**ID197**

# NON-LINEAR ALGORITHM ANALYSIS OF LARGE SCALE TRAFFIC DATA FOR SYSTEMATIC PATTERN DETECTION

N. Idris[1], K. Ambak[2], M. E. Sanik[3] and N. Arbaiy[4]

[1,2,3]Department of Infrastructure and Geomatic Engineering, Faculty of Civil And Environmental Engineering (FKAAS), University Tun Hussein Onn Malaysia (UTHM)

*E-mail:*[1]idmah@gmail.com,[2]kamardin@uthm.edu.my, [3] erwans@uthm.edu.my

[4]Department Software Engineering, Fakulty of Computer Science and Information Technology (FSKTM), University Tun Hussein Onn Malaysia (UTHM)

*E-mail:* [3]nureize@uthm.edu.my

Large-scale sensor data of are well known to transportation researchers, but have resisted systematic analysis due to the significant challenges of dealing with missing data. Missing traffic data is a common problem in road traffic management when electronic devices do not work during some part of daily time. When some of traffic data has been lost, the continuity of traffic data cannot been achieved, so analyzing data will be face of lacking data. In this study, an artificial neural network (ANN) model has been choose to estimate the data, which have been lost in the Federal Highway Kuala Lumpur when Automatic Incident Detection (AID) devices had not been working during in the parts of daily time. Volume, vehicles speed, occupancy and vehicles gap are the major parameters will be consider for modeling. The present model will be evaluate with ANN. The NeuroXL predictor software will be use in this study. The prediction data will be compare with actual data for model performance comparison. Expected result from this study will show that ANN model successfully predict missing data based on some independent variables, which can be gather more easily.

*Keywords: large scale data; artificial neural network; missing data; automatic incident detection; prediction.*

## Introduction

Large-scale sensor instrumentation is now common in a variety of applications including environmental monitoring, industrial automation, surveillance and security. Real-world sensor time series are often significantly noisier and more difficult to work with than the relatively clean data sets that tent to be used as the basis for experiments in many researches (Ihler, Hutchins, and Smyth., 2008).Large-scale sensor data of are well known to transportation researchers, but have resisted systematic analysis due to the significant challenges of dealing with noisy real-world sensor data at this scale. Bickel *et al.* (2007) outline some of the difficulties in a surveypaper; bad and missing samples present problems for any algorithm that uses the data for analysis. Therefore, the need to detect the missing data is important.

The data of traffic volume and their continuity are very important in management and traffic planning, so they are known as major concern for people who are working or responsible in traffic management. Some electronic devices such as radar detection systems and conductive loops mostly store traffic data, so it is a common problem when electronic devices fault and some of data are being lost.

According to the Alberta Transportation Department, over seven years, more than half of total counts have missing values (Zhong et al., 2003). During some years the percentage is as high as 70-90%. A year data from Minnesota Department of Transportation (MnDot) show more than 40% counts having missing values.For the traffic counts with missing values, highway agencies usually either retake the counts orestimate the missing values. Estimating missing values is known as data imputation. Since sometimes retaking counts was impossible due to limited resources and time, imputing the data became a popular method (Albright, 1991). For example, it was reported that many highway agencies in the United States estimated missing values for their traffic counts (New Mexico, 2000). In Europe, highway authorities in Netherlands, France, and the United Kingdom all used some computer programs for data validation routines. Usually missing or invalid data was replaced with historical data from the same site during the same period (FHWA, 2007).

Expert need estimate the lost data by using statistical method such as time series and regression model to repair data use for their analysis. It's common obsession that an accurate model with high performance of result could be used in prediction of data particularly when they are dependent on the other factors, which are simply to be gathered. On the other hand, during the recent years, artificial neural networks have been widely used for estimating traffic parameters and it can be conclude that Artificial Neural Network (ANN) technique may be used to estimating missing data based on some independent variables, which can be gather more easily(Mahmoudabadi and Fakharian,2010). Vlahogianni et al., 2005 state that although many different methodologies have been used for short-term prediction, literature suggests neural networks as one of the best alternative for modeling and predicting traffic parameters. Dia, 2001 develop an object-oriented neural network model for predicting short-term traffic conditions on a section of the pacific highway between Brisbane and the Gold Coast in Queensland, Australia. Mahmoudabadi (2010) develop and artificial neural network model to estimate the frequency of rural road accidents in two separate freeways in Iran and showed that ANN model can estimate road traffic safety measures based on the effective parameter, with high performance of mean square errors comparing with simple and weighted linear regression models.

**Prediction Models**

There are various kinds of models that are used for estimating parameters in road safety. Regression models are known as common, because of simplicity. Recently, artificial neural network models have been used widely as prediction models of accidents. In this section a brief discuss of ANN. An ANN is statistical model comprised of simple, interconnected processing elements that are configured through iterative exposure to sample data. ANN was originally developed as mathematical theories of the information processing activity of biological nerve cells ( Dash, Kajiji and Forman, 2010).

The neural networks used in this study consist of three layers: input, hidden, and output. Theinput layer receives data from the outside world. The input layer neurons send information to thehidden layer neurons. The hidden neurons are all the neurons between the input and output layers. They are part of the internal abstract pattern, which represents the neural networks solution to the problem. The hidden layer neurons feed their output to the output layer neurons, which provide the neural networks response to the input data. For example, the network in Figure1.0receives a single input from the external environment. The remaining nodes in the input layer get their input from the neuron on the left by one time interval. The input layer at any time will hold a part of the time series. Neurons process input and produce output. Each neuron takes in the output from many other neurons. Actual output from a neuron is calculated using a transfer function.

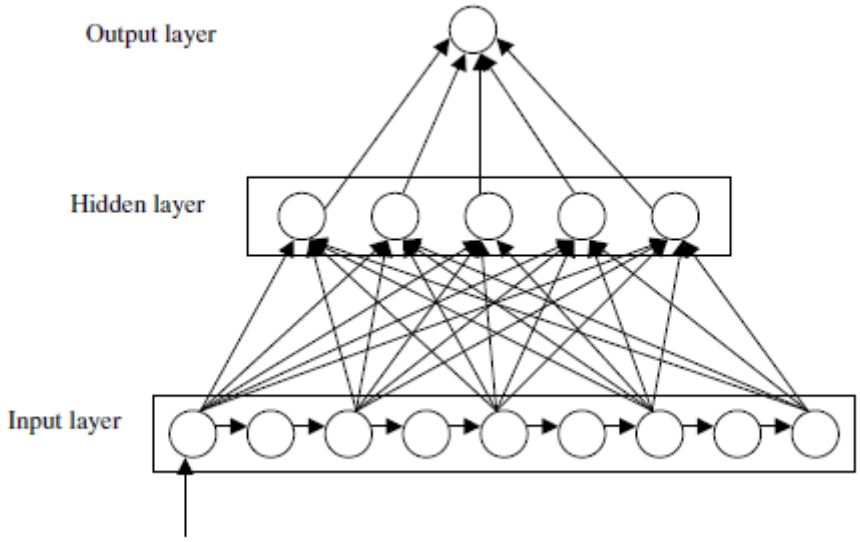


**Figure 1: Flow from input to output from Zhong, M., Ligras, P., Sharma, S. (2004).**

In this study, a sigmoid transfer function is chosen because it produces a continuous value in the range [0, 1]. A neuron in a given layer is connected to neurons
(n1,n2, . . .,nm) in the previous layer. The connectionfrom nj to ni has the weight wji. The weights of the connections are initially assigned an arbitraryvalue between 0 and 1. The appropriate weights are determined during the training phase.Input to the ni is obtained using the following equation:

$$\text{input}_i = \sum_{j=1}^{n} w_{ji} \times \text{output}_j \qquad (1)$$

Output fromthe $n_i$ is calculated using a sigmoid transfer function as:

$$\text{output}_i = f(\text{input}_i) = 1/(1+e^{\text{gain} \times \text{input}_i}) \qquad (2)$$

Today, neural network really useful to organizational research. Scarborough and Somers (2002), the most significant departure of neural network analysis from conventional analysis is that neural model development is relatively unconstrained by researcher expectations

compared with the defined parameters of anticipated functional relationships inherent of hypothesis testing. Neural network analysis does not required or yield individual hypothesis confirmation. A trained neural network's output and structure are used to make inferences about associations, interaction, nonlinearities, and other characteristics of the data. If such inferences are accurate, they can be replicated across multiple networks and samples and confirmed using conventional procedures. The important points is, ANNs can help uncover structural elements in research data that we may not have known of or thought to look for.

**Study Area**

This study take several location at Federal Highway in Kuala Lumpur as the research area. Transportation Management Centre in Bukit Jalil maintains an extensive network of the Automatic Incident Detection Device (AID) on Federal Highway in Kuala Lumpur and Klang Valley. AID recovery for wide area of traffic network that has two elements: roadway surveillance system that provides the real-time traffic data for detection and incident detection algorithm that interprets the data and determines the presences of a capacity-reducing incident. Every 3 minutes each of these traffic sensors reports a count of the number of vehicles that passed over the sensor as well as average speed and several other parameters. The data are continuously archived providing a potentially rich source from which to extract information about urban transportation patterns, traffic flow, accidents and human behavior in general. Same with other electronic devices, during some parts of daily time AID do not work effectively. Raw data from AID sometime has missing value.Large-scale data from AID are very useful for the transportation researchers especially for the Federal Highway Kuala Lumpur as a main route. To solve this problem, this study describes the application of probabilistic modeling to this data set and illustrates how these approaches can successfully detect underlying or fundamental systematic pattern even in the presence of substantial noise and missing data.

**Methodology in NeuroXL Software**

An artificial Neural Network (ANN) is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. In more practical terms, neural networks are non-linear statistical data modeling tools. They can be use to model complex relationships between inputs and outputs or to find patterns in data.

NeuroXL is one of the software that related to the ANN which will be use in this study. Advantages of NeuroXL Predictor it is easy to learn and use, no prior knowledge of neural networks required, integrates seamlessly with MicrosoftExcel, provides proven neural technology for highly accurate forecast and lowest cost neural production the market.

**i. Define inputs and outputs detail**

Variable has been chosen to be used as network inputs and outputs, and to either specify or compute the minimum and maximum value for each value for each variable. Variables are included as columns and patterns are rows in the NeuroXL internal file format.

### ii. Test set and production Set

Test set and production set of data from the training patterns will be extract. The test set will be used with calibration which prevents overtraining networks, so they will generalize well on new data. Calibration is use with back propagation networks. The production set may be use to test the network's results with the network data that never see before.

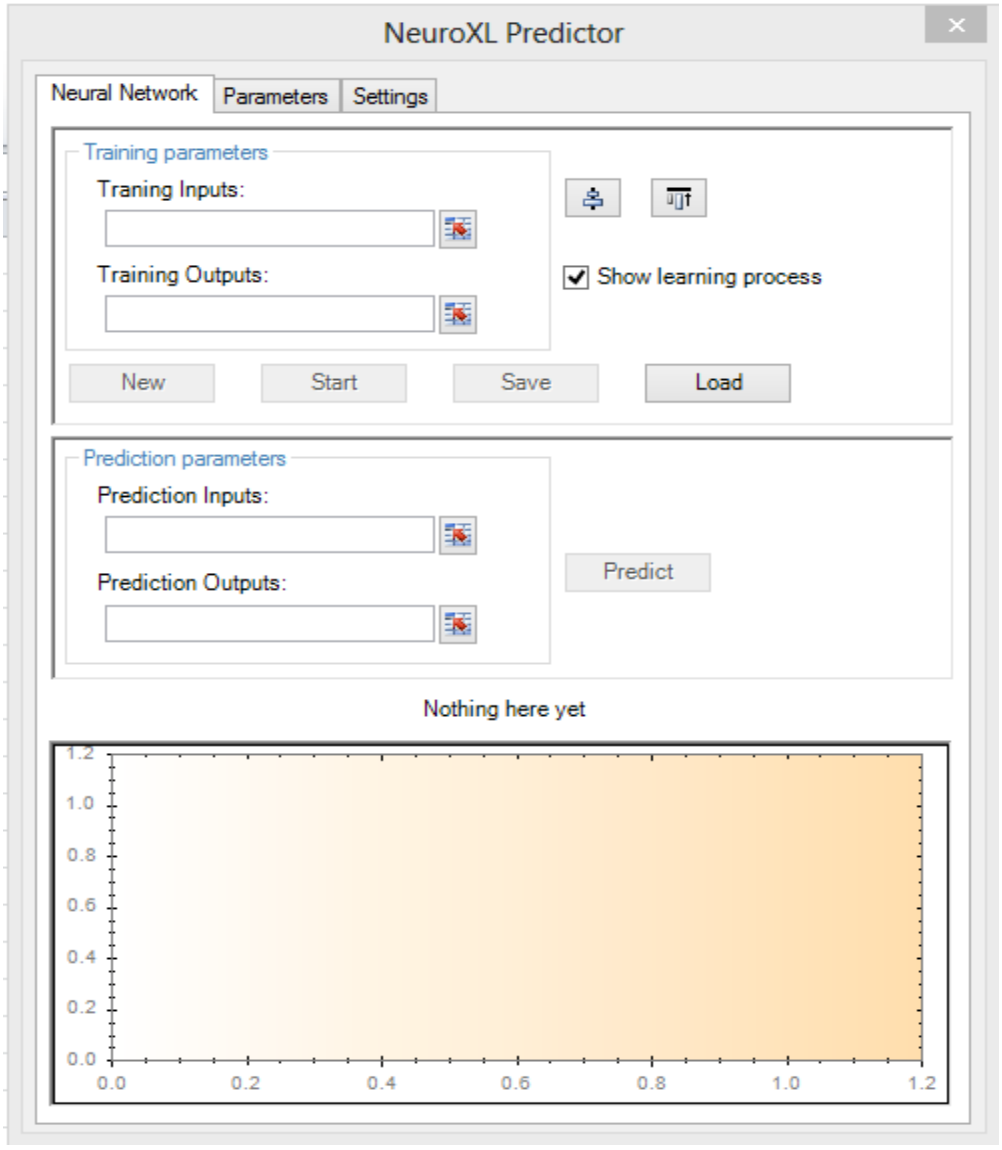


**Figure 2: Training and prediction set**

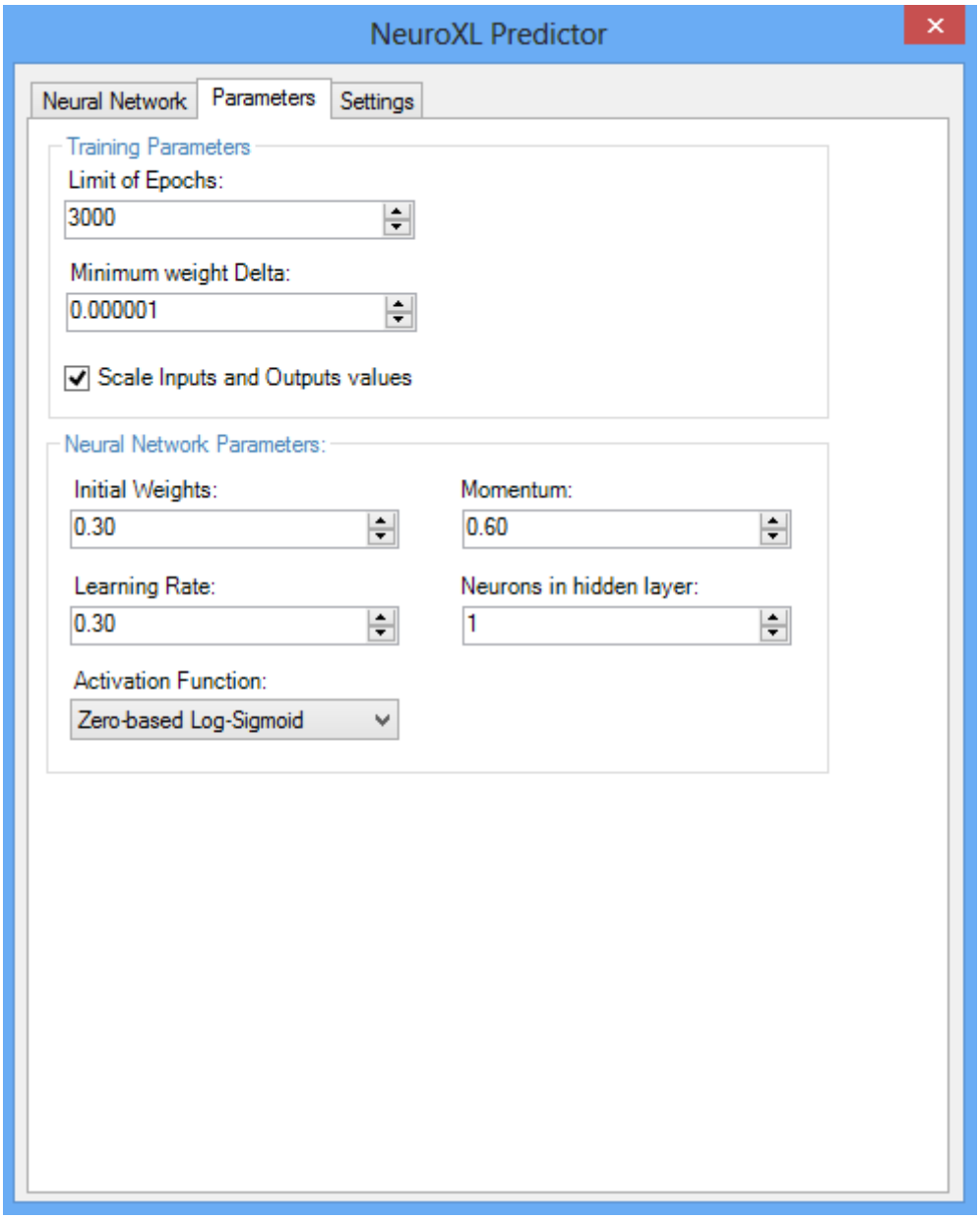


**Figure 3: Parameters set**

**iii. Network architecture determination**

The objective of this part is to find a suitable neural network topology to predict on accident. Back propagation networks are known for their ability to generalize well on a wide variety of problems. That is why they are used for the vast majority of working neural network applications. Depending upon the number of patterns, training may be slow but worth it because they are such robust and global algorithms (although if you use calibration training is usually much shorter). When using back propagation networks, the precision of the network can be improve by creating a separate network for each output if the outputs are not categories. ANN offers several different variations of back propagation networks.

**iv. Neural network structure**

Three slabs with different activation function have been chosen as the structure of ANN. This ward network is a back propagation network that adds

a third slab to the hidden layer. When each slab in the hidden layer has a different function applied to hidden layer slabs detect different features in a pattern processed through a network.

**v. Training and stop training the network criteria**

Once a network has been structure for a particular application, that network is ready to be trained. To start this process, the initial weights are chosen randomly. Then the training or learning begins.Several different parameters have to be set, which to determine how the training to proceed and under what conditions the training process need to be terminated. There are different criteria depending upon the chosen network type: To stop training a network, it is recommended that author train until the events since the minimum error factor is greater than 20,000 to 40,000 events (higher for recurrent nets). Don't set any other stopping criteria.Hidden layers in a neural network are known as feature detectors. Different activation function applied to hidden layer slabs detect different features in a pattern processed through a network. Thus, the output layer will get different views of the data.

**vi. Testing the networks**

If a network is well trained, then it will also test well. This means that when presented with inputs that it has never seen, (ones that were not used to train) it will still perform within the define percentage of the margin error. If the network does not test well, then there may be too many connections in the systems, which caused the network to memorize each of the training facts as separate items. In this case, there are two options:

a. Removing hidden neurons from the system will reduce the number connections, which will help the network to train better. This requires the network to be rerandomized.
b. If there are not enough examples of each pattern to classify, the network might not be able to find differences to test between pattern to test well. Collecting additional training facts may also help the network to test better. These additional facts can be added to the original one and training can continue without starting.
c. When the network test within the guideline of the project definition, the the network training is complete.

**Conclusion**

It is important to conduct this study to predict all missing data as many as possible. Artificial Neural Networks (ANN) model is usually used for prediction cases. By using ANN model, these missing data can be determined by collecting the input data. The input data can be process by ANN applied software to get the predict result for the forecasting purpose for the highway. This ANN applied software is also easy and ready to use for any level of users which they can implement or analyze all the parameters and accident data for the future prediction. ANN will be the analyzing tool which it is one of the highly performance tool in developing prediction model.

**Acknowledgement**

Thanks to the Ministry of Higher Education for giving support and sponsor to this research through the Fundamental Research Grant Scheme.

**References**


Bickel, P., Chen, C., Kwon, J., Rice, J., Van Zwet, E., &Varaiya, P. (2007). Measuring traffic statistical Science.

Albright, G. (1991). History of estimating and evaluating annual traffic volume statistic.

Dia, H. (2001). An object-oriented neural networks approach to short-term traffic forecasting.European Journal of Operational Research.

Dash, G. H., Kajiji Nina.& Forman, J. (2010).Combining artificial neural networks with high-frequency real-time data to achieve efficient trading opportunities.

FHWA's Scanning Program. (2007). FHWA study tour for European traffic monitoring programs & technologies.Federal Highway Administration, U.S Department of Transportation, Washingtn, DC.

Hutchins, J., Ihler, A., Smyth, P. (2008). Probabilistic analysis of large-scale urban traffic sensor data set. University of California, Irvine.

Mahmoudabadi, A. (2010). Comparison of weighted and simple linear regression and artificial neural network models in freeways accident prediction (Case study : Qom &Qazuin freeways in Iran).

New Maxico State Highway & Transportation Department.(2000). Survey of traffic monitoring practices among state transportation agencies of the United States.

Scarborough, D., & Somers, M. J. (2002). Neural networks in organizational research : Applying pattern recognition to the analysis of organizational behavior.

Zhong, M., Ligras, P., Sharma, S. (2003). Estimation of missing counts using factors, genetic, neural, and regression techniques. University of Regina, S K, Canada.