Short-term prediction of air pollution levels using neural networks

G. Ibarra-Berastegi
University of the Basque Country, Bilbao, Spain

Abstract

This paper focuses on the prediction of hourly levels up to 8 hours ahead for five pollutants (SO_2, CO, NO_2, NO and O_3) in the area of Bilbao. Traffic, meteorological and air pollution network data corresponding to the years 2000 and 2001 have been used. 216 specific models based on different types of neural networks have been built using data for the year 2000. For each of the 216 cases, the choice of the best model has been made under the criteria of simultaneously having at a 95% confidence level the best values of R^2, *d, FA2* and *RMSE* when applied to the data for the year 2001. Depending on the pollutant, location and number of hours ahead the prediction is made, different architectures have been selected. In the case of SO_2 and CO, in most cases persistence of levels or linear models outperformed those based on neural networks. Predictions of NO_2 and O_3 hourly levels required in most cases linear models while MLP, RBF or GRNN architectures were needed in few predictions. For the predictions of NO, linear models in some cases and MLP, RBF or GRNN based models in others, were the major options. In spite of the different architectures and also the different explanatory mechanisms involved the performance of the selected models is very similar.
Keywords: air pollution forecasting, neural networks, MLP, RBF, GRNN.

1 Introduction

Air quality networks are usually designed for diagnosis purposes, being the most important feature of a good network, that it has enough time and space resolution to follow the evolution of the most important fields of concentrations of pollutants. Very often, the same network also measures meteorological variables.

WIT Transactions on Ecology and the Environment, Vol 86, © 2006 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR06003

Bilbao is located North-central Spain (Europe) and as many other urban environments in the world has air pollution problems, mainly due to photochemical smog [1, 2].

This work describes the results of a study carried out in Bilbao (Spain) corresponding to years 2000 and 2001, in which data from the three existing networks (air quality, meteorological and traffic) in this city have been analyzed jointly to see if short-term real time hourly forecasts can be obtained for ozone, NO, NO_2, SO_2 and CO.

The underlying assumption for this work was that if the system formed by the three existing networks can properly describe the joint evolution of air pollution, meteorology and traffic, an analysis of their historical records to detect and recognize patterns and relationships among them, can lead to the prediction of future air pollution levels. These patterns and relationships relating overall inputs (current and past values of air pollutants, meteorology and traffic) and outputs (future values of air pollutant) can be described using statistical techniques. The statistical models obtained in this way will be used to forecast future levels of air pollutants. Due to the highly non-linear effects known to be involved, different types of neural networks (NN) were used to build the models.

2 Methodology

The most popular type of NN used in air pollution has been the MLP. A MLP network with one single intermediate layer and a sigmoid activation function is at least theoretically, able to approach any function if correctly trained [3]. However, if the network must learn a function which shows discontinuities two hidden layers may be required [4]. In general, using more than one hidden layer provides greater flexibility and enables approximation of complex functions with fewer connection weights [4]. The main features of MLP's have been widely explained in the literature [5, 6]

For this work two more types of NN were also used as candidate techniques to model the complex relationships that exist among air pollution, meteorology and traffic: Radial Basis Functions (RBF), and Generalized Regression Neural Network (GRNN).

RBF networks represent another type of NN with an input layer, an output layer and a hidden layer of radial units each actually modeling a Gaussian response surface. The network outputs are then calculated as a weighted sum of the Gaussian outputs. The standard RBF has an output layer containing dot product units with identity activation functions and one single layer is in principle, enough to model any non-linear function [5, 7, 8]. A comparison of the performance between MLP and RBF models to predict daily concentrations of $PM_{2.5}$, suggests that RBF networks show the best behaviour and stability and shortest training times [9].

GRNN's are intended for regression purposes [8, 10, 11] and have two hidden layers. The first hidden layer in the GRNN contains radial units and the second hidden layer consists of neurons that help to estimate the weighted average. The second hidden layer always has exactly one more unit than the output layer. In

regression problems, typically only a single output is estimated, and so, the second hidden layer usually has two units [8].

One of the problems when building neural networks is when overfitting takes place. If at the training stage, the network parameters are calculated after too many cycles (epochs) the network may reproduce various idiosyncrasies associated to the random noise variation of the particular data from which the parameters of the model are estimated [8] instead of capturing the main mechanism the network is trying to describe. The most mathematical and practical aspects of NN's have been widely explained in the literature [4, 5, 8].

After building the models, an important aspect is the evaluation of their performance when faced with the new data belonging to the test data set. The most widely used statistical indicators of the goodness of fit for a model is the Pearson correlation coefficient R and its square R^2, which represents the proportion of the observed variance explained by the model. However, several works [12–14] have shown the shortcomings and limitations of this indicator, though for comparison purposes are still used. Some proposals have been made for more meaningful statistical indicators [12–15] and, apart from the classical R^2, in the last years the following indicators are being used widely [12–15].

1. The index of agreement d (1) varies between 0 and 1 and is a dimensionless measure of the degree to which a model's predictions (P_i) are error free when compared with the observations (O_i).

$$d = 1\text{-}[\Sigma|P_i\text{-}O_i|][\Sigma|P_i\text{-}\bar{O}|+|O_i\text{-}\bar{O}|]^{-1} \quad (1)$$

   If the value of d is 1 indicates perfect agreement between the observed and predicted observations while 0 connotes a complete disagreement.

2. Fraction of two (*FA2*) which represents the proportion of the ratio between observed and predicted values that falls in the range 0.5-2.
3. Total root mean squared error RMSE [12].

For this study, historical hourly records of traffic, air pollution and meteorology corresponding to years 2000 and 2001 were available. The objective was to build short-term prognostic models for SO_2, CO, NO_2, NO and O_3 in the area of Bilbao. The analysis was carried out for 6 locations in the area and predictions from 1 to 8 hours ahead. That made 216 NN's which were selected following the next steps:

1. For each of the 216 predictions, 100 NN's were built using data of year 2000. The 100 NN's included MLP of 1 and 2 hidden layers, RBF's, GRNN's and also linear networks. In general, a linear regression analysis can be understood as particular case of MLP with one hidden layer and a linear transfer function.
2. The 5 networks with minimum error in the validation set were chosen.

3. These 5 networks plus persistence of levels constituted the six candidate networks to be tested for each of the 216 predictions with data of year 2001.
4. The performance of the models for the predictions of the different pollutants was calculated after applying them to year 2001. The best model out of the six candidates was chosen under the criteria of simultaneously having the best values for the four statistical indicators (R^2, *d, FA2* and *RMSE*) at a 95% confidence level.

3 Results

For the prediction of SO_2 levels up to 8 hours ahead, out of the 48 predictions in most cases (33), persistence of levels is either the best option or is not outperformed by any other model. In 13 cases, linear models are either as good as different neural networks or perform better. Only in two cases, the use of the more complicated MLPs are justified. Depending on the sensor, persistence of levels tend to be the best model for predictions up to K=4 hours ahead while for predictions from H+5 to H+8, linear models and/or different types of neural networks perform better.

In the case of the predictions of CO, at H+K with low values of K, persistence of levels and simple linear models are for most sensors, the best options. From H+5 to H+8 predictions tend to be obtained best using linear models and MLP or RBF's.

For the predictions of NO_2 in most cases (31), linear models are enough to launch forecasts. For the predictions up to 2 hours ahead persistence of levels tend not to be outperformed, while linear models are the most usual for the rest of predictions. More sophisticated models like MLP and RBF networks need to be built for a few predictions from K=4 to 8 hours ahead.

In the case of NO, persistence is the best model in four sensors for K=1 while for higher values of K in approximately half the predictions, linear models are enough. In the rest of the cases (18 out of 48) MLP, RBF or GRNN models have to be built.

Ozone predictions are obtained at 3 locations and only in 4 cases out of 24 - for values of K below 2- persistence of levels is not outperformed by any other type of model. In 13 cases, linear models work better than any other and in 6 cases as well as non-linear networks. Only in one case (K=7) it is necessary to build a RBF model.

For each prediction, the best model has been chosen under the criteria that its R^2, *d, FA2* and *RMSE* values at a 95% confidence level were simultaneously the best when compared with the rest of the models. For each pollutant, location and number of hours ahead, the best prediction has been obtained using different types of models. In figures number 1 and 2 it can be seen the maximum and minimum values of the index of agreement, *d* and R^2 corresponding to those obtained with the best models to forecast ozone at the six locations of the area studied.

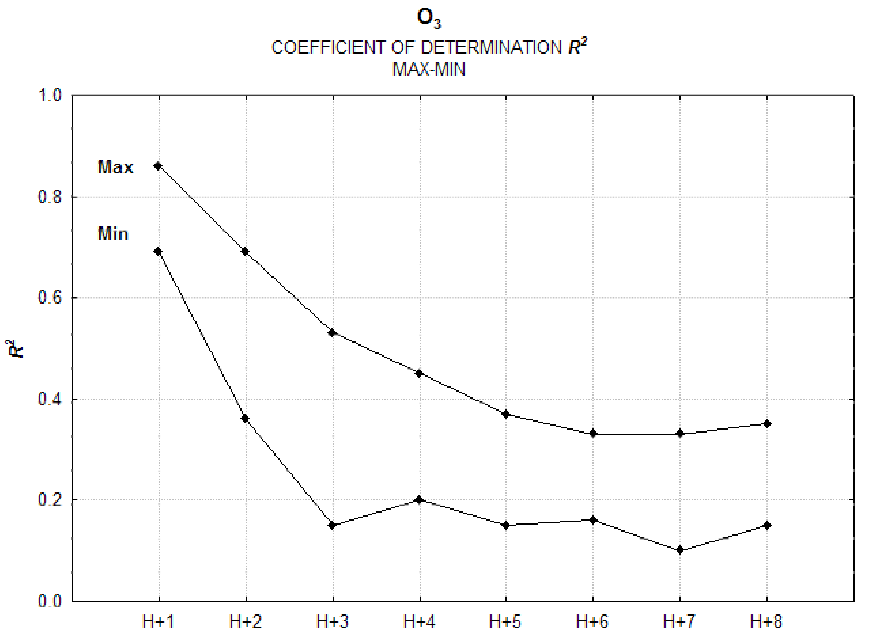

Figure 1: Maximum and minimum values of R^2 obtained up to 8 hours ahead.

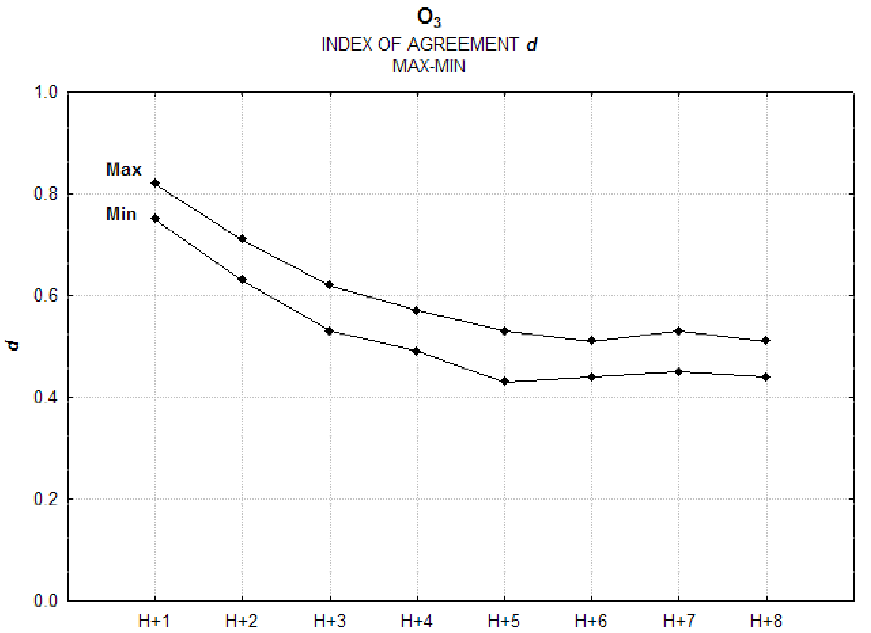

Figure 2: Maximum and minimum values of *d* obtained up to 8 hours ahead.

In the case of SO_2 and CO, persistence of levels is quite often the best option, followed by linear models. However, in the predictions of NO_2, NO and O_3, linear models are usually enough and only in a few cases more sophisticated neural networks are needed.

In the time scale of this study, SO_2 and CO do not suffer important chemical transformations since emitted until measured, while for NO_2, NO and O_3, the photochemical reactions that take place since precursors are emitted are crucial to explain the measured levels. Furthermore, some pollutants (CO, NO_2, NO and O_3) are related to traffic emissions while others are not (SO_2). However, the performance of the models for the different pollutants, including persistence, is quite similar with a trend to a higher error the higher the value of K is.

Though certain trends in the type of model can be detected, the local conditions of the sensor seem to be the most important factor explaining why a certain type of model works best at a given location for a certain pollutant and value of K.

4 Conclusions

This work focuses on the prediction of hourly levels up to 8 hours ahead for five pollutants and six locations in the area of Bilbao. This represents the need to build 216 specific models to launch hourly forecasts of the forthcoming levels. To that end, historical records of the traffic, meteorological and air pollution networks corresponding to years 2000 and 2001 have been analyzed jointly and for each of the 216 predictions, 5 statistical models have been built. Their performance has also been compared with the simplest prediction, persistence of levels, according to four statistical indicators: R^2, *d, FA2* and *RMSE*. Depending on the pollutant, location and number of hours ahead the prediction is made, different types of models have been chosen. In the case of SO_2 and CO, in a great number of cases persistence of levels outperform linear models and also those based on neural networks. Predictions of NO_2, NO and O_3 hourly levels require in most cases linear models and only in a few cases MLP, RBF or GRNN based models. However, despite the different architecture of the models and also the explanatory mechanisms describing the emissions, transport and chemical transformations of each pollutant, the performance of the chosen models – including persistence – is very similar.

Therefore the chosen models for each case represent the maximum forecasting capability that can gain the network, regardless the type of model used for a given prediction.

Instead of this statistical approach, a second modeling strategies intended to describe every physical and chemical mechanism involved relating emissions and inmissions, have not been applied in the area, and comparison of results is not possible. However, in the last years, this second modeling strategy is being applied in urban areas at different time scales being their objective to describe future air quality scenarios rather than yield short-term predictions of a given pollutant.

For the time scale of this study, in the period since SO_2 and CO emissions take place until inmission levels are measured, these two pollutants do not experiment important chemical reactions. Therefore, persistence or linear models seem logical. In the case of NO_2, NO and O_3 complicated transport mechanisms and highly non-linear photochemical reactions relating traffic, meteorology and air pollution are known to be involved.

However, for these three pollutants and at the time scale of this study, in many cases simple linear models work as well as more complicated neural networks or slightly worse. This is something that has also been largely reported [16] and pointed out in previous works carried out in other urban areas. The explanation may be that the combination of the great number of highly non-linear mechanisms associated to photochemical smog result in the linearization of the overall effect.

The air pollution network in the area of Bilbao was originally designed as a diagnosis tool to follow in real time the evolution of several pollutants and it also measured some meteorological parameters. The traffic network was also intended to follow the evolution of traffic flow in the area of Bilbao. Bringing together the information from these networks it is possible to build statistical models that can yield short-term forecasts of air pollution levels. The performance of the models represent the boundaries in the prognostic capabilities of the network for the different pollutants measured in the area. Most air pollution networks are originally designed for control purposes and many times, meterological variables are also measured. If traffic data are measured jointly along with air pollution and meteorological values, it is possible to expand the capabilities of the original air pollution network if an integrated approach is done, shifting towards a joint management of the existing networks.

Though this group of models have been built for Bilbao, the same methodology can be applied to urban environments if like in this case, air pollution, meteorological and traffic data are available simultaneously. The models obtained can be quite easily built, run on a simple PC and can be incorporated into the daily network managing activities. The model needs to be updated every year and a good training level is required for the staff in charge of the air pollution network(s). Including in the network management activities an intensive and comprehensive program of data processing can expand the capabilities of a diagnosis air pollution network and provide the same with a range of prognosis capabilities. Many cities in the world are currently making serious efforts towards sustainability. The present approach can be incorporated into the overall management activities and strategies (like Agenda 21) towards a cleaner air and a better environment in many urban areas. Being air pollution in cities nowadays an issue of major concern, this approach can constitute the core of an alert system in real time. Its nature is modular and information from more sources of pollutants can be easily incorporated in the future.

Acknowledgements

This work is part of a research project financially supported by the University of the Basque Country UPV-EHU (Spain) under contract n# 1/UPV 00149.345-E-

15398/2003. The author wishes to thank the Environmental Department of the Basque Government and the Traffic Department of Bilbao Municipality for providing with data for this study.

References

[1] Ibarra-Berastegi G.; Elias, A.; Agirre, E.; Uria, J. Long-term changes in ozone and traffic in Bilbao; *Atmos. Environ.* **2001**, *35,* 5581-5592.
[2] Ibarra-Berastegi G.; Elias, A.; Agirre, E.; Uria, J. Traffic congestion and ozone precursor emissions in Bilbao (Spain); *Environ. Sci. & Pollut. Res.* **2003**, *10,* 361-367.
[3] Hornik, K.; Stinchcombe, M.; White, M. Multilayer feedforward networks are universal approximators; *Neural networks*, **1989,** *2*, 359-366.
[4] Masters, T. *Practical Neural Network Recipes in C++*; Academic Press, 1993.
[5] Bishop, C. *Neural Networks for pattern recognition*; Oxford University Press, 1995.
[6] Gardner, M.W.; Dorling, S.R. Statistical surface ozone models: an improved methodology to account for non-linear behaviour; *Atmos. Environ.* **2000***, 34*, 21-34.
[7] Haykin, S. *Neural Networks: A comprehensive foundation*; New York. McMillan Publishing, 1994.
[8] Statistica 7.0. User's manual. http://www.statsoft.com/ 2005.
[9] Ordieres, J.B.; Vergara, E.P.; Capuz, R.S.; Salzar, R.E.; Neural network prediction model for fine particulate matter $PM_{2.5}$ on the US-Mexico border in El Paso (Texas) and Ciudad Juarez (Chihuahua); *Environ. Model. & Softw.* **2005**, *20*, 547-559.
[10] Speckt, D.F. A Generalized Regression Neural Network. *IEEE Transactions on Neural Networks.* **1991***, 2 (6)*, 568-576.
[11] Patterson, D. *Artificial Neural Networks.* Singapore. Prentice Hall, 1996.
[12] Wilmott, C.J. On the validation of models; *Phys. Geogr.* **1981**, *2*, 184-194.
[13] Wilmott, C.J. Some comments on the evaluation of model performance; *Bull. Am. Meteo. Soc.* **1982**, *63, 11*, 1309-1313.
[14] Wilmott, C.J.; Ackleson, S.G.; Davis, R.E.; Feddema, J.J; Klink, K.M.; Legates, D.R.; O'Donnell, J.; Rowe, M.C. Statistics for the evaluation and comparison of models; *J. Geo. Res.* **1985**, *90*, 8995-9005.
[15] Hanna, S.R.; Strimaitis, D.G.; Chang, J.C. *User's guide for software for evaluating hazardous gas dispersion models.* 1991. American Petroleum Institute. 1220 L. Street, Northwest. Washington. D.C. 20005.
[16] Comrie, A. Comparing neural networks and regression models for ozone forecasting; J *Air & Waste Manage Assoc.* **1997**, *47,* 653-663.
[17] Ibarra-Berastegi, G; Madariaga, I; Agirre, E; Uria, J. Short-term real time forecasting of hourly ozone, NO2 and NO levels by means of multiple linear regression modelling; *Environ. Sci. & Pollut. Res.* **2001**, *8*, 250.
[18] Ibarra-Berastegi, G; Madariaga, I; Agirre, E; Uria, J. Short-term forecasting of ozone and NO_2 levels using traffic data in Bilbao (Spain). *Urban Transport IX*. pp. 235-242. WIT Press. Southampton. UK. 2003.

[19] Ibarra-Berastegi, G; Madariaga, I. Identification of joint targets for traffic and ozone in Bilbao (Spain)). *Air Pollution XII*. pp. 519-528. WIT Press. Southampton. UK. 2004.
[20] Ibarra-Berastegi, G; Madariaga, I. Impact assessment of Bilbao's metro. *Advances In City Transport: Case Studies* pp. 31-42. WIT Press. Southampton. UK. 2005.