Multilayer perceptron and regression modelling to forecast hourly nitrogen dioxide concentrations

C. Capilla
Department of Applied Statistics and Operations Research and Quality, Polytechnic University of Valencia, Spain

Abstract

This paper presents the application of feed-forward multilayer perceptron networks and multiple regression models, to forecast hourly nitrogen dioxide levels 24 hours in advance. Input data are traffic and meteorological variables, and nitrogen dioxide hourly levels. The introduction of four periodic components (sine and cosine terms for the daily and weekly cycles), and nitrogen oxide hourly levels was analyzed in order to improve the prediction power. The data were measured for three years at two monitoring stations in Valencia (Spain). The model evaluation criteria were the mean absolute error, the root mean square error and the mean absolute percentage error. The multilayer perceptron networks performed better than the regression models in nonlinear relationships like that involving nitrogen oxides, meteorological and traffic variables. Comparisons of the multilayer perceptron-based models proved that the insertion of the four additional seasonal input variables improved the ability of obtaining more accurate predictions, which emphasizes the importance of taking into account the seasonal character of nitrogen dioxide. The advantages of neural networks were that they did not require very exhaustive information about air pollutants, reaction mechanisms, meteorological parameters or traffic flow, and that they had the ability of allowing nonlinear relationships between very different predictor variables in an urban environment.

Keywords: air quality, nitrogen dioxide concentration, urban atmospheric pollution, multilayer perceptron, multiple regression model.

WIT Transactions on Ecology and The Environment, Vol 183, © 2014 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR140041

1 Introduction

Nitrogen dioxide (NO_2) is one of the most relevant air pollutants in Valencia (Spain) [1, 2]. Atmospheric pollution in Valencia is mainly a consequence of motor vehicle emissions. Tenias *et al*. [3] have reported a significant connection between a 10 $\mu g/m^3$ increase in NO_2 level and asthma, measured as a relative risk of emergency visits. Daily levels of NO_2 in Valencia are also associated with cardiovascular admissions [4, 5].

It is well known that other secondary pollutants, such as ozone, are produced owing to the interaction of meteorology, NO_2 and volatile organic compounds [6]. These secondary pollutants are related to photochemical smog and acid rain. Ambient air NO_2 is in large part originated by the oxidation of nitric oxide (NO). NO is emitted primarily by motor vehicles, making it a strong indicator of vehicle emissions [7]. The link between climate and pollutants plays an important role on the variability of NO and NO_2. This link has to be taken in account when selecting optimal pollutant reduction strategies to avoid exceeding emission directives.

There is nowadays a considerable challenge to air quality forecasting. Tools to forecast pollution levels can be used in different ways. One approach to predict future concentrations is to use detailed atmospheric diffusion models. Such models aim to resolve the underlying physical and chemical equations controlling pollutant concentrations and therefore require detailed emissions data and meteorological fields. 3-dimensional air quality models can be applied to integrate chemistry, transport and dispersion. This approach is time-consuming and requires very large databases to initialize and run the model. As the complexity of a problem increases, the theoretical understanding decreases due to ill-defined interactions between systems, and statistical approaches are required.

Statistical models generally directly connect meteorological conditions to level of pollutants. Neural networks have been shown to be effective alternatives to more traditional statistical techniques [8]. The neural network models can be trained to approximate virtually any smooth, measurable function [9]. Unlike other statistical techniques the neural network models make no prior assumptions concerning the data distribution. They can model highly non-linear functions and can be trained to accurately generalize when presented with new, unseen data [10]. These features of neural networks make them an attractive alternative to developing numerical models, and also when choosing between statistical approaches.

During the last decades the use of neural networks, and in particular the multilayer perceptron, has been developed to forecast pollutant concentrations. Oxides of nitrogen (NO_x) and NO_2 levels were forecasted using a multilayer perceptron model and other statistical techniques, and the comparison of results showed that the multilayer perceptron had advantages [11]. Ibarra-Berastegi *et al*. [12] focused on the prediction of hourly levels up to 8 h ahead for five pollutants and six locations in the area of Bilbao (Spain) using multilayer perceptron models. The performance of these models at the different sensors in

the area range from a maximum value of R^2=0.88 for the prediction of NO_2 1 hour ahead to a minimum value of R^2=0.15 for the prediction of ozone 8 hour ahead.

The objective of this study is to investigate the forecasting capability of feed-forward multilayer perceptron networks and multiple regression models. Based on these techniques, several models are designed, and comparisons between them establish the most efficient performer as a forecasting tool. The primary goal of the work is to predict NO_2 concentrations 24 hours ahead in the urban area of Valencia (Spain) at two different locations. Pollutant concentrations, traffic, meteorological variables and seasonal components are used as predictors to develop the models.

2 Data set and methods

The study area is in the city of Valencia (Spain). Valencia has around one million inhabitants, and its climatology and structure are typically Mediterranean. An automatic air pollution network with five monitoring stations is operated in the whole urban area. The network is managed by the Environment Department of the local government. Its monitoring and design criteria follow the Air Quality Framework Directive and subsequent Daughter Directives (1996/62/EC, 1999/30/EC and 2000/60/EC). The reliability of data is ensured by the application of quality assurance and quality control procedures. The monitoring sites provide hourly measurements of air pollutants levels in locations with high traffic density. Meteorological observations are available in two background monitoring stations: P.Silla and Viveros. The P.Silla station is in a roadside site located a few meters from a motorway, and Viveros is in an avenue close to the city centre. The meteorological variables are: wind speed (WS, m/s), wind direction (WD, degrees), temperature (T, ºC) and solar radiation (SR, W/m^2). In P.Silla, relative humidity (RH, %) and pressure (P, mbar) are also measured. The traffic monitoring network managed by the Department of Transport of the Local Municipality, provides traffic data (hourly number of vehicles).

The study period is 1st January 2003–31st December 2005 in P.Silla, and 1st January 2002–31st December 2004 in Viveros. The hourly means of NO_2 have not exceeded at both stations during the study period, neither the limit value nor the alert threshold set by the European Council Directive 1999/30//EC. However the limit value of this pollutant for the protection of human health in a calendar year has been exceeded at P.Silla, in 2003, 2004 and 2005. At this site the highest annual NO_2 mean was observed in 2003. Mass concentrations of nitrogen oxides are determined using chemiluminescence method. Concentrations are expressed in μg/m^3. The volumes are standardized at a temperature of 293ºK and a pressure of 101.3 kPa. Table 1 gives the means, standard deviations and maximum values of the pollutants and meteorological variables at both stations, during the period analyzed. Mean levels of pollutants have been higher at P.Silla.

In this work, applications of artificial neural network and multiple regression models are presented to predict hourly NO_2 concentrations 24 hours in advance,

from local pollutants concentrations, traffic, meteorological data, and periodic components (sine and cosine terms for the daily and weekly cycles). Since the factors mainly contributing to the NO_2 concentration are connected with the source activity (e.g. traffic) and periodic variations in nature (e.g. photochemical reactions in the atmosphere), it is natural to expect periodic components to be found in the NO_2 time series. These are expected at the week level and in the form of daily variations.

Table 1: Descriptive statistics of the variables.

Station	Variable	Mean	Stand. deviation	Maximum
P.Silla	NO_2	58.8	30.2	249
	NO	52.0	59.2	624
	WS	1.1	0.9	8.6
	WD	187.8	107.0	360
	T	18.7	6.8	38.2
	RH	60.8	15.5	92
	P	1022.2	6.8	1044.7
	SR	153.3	246.4	947
Viveros	NO_2	36.72	24.6	238
	NO	19.6	37.8	596
	WS	1.8	1.3	11.9
	WD	168.3	117.2	360
	T	18.9	6.4	38.2
	SR	170	260	1033.3

Figure 1 represents the daily cycle of NO_2 and NO mean levels in P.Silla. Figure 2 gives the weekly cycle at the same station for the two pollutants. Similar patterns were observed at Viveros station.

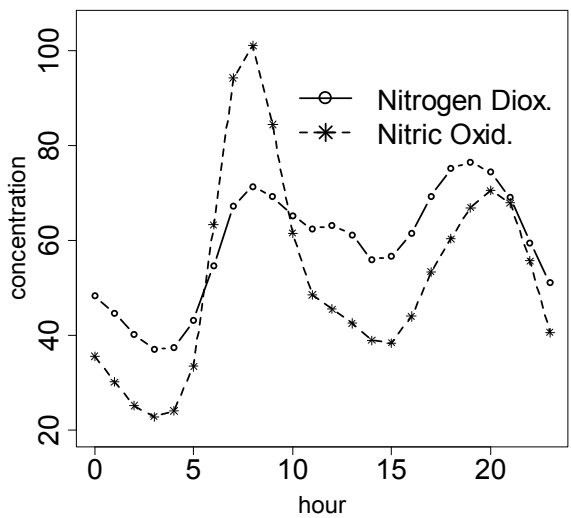

Figure 1: Daily cycle at P.Silla.

The predictions produced by two different methods are compared: feed-forward multilayer perceptron networks (MLP) and multiple regression (MR) models. The number of neurons in the hidden layer for the MLP networks is the optimum found by experimentation. The transfer functions selected are the hyperbolic tangent for the hidden layer, and linear for the output layer. The MLP model is applied using the Levenberg-Marquard algorithm. These parameters were also selected by experimentation. In the multiple regression models, the logarithm transformation was employed in order to normalize pollutants data and so stabilize the variance.

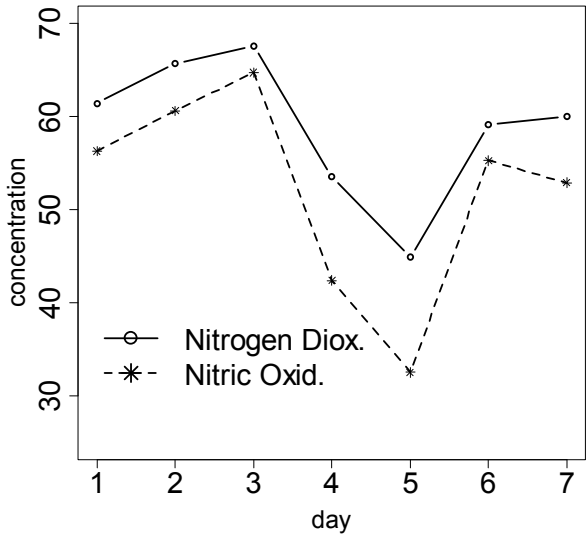

Figure 2: Weekly cycle at P.Silla.

Table 2 shows the different models that have been analyzed. They differ in the number of input or predictor variables that are included. In all of them the output variable is the prediction of NO_2 24 hours in advance.

When the MLP memorizes the patterns introduced to it and it is not capable of identifying new situations overtraining does occur. The early stopping technique can be used to avoid overtraining. In the early stopping technique the available data are separated into three sets: the training set, the validations set and the test set. The training set is used to update the network weights and biases. During the training, the validation set is used to guarantee the generalization capability of the model. Training should stop before the error on the validation set begins to rise. Finally, the test set is a new set used to check the generalization of the MLP. In this work, the models are trained on data from 2003 in P.Silla, and from 2002 in Viveros. Data from 2004 are used as the validation set and observations from 2005 are the test data set in P.Silla. In Viveros the validation set was year 2003 and the test set was year 2004. The

model evaluation criteria are the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE) and the correlation coefficient (r) between the pollutant observations and the predictions. The *Neural Network Toolbox* of MATLAB is used.

Table 2: Models analyzed in the paper.

Model	Output variable	Input variables
MLP1	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, $(NO_2)_t$
MLP2	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, $Seasonality_{t+24}$,$(NO_2)_t$
MLP3	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, $(NO_2)_t$, NO_t
MLP4	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, $Seasonality_{t+24}$,$(NO_2)_t$, NO_t
MR1	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, NO_t
MR2	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, $Seasonality_{t+24}$,NO_t
MR3	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, NO_t, $(NO_2)_t$
MR4	$(NO_2)_{t+24}$	$Meteorology_t$, $traffic_t$, $Seasonality_{t+24}$, NO_t, $(NO_2)_t$

3 Results and discussion

Table 3 gives the values of the four performance criteria for the 8 models at P.Silla. It is included the number of neurons in the hidden layer that had the best results.

Table 3: Performance criteria results at P.Silla.

Model	n_h	MAE	RMSE	MAPE	r
MLP1	5	16.5067	20.9563	0.5261	0.5780
MLP2	10	16.6987	20.7523	0.5110	0.6367
MLP3	14	16.8205	20.8091	0.5818	0.5367
MLP4	10	16.5598	20.5803	0.5058	0.6610
MR1		17.2526	21.0154	0.5535	0.6088
MR2		17.2975	21.1004	0.5510	0.6042
MR3		17.2719	21.0401	0.5544	0.6086
MR4		17.3177	21.1212	0.5520	0.6040

The best results are obtained with the multilayer perceptron, including as input variables all the meteorology parameters, traffic, the daily and weekly cycles, and the NO_2 and NO concentrations (MLP4). The multiple regression models performed worse than the neural networks. Among the multiple regression models the best results were obtained including only meteological

variables, traffic and NO_2 concentrations (MR1). Figure 3 represents NO_2 observations at P.Silla for the first 100 days of the test period (2005). Predictions obtained with MLP4 and MR1 are included.

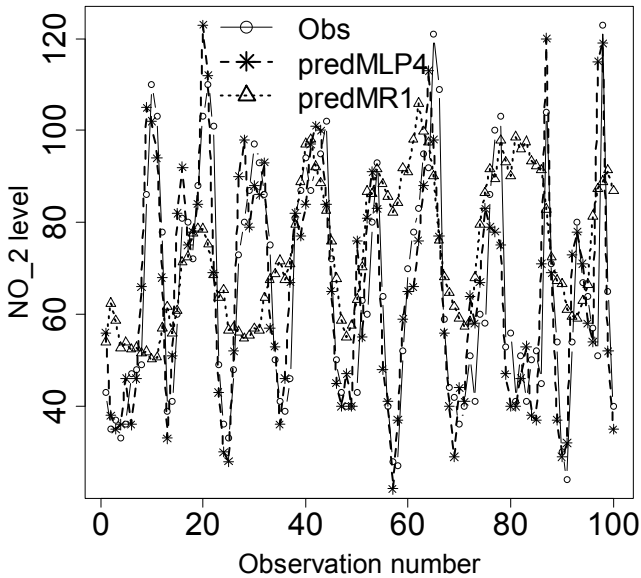

Figure 3: Observations and predictions at P.Silla.

The results at Viveros can be found in Table 4.

Table 4: Performance criteria results at Viveros.

Model	n_h	MAE	RMSE	MAPE	r
MLP1	7	18.3393	23.3951	0.9343	0.118
MLP2	12	19.0325	24.1552	0.9818	0.1065
MLP3	20	18.5432	23.6416	0.9200	0.1189
MLP4	14	19.1380	24.3809	0.9974	0.1278
MR1		19.4308	24.7099	1.0031	0.0799
MR2		19.6402	25.0202	1.0052	0.0490
MR3		19.4622	24.7764	1.0022	0.0735
MR4		19.6646	25.0797	1.0051	0.0442

The optimum number of hidden neurons for the multilayer perceptros is indicated. In this case, the eight models had a worse performance than at P.Silla. Including seasonality did not improve predictions accuracy. The best values of the performance criteria were obtained for MLP1. Regression models had worse

results than the multilayer perceptron. Figure 4 plots observations and predictions with MLP1 and MR1 for the first 100 days of the test period (2004) at Viveros.

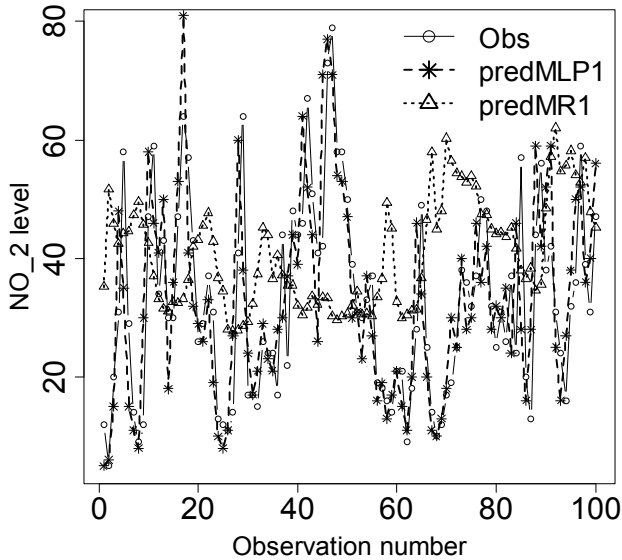

Figure 4: Observations and predictions at Viveros.

4 Conclusions

The aim of this research is to develop a predictive model to forecast hourly pollutant values 24 hours in advance in the urban area of Valencia (Spain). Air quality forecasting is an adequate method to plan a health warning system. The neural network models have become an alternative to conventional methods, because the relationship between air pollutants and meteorology is complex and extremely non-linear. They are an important instrument to model air pollution distribution. In the last decade the feedforward multilayer perceptron networks have been applied to analyze different pollutant levels. In this work a comparison of this method with multiple regression models is developed. The study is based on data obtained at an official monitoring station in the city of Valencia, in an area with high traffic density.

The results showed that good forecast estimates of air quality can be achieved by applying neural networks methods to the prediction of time series of NO_2 concentrations. As the signal studied shows periodicity, a reasonable estimate can be recovered by using periodic components as estimators. Meteorological, traffic and nitrogen monoxide observations are also used as input variables to the models. The relative importance of meteorological and vehicle emission

variables on the surface pollutant prediction is of great interest to establish the legislative measures that permit to reduce the pollutant levels.

The parameters of the models were chosen by experimentation, to optimize the values of several criteria. The forecasting capability of the models was evaluated using the mean absolute error, the root mean square error, the mean absolute percentage error and the correlation coefficient between the pollutant observations and the predictions. The multilayer perceptron networks performed better than the regression models, giving satisfactory result for the prediction of pollutant 24 h in advance. The developed model is a potential tool for predicting air quality parameters inside the city, making the proposed forecaster a powerful tool for pollution management systems.

References

[1] European Communities, *Europe´s Environment: The Second Assessment*. European Environment Agency, Office for Official Publications of the European Community, 1998.

[2] Capilla, C., Modelling temporal changes of nitrogen dioxide concentrations in an urban area. *WIT Transactions on Ecology and The Environment*, **157**, pp. 71-79, 2012.

[3] Tenias, J.M, Ballester, F. & Rivera, M.L., Association between hospital medical emergency visits for asthma and air pollution in Valencia, Spain, *Journal of Occupational and Environmental Medicine,* **55**, pp. 541-547, 1998.

[4] Ballester, F., Tenias, J.M. & Pérez-Hoyos, S., Air pollution and emergency hospital admissions for cardiovascular diseases in Valencia, Spain, *Journal of Epidemiology and Community Health,* **55**, pp. 57-65, 2001.

[5] Ballester, F., Rodríguez, P., Iñíguez, C., Sáez, M., Daponte, A. et al., Air pollution and cardiovascular admissions association in Spain: result within the EMECAS project, *Journal of Epidemiology and Community Health,* **60**, pp. 328-336, 2006.

[6] Saunders, S.M., Jenkin, M.E., Derwent, R.G. & Piling, M.J., WWW site of a Master Chemical Mechanism for use in tropospheric chemistry models, *Atmospheric Environment*, **31**, pp. 1249, 1997.

[7] World Health Organization, *Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide,* report on a WHO Working Group, Bonn, Germany, 2003.

[8] Shalkoff, R., *Pattern Recognition: Statistical Structural and Neural Approaches*, Wiley, New York, 1992.

[9] Hornik, K., Stinchcombe, M. & White, H., Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, pp. 359-366, 1989.

[10] Bishop, C.M., *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[11] Gardner, M.W. & Dorling, S.R., Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London, *Atmospheric Environment*, **33**, pp. 2627-2636, 1999.
[12] Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A. & Diaz de Argandoña, J., From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao, *Environmental Modelling & Software*, **23**, 622-637, 2008.