Predicting Fine Particulate Concentrations near a Busy Intersection in Sydney using Artificial Neural Networks

Tharit Issarayangyun and Stephen Greaves
University of Sydney, Sydney, NSW, Australia

1 Introduction

In July, 2007, medical researchers at the University of California at Los Angeles uncovered for the first time a direct genetic link between exposure to vehicle exhaust pollutants and arteriosclerosis, a primary precursor of cardiovascular diseases (Gong et al., 2007). This finding adds to a growing body of scientific evidence suggesting that the characteristics of roadway microenvironments coupled with when and for how long people are in these microenvironments could be particularly relevant in terms of overall health impacts (Burnett, 2003). While pollution concentrations in roadway microenvironments have been shown to be consistently higher than ambient levels on which air quality standards are assessed, there is evidently great variability across space and time due to a range of meteorological, traffic, modal, and personal factors (Kaur et al., 2007). The implications are that to deepen our understanding of exposure, we must monitor and predict pollution concentrations at increasingly disaggregate levels of temporal and spatial resolution (Greaves, 2006).

While methods to measure air pollution have become increasingly refined, prediction remains a challenge despite the development of sophisticated vehicular exhaust dispersion models. This is largely due to the complexity, non-linearity and unknown distributional qualities of air pollution data (Zamurs and Conway, 1991). In response, there is growing interest in using data-driven machine-learning techniques, such as Artificial Neural Networks (ANN) to model air quality data (Perez, 2000). The appeal of ANNs is that they are capable of modelling highly non-linear functions and can be trained to accurately generalise from a new independent data set. ANNs are also good at detecting the underlying pattern masked by noisy factors in a complex, highly disaggregate, system (Zhang et al., 1998).

Despite the potential, ANN-based approaches have largely been applied to the problem of predicting regional or city-wide pollution (Perez et al., 2000, Grivas and Chaloulakou, 2006). Relatively few applications have focused on roadside exposures (Moseholm et al., 1996; Nagendra and Khare, 2004). The current paper reports on the development and application of ANN-based methods to address the problem of temporally disaggregate-level prediction of $PM_{2.5}$[1] near a busy intersection in Sydney, Australia. Following details of the data collection required, the paper explains the rationale for the ANN structure used for this application. We then apply the ANN and compare to other modelling approaches before drawing conclusions on the merits of the approach.

2 Methodology

2.1 Study Area and Data

To develop and test the approach, $PM_{2.5}$ concentration levels were collected on a minute-by-minute basis over two weeks (25/05/2007 – 06/06/2007) at the intersection of Military Road

[1] *$PM_{2.5}$ refers to particulate matter with an aerodynamic diameter of less than 2.5 microns. It is associated with an increased risk of cardiopulmonary and lung cancer mortality, reduced lung function, and as a potential trigger for existing respiratory problems such as asthma (Kappos et al., 2004).*

and Wycombe Road in Sydney, Australia (FIGURE 1). Military Road is a major traffic route in Northern Sydney that carries approximately 77,000 vehicles per day. The equipment comprised the AM510 SidePak™ personal aerosol monitor (also shown in FIGURE 1), which uses nephelometric (light-scattering) techniques to estimate $PM_{2.5}$ concentrations (see Greaves, 2006 for more details). The monitor was placed in an apartment overlooking the intersection. Ambient $PM_{2.5}$ concentration levels were collected at the same time using the same model device at a location approximately 300 metres south where effects from primary particulates originating from traffic were believed to be minimal (Zhu, 2002). One minute temperature, wind speed and direction, relative humidity, and mean-sea level pressure were obtained from the closest fixed site stations of the Bureau of Meteorology – note precipitation levels were zero during the data collection period. Fifteen-minute traffic volumes contiguous with the monitoring period were computed from intersection counts – these are automatically collected and stored for the vehicle-actuated signal timing system maintained by the Roads and Traffic Authority (RTA) of New South Wales.

FIGURE 1: The Monitoring Site and Portable Aerosol Monitor

2.2 Data Analysis and Development of ANNs

The data were screened resulting in 1,200 valid-data point. The data were then analysed with ANNs using the 'back propagation' technique with momentum term algorithm. The neural network architecture used was the fully-connected feed forward multi-layer perceptron

(MLP) with one hidden layer. This setup is considered able to approximate almost every measurable function between input and output vectors by selecting a suitable number of neurons, connecting weights and transfer functions (Gardner and Dorling, 1998). NeuroSolution software was used for the analysis.

While readers are referred to texts such as Haykin (1999) for more details on ANNs, there were important considerations for the application detailed here. First was the choice of a suitable transfer function. Previous studies have shown the logistic sigmoid transfer function (Perez et al., 2000) and the hyperbolic sigmoid transfer function (Nagendra and Khare, 2004) are among the most efficient functions in mapping the input and output patterns for atmospheric pollution. This study assessed both transfer functions for the most accurate results. The formula for the proposed transfer functions and the back propagation algorithm are presented below:

(a) Logistic function: $f(u) = \dfrac{1}{1+e^{-u}}$

(b) Hyperbolic tangent: $f(u) = \dfrac{1-e^{-u}}{1+e^{-u}}$

(c) Back propagation algorithm: $w_{ij}(t+1) = w_{ij}(t) + \eta e_j x_i + \mu\left[w_{ij}(t) - w_{ij}(t-1)\right]$

where u is the sum of the adjusted input signals, w_{ij} is the connecting weight between neuron *i* and neuron *j*, x_i is the output from neuron (i), η is the learning rate, μ is the momentum factor to ensure network stability and *t* is a set of input data.

A second consideration was the selection of the optimal number of neurons in the hidden layer. Networks with fewer hidden neurons are preferable since it is easier to generalise but networks with too few hidden neurons have limited power in mapping and predicting data. The number of neurons in the network is directly proportion to the number of weight needed to be estimated. Here we followed an empirical rule suggested by Haykin (1999) to restrict the number of weight needed to be estimated to not more than ten percent of the total number of data point used to train the network. The final procedural issue related to how to split the sample for training, cross-validation[2] and testing. In this case, we decided to use 720 data point (60 percent of the sample) to train the network, 240 data point (20 percent) to cross-validate, and 240 data point (20 percent) to test the network.

The performance of the ANN models were evaluated using the root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2). The better the model performance means the smaller RMSE and MAE and the closer R^2 is to 1. The formulas for these evaluation criteria are presented below:

(d) $RMSE = \sqrt{\dfrac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}}$

[2] *Cross-validation helps prevent overfitting of the data and determines the stopping point of the training process. In this study, the training networks were trained 20 times with a maximum epoch of 3,000 each time. The network will be stopped if the performance of the cross-validating set is not improved after 300 repetitive runs.*

(e) $$MAE = \frac{\sum_{i=1}^{N} |O_i - P_i|}{N}$$

(f) $$R^2 = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (O_i - \overline{O})^2}$$

where N is the total number of observations, O_i is the observed value, P_i is the predicted value and $\overline{O}$ is the mean of the observed value.

3 Results and Discussion

Initially the minute-by-minute data were used as this captures the most information. However, the traffic data were available at 15-minute intervals necessitating use of this averaging interval. FIGURE 2 shows a time-series plots of the roadside (i.e., collected at the intersection) $PM_{2.5}$ concentration, ambient $PM_{2.5}$ concentration and the traffic volume on Military Road over the two week sampling period. The plot shows the roadside $PM_{2.5}$ concentrations (R_$PM_{2.5}$) generally track the ambient $PM_{2.5}$ concentrations (A_$PM_{2.5}$) at a higher level, with more pronounced differences observed during the peak traffic periods. The statistics in TABLE 1 confirm this with the average for the roadside $PM_{2.5}$ (21.09 $\mu g/m^3$) being 45% higher than the ambient average (14.54 $\mu g/m^3$).

Looking in more detail at the results both the average ambient and roadside $PM_{2.5}$ concentrations were below the proposed Australian standard of 25 $\mu g/m^3$ (Environmental Protection and Heritage Council, 2007). However, this hides the fact that for 385 out of 1,200 data points (around one-third of the monitoring time), levels exceeded this value. Four of the days had substantial periods of time above 50 $\mu g/m^3$ and the maximum observed concentration was 72.67 $\mu g/m^3$.

A natural logarithm transformation was applied to the $PM_{2.5}$ concentrations and the traffic variables to reduce fluctuations in data. Several experiments were performed to determine the best combination of network parameters. A fully-connected feed-forward network, with seven neurons in the input layer, night neurons in the single hidden layer and one neuron in the output layer, using the hyperbolic tangent as a transfer function yielded the best prediction on the test data set. The total number of weight needed to be estimated were 72 which was sufficient achieved by the training data set. The architecture of the network with input labels is shown in

FIGURE 3.

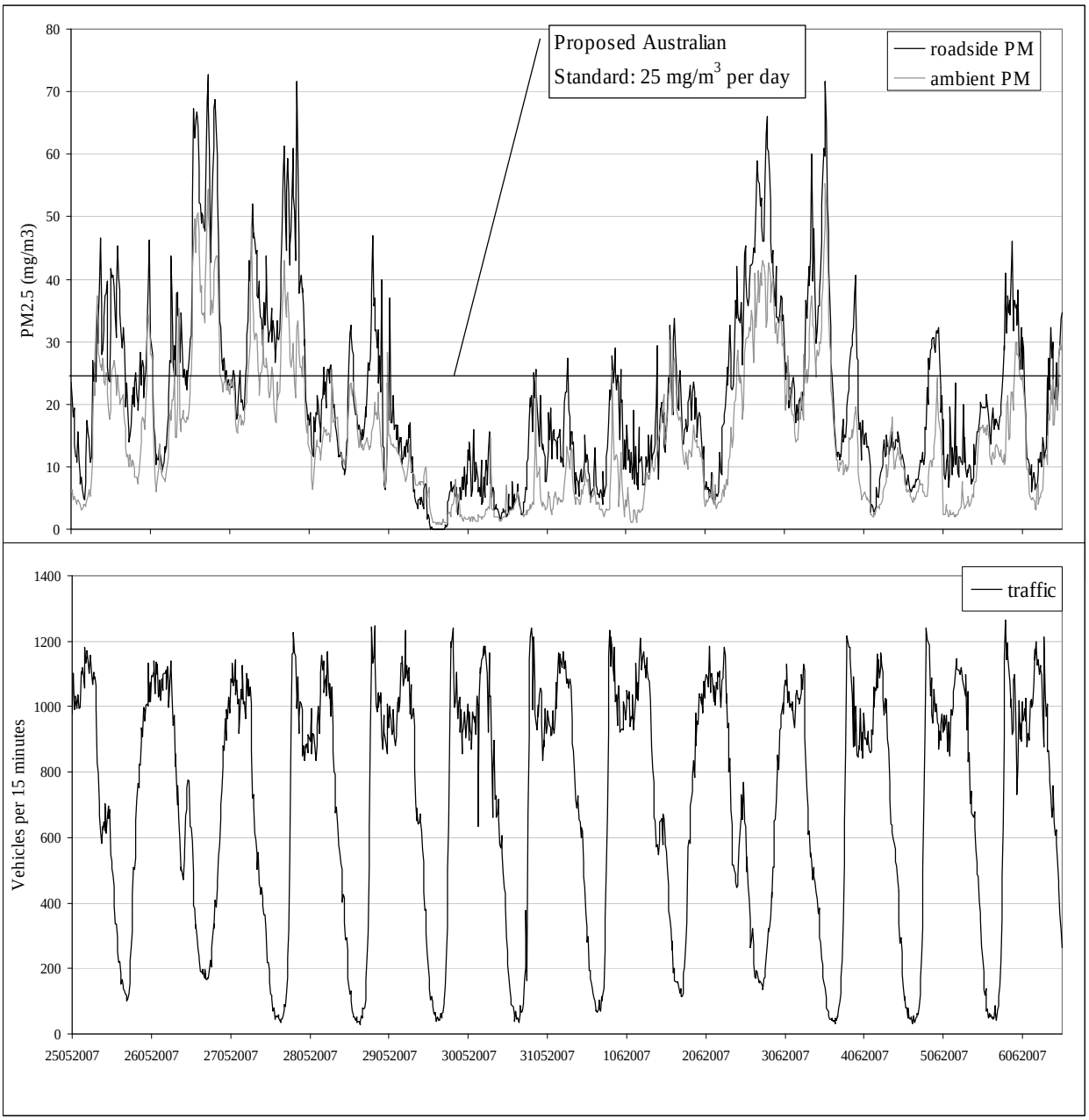

FIGURE 2: Time-Series Plot of $PM_{2.5}$ and Traffic Volume over the Sampling Period

TABLE 1: Descriptive Statistics for the Two Week Monitoring Period

	Mean	**Std. Deviation**	**Min.**	**Max.**
Roadside $PM_{2.5}$ ($\mu g/m^3$)	21.09	14.31	0.00	72.67
Ambient $PM_{2.5}$ ($\mu g/m^3$)	14.54	10.99	0.67	55.33
Traffic (vehicles/15min)	698.89	380.75	28.00	1266.00
Temperature (Celsius)	14.71	3.52	7.81	23.93
Relative Humidity (%)	68.73	15.54	24.93	94.37
MSL Pressure (hPa)	1023.42	2.89	1015.22	1029.30
Wind Speed (km/hr)	13.06	6.36	0.00	33.60
	Median	**Mode**	**Std. Deviation**	
Wind Direction (degree)	273	281	90.40	

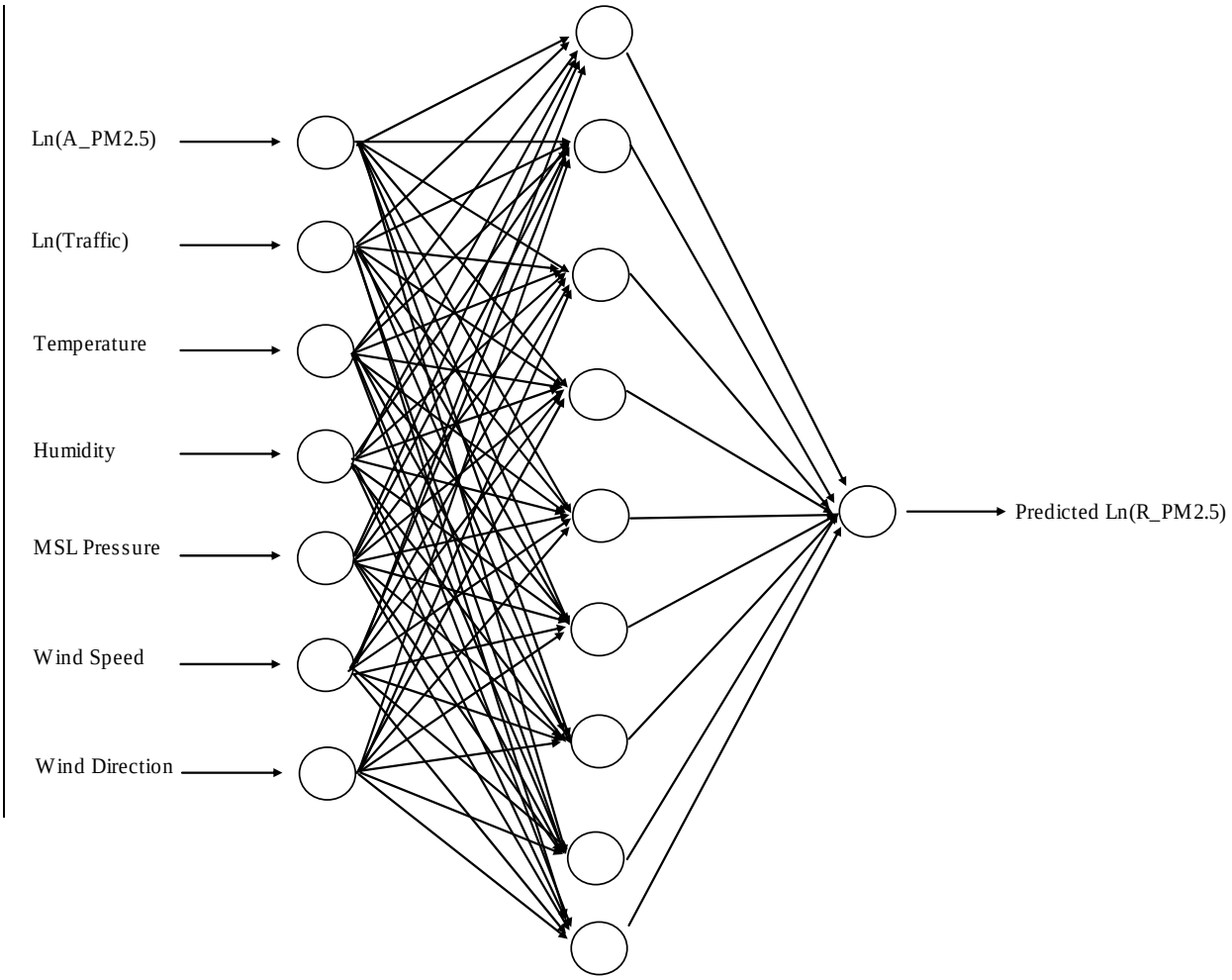

FIGURE 3: Structure of 7:9:1 ANN based roadside $PM_{2.5}$ model

TABLE 2 compares the performance of the ANN with the multivariate ARIMA model on the training data set and the testing data set. The multivariate ARIMA model considers the effects of past values of $PM_{2.5}$ levels (autocorrelation) in addition to the input variables on predicted levels of $PM_{2.5}$ and has been shown to be particularly useful for drawing insights from the type of data with which we are dealing here (Issarayangyun and Greaves, 2007). Note that for the comparison presented here, time-lag data (i.e., previous $PM_{2.5}$, traffic volumes etc) were used for the development of the ARIMA model. This may be practical during the development of the model but is unlikely to be available in a real-time sense for prediction.

During training, the multivariate ARIMA (RMSE = 0.196 $\mu g/m^3$, MAE = 0.145 $\mu g/m^3$) performed slightly better than the ANN (RMSE = 0.213 $\mu g/m^3$, MAE = 0.157 $\mu g/m^3$). Both techniques explained approximately 95 percent of the total variation in the training set. After training, both techniques were then presented with the test data. The trained ANN (RMSE = 0.282 $\mu g/m^3$, MAE = 0.207 $\mu g/m^3$), however, outperformed the multivariate ARIMA (RMSE = 0.308 $\mu g/m^3$, MAE = 0.245 $\mu g/m^3$). Without significantly losing its ability to predict 15-minute $PM_{2.5}$ concentrations (which reflects the absence of an overfitting problem), the trained ANN explained 71 percent of the total variation in the test set while the multivariate ARIMA only explained 65 percent of the total variation.

TABLE 2: Model Comparison

Dependent Variable: Ln(R_PM2.5) ($\mu g/m^3$)				
Technique	**Data Set**	**Performance Index**		
		RMSE	**MAE**	**R^2**
Neural Network – MLP (7:9:1)	Training	0.213	0.157	0.94
	Testing	0.282	0.207	0.71
Multivariate ARIMA[1]	Training	0.196	0.145	0.95
	Testing	0.308	0.245	0.65

Note: 1. The significant multivariate ARIMA model was ARIMA (1, 1, 13) with Ln(A_PM2.5), Ln(traffic), Pressure and wind speed as statistically significant input variables. The study employed TSMODEL_EM with automatic outliers detection incorporated in SPSS version 14 to do the analysis.

Sensitivity analysis about the mean was performed on the trained network to gain insight into the correlation and the relative importance among the input variables to the output variable. Each input was varied between its mean ($\pm$ one standard deviation) while all other inputs were fixed at their respective means. The sensitivity plots of each input are shown in FIGURE 4. The plots confirm the ambient $PM_{2.5}$ concentration was the most important factor in predicting the roadside $PM_{2.5}$ concentration. The plots show the roadside $PM_{2.5}$ concentration decreased when either wind speed or temperature increased. Relative humidity, pressure and wind direction were marginally important in predicting the $PM_{2.5}$ concentration. Even though it may be argued these meteorological variables are redundant in the network, we prefer to keep them in because from the sensitivity analysis we can never be completely sure what impact omitting these variables will have on the network. As expected, the roadside $PM_{2.5}$ concentration has a positive correlation with the traffic volume, the higher traffic volume the more $PM_{2.5}$ concentration measured. However, the absolute change in $PM_{2.5}$ concentration due to the change in traffic volume was low.

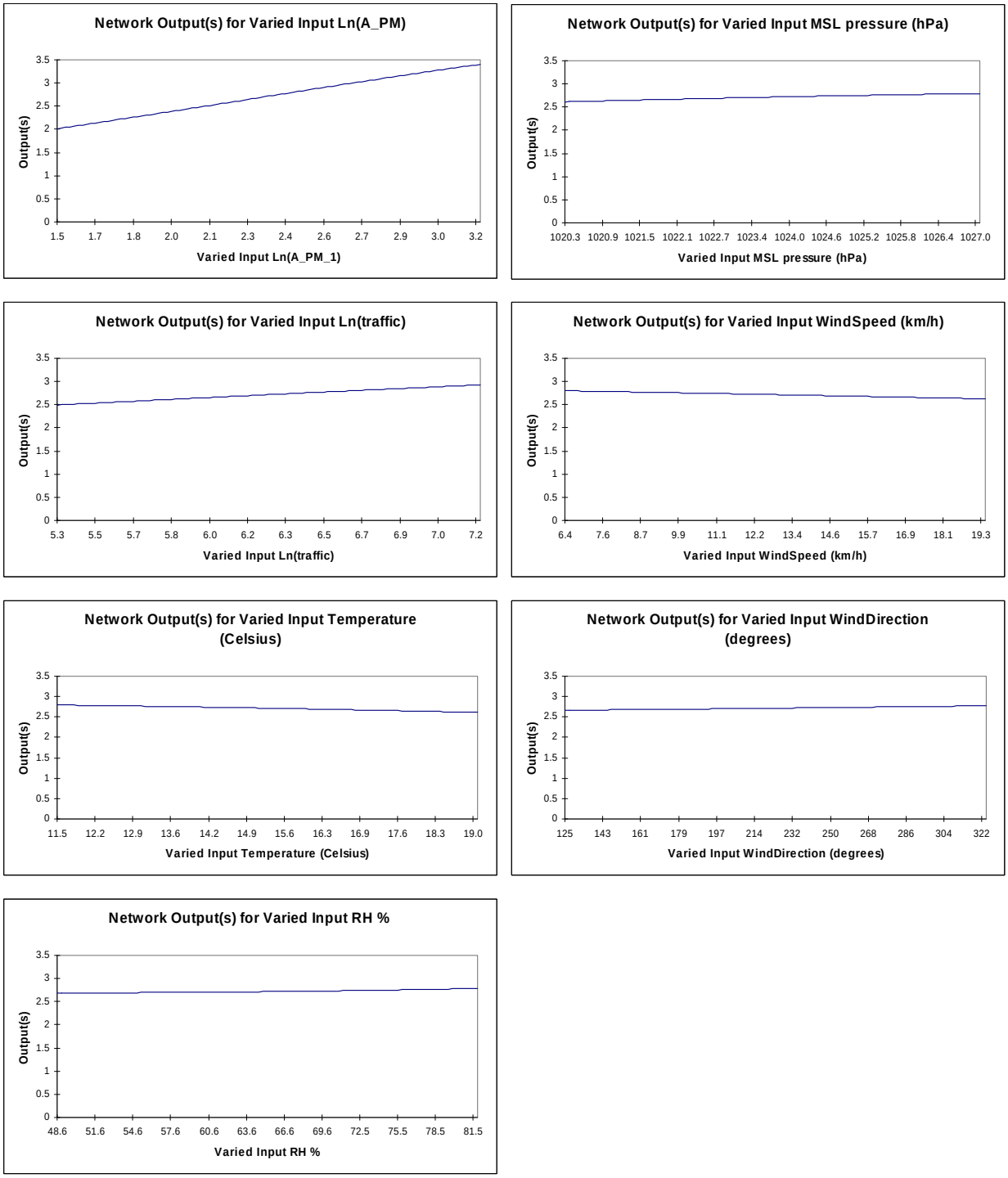

FIGURE 4: Sensitivity Analysis Plots for Input Variables in the ANN

4 Conclusions

This study investigates the potential for applying Artificial Neural Network to the problem of predicting roadside $PM_{2.5}$ concentrations near a busy intersection at a highly disaggregate temporal level (15-minute interval). The MLP (7:9:1) network was trained and cross-validated using the back propagation algorithm. The network captured the complex correlation between the observed variation in ambient PM concentration, traffic and weather conditions. The ANN outperformed the traditional statistical techniques by explaining 71 percent of the total variation in the pollution data on the testing data set.

While these evaluation measures seem impressive, there are several issues that need to be raised. First, the strength of performance is largely down to the availability of contiguous ambient readings, which as the plot in Figure 2 shows are highly correlated (as logic dictates). In a practical application, these data would not be available in a timely manner to predict roadside concentrations raising the question of how close we could get with say the previous day(s) ambient readings. Second, while the sensitivity analysis showed the expected reaction of $PM_{2.5}$ to changes in levels of input variables such as wind speed and traffic, the magnitude of this change was in reality, marginal. This could be down to the use of the fifteen minute averaging interval (dictated by available traffic data) or simply that other traffic parameters (such as proportion of trucks) are more critical than volume per se. Third, while ANNs are designed to (and invariably find) patterns and meaning in the data, often it is difficult to interpret or explain results as one might from classical statistical methods. Finally, the real potential of ANNs appears to lie in prediction. In the study presented here, given the trained ANN was applied to test data collected at the same location under similar conditions we would expect prediction to be good. The real test will come when an ANN developed at one time and location is applied elsewhere.

5 References

Burnett, R. T., Dewanji, A., Dominici, F., Goldberg, M. S., Cohen, A. and Krewski, D. (2003) On the Relationship between Time-Series Studies, Dynamic Population Studies, and Estimating Loss of Life due to Short-Term Exposure to Environmental Risks. *Environmental Health Perspective*, 111, pp. 1170-1174.

Environmental Protection and Heritage Council (EPHC) http://www.ephc.gov.au [accessed 17/05/2007]

Gardner, M. W. and Dorling, S. R. (1998). Artificial Neural Networks (The Multilayer Perceptron) – A Review of Applications in the Atmospheric Sciences. *Atmospheric Environment*, 32, 2627-2636.

Gong, K. W. Zhao, W., Li Ning, Barajas, B., Kleinman, M., Constantinos, S., Horvath, S., Lusis, A., Nel, A. and Araujo, J (2007) Air-pollutant chemicals and oxidized lipids exhibit genome-wide synergistic effects on endothelial cells. *Genome Biology* 2007, **8:**R149.

Greaves, S. P. (2006) Variability of Personal Exposure to Fine Particulates for Urban Commuters inside an Automobile. *Transportation Research Record* 1987, 161-170.

Grivas, G. and Chaloulakou, A. (2006). Artificial Neural Network Models for Prediction of PM_{10} Hourly Concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment*, 40, 1216-1229.

Haykin, S., (1999). *Neural Networks A Comprehensive Foundation*, 2nd Edition, Prentice Hall, Upper Saddle River, New Jersey.

Issayarayangyun, T. and Greaves, S. P. (2007). Analysis of Minute-by-Minute Pollution Exposure inside a Car – a Time-Series Modelling Approach. *Transportation Research Part D,* 12(5), 347-357.

Kappos, A. D., P. Bruckman, and T. Eikmann (2004). Health Effects of Particles in Ambient Air. *International Journal of Hygiene and Environmental Health,* 207, 399-407.

Kaur, S., Nieuwenhuijsen, M. J., and Colvile, R. N. (accepted 05/02/2007) Fine Particulate Matter and Carbon Monoxide Exposure Concentrations in Urban Street Transport

Microenvironments. *Atmospheric Environment*.

Moseholm, L., Silva, J. and Larson, T. (1996). Forecasting Carbon Monoxide Concentrations near a Sheltered Intersection using Video Traffic Surveillance and Neural Networks. *Transportation Research Part D*, 1, 15-28.

Nagendra, S. M. S. and Khare, M. (2004). Artificial Neural Network based Line Source Models for Vehicular Exhaust Emission Predictions of an Urban Roadway. *Transportation Research Part D*, 9, 199-208.

Perez, P., Trier, A. and Reyes, J. (2000). Prediction of PM2.5 Concentrations several Hours in Advance using Neural Networks in Santiago, Chile. *Atmospheric Environment*, 34, 1189-1196.

Zamurs, J. and Conway, R. (1991). A Comparison of Intersection Air Quality Models' Ability to Simulate Carbon-monoxide Concentrations in an urban Area. *Transportation Research Board*, 70th Annual Meeting, Washington, DC.

Zhang, G., Patuwo, B. E. and Hu, M. Y. (1998). Forecasting with Artificial Neural Networks: The State of the Art. *International Journal of Forecasting*, 14, 35-62.

Zhu, Y., Hinds, Kim, S. Shen, S. and Sioutas, C. (2002) Study of ultrafine particles near a major highway with heavy-duty diesel traffic *Atmospheric Environment*, 2002, 36, 4323-4335.