatory control is both intuitively and theoretically preferred over reactive control. Nevertheless, the appeal of anticipatory control is usually discounted because of the inaccuracy of the essential 'traffic forecasting' component. Forecasting traffic conditions, even a few minutes into the future, has proven to be a challenging task that needs more research and attention.

In this research, the problem of short term forecasting of traffic variables is studied and a new Artificial Intelligence (AI) based model using a combination of Genetic Algorithms (GA) and Neural Networks (NN) is presented.

EXISTING MODELS VS. THE PROPOSED METHODOLOGY

Evolution of traffic patterns at a particular location x is essentially a spatio-temporal process. If both space and time are discretized, traffic patterns at a location x at time t depend on traffic patterns at locations x, (x-i) and (x+i), i= 1,2,…n, at times (t-j), j=1,2,….,m. For such a relatively confined environment as a freeway stretch, upstream sections send traffic to the location under consideration and downstream sections may send backward propagating shockwaves as well, as shown in Figure 1. Although the process is intuitively simple, its modeling is not. Several factors interact in a complex manner, including the levels of traffic on both the affecting and the affected sections, as well as the less well-understood effects of driver behavior. A model is needed, the inputs to which are the measured traffic parameters up to the current interval at the section under consideration, as well as from both the upstream and the downstream chain of

sections, and the output of which would be the anticipated traffic conditions at that location in the near future. The model should also be adaptive to tailor itself to the dynamic traffic environment.

Research in this area has been active in the past few years, but unfortunately not comprehensive enough. One common approach to the problem of forecasting traffic parameters is based on time series models. However, forecasts using time series have been found to be *over-predictive and lagging* (Smith and Demetsky, 1995), which makes the prediction itself reactive in some sense, as it follows the current measurement with some time lag. Davis and Nihan (1991) attempted to replace the time series approach by a non-parametric regression approach, but they concluded that the performance of their k-nearest neighbor approach "*performed comparably to, but not definitely better than, the time series approach*". Several exploratory attempts have been made to use Neural Networks (NN) in replacement of the more traditional regression and time-series approaches (Smith and Demetsky, 1995, Doughetry and Lechevallier, 1995, Dougherty and Kirby, 1993). Common to all are a conclusion of potential superiority of NN and a recommendation for further in-depth investigation under different scenarios and using larger real databases.

Existing conventional macroscopic traffic flow models also are numerous and vary from very simple to complex. A conceptually plausible set of models is the Payne's model and its off springs. In a series of publications, Payne (1971, 1978) introduced a traffic model that includes a momentum equation in addition to the continuity equation characterizing the fairly old continuum model. The momentum model was derived from car-following theory concepts. Payne's model has the form:

$$\frac{\partial \mathrm{u}}{\partial \mathrm{t}} = -u\frac{\partial \mathrm{u}}{\partial \mathrm{x}} - \frac{1}{T}\left[u - u_e(k)\right] - \frac{\nu}{k}\frac{\partial \mathrm{k}}{\partial \mathrm{x}} \qquad \ldots\ldots\ldots (1)$$

where: u, k, x, t and u_e are speed, density, distance, time and equilibrium speed respectively.

The second term on the right-hand-side of equation (1) represents relaxation to equilibrium, that is, the effect of drivers adjusting their speeds towards the equilibrium speed-density relationship. The third term represents anticipation, which is the effect of drivers reacting to downstream traffic conditions (for example, the tendency to decrease speed if downstream density is higher due to congestion and vice versa). Discretization of the model by finite differences in time and space gives the following equations:

$$k_j^{n+1} = k_j^n + \frac{\Delta t}{l_j \Delta x_j}\left[q_{j-1}^n + g_j^{on,n} - q_j^n - g_j^{off,n}\right] \qquad \ldots\ldots\ldots (2a)$$

$$u_j^{n+1} = u_j^n - \Delta t\left[u_j^n\left(\frac{u_j^n - u_{j-1}^n}{\Delta x_j}\right) + \frac{1}{T}\left(u_j^n - u_e^n\left(k_j\right)\right) + \nu\left(\frac{k_{j+1}^n - k_j^n}{k_j^n \Delta x_j}\right)\right] \qquad \ldots\ldots\ldots (2b)$$

where superscripts denote time step, subscripts denote space step, and g^{on} and g^{off} indicates on- and off-ramp flows.

Conceptually, the model attempts to capture shorter term dynamic deviations from equilibrium values of traffic flow variables. Also, it attempts to capture the effects of downstream conditions on the section under consideration. Although it has been reported that the application of this model has presented several problems, including instability (see for instance Rathi et al,1987), the conceptual formalization of the model is appealing. However, the mathematical formulation might be limiting and unjustified; Artificial Neural Networks (ANNs) may better capture the overall concept underlying the model, without the limitation of pre-specifying the model structure or the level of nonlinearity involved.

The proposed model draws heavily of the concepts embedded in traffic flow theories and models, as well as forecasting models. However, the model structure itself is new, adaptive and dynamic as it is capable of tailoring itself to changes in the traffic environment and the levels of the model's sophistication and non-linearity evolve during the training process itself. Genetic Algorithms (GA) are used in this research to 'evolve' several advanced Time-Delay-based Neural Networks (TDNN). This research also seeks to investigate and optimize all of the key variables in the prediction problem, as opposed to arbitrary fixing of some of them. These are:

1. The extent of the look-back interval in time,
2. The extent of prediction in the future or prediction 'horizon',
3. The extent of spatial contribution from neighboring freeway,
4. The resolution of the data used for prediction (30 sec., 1 min., 5 min. ..etc.)

This research also utilizes:

1. Real as well as simulated freeway data.
2. Inputs from both upstream and downstream sections, to capture the effects of the incoming traffic as well as the back propagating shockwaves.
3. Different freeway sites with different geometrics and on and off ramps.
4. Peak and around near-peak traffic conditions to capture different levels of congestion.
5. Different NN architectures.

GENETICALLY OPTIMIZED NEURAL NETWORKS

Genetic algorithms, from artificial intelligence, are defined by a problem-solving methodology that uses genetics as its model for problem solving, applying the rules of

reproduction, gene crossover, and mutation to a population of candidate solutions or pseudo-organisms. Those organisms can pass beneficial and survival-enhancing traits to new generations (Chambers 1996). GA are known to be a powerful new technology for searching through large and complex solution spaces featuring large numbers of local minima.

Alternatively, Artificial Neural Networks (ANNs), also from artificial intelligence are mathematical models inspired by the human brain structure. ANNs prove to be superior to conventional techniques in the particular area of traffic pattern recognition and classification and are capable of learning from exemplar patterns (see for instance Abdulhai, 1996). The choice of neural networks structure and parameters, however, is an empirical-artistic exercise that relies on "rules of thumb" derived from past development experiences. The space of possible architectures and parameter combinations is extremely large. As a consequence, some significant amount of trial-and-error experimental hand-crafting is necessary before an adequate solution is achieved. It is impractical to rely on such "guesstimation" and trial and error to design networks for serious real-world problems. The empirical approach does not always produce a near-optimal network. Additionally, a good solution might be data-dependent, requiring re-optimization after every significant change in the application environment. The search for the best attainable network structure and parameter-setting combination is therefore a logical application for genetic algorithms.

Genetic algorithms have been applied to the problem of NN design in several ways. For instance, Montana and Davis (1989) have explored the use of GA in training a NN of known structure. Belew et al (1990) used GA to set the learning and momentum rates for feed forward NN. Chang and Lippmann (1991) used GA to preprocess data in order to reduce the inputs to a NN without degrading performance. Harp and Samad (1991) explored using GA to discover the size, structure and parameters of a network to be trained by a separate NN learning algorithm. Koza and Rice (1992) looked at GA as a tool for developing architectures and weights together. In 1997 BioComp Systems released a Neuro-Genetic Optimizer for the architectural optimization of neural networks. More details on the subject of using GA for NN development can be found in Chambers (1996) and Winter et al (1995).

THE TIME DELAY NEURAL NETWORK MODEL

The Time Delay Neural Network (TDNN), schematically shown in Figure 2, features multiple connections between the individual neurons, as opposed to single connections as in the more basic NN. The multiple connections look-back over time to capture the temporal evolution of

patterns in the data; i.e., each neuron is provided with a memory in order to remember previous layer outputs for N periods of time. This is different from just lagging inputs N periods of time, as the look-back period of the TDNN affects the hidden layer and output layer as well, causing it to remember previously developed patterns and not just inputs. Adaptive Time NN (ATNN) extends the TDNN by making the look-back intervals automatically adapt and change as learning progresses, seeking phase relationships that produce higher correlation over history and optimizing the look-back interval. Both TDNN and ATNN promise potential higher accuracy than the commonly used Multi-Layer Feed Forward Neural Network (MLF), also known as Back Propagation (BP).

Non time-delay NN such as the MLF can only learn an input-output mapping that is *static*. This form of mapping is well suited for cases where both the input vector and the output vector represent *spatial* patterns that are independent of time. It can be used to perform nonlinear predictions on a stationary time series i.e., when its statistics do not change with time. However, the time dimension is important for traffic flow predictions as traffic flow patterns evolve and change with time. Therefore, TDNN is expected to outperform MLF-like models and can be considered a more general form of the MLF. Similar to the MLF, the TDNN employs back propagation techniques for setting weights between neurons.

To train the TDNN network, the actual response of each neuron in the output layer is compared with a desired target response at each time instant. Assume that neuron j lies in the output layer with its actual response denoted by $y_j(n)$ and that the desired response for this neuron is denoted by $d_j(n)$, both of which are measured at time n. The instantaneous value for the sum of squared errors produced by the network is as follows:

$$\xi(n) = \frac{1}{2}\sum_j e_j^2(n) \qquad \text{. . . . . .(3)}$$

where the index j refers to the neurons in the output layer only, and $e_j(n)$ is the error signal, defined by

$$e_j(n) = d_j(n) - y_j(n) \qquad \text{. . . . . (4)}$$

The goal is to minimize the cost function defined as the value of $\xi(n)$ computed over all time:

$$\xi_{total} = \sum_n \xi(n) \qquad \text{. . . . . .(5)}$$

Differentiating the cost function with respect to the weight vector w_{ji}

$$\frac{\partial \xi_{total}}{\partial w_{ji}} = \sum_n \frac{\partial \xi(n)}{\partial w_{ji}}$$

using chain rule to express the derivative of cost function ξ_{total} with respect to the weight vector,

$$\frac{\partial \xi_{total}}{\partial w_{ji}(n)} = \sum_n \frac{\partial \xi_{total}}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)} \qquad \ldots\ldots\ldots (6)$$

where the time index n runs over $v_j(n)$ and not $\xi(n)$. The partial derivative $\partial \xi_{total} / \partial v_j(n)$ is the change in cost function ξ_{total} produced by a change in the internal activation potential v_j of neuron j at time n. Moreover we recognize that

$$\frac{\partial \xi_{total}}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)} \neq \frac{\partial \xi(n)}{\partial w_{ji}(n)} \qquad \ldots\ldots\ldots (7)$$

It is only when the expression is summed over all n that the equality holds. From (6) and using the idea of gradient descent in weight space, the updating of the weights is done as follows:

$$w_{ji}(n+1) = w_{ji}(n) - \eta \frac{\partial \xi_{total}}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)} \qquad \ldots\ldots\ldots (8)$$

where η is the *learning-rate parameter*. For any neuron j in the network, the partial derivative of the activation potential $v_j(n)$ with respect to the weight vector $w_{ji}(n)$ is given by $\frac{\partial v_j(n)}{\partial w_{ji}(n)} = x_i(n)$, where $x_i(n)$ is the input vector applied to neuron j. The local gradient for neuron j is

$$\delta_j(n) = -\frac{\partial \xi_{total}}{\partial v_j(n)} \qquad \ldots\ldots\ldots (9)$$

Accordingly, we may rewrite (8) as

$$w_{ji}(n+1) = w_{ji}(n) - \eta \delta_j(n) x_i(n) \qquad \ldots\ldots\ldots (10)$$

The explicit form of the local gradient $\delta_j(n)$ depends on whether neuron j lies in the output layer or a hidden layer of the network. The local gradients for the two cases are given below as

$$\delta_j(n) = \begin{cases} e_j(n)\varphi'\left(v_j(n)\right), & \textit{neuron } j \textit{ in the output layer} \\ \varphi'\left(v_j(n)\right)\sum_{m \in A} \Delta_m^T(n) w_{mj} & \textit{neuron } j \textit{ in the hidden layer} \end{cases}$$

where,

$$\varphi'\left(v_j(n)\right) = \frac{\partial y_j(n)}{\partial v_j(n)}$$

$\Delta_m^T(n) w_{mj}$: inner product of the vectors $\Delta_m(n)$ and w_{mj} both of which have dimensions $(m+1)$.

$\Delta_m(n) = \left[\delta_m(n), \delta_m(n+1), ...\delta_m(n+M)\right]^T$, a vector.

A : set of all neurons whose inputs are fed by neuron j, located in a hidden layer, in a forward manner.

$v_j(n)$: internal activation potential of neuron j that belongs to set A.

A modification to TDNN is the *Adaptive Time Delay neural network* (ATDNN) and the slightly different *Continuous Adaptive Time neural network* (CATNN). The latter two have look-back intervals that adapt automatically as learning progresses, seeking phase relationships that produce higher correlation over history.

Both TDNN and CATNN have been employed in this research to develop traffic flow prediction models. Their look-back features make them particularly appropriate for learning spatial patterns that change in time (*i.e., spatio-temporal*). Therefore, they tend to be superior to time series approaches which ignore the spatial component of the spatio-temporal patterns and also superior to static artificial neural networks that ignore the temporal component.

FREEWAY SITE DATA DESCRIPTION

A section of interstate 5 (I-5) freeway in Orange County, California, was selected for this research. The length of the section is about 5 miles with 9 loop detector stations along the main line between the intersection of the I-5 and the I-405 freeways (the El-Toro “Y”) and the intersection of Jeffrey Rd. and I-5 in the city of Irvine. This section also includes 2 off-ramps and 5 on-ramps. Therefore, flow and density data from a total of 16 loop detector stations were used. Real-time on-line data were available via the Caltrans Advanced Traffic Management (ATMS) Testbed headquartered the University of California Irvine (UCI). In addition to the real data, a comprehensive set of simulated data was produced using Paramics, a state of the art, ATMS-ready microscopic simulator (Paramics, 1998).

A dynamic Origin-Destination matrix available for the Testbed network that includes the section noted above was used to drive the simulator; O-D Data for the evening peak of April 2, 1997, from 16:00 hr. to 18:00 hr. were arbitrarily selected. The whole Irvine network including the

same section of the I-5 freeway was coded into Paramics with the exact geometry and loop detector station layout as in the real world. At each detector station, flow and density values were collected, averaged across lanes and used to develop the NN models. The finest resolution of the data used was 30 seconds.

The simulation data set was used for training, testing and validating the TDNN. The real data were reserved for real-world validation of the model and were not used in the development phases due to absence of some ramp data. The neural networks were trained to predict the flow and density in the immediate future for a middle location, based on the flow and density in the past at the same location as well as at the neighboring upstream and downstream locations. Of the 9 main line loop stations, numbered consecutively, station number 5 was used as the location at which prediction takes place, given the input from the other stations (1 to 4, and 6 to 9 and the on/off ramp stations).

The "*walk forward*" method was used for developing the network, training on the first number of records defined as the training set, testing on the next number of records defined as the test set and validating on the following number of records defined as the validation set. This is called the first "*fold*". Once that is complete, the process walks forward in the data by the number of records defined as the validation set and retrain. The process continues until the end of the file. The "walk forward" parameters used are: 180 training records, 25 testing records and 25 validation records.

For the genetic algorithm used to optimize the neural network structure, 30 generations and population size of 300 were settled upon, as will be discussed shortly. The well-known *roulette wheel* method was used for selection, which gives chromosomes with the highest fitness greater probability of being selected and, hence, producing better generations.

Different sets of TDNN and CATNN were developed in subsequent phases as follows:

- Investigate the effect of the extent of prediction in future on the prediction accuracy, using 30 second data resolution,
- Optimize the spatial contribution from neighboring detector stations,
- Optimize/select data resolution that minimizes prediction errors.

PRELIMINARY INVESTIGATION AND CHOICE OF OBJECTIVE FUNCTION:

Since the number of generations and the population size used by the genetic algorithm has a direct bearing on the optimality of the structure of the resulting neural network, a phase of preliminary investigation was necessary. In this phase, 30-second flow and densities from all

stations were used to train TDNN and CATNN to predict flow and density values at the middle station number 5.

The extents of prediction in future were: 30 seconds, 1, 2, 4, 5, 10 and 15 minutes. An optimization run using the GA to produce a winner TDNN or CATNN was made for 'each' prediction extent. The objective (fitness) function used was the Average Absolute Error *(AAE)*, defined as the absolute of the difference between the actual value and the predicted neural output and is averaged across all records as follows:

$$AAE = \frac{\sum \left| \left(y_{actual} - y_{predicted} \right) \right|}{n}$$

where:

y_{actual} = actual value of the output in the data set;

$y_{predicted}$ = predicted output value, and

n = number of records in the data set.

For each of the prediction horizons, the population size and the number of generations used by the GA to optimize the NN were incremented and the effect of the results observed.

The population size was incremented from 30 to 300 and the number of generations was incremented from 10 to 30. It was observed that as both numbers increased the prediction errors decreased. However, higher numbers of generations and population sizes were found not warranted due to minimal improvements and considerable increase in run times. Therefore, the number of generation was set to 30 and the population size to 300 in the remainder of the research.

EFFECT OF EXTENT OF PREDICTION ON PREDICTION ACCURACY

To examine the effect of the extent of prediction on prediction accuracy, data resolution was fixed to 30 seconds and the spatial contribution was also fixed to full contribution from all loop stations. The TDNN and CATDD both optimize the look-back interval (temporal contribution). Although the TDNN has a fixed look-back interval (as opposed to the CATNN), the GA optimizes its look-back interval. The GA examines several TDNNs, as chromosomes in the gene pool during optimization, with varying look-back interval settings; and, hence, the look-back interval gets optimized as well. However, it should be made clear that each TDNN in the population has a fixed look-back interval. In the case of the CATNN, each network varies the look-back interval during training, together with the variety of networks in the population processed by the GA, resulting in

an optimized look-back interval. An optimization scenario using the GA is implemented for each prediction horizon of 30 seconds, 1, 2, 4, 5, 10 and 15 minutes. The resulting optimal network in each case is validated on the validation data set and the error reported. Figures 3.a, for example, show the predicted and actual flow and densities for the case of a 30-seconds prediction. Figures 3.b show the same for the other end of the spectrum for a 15 minutes prediction.

From the plots of predicted value versus actual values (not all are shown in order to save space), it can be seen that the predicted flow and density values are quite close to the actual values for up to 2-minutes of prediction. After two minutes, the predicted values tend to become closer to the mean of the actual values, which becomes very evident at 15-minutes extent of prediction. This is in agreement with expectations; as the extent of prediction increases it becomes increasingly difficult to predict far ahead using 30-second dynamics, and the model resorts to guessing the 'average'.

Figure 4 shows a summary of the extent of prediction versus the average absolute percentage error, defined as:

$$Average \quad Percentage\ error\ value = \frac{\left(\left| \left(y_{actual} - y_{predicted} \right) \right| * 100 \right)}{y_{actual}} / N$$

where N is the total number of records for which predictions are made.

It can be seen from the figure that the average percentages of error are less than 10% for both flow and density up to 4 minutes of prediction extent. After 2 minutes, the average percentage error values exceed 10% and increase gradually to 15% for 15 minutes prediction. The best neural network that survived the genetic evolution over the 300 generations in this case was the TDNN (and not the CATNN).

The GA also reports the optimal "*look-back*" interval for each extent of prediction. The look-back interval is the number of time steps in the past that affected the prediction the most. An interesting pattern of look-back intervals was observed. The "look-back" interval was found to decrease as the extent of prediction increased, indicating that the temporal history has less bearing on far predictions--the best guess for which is around the mean values. Figure 5 shows the "*look-back*" interval vs. the extent of prediction.

EFFECT OF SPATIAL CONTRIBUTION ON PREDICTION ACCURACY

In this section, the effect of the extent of upstream and downstream spatial contribution on the prediction accuracy is examined. Spatial contribution is defined as how many loop detector stations, upstream and downstream the subject station #5, are included in the training. For any

given extent of prediction, four data files were prepared. The first data file had the spatial contribution from all the detector stations except the farthest two mainline stations. In the second data file the penultimate stations were dropped as well as any on/off ramp stations in between the farthest and the penultimate station pairs, and so on. The fourth data included the subject station #5 only (*i.e.,* no spatial contribution and predictions are based on temporal history only). The extents of prediction used were 30 second, 1 minute and 15 minutes only to keep the number of optimization runs reasonable. Figure 6 shows the average absolute error versus the extent of spatial contribution for the three prediction extents respectively. Figure 7 shows the relationship between the extent of prediction and the error for the cases of full spatial contribution and no spatial contribution at all for contrast. The following observations can be made:

- The less the spatial contribution the higher the error as shown in Figure 6,
- Three stations on both sides of the subject loop station are probably sufficient. It should be noted that its extremely difficult to obtain good loop data from a large number of consecutive loop stations, in practice,
- The longer the extent of prediction (towards 15 minutes) the less pronounced the effect of the spatial contribution. Figure 7 shows that the benefit from full spatial contribution as opposed to no contribution at all is much more evident in the case of 30-second predictions.

EFFECT OF DATA RESOLUTION ON PREDICTION ACCURACY

In this section, the effect of data resolution itself on the accuracy of prediction is examined. The resolutions considered were 30 seconds (the original data), and 1 minute, 2 minutes, 5 minutes and 15 minutes aggregations of the original data. The extents of prediction were multiples of the resolution of data used. For 2 minutes resolution, the extents of predictions were 2, 4, 6 , 10 and 14 minutes. Similarly, for 5 minutes resolution the extents of prediction were 5,10 and 15 minutes. All of the detector stations were considered to provide full spatial contribution.

Figure 8 summarizes the errors versus the extent of prediction for all resolutions used. Figure 9 summarizes the effect of data resolution on prediction error for the case of 15 minutes prediction, taken as an example. It can be seen that the *higher* the level of aggregation (the lower the resolution), the *lower* the prediction errors in general for all prediction horizons. This is due to the disappearance of erratic dynamics in the values of flow and occupancy, common at 30 second readings and due to closer fit of the predicted values to the actual ones. This finding should be carefully interpreted, however. It does not mean that higher levels of aggregation and coarser data are always better, but rather that *higher levels of aggregation are better only for longer*

prediction horizons. For instance, if 10 minute predictions are desired, 10 minute resolution is best, and so on. That is, *the level resolution should be the same as the prediction horizon*. This is significant because it has been thought that finer data should lead to better results.

VALIDATION OF TRANSFEREABILITY TO REAL TRAFFIC DATA

To validate the models transferability and verify that the above findings from simulated traffic scenarios are applicable to the real world, a validation phase using real freeway data was conducted. Real 30-second flow and occupancy data were obtained from the ATMIS testbed, headquartered at the University of California Irvine (UCI). The Testbed is connected to the Caltran's District 12 *(D12*) Traffic Management Center over fiber optic lines, giving access both to real time data as well as historical data. The data collected for validation were for the two evening peak hours, i.e., 1600 hr. to 1800 hr. on the 15th of November, 1997. This date was selected after thorough search for good data from consecutive loop stations on different dates and different sections of the freeway. The selected site was on the I-5 North in San Clemente and San Juan Capistrano in Orange county, California. The selected freeway stretch included the section between South of Vaquero up to San Juan Creek. Only the data for mainline stations were used, due to the absence of on/off ramp data from several consecutive stations. Holes in the data were filled using interpolation. Data files were prepared similar to the case of simulated scenarios. The extents of prediction used were kept the same i.e., 30 sec, 1 min, 2 min, 4 min, 5min, 10min and 15 minutes in future. The genetic and neural parameters for training and testing were also kept the same.

Figure 10 shows the comparison of the average percentage errors for flow prediction for both real data and simulated data. It can be seen that the model behavior in the real world follows the same trends observed with simulated data, therefore, the previous findings are validated. The average percentage errors were found slightly higher for the case of real data than for simulated data for many extents of prediction. Also, the difference between the average percentage error values for real data and simulated data seem to increase as the extent of prediction increases. This could be attributable to more dynamics in the real data as well as less accuracy due to absence of ramp data and holes in the mainline data.

COMPARISON TO THE MLF MODEL (ZHANG1997)

In this section, the variety of TDNNs developed in this research is compared to the widely used Multi Layer Feed Forward model (see for instance Zhang & Ritchie 1997). The prime differences in their effort are:

- The spatial contribution was limited to only one neighboring station,
- The model used was the MLF NN,
- Extent of prediction was set to 15 seconds only,
- Data resolution was also set to 15 seconds only,
- Inputs used were speed, density and ramp volumes,
- Only simulated data were used.

To facilitate cross comparison, Zhang's model was replicated using the same data from this research as well as 30-seconds resolution instead of 15. Figure 11 shows the relative performance of the two models.

It can be seen that the average percentage error for TDNN with no spatial contribution is the highest. However, the average percentage error for the MLF is higher than the TDNN with same spatial contribution. This shows that TDNN is superior to the MLF for traffic flow modeling and prediction, mainly because of its ability to “look-back” over time and select the optimal temporal contribution. Also, the percentage error for TDNN with full spatial contribution is the lowest. This clearly indicates the significance of both the temporal and spatial contributions to capture and predict spatio-temporal traffic patterns.

SUMMARY AND CONCLUSION

In this paper, we presented a new short term traffic flow prediction study and produced a system based on an advanced Time Delay Neural Network (TDNN) model synthesized using Genetic Algorithms (GA). The model structure was presented and its performance validated using both simulated and real traffic flow data obtained from the California Testbed in Orange County. The model predicts flow and occupancy values based on their recent temporal profile at a given freeway site during the past few minutes as well as the spatial contribution from neighboring sites. Both temporal and spatial effects were found essential for proper prediction. An in-depth investigation of the variables pertinent to traffic flow prediction was conducted; the extent of the look-back interval, the extent of prediction in the future, the extent of spatial contribution, the resolution of the input data, and their effect on the prediction accuracy. Obtained results indicated

that the less the spatial contribution the higher the prediction errors, and that the inclusion of three loop stations in both directions of the subject station is sufficient for practical purposes. Also, it was found that the longer the extent of prediction, the more the predicted values lean towards the mean of the actual ones for a given data resolution. The optimal look-back interval also shortens due to becoming increasingly irrelevant. Interestingly, results revealed that coarser data resolution is better for *longer* extents of prediction. The implication is that the level of data aggregation/resolution should be comparable to the prediction horizon for best accuracy. The model performed acceptably using both simulated and real data. The model also showed potential to be superior to other well-known neural network models such the MLF.

ACKNOWLEDGMENT

This research was funded by the California Partners for Advanced Transit and Highways (PATH) headquartered at the University of California Berkeley, and the California Department of Transportation (Caltrans). It was also facilitated for by the California ATMIS Testbed, headquartered at the University of California Irvine. The main author was a Visiting Postdoctoral Research Engineer at the PATH Center for ATMIS Research while conducting this research.

REFERENCES

Abdulhai, B. 1996, “A Neuro-Genetic-Based Universally Transferable Freeway Incident Detection Framework”, Ph.D. Dissertation, University of California Irvine.

Beale and Jackson. Neural Computing : An Introduction. *Institute of Physics publishing*, Bristol and Philadelphia, 1992.

Chakroborty, P., Deb, Kalyanmoy and Subrahmanyam (1995). Optimal scheduling of urban transit systems using genetic algorithms. Journal of transportation Engineering, Nov./Dec. 1995, pp. 544-553.

Daganzo, Carlos. F. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research, Vol. 28B, pp. 269-287.

Hinton, G. and Waibel, Alexander et al (1989). Phoneme recognition using time-delay neural network. IEEE transactions on acoustics, speech, and signal processing, Vol. 37, pp. 328-339.

Kunhe, R. D. (1989). Freeway control and incident detection using a stochastic continuum theory of traffi cflow, Proc. 1st Int. conference on applied advanced technology in transportation engineering, San Diego, CA, pp. 287-292, ASCE, New York.

Lighthill, M.J., and G.B. Whitham (1955). On Kinematic waves: II. A theory of traffic flow on long crowded roads. *Royal society, London series A*, Vol. No. 229, 1178, pp. 317-345.

Michalopoulos, P. G. Yi, P., and Lyrintzis, Anastasios. Development of an improved higher-order Continuum traffic flow model for congested freeways. Presented at the 70th annual conference of TRB, Jan. 12-16, 1992.

Paramics Traffic Simulator 1998, "user manuals", Quadstone Limited, U.K.

Payne, H.J (1971). Models of freeway traffic and control. Simulation councils proceedings series, pp. 51-60.

Phillips, W. F. (1978). A new continuum model obtained from kinematic theory, IEEE *Trans. Autom. Control* AC - 23, pp. 1032-1036.

Ritchie, S.G., Zhang, H. and Lo, Zhen-Ping (1997). Macroscopic modeling of freeway traffic using an artificial neural network. *Transportation Research Record* 1588, pp. 110-121.

Ross, P. (1989). Traffic dynamics, Transportation Res. B, Vol. 22B, No. 6, pp. 421-435.

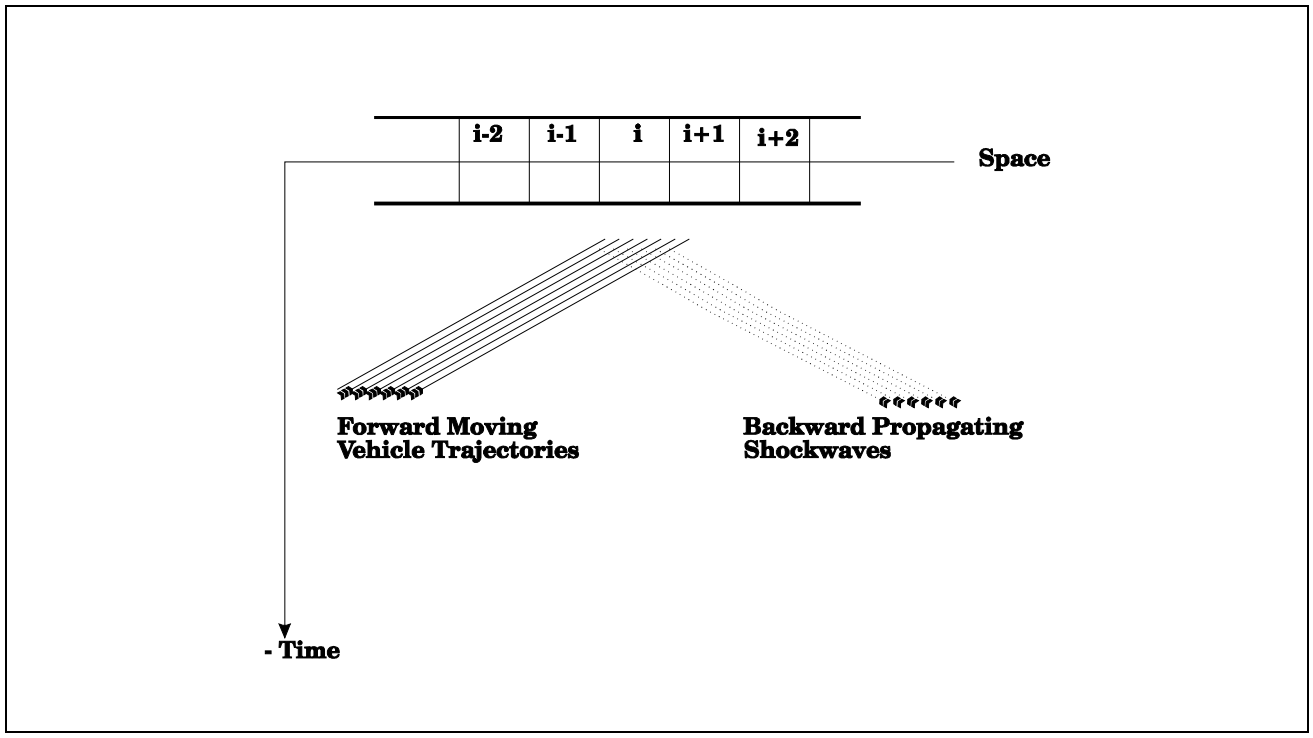


Figure 1 Schematic Illustration of the Traffic Forecasting Problem

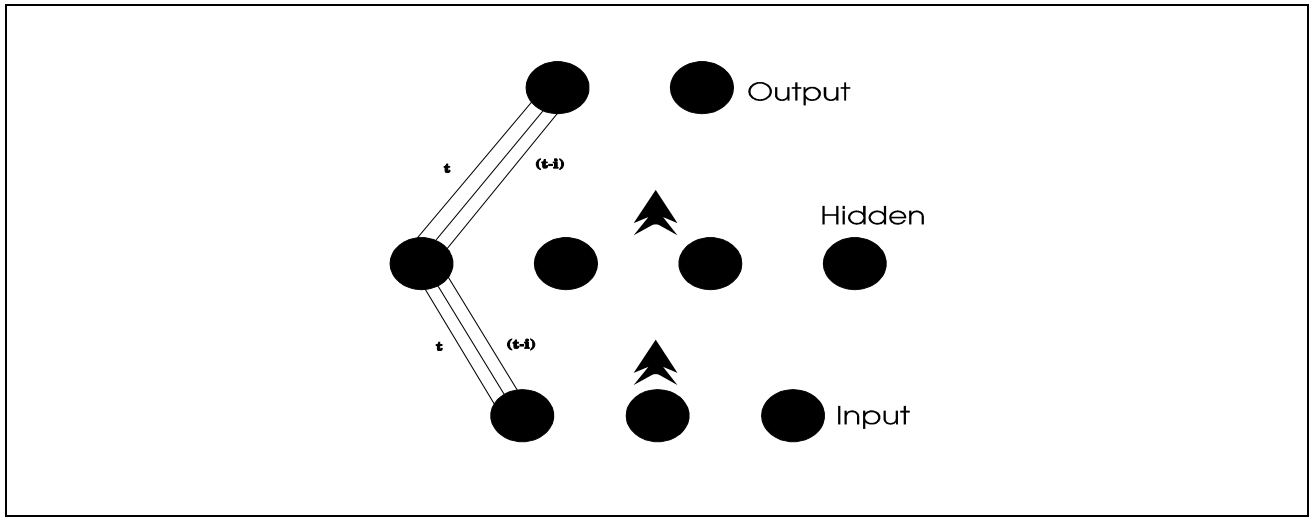


Figure 2. Architecture of the Time Delay Neural Network

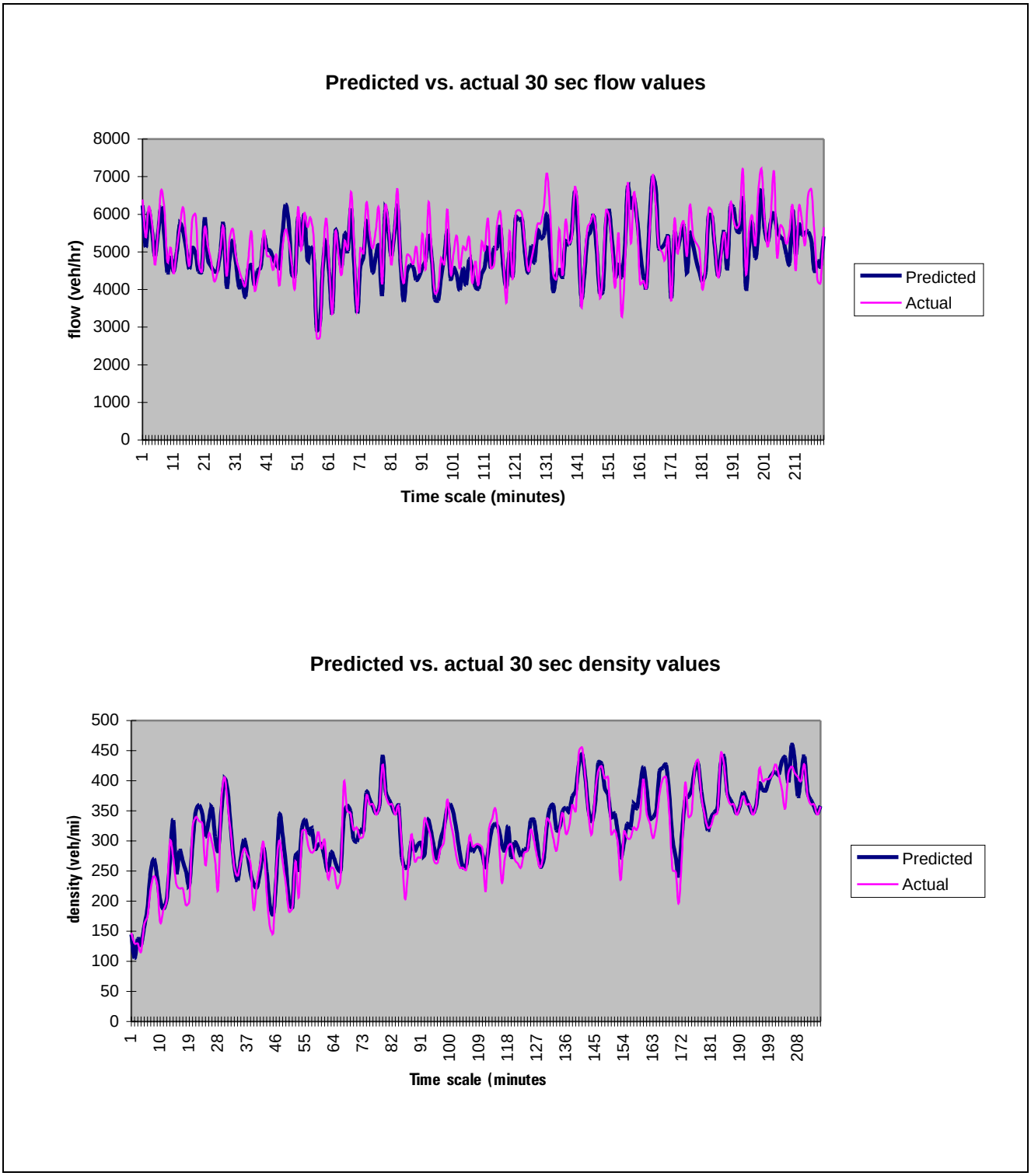


Figure 3.a. 30 sec. predictions using 30 second data resolution

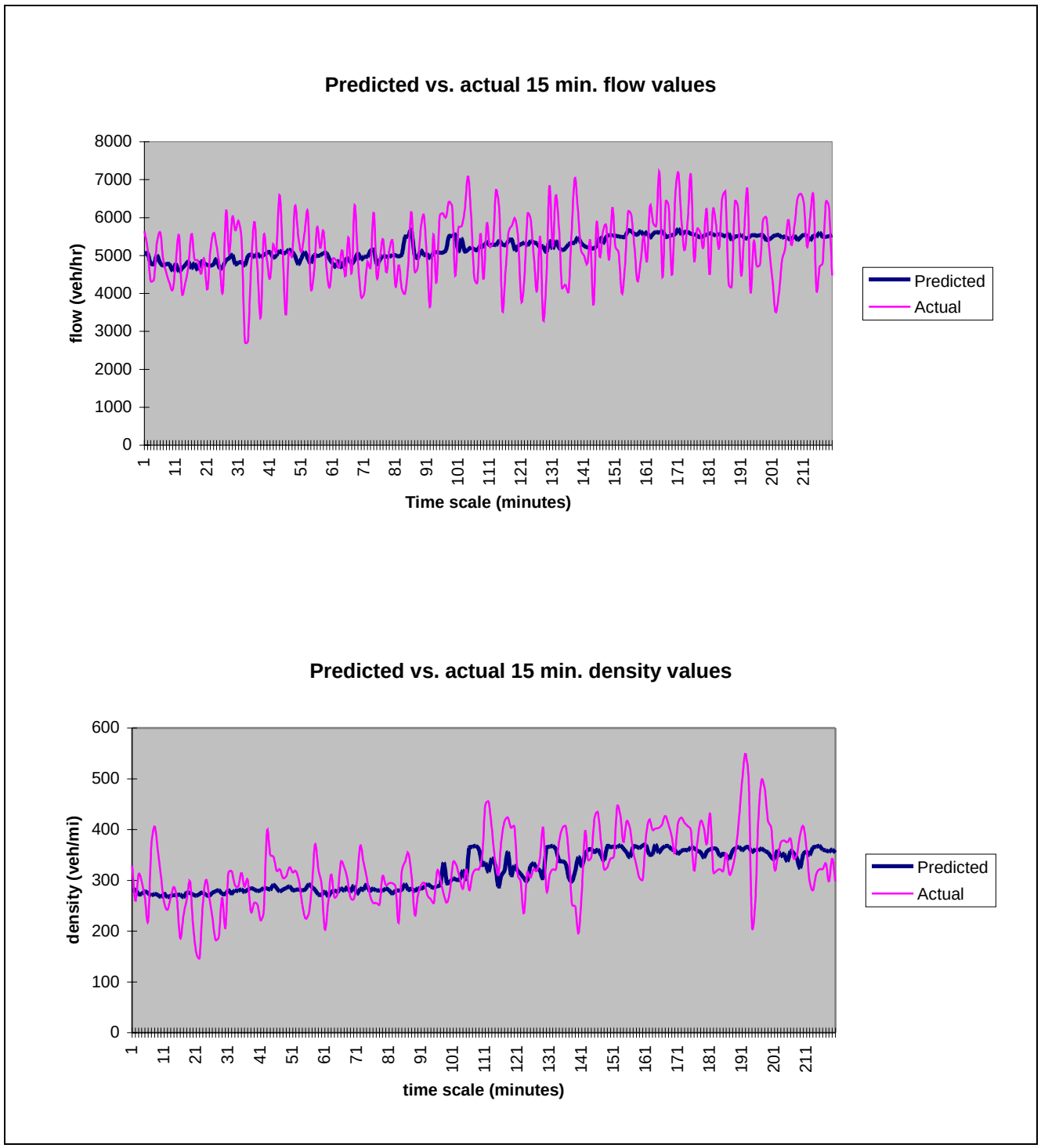


Figure 3.b. 15 min. predictions using 30 second data resolution

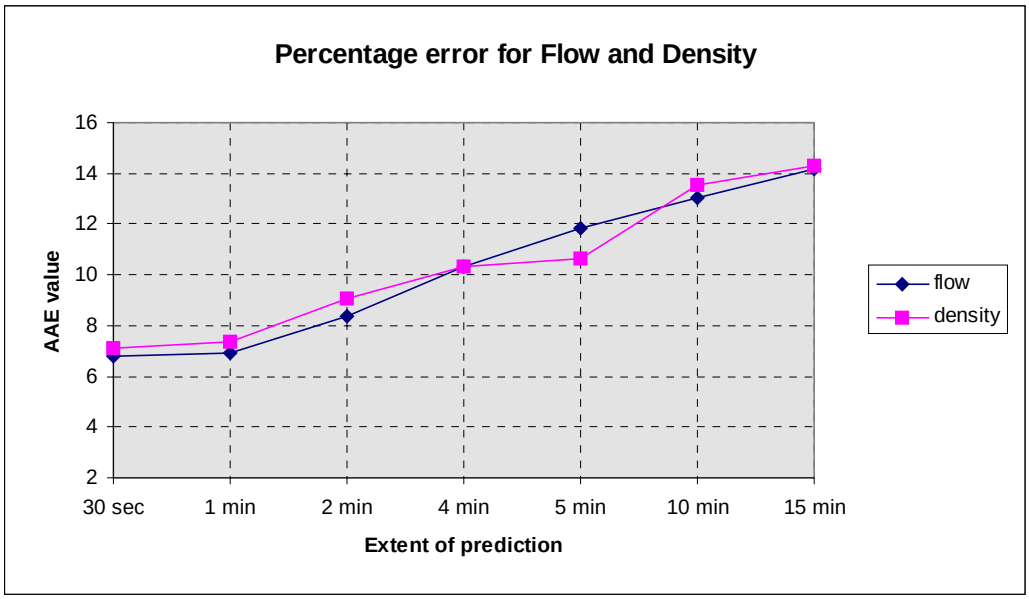


Fig. 4 Average percentage errors for various extents of prediction

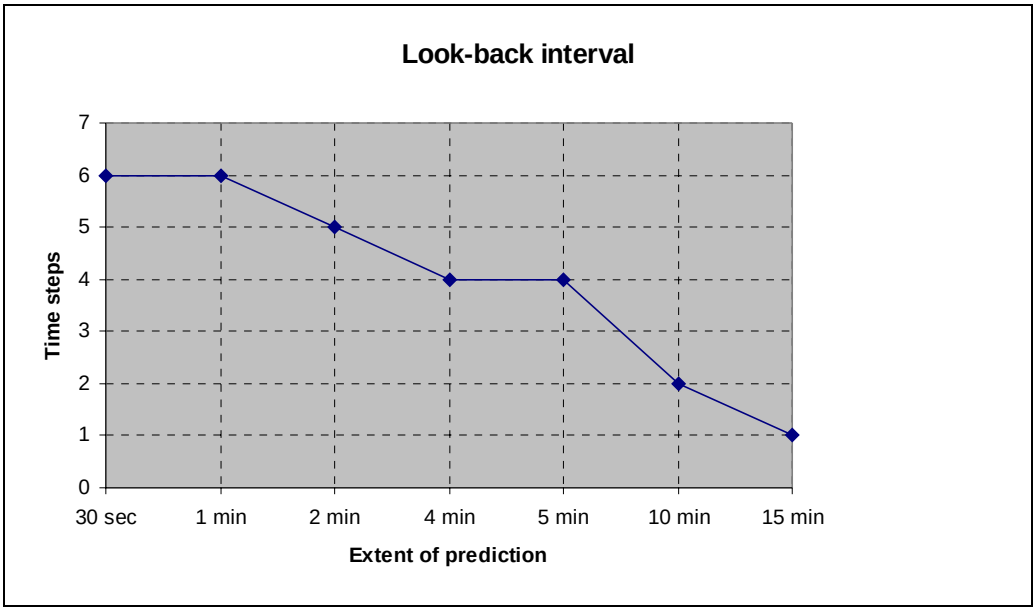


Fig. 5 Look-back intervals for various extents of prediction

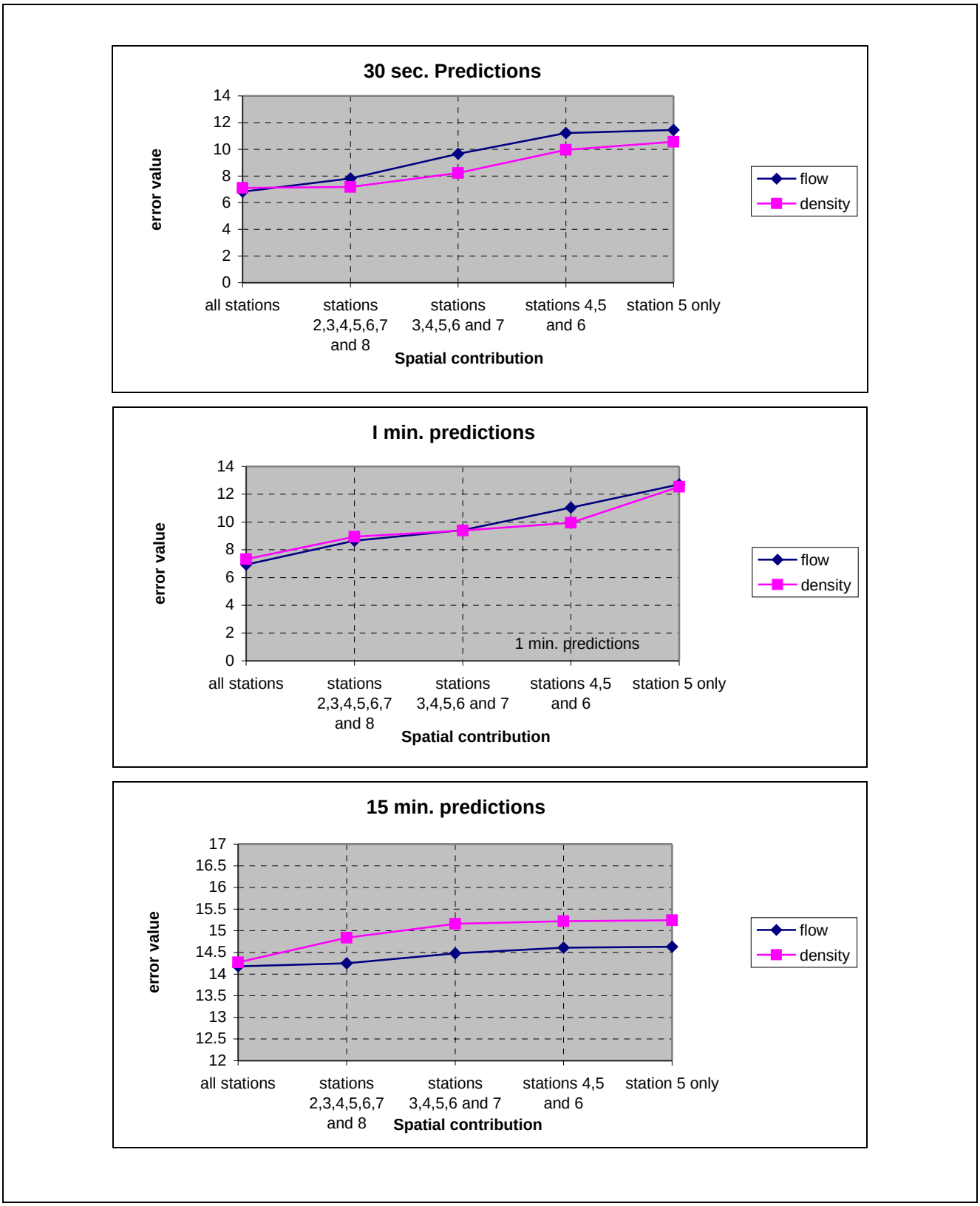


Fig 6. Average percentage errors vs. spatial contribution

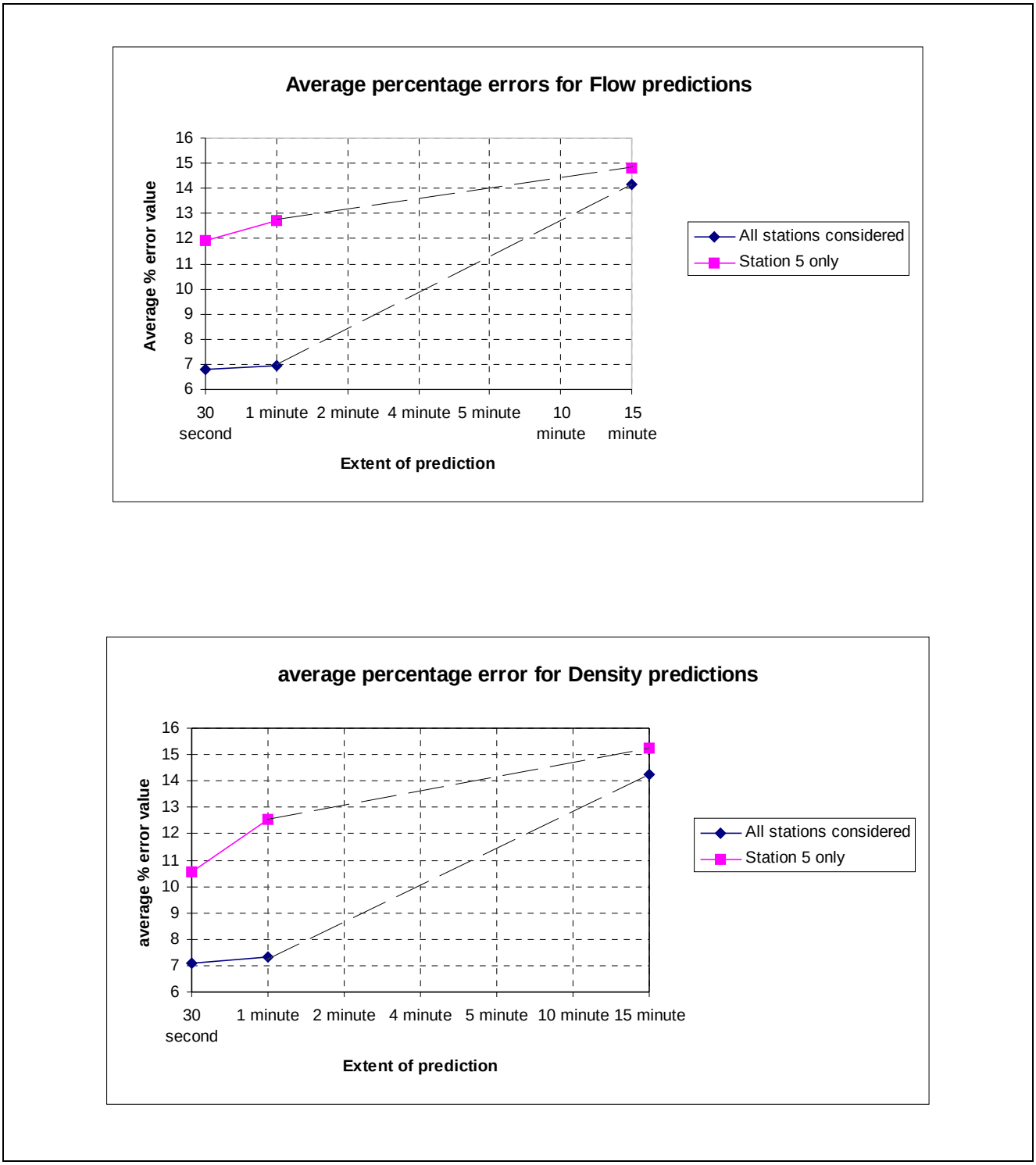


Fig. 7 Benefit of spatial contribution

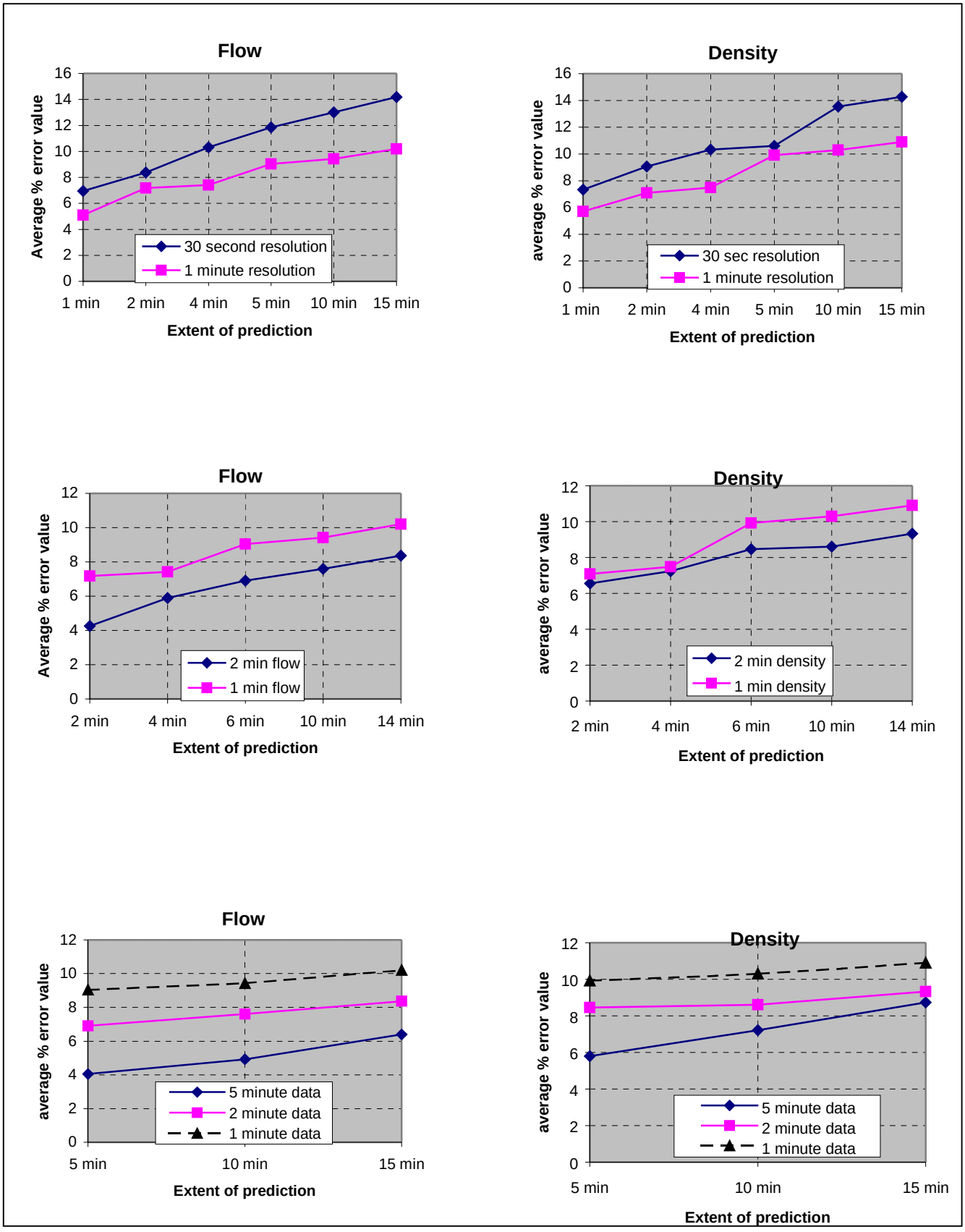


Fig. 8 Effect of data resolution on prediction accuracy

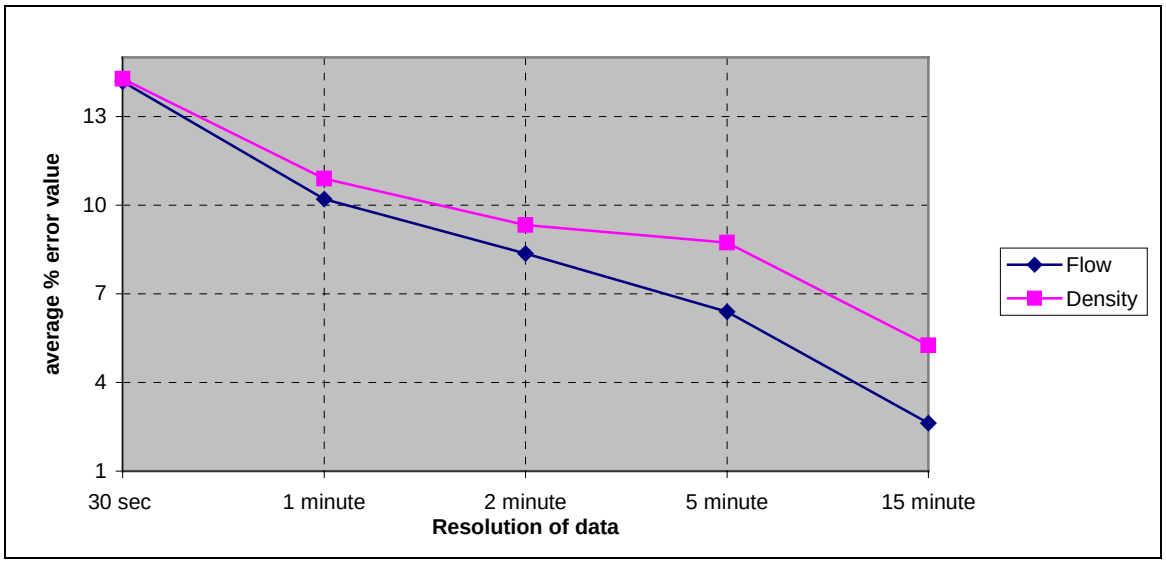


Fig. 9 Reduction in error with higher data resolution for 15 min. prediction

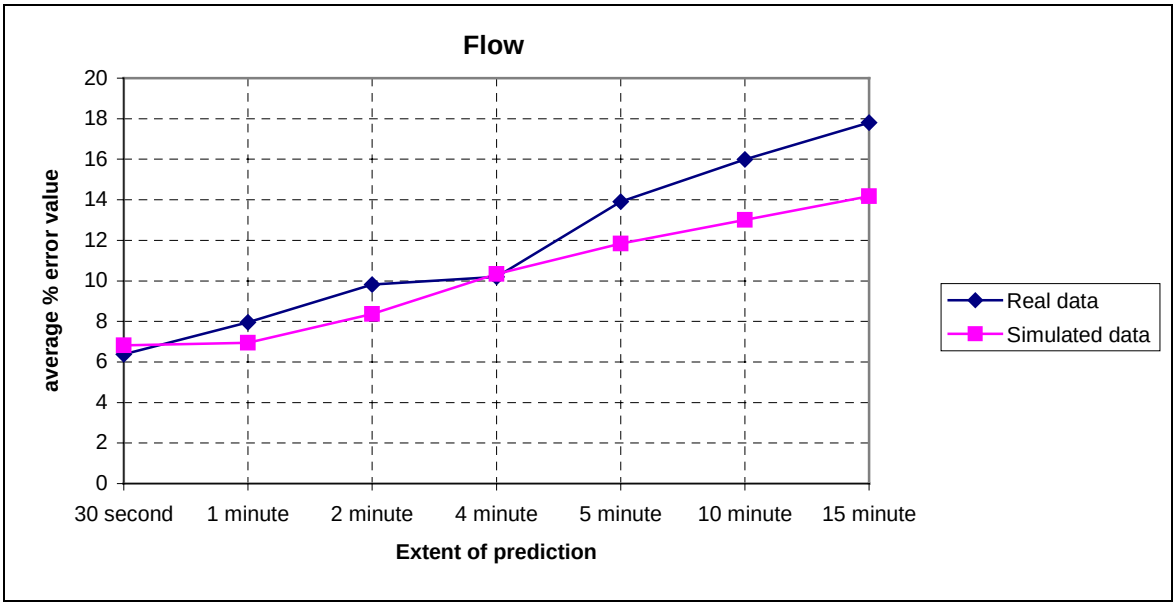


Fig. 10 Performance using real data compared to simulated data

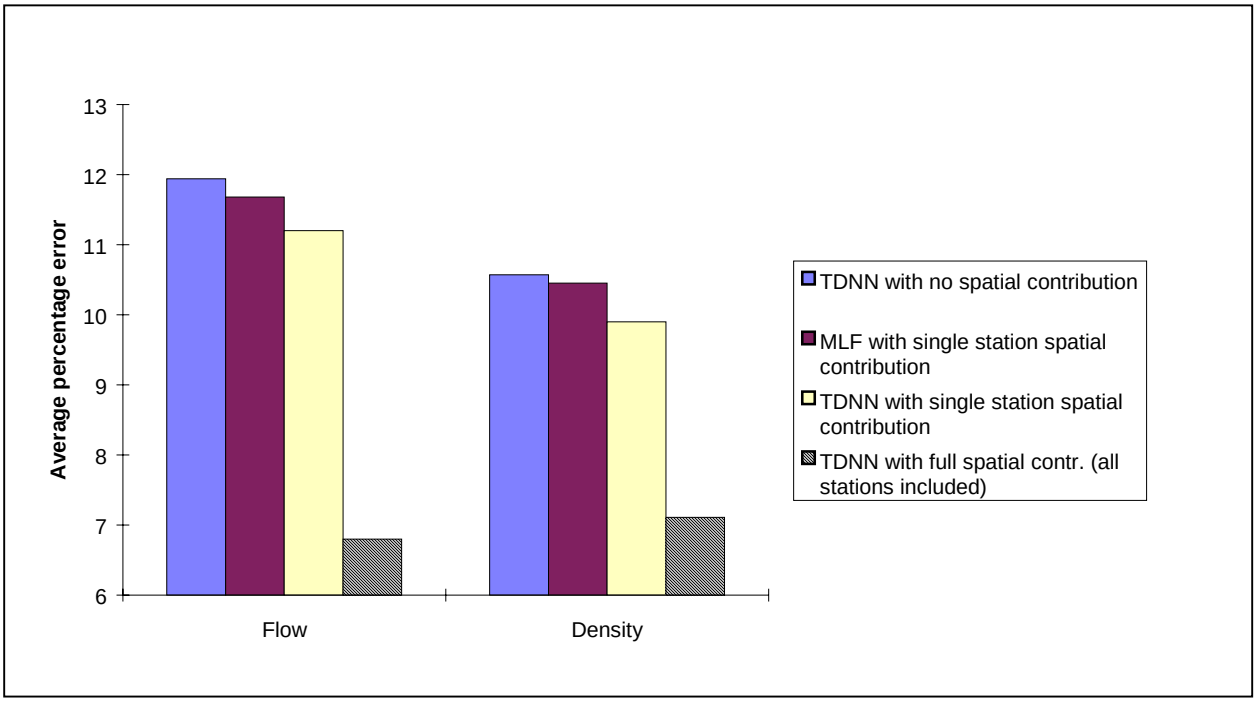
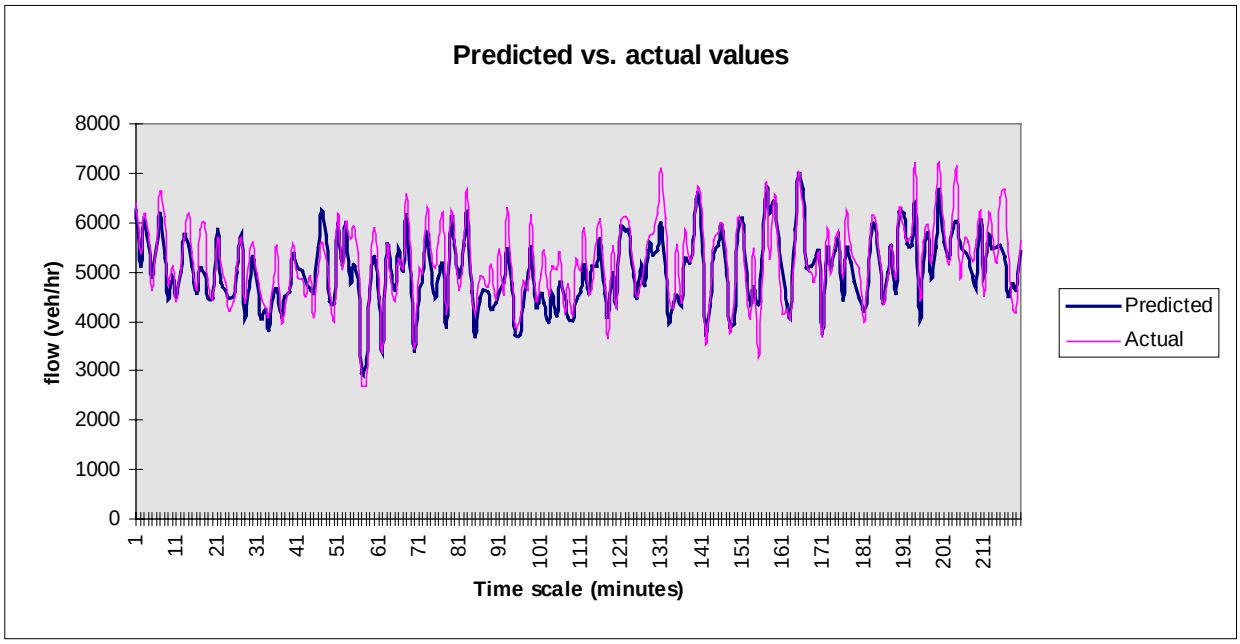


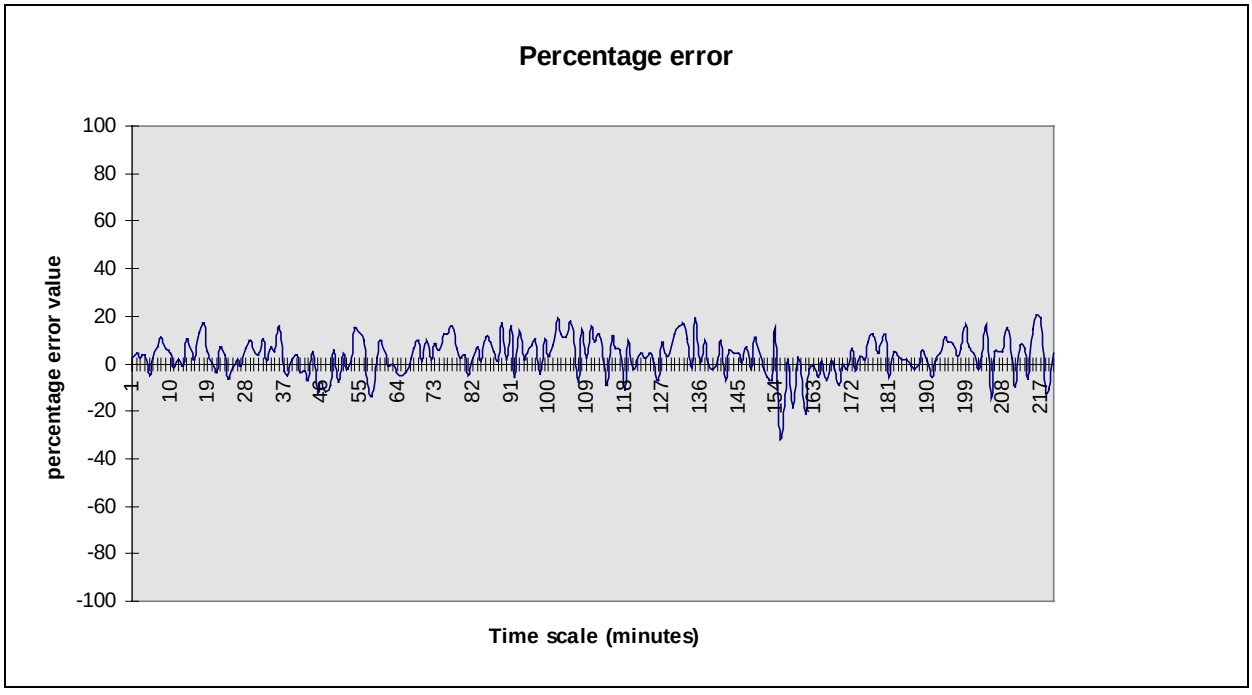
Fig. 11 comparison of the TDNN to the MLF

APPENDIX A

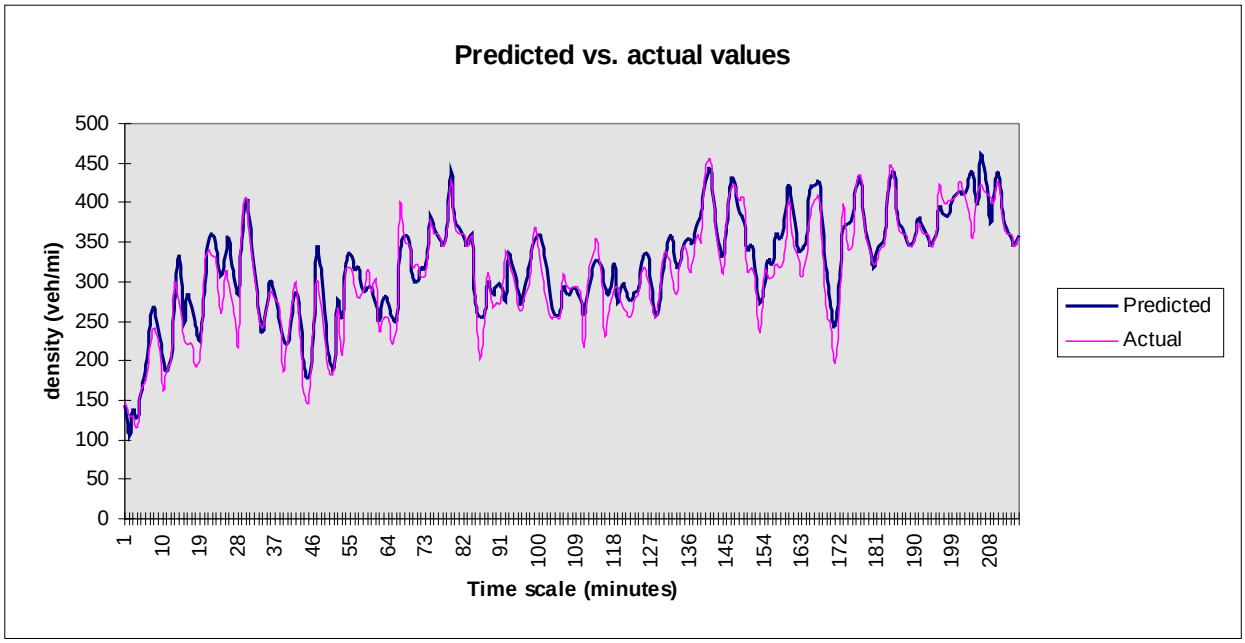
Prediction accuracy for different prediction horizons

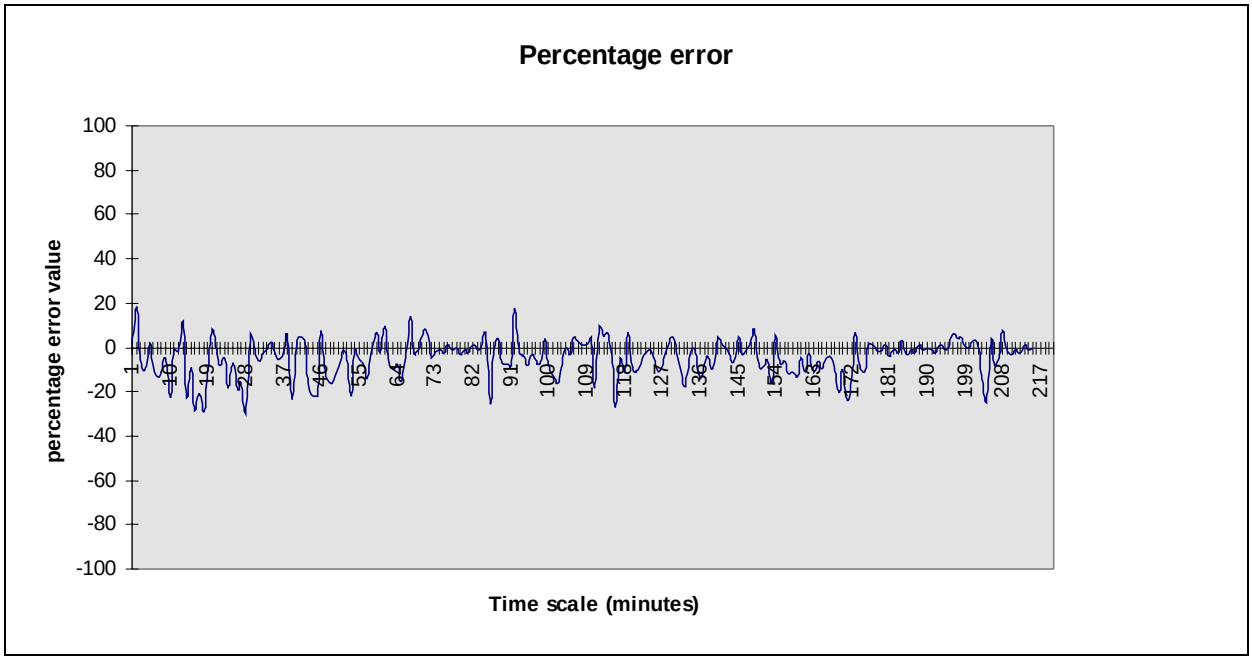
30 second: Flow



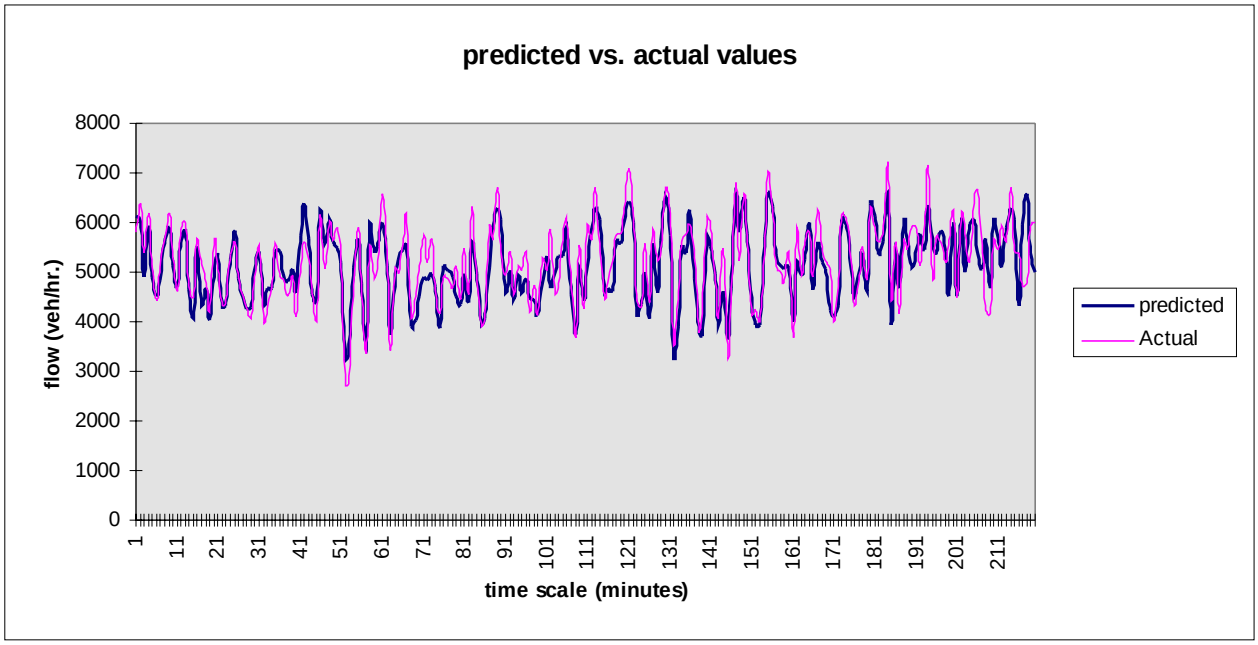


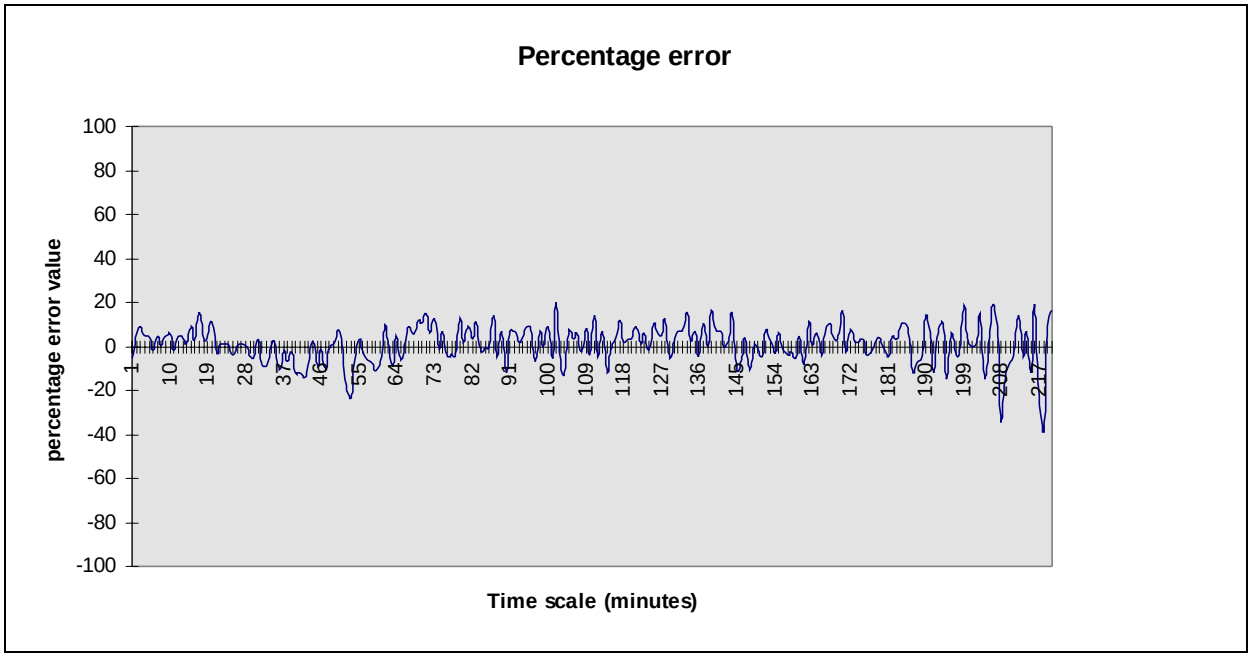
30 second: Density



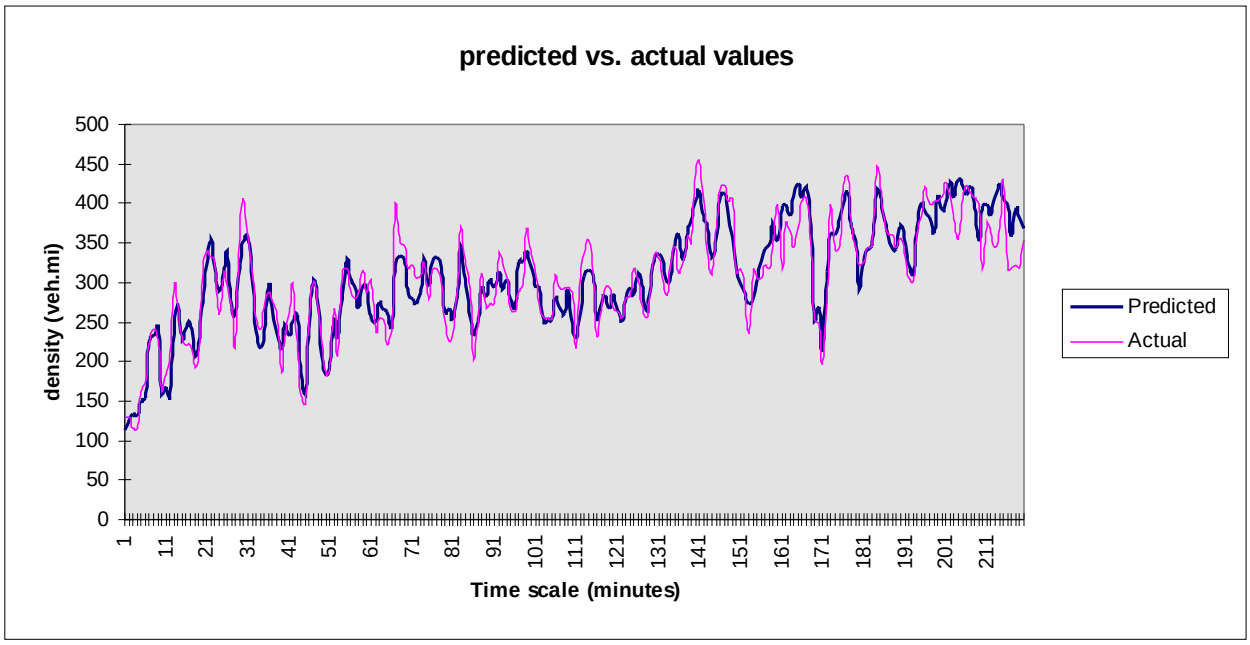


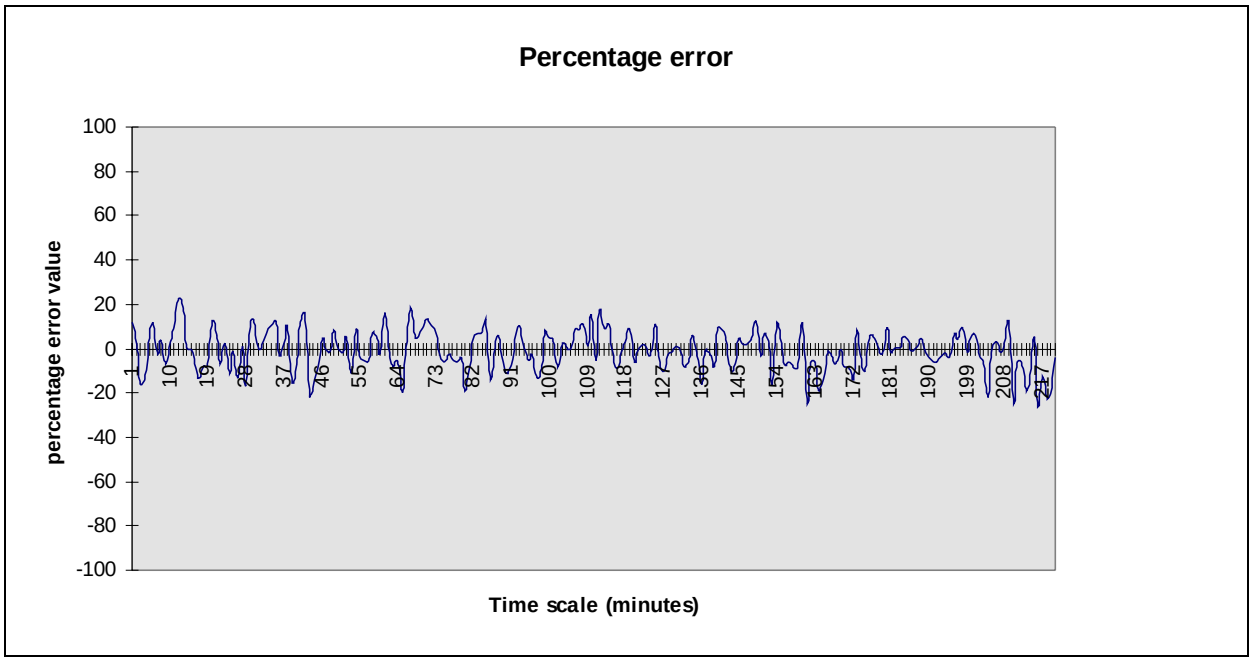
1 minute: Flow



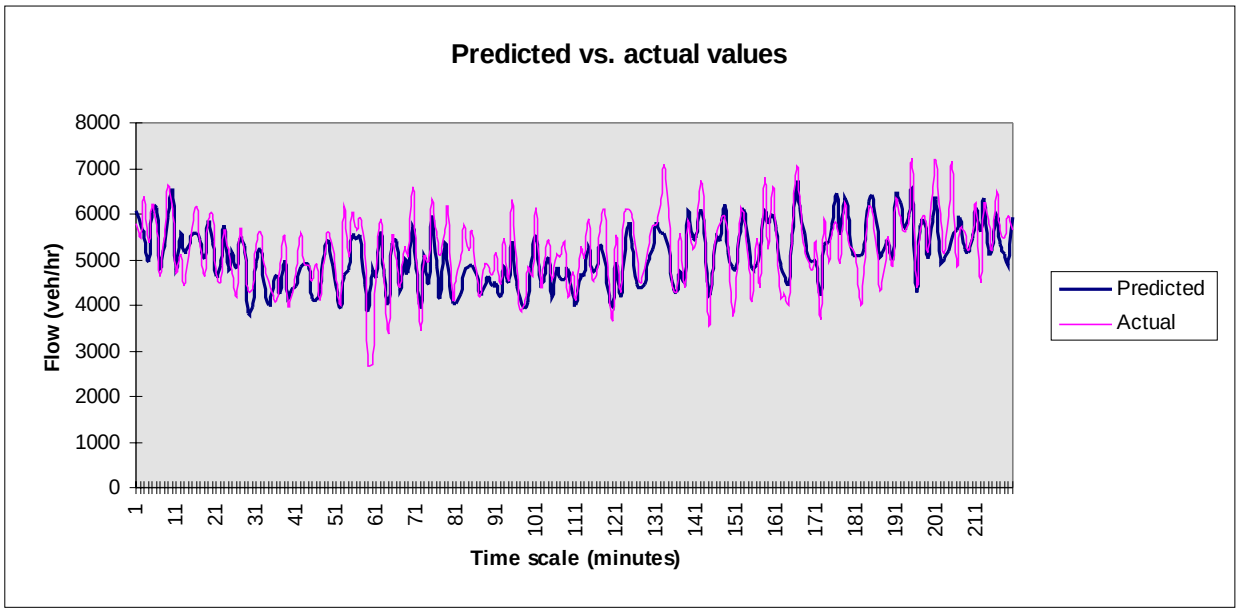


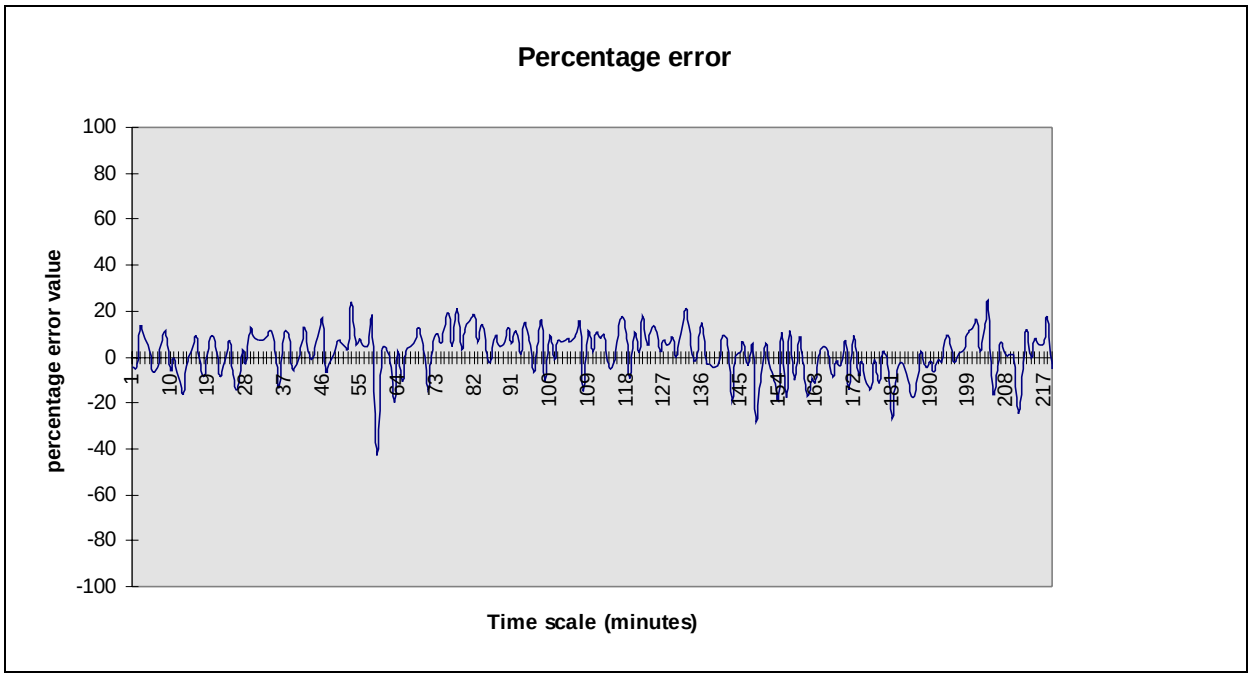
1 minute: Density



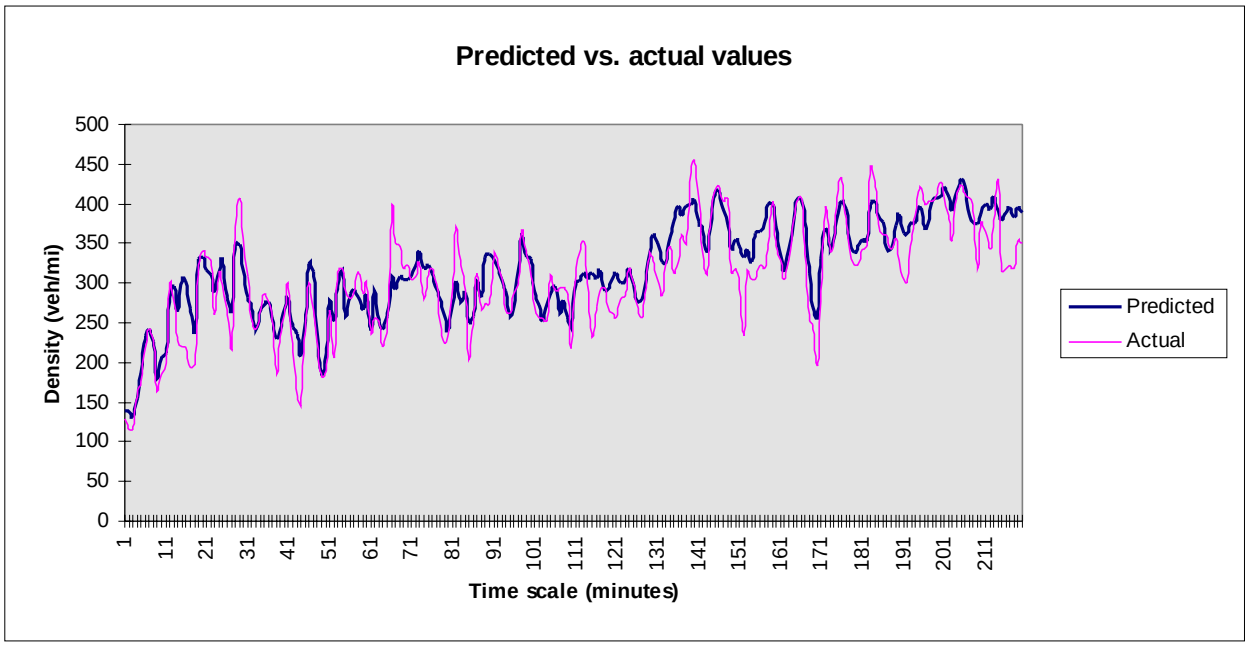


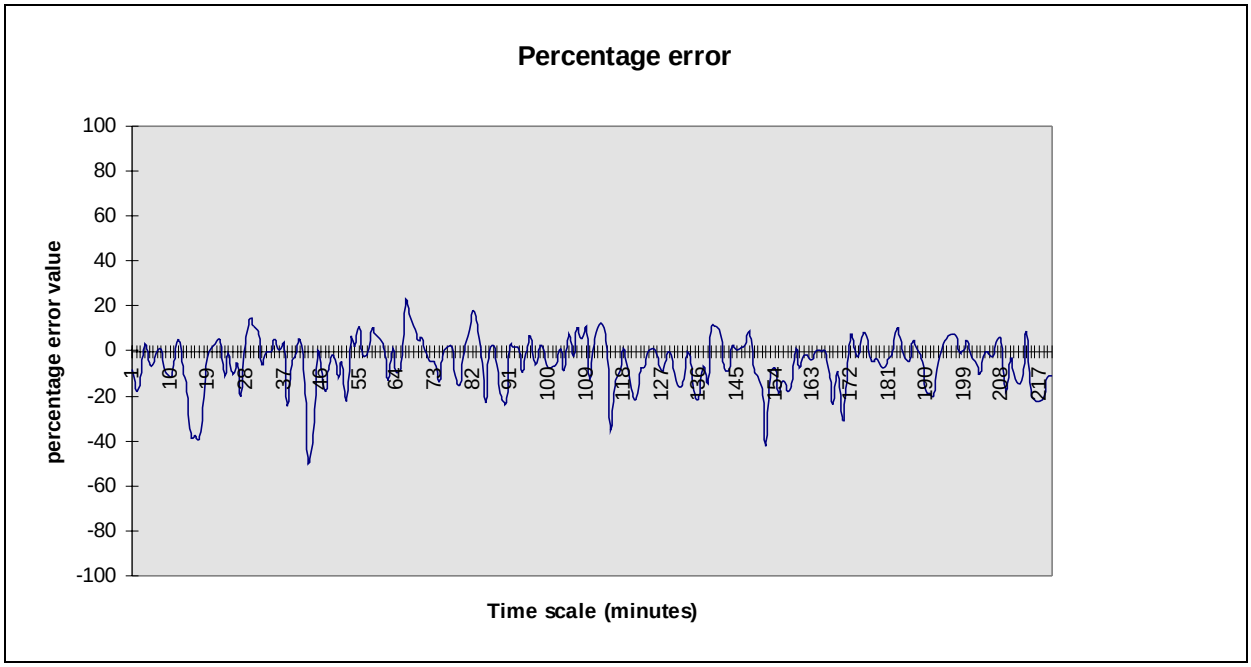
2 minute: Flow



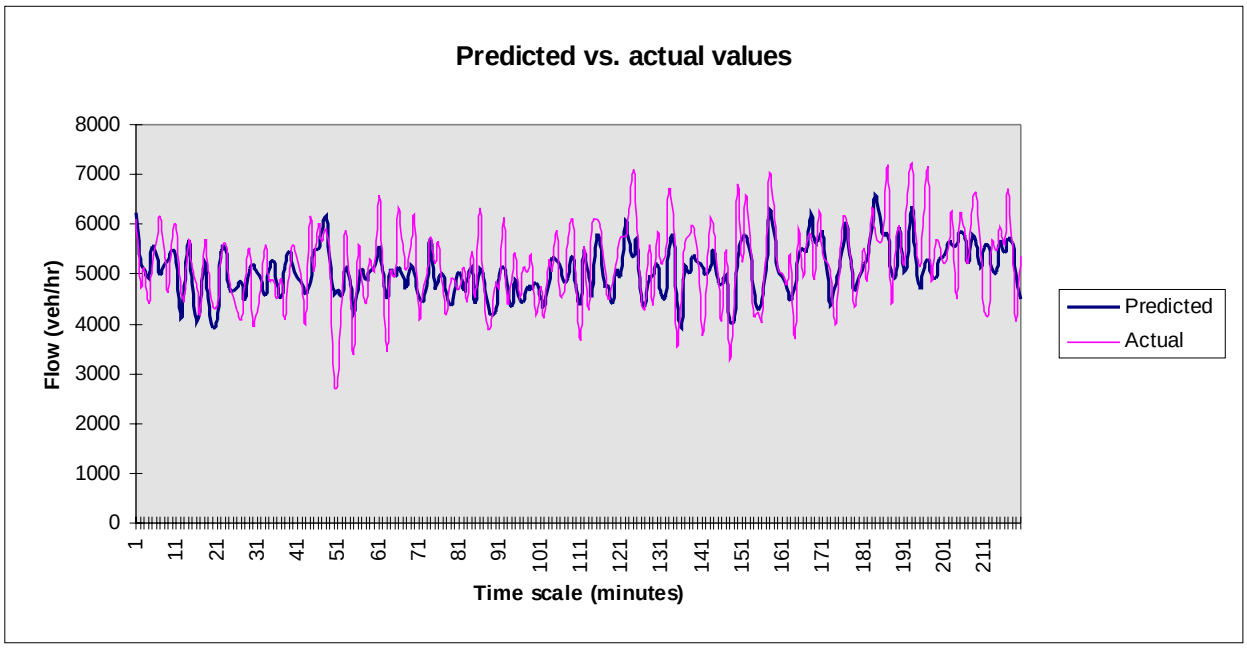


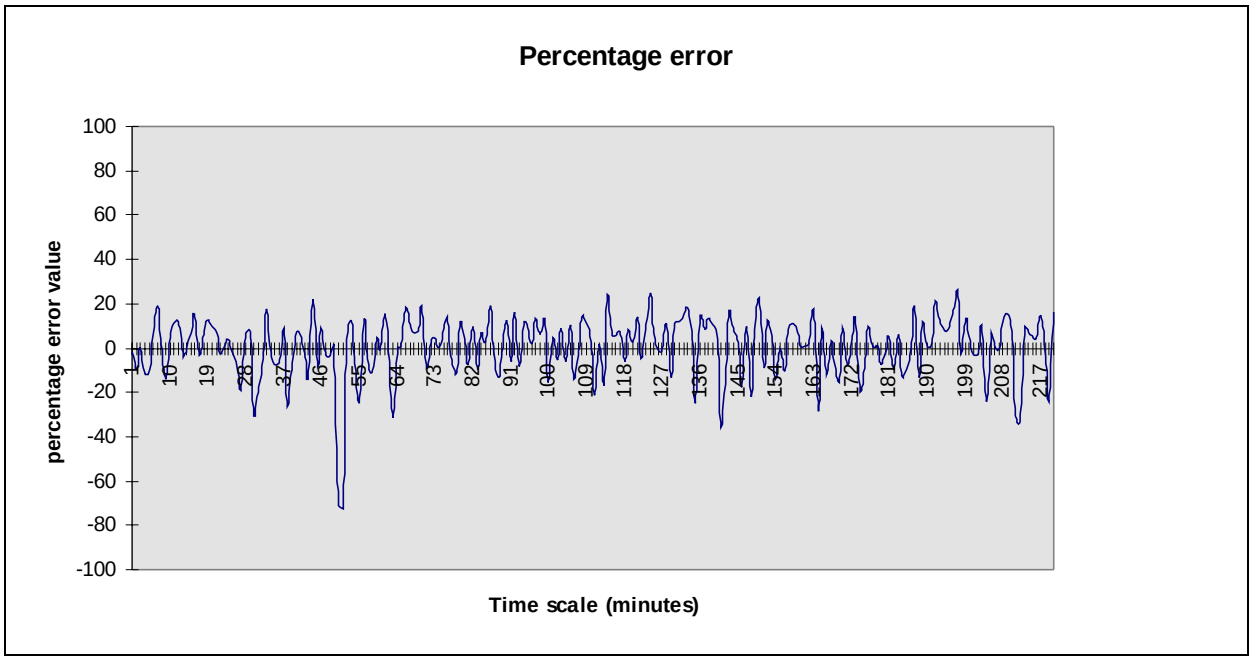
2 minute : Density



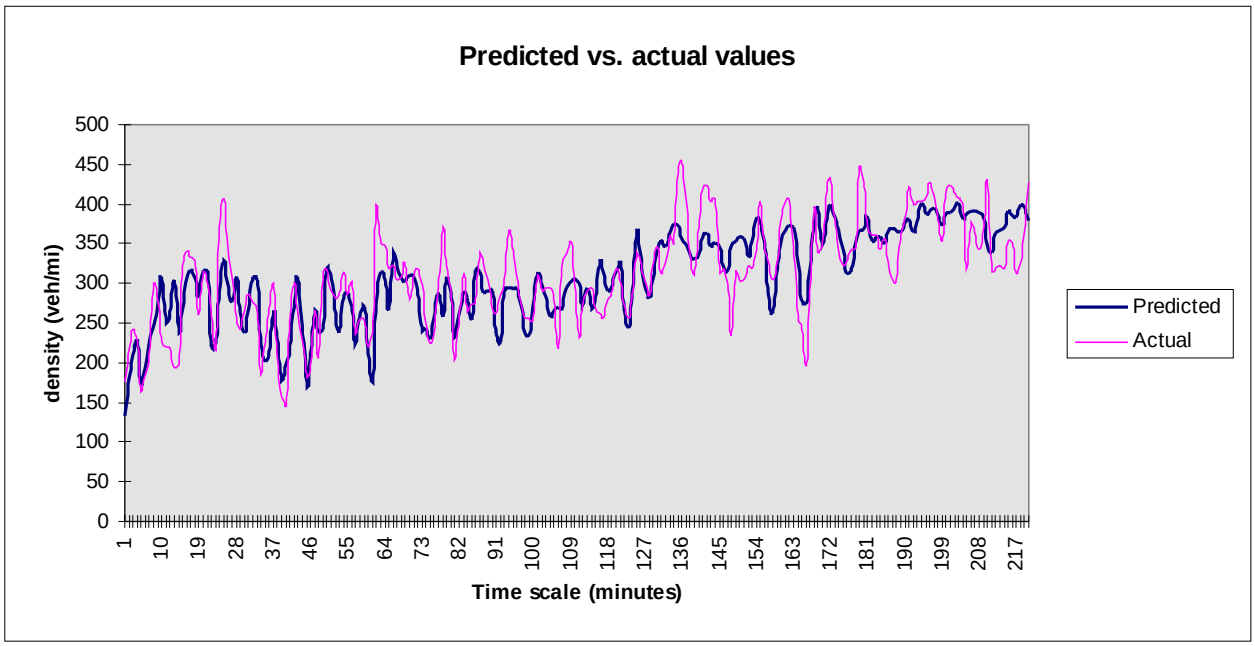


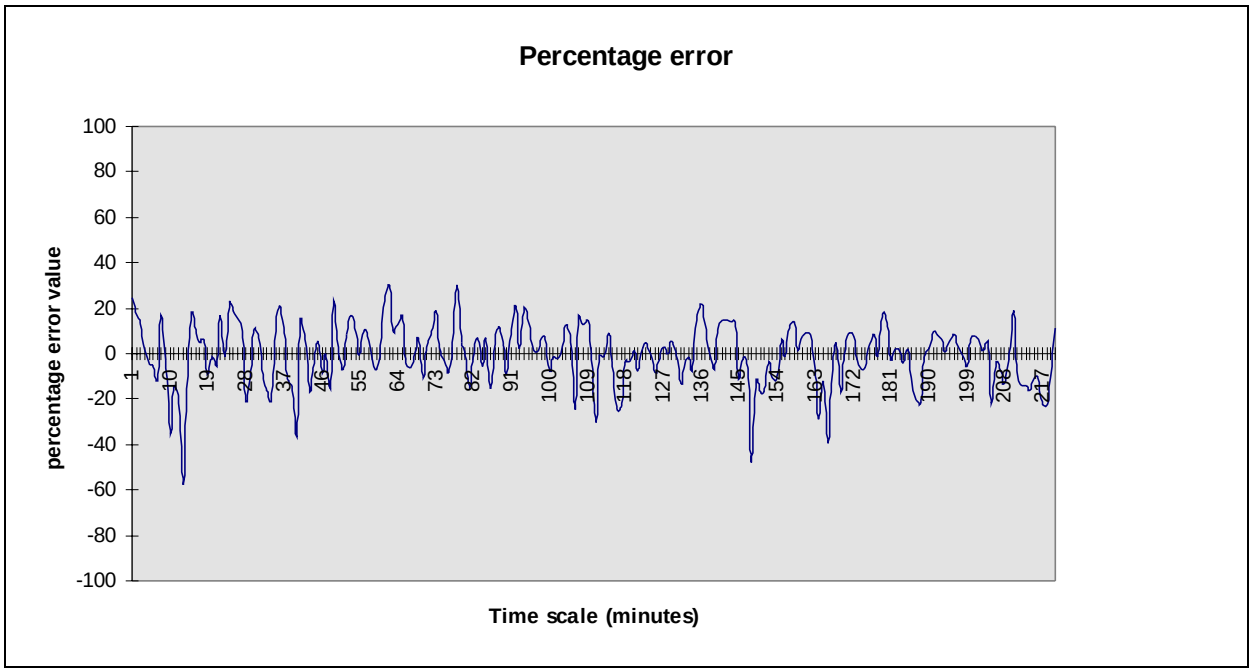
4 minute: Flow



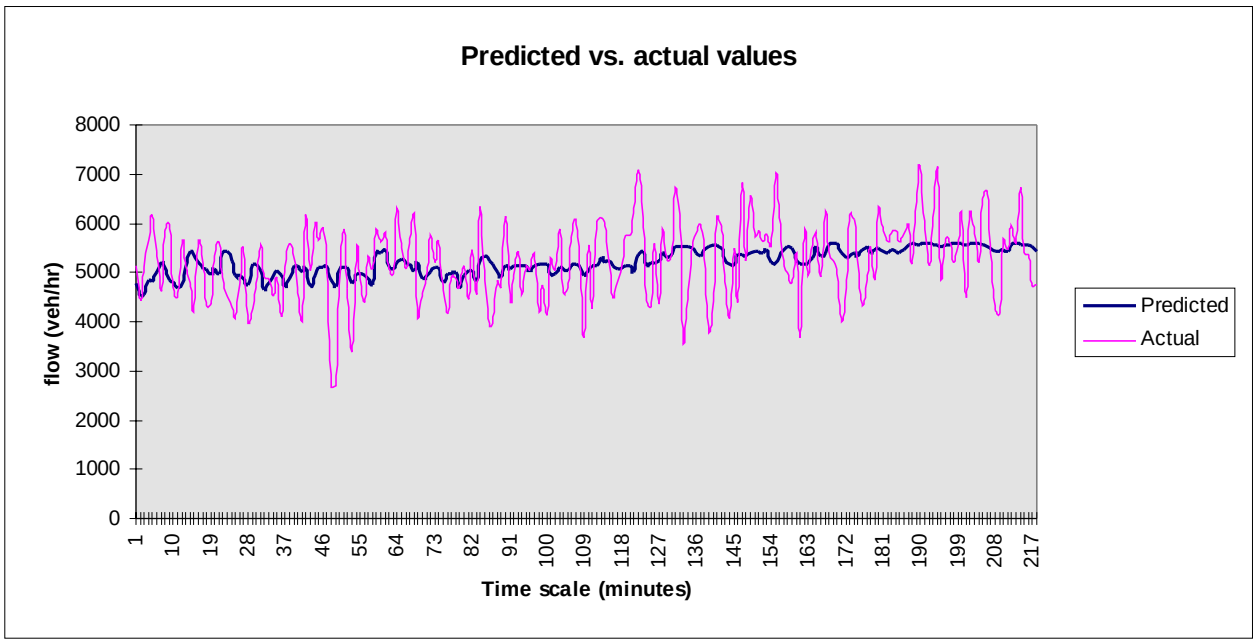


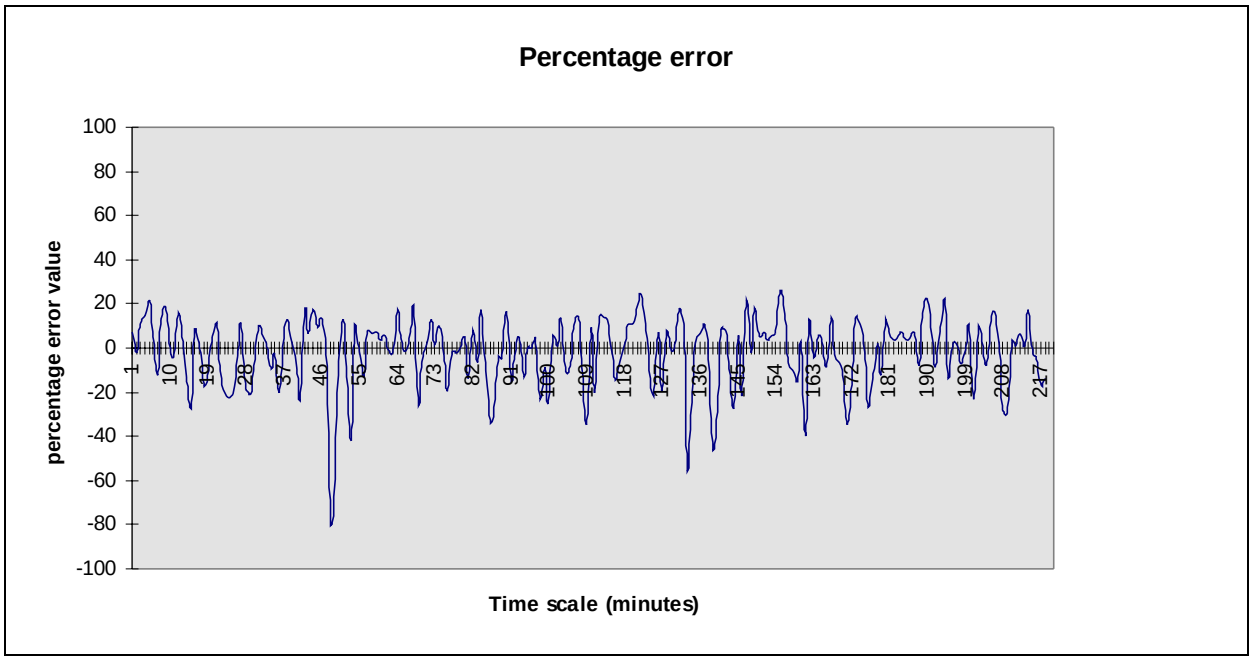
4 minute: Density



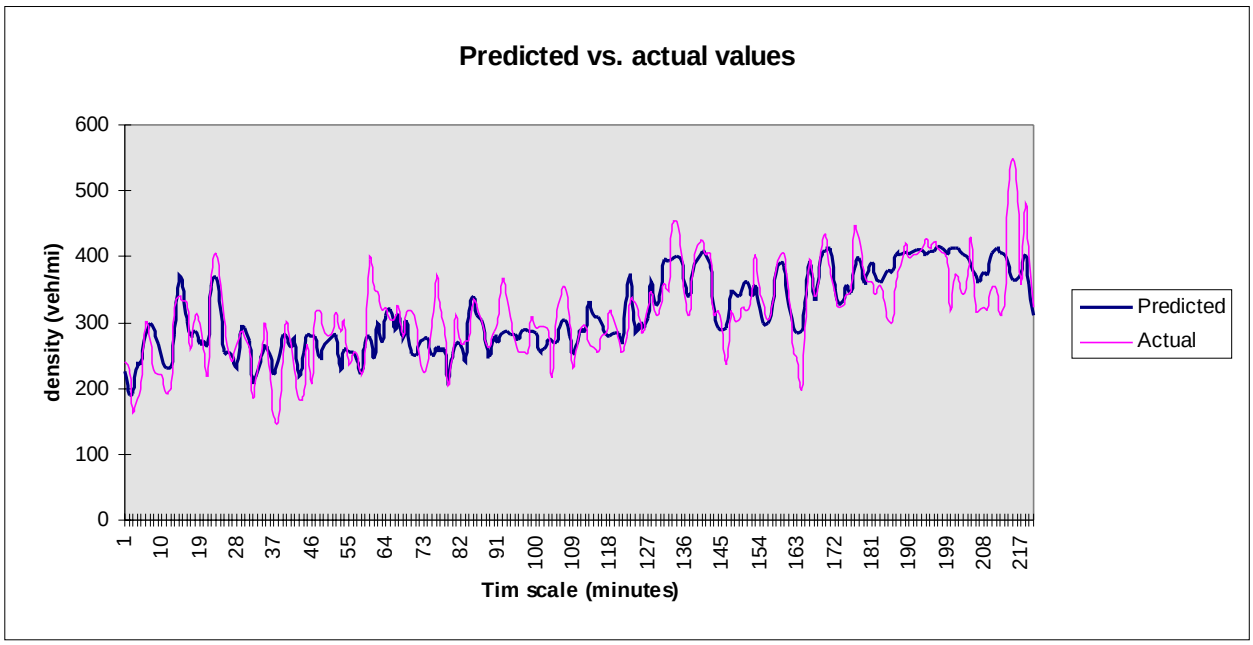


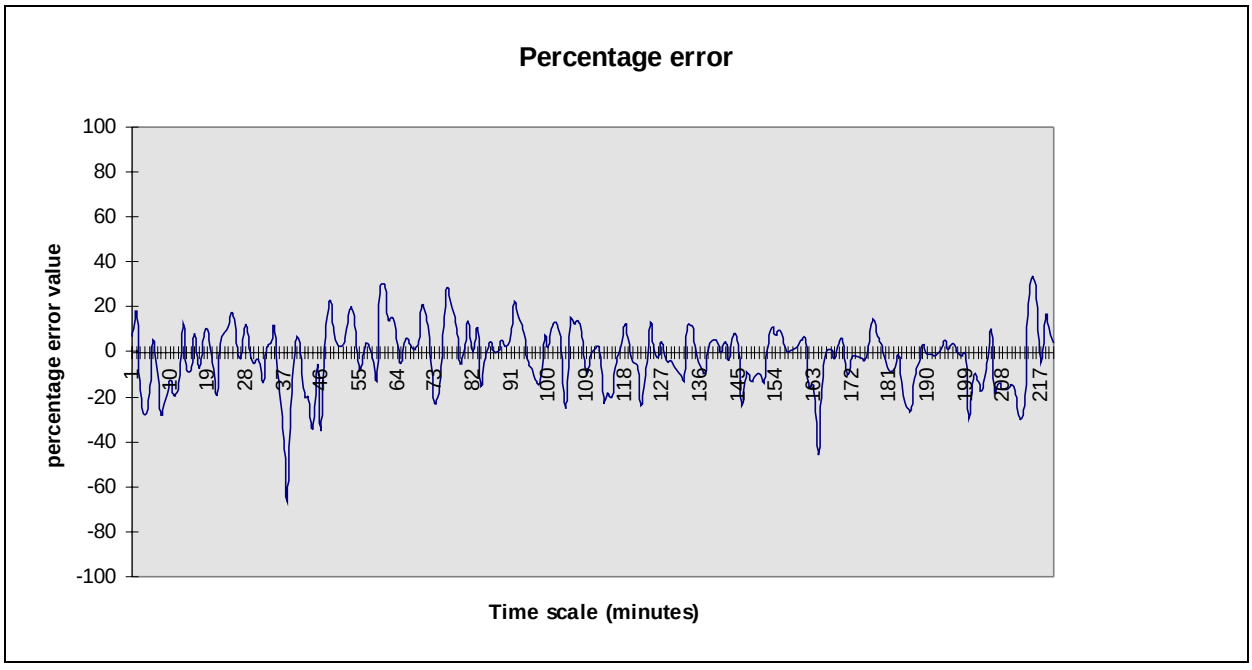
5 minute: Flow



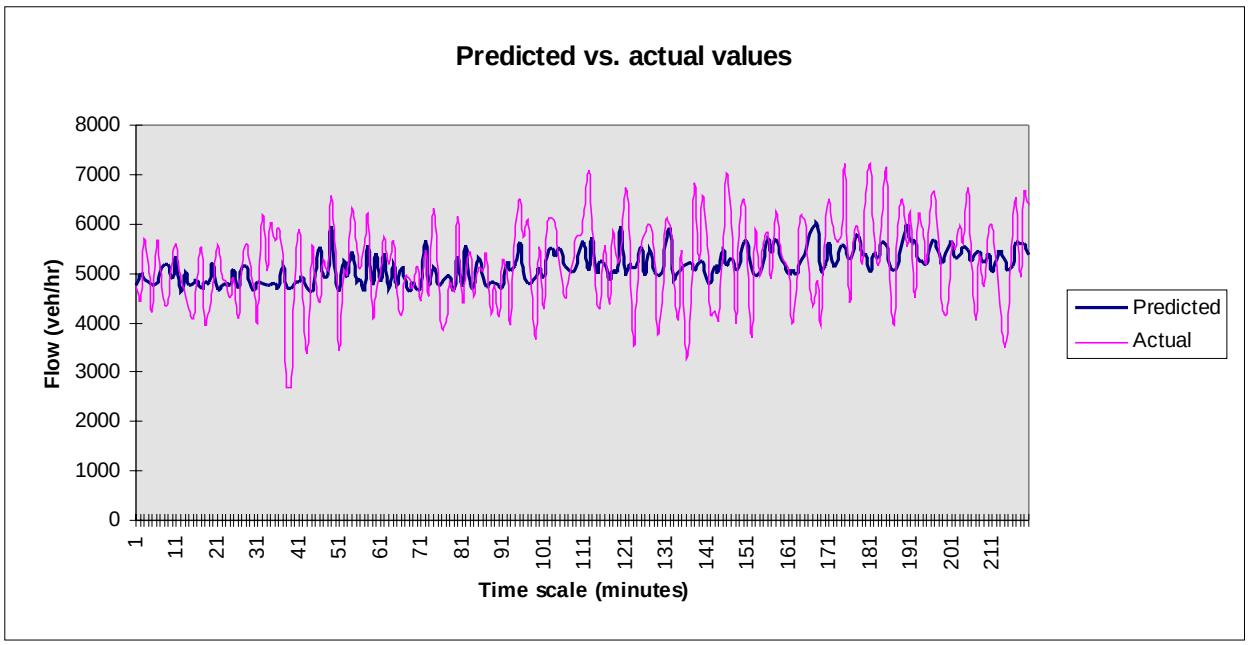


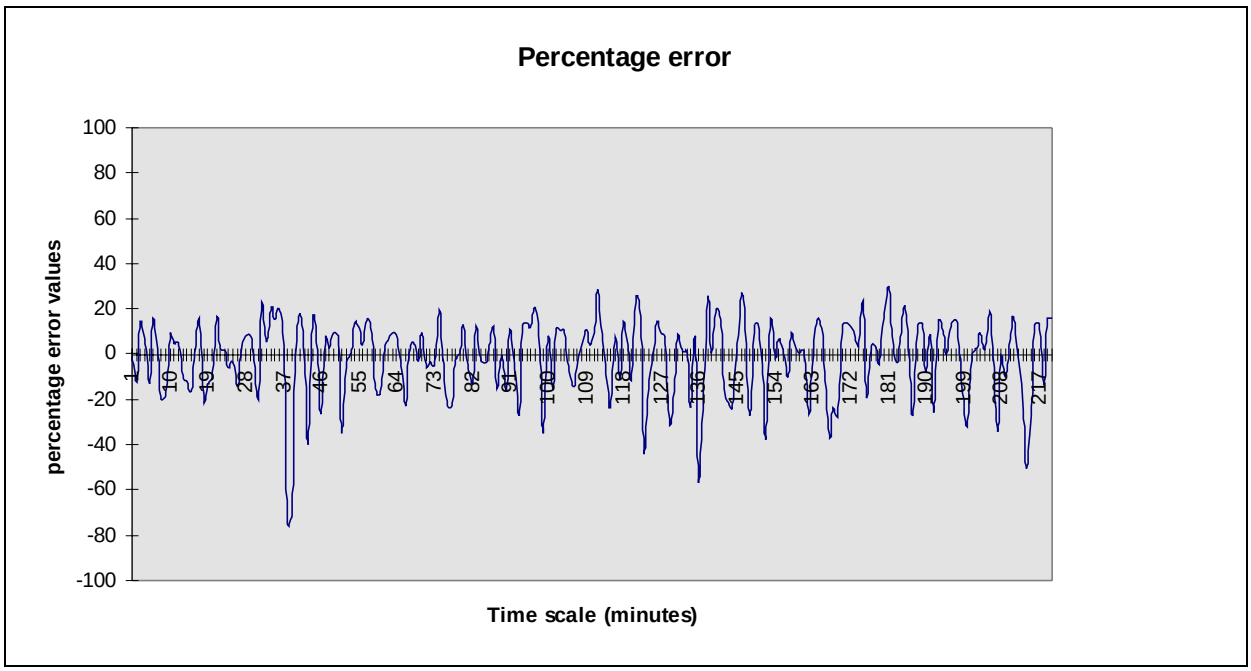
5 minute: Density



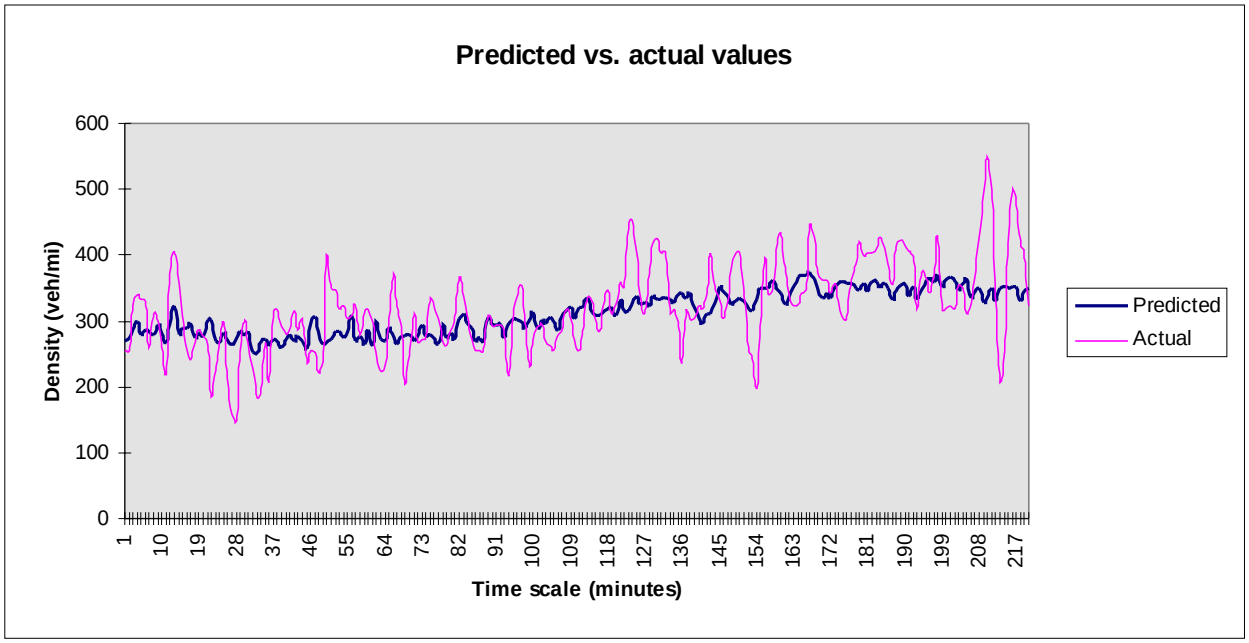


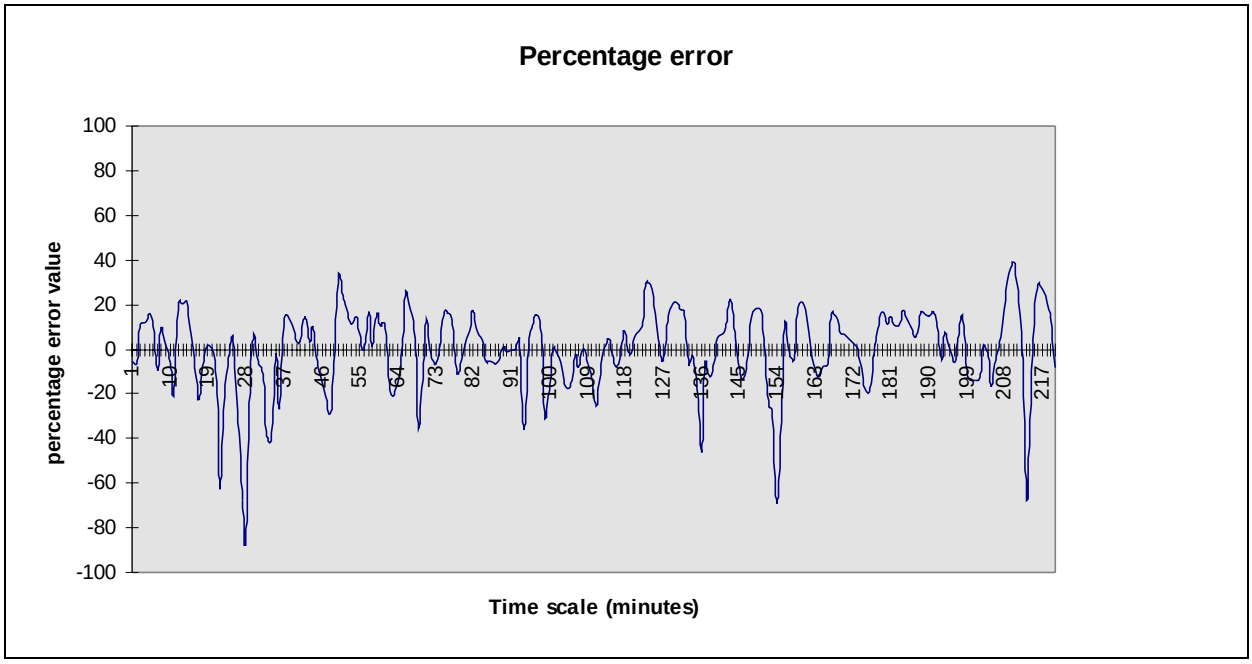
10 minute: Flow



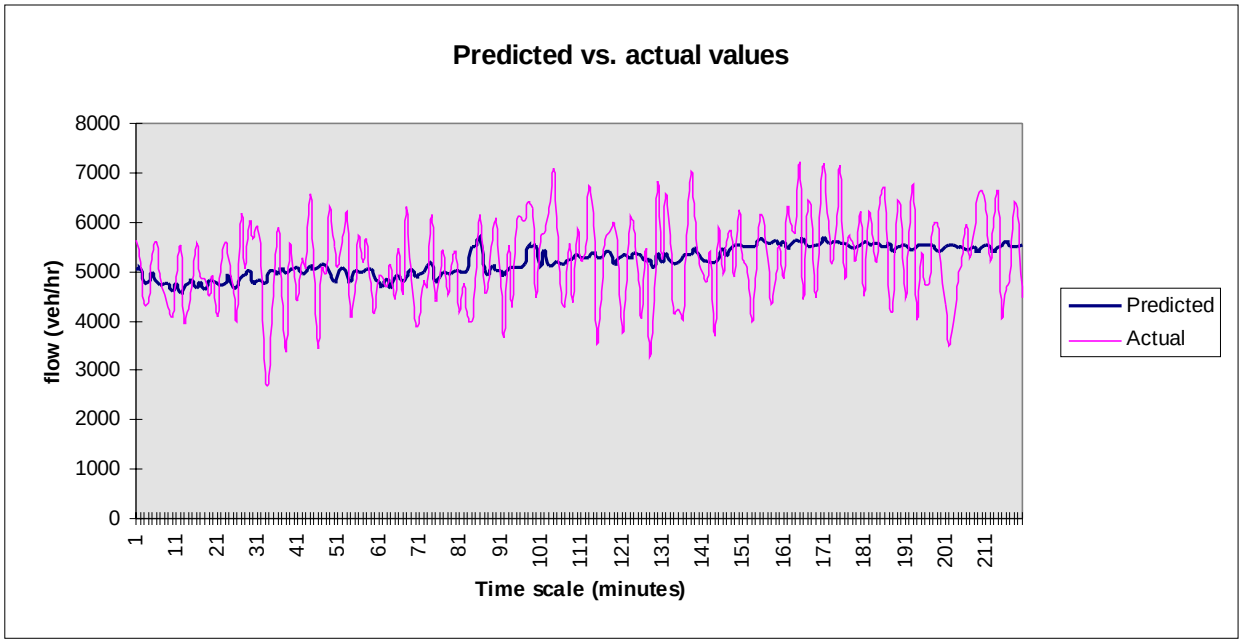


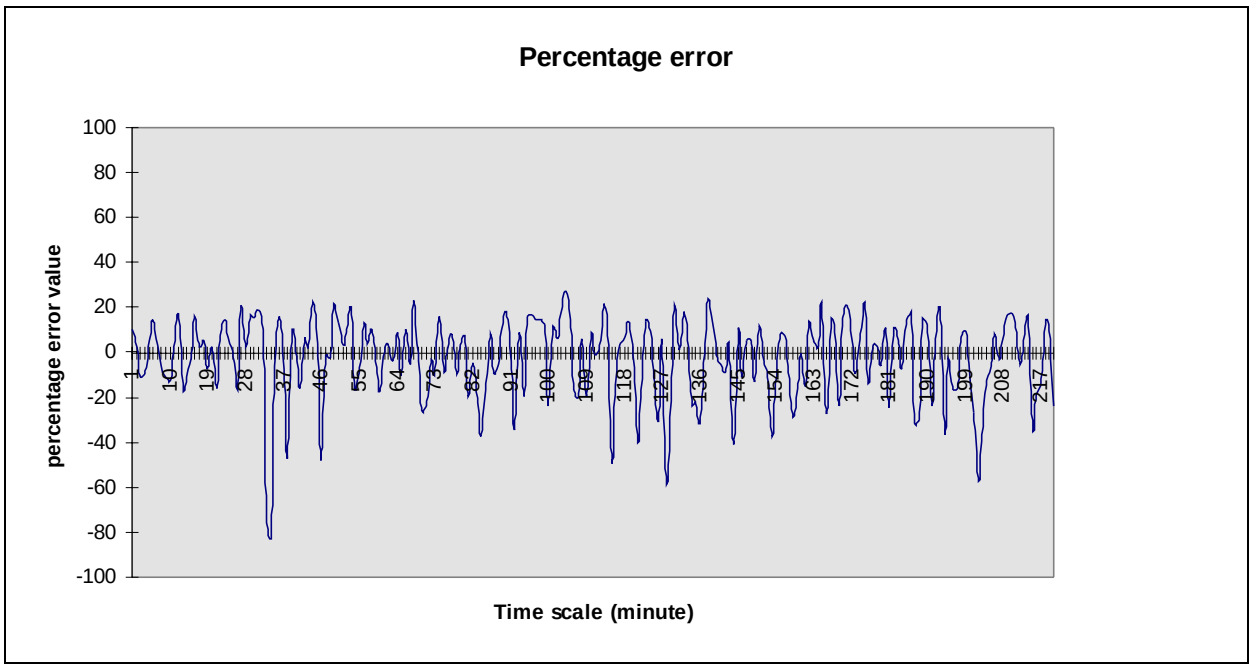
10 minute : Density



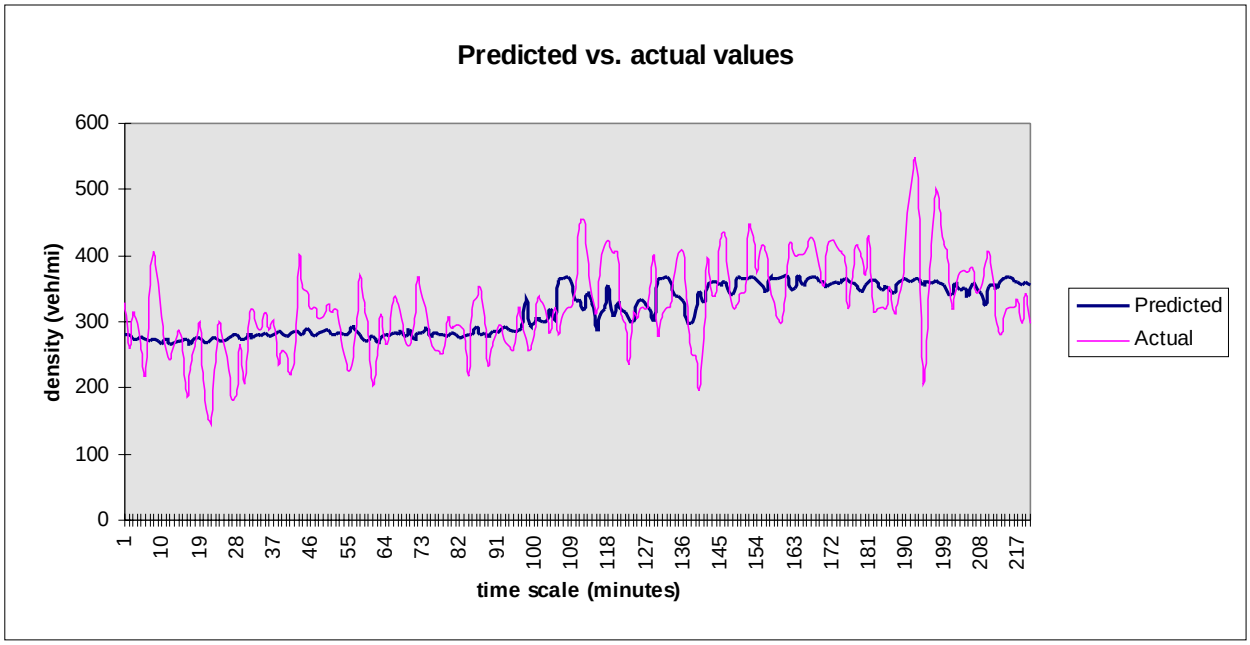


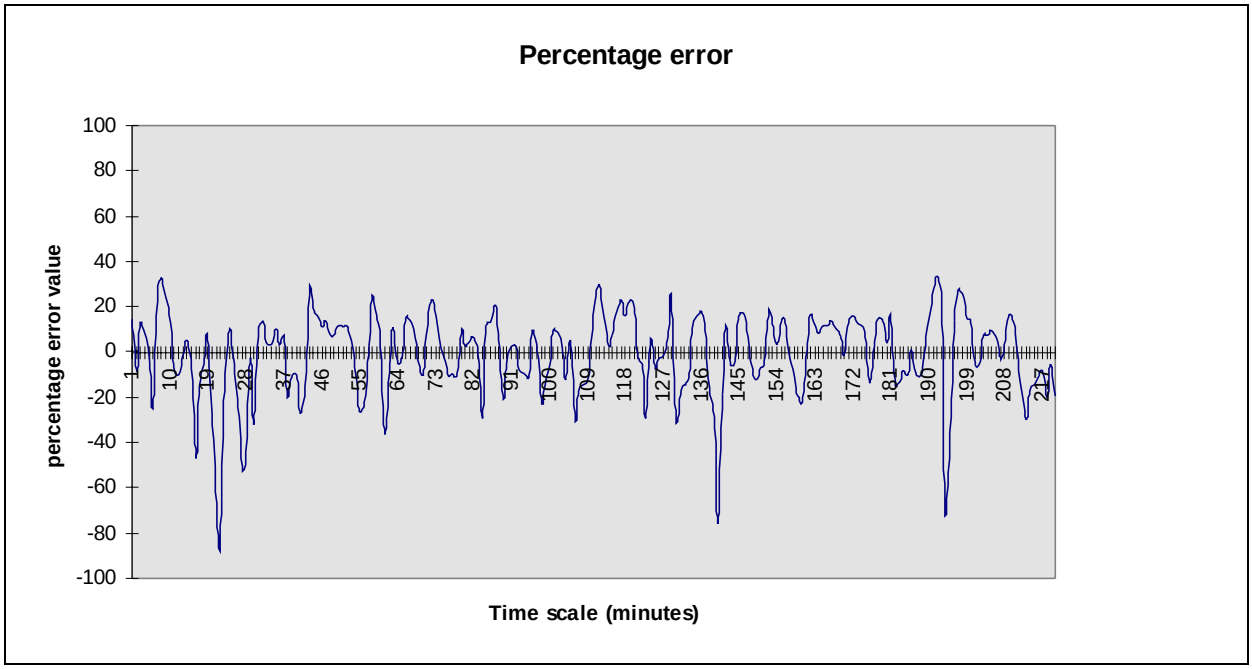
15 minute: Flow





15 minute: Density


Predicted vs. actual values
density (veh/mi)
600
500
400
300
200
100
0
1
10
19
28
37
46
55
64
73
82
91
100
109
118
127
136
145
154
163
172
181
190
199
208
217
time scale (minutes)
Predicted
Actual


Percentage error
percentage error value
100
80
60
40
20
0
-20
-40
-60
-80
-100
1
10
19
28
37
46
55
64
73
82
91
100
109
118
127
136
145
154
163
172
181
190
199
208
217
Time scale (minutes)