A neural network based model to forecast hourly ozone levels in rural areas in the Basque Country

E. Agirre[1], A. Anta[2], L. J. R. Barrón[3] & M. Albizu[4]
[1]*University of the Basque Country, Department of Applied Mathematics, Bilbao, Spain*
[2]*Systems analyst, Vitoria-Gasteiz, Spain*
[3]*University of the Basque Country, Department of Food Technology, Vitoria-Gasteiz, Spain*
[4]*Basque Government, Environmental Department, Bilbao, Spain*

Abstract

The goal of this work is to build and evaluate a multilayer perceptron based model to forecast tropospheric ozone (O_3) levels, in real-time, up to eight hours ahead at two rural stations located in the Autonomous Community of the Basque Country (North Central Spain). Current and historical hourly concentrations of ozone, nitrogen dioxide (NO_2) and meteorological variables were used to determine the input variables of the model. The designed basic model established sixteen multilayer perceptrons, which were trained using the scaled conjugate gradient algorithm. The performance of the model was evaluated using the statistics of the Model Validation Kit. The study proved the capability of artificial neural networks to forecast efficiently ozone concentrations at rural stations in the Basque Country.
Keywords: multilayer perceptron, artificial neural networks, air quality modelling, ozone.

1 Introduction

The pollution caused by photochemical oxidants is one of the main problems in air quality. In this way, the tropospheric ozone (O_3) must be considered as a relevant air pollutant. The tropospheric ozone is a secondary pollutant, originated

WIT Transactions on Ecology and the Environment, Vol 101, © 2007 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR070111

as a consequence of the reactions produced among the nitrogen oxides (NOx) and the volatile organic compounds (VOCs) under the solar radiation. Based on the data registered in the Air Quality Monitoring Network of the Basque Government, it is known that high O_3 concentrations were recorded at rural stations such as Pagoeta and Valderejo, and high ozone concentrations have adverse effects on human health and the environment. Consequently, the short-term prediction of O_3 would be very helpful. Therefore, the goal of this work is the elaboration and validation of a prediction model to forecast, in real time, hourly O_3 concentrations up to eight hours ahead at rural stations.

In the last decades cause/effect models and statistical models have been developed with the purpose of forecasting O_3 hourly levels [1]. Recently, the artificial neural networks have become very useful in the elaboration of prognostic models to forecast air quality levels [2, 3]. The artificial neural networks have proved their efficiency to describe non-linear relationships such as those involved in ozone formation, and they have generally provided better results than linear methods [4, 5].

Our research team has elaborated and evaluated a prognostic model, based on the use of artificial neural networks, to forecast in real time hourly ozone concentrations up to eight hours ahead at several stations of the Air Quality Monitoring Network of the Basque Country. This paper shows the results obtained at two rural stations named Pagoeta and Valderejo.

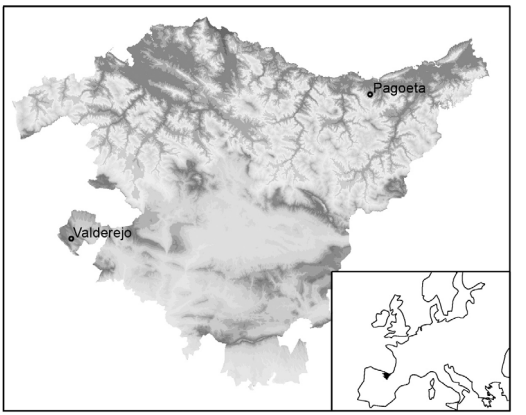

Figure 1: Pagoeta and Valderejo stations.

2 Database

The Air Quality Monitoring Network of the Autonomous Community of the Basque Country measures hourly several meteorological parameters and air pollution variables at each station. The database considered in this work is formed by the hourly measures of tropospheric ozone (O_3), nitrogen dioxide (NO_2), temperature, relative humidity, pressure, solar radiation, wind speed and

wind direction registered during the period 2001-2004 at Pagoeta (lat.: 43º15´2´´, long.: 2º9´18´´, alt.: 215) and Valderejo (lat.: 42º52´31´´, long.: 3º13´53´´, alt.: 911). Pagoeta is located on the coast of the Basque Country and Valderejo is situated in the SW of the Basque Country, approximately 97 km away from Pagoeta (Figure 1).

3 Methodology

3.1 The multilayer perceptron

The artificial neural networks are structures similar to the nervous human system, where the neuron is the fundamental element. Depending on the structure and connections, the characteristics of the neurons and the learning algorithm of the artificial neural network, there are different types of artificial neural networks. The multilayer perceptron (MLP) is the artificial neural network with the biggest number of practical applications.

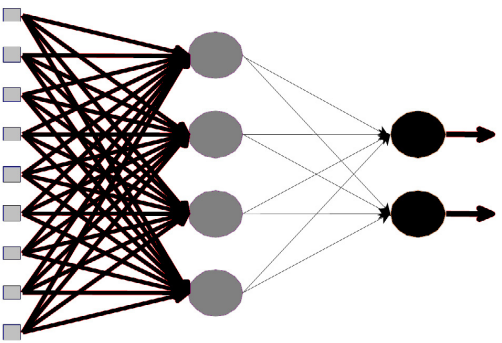

Figure 2: The 9-4-2 multilayer perceptron.

Figure 2 shows a multilayer perceptron with 9 neurons in the input layer, 4 neurons in the hidden layer and 2 neurons in the output layer. The multilayer perceptron consists of at least three layers: the input layer, the hidden layer(s) and the output layer. The input layer receives the information that enters from the outside to the artificial neuronal network. This information propagates towards ahead, so that each input is multiplied by the synaptic weights and the total sum of the products is connected to each neuron of the hidden layer. A transfer function is applied to this sum and the result becomes the input of the following layer. The multilayer perceptron could be formed by one or more hidden layers. Finally, the output layer produces the output of the multilayer perceptron.

The artificial neural networks, and in particular the multilayer perceptron, possess the aptitude to learn from the patterns introduced to them and the error

measured in the learning process, so that finally they are capable of identifying a pattern never seen previously. Consequently, it is said that an adequately trained artificial neural network has high generalization capability. The learning procedure is equivalent to the minimization process of the error

$$E = \frac{1}{S}\sum_{k=1}^{S}(t_k - y_k)^2 \tag{1}$$

observed between the target $(t_1, t_2, \ldots, t_S)$ and the output $(y_1, y_2, \ldots, y_S)$ of the neural network, where S is the number of training patterns. The output of the multilayer perceptron is compared to the target, and the error is propagated backward through the network to produce an adjustment in the weights and biases of the network, so that the difference between the output of the network and the target is minimized. Once the minimum of the difference has been reached, the learning finishes. This method is known as backpropagation.

3.2 Building the model

The multilayer perceptron based models built in this work have the structure N-L-1, with N neurons in the input layer, L neurons in the unique hidden layer and one neuron in the output layer, which corresponds to the prediction of ozone at time $t+k$ or output of the model $O3(t+k)$, being $k = 1, \ldots, 8$.

The inputs of the model were determined by stepwise regression and tolerance filtering using data from 2001-2002. In this manner, the number of input variables was reduced, being the ozone concentration at the prediction time, $O3(t)$, the variable that explained the biggest percentage of the variance in the prediction models to forecast ozone concentrations at time $t+k$, ($k = 1, \ldots, 8$) at Pagoeta and Valderejo. Past values of the solar radiation and wind direction variables played also a significant role as input variables in these prognostic models. In the same way, taking into account the importance of the utilization of seasonal components as predictors to forecast hourly ozone levels, the variable $\cos(2\pi h/24)$ was considered as another important input variable. Therefore, the general structure of the prognostic model was N-L-1, with $N = 3, 4, 5$. The number of neurons in the hidden layer L was calculated by a generalization rule [6] applied in a trial and error procedure.

The training algorithm used was the scaled conjugate gradient (SCG) algorithm; it is a variation of backpropagation that provides generally better results and a faster convergence [7]. Moreover, in order to avoid overtraining, the early stopping technique was applied, by dividing the whole database into three subsets: data from the period 2001-2002 formed the training set, data from 2003 formed the validation set and the test set was formed by data from year 2004. The validation set is used to guarantee the generalization capability of the model. It indicates the stop of the training, before the error on the validation set begins to rise.

The hyperbolic tangent function *tansig(x)* was considered as the transfer function between the input layer and the hidden layer and the linear function *lin(x)* connected the hidden layer and the output layer.

$$tansig(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2}$$

$$lin(x) = x \tag{3}$$

Finally, eight multilayer perceptrons were set to forecast the values of the variables *O3(t+k)* at each station of the study, being $k = 1, 2, \ldots, 8$.

3.3 Goodness of the fit

Once the prognostic model was built, the statistics of the Model Validation Kit were chosen to determine the goodness of the fit of each ozone prediction [8]. These are the proposed measures in the Model Validation Kit: (i) the correlation coefficient

$$R = \frac{\overline{(C_o - \bar{C}_o)-(C_p - \bar{C}_p)}}{(SC_p)(SC_o)};$$

(ii) the Normalized Mean Square Error,

$$NMSE = \frac{\overline{(C_o - C_p)^2}}{\overline{(C_o)}\,\overline{(C_p)}};$$

(iii) the factor of two, *FA2,* which explains the ratio $0.5 \le C_o/C_p \le 2$;
(iv) the Fractional Bias,

$$FB = 2\frac{\overline{C_o} - \overline{C_p}}{\overline{C_o} + \overline{C_p}};$$

and (v) the Fractional Variance,

$$FV = 2\frac{SC_o - SC_p}{SC_o + SC_p}.$$

The notation in use in these terms is the following one: C_o is the observed concentration of ozone, C_p is the prediction, $\overline{C_o}$ and $\overline{C_p}$ are the mean values and SC_o and SC_p are the standard deviations of C_o and C_p respectively.

Based on the values of the statistics of the Model Validation Kit, the best forecast is that whose *NMSE, FV* and *FB* values are zero and the corresponding values of *R* and *FA2* are the unit.

4 Results

In this study, the calculation of the statistics of the Model Validation Kit on the test set (year 2004) determined the goodness of the fit of the prognostic model designed to forecast ozone concentrations up to eight hours ahead. The values

presented in Table 1 prove the accuracy of the ozone predictions obtained as outputs of the designed models at Pagoeta and Valderejo.

Furthermore, the accuracy of the model could be graphically observed as in the example presented from Figure 3 to Figure 10, where the ozone forecasts up to eight hours ahead were calculated at time t (t = 1, 2, …, 24) with data of 4 August 2004 at Pagoeta. In these figures the solid lines represent the real ozone concentrations (observations) and the dotted lines depict the ozone forecasts (outputs) at time $t+k$, being k = 1, 2,…,8.

Table 1: Statistics of the Model Validation Kit at Pagoeta and Valderejo on the test set (year 2004).

	NMSE	*R*	*FA2*	*FB*	*FV*
O3pago(t+1)	0.0007	0.9969	0.9978	-0.0023	0.0468
O3pago(t+2)	0.0032	0.9840	0.9964	-0.0007	0.1061
O3pago(t+3)	0.0046	0.9792	0.9932	-0.0036	0.1556
O3pago(t+4)	0.0077	0.9690	0.9926	-0.0024	0.2376
O3pago(t+5)	0.0090	0.9611	0.9905	-0.0003	0.2487
O3pago(t+6)	0.0104	0.9599	0.9912	0.0049	0.2953
O3pago(t+7)	0.0117	0.9543	0.9895	0.0006	0.3194
O3pago(t+8)	0.0138	0.9360	0.9901	0.0030	0.3166
O3valde(t+1)	0.0010	0.9965	0.9998	0.0100	0.0408
O3valde(t+2)	0.0045	0.9851	0.9884	0.0211	0.1131
O3valde(t+3)	0.0183	0.9216	0.9926	0.0384	0.1530
O3valde(t+4)	0.0198	0.9237	0.9926	0.0430	0.2318
O3valde(t+5)	0.0293	0.8751	0.9862	0.0403	0.3057
O3valde(t+6)	0.0221	0.9510	0.9871	0.0248	0.4571
O3valde(t+7)	0.0491	0.8729	0.9770	0.0665	0.7184
O3valde(t+8)	0.0736	0.7592	0.9582	0.0773	0.9800

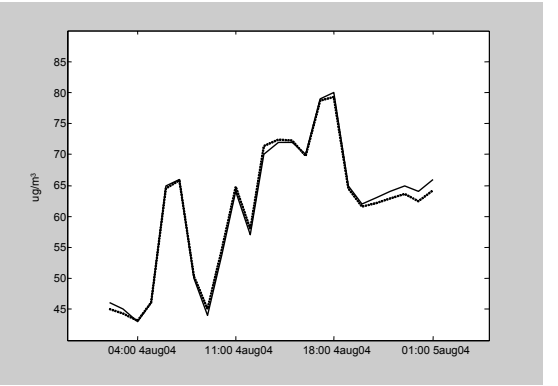

Figure 3: Observation and output *O3(t+1)* at Pagoeta with data of 4 August 2004.

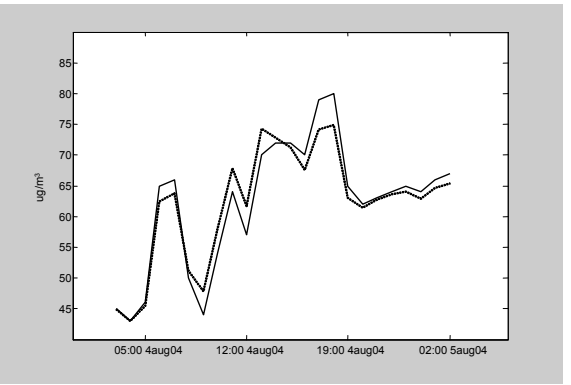

Figure 4: Observation and output *O3(t+2)* at Pagoeta with data of 4 August 2004.

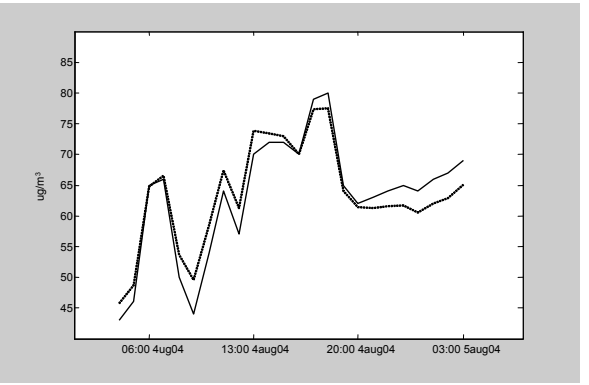

Figure 5: Observation and output *O3(t+3)* at Pagoeta with data of 4 August 2004.

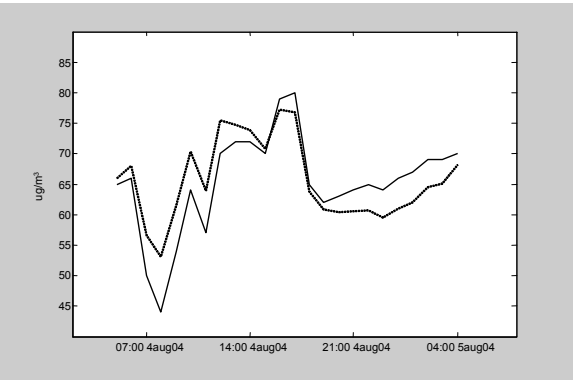

Figure 6: Observation and output *O3(t+4)* at Pagoeta with data of 4 August 2004.

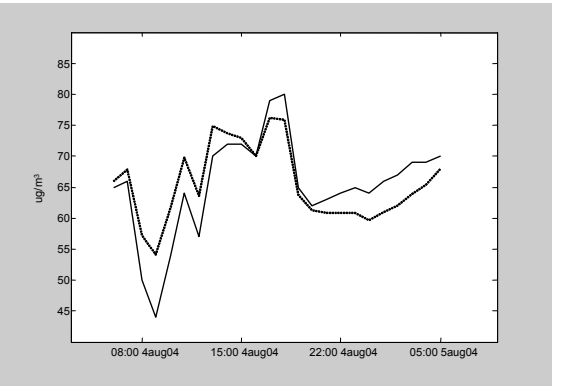

Figure 7: Observation and output *O3(t+5)* at Pagoeta with data of 4 August 2004.

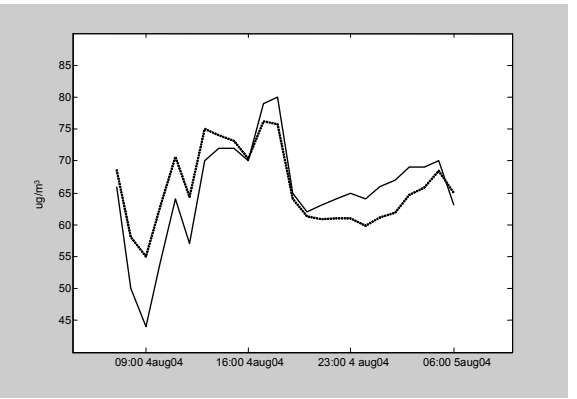

Figure 8: Observation and output *O3(t+6)* at Pagoeta with data of 4 August 2004.

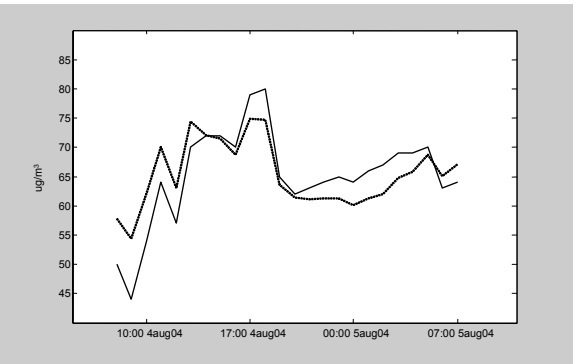

Figure 9: Observation and output *O3(t+7)* at Pagoeta with data of 4 August 2004.

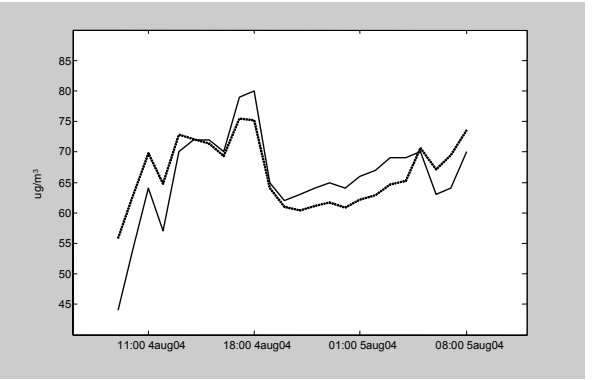

Figure 10: Observation and output *O3(t+8)* at Pagoeta with data of 4 August 2004.

5 Conclusions

In the present work, a multilayer perceptron based model was used to forecast hourly ozone concentrations up to eight hours ahead at two rural stations in the Basque Country, named Pagoeta and Valderejo. A joint study of the values of the statistics of the Model Validation Kit proved the efficiency of the designed model, which performed in the same way at these two rural stations, independently of its location. The model was also successful in other rural and urban stations on the coast of the Basque Country.

Acknowledgement

The authors in this study, the calculation of the statistics of the Model Validation Kit on the test set determined the goodness of the fit of the LR, MLP1 and MLP2 models in a quantitative manner, wish to acknowledge the Environmental Department of the Basque Government for providing data from the air pollution network and for their financial support to realize the present study.

References

[1] Lissens, G., Debruyn, W., Dumont, G., *Forecasting maximum hourly ozone concentrations on a daily basis in Belgium by means of the model SMOGSTOP,* http:// www.vito.be.

[2] Gardner, M.W. & Dorling, S.R., *Statistical surface ozone models: an improved methodology to account for non-linear behaviour,* Atmospheric Environment, 34, 21-34, 2000.

[3] Elkamel, A., Abdul-Wahab, S., Bouhamra, W., Alper, E., *Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach.* Advances in Environmental Research 5, 47-59, 2001.

[4] Cobourn, W.G., Dolcine, L., French, M., Hubbard, M.C., *A comparison of nonlinear regression and neural network models for ground-level ozone forecasting* Journal of the Air and Waste Management Association, 50, 1999-2009, 2000.
[5] Agirre, E., Ibarra, G., Madariaga, I., *Regression and multilayer perceptron based models to forecast hourly O_3 and NO_2 levels in the Bilbao area.* Environmental Modelling and Software 21, 430-446, 2006.
[6] Amari, S-I., Murata, N., Müller, K.R., Finke, M., Yang, H.H., *Asymptotic statistical theory of overtraining and cross-validation*, IEEE Transactions on Neural Networks, 8, 985-996, 1997.
[7] Moller, M.F., *A scaled conjugate gradient algorithm for fast supervised learning,* Neural Networks, 6, 525-533, 1993.
[8] European Commission, *The Evaluation of Models of Heavy Gas Dispersion. Model Evaluation Group Seminar*, Office for Official Publications of the European Communities, L-2985, Luxemburg, 1994.