Environ Monit Assess (2008) 143:131–146
DOI 10.1007/s10661-007-9963-0

Prediction models of CO, SPM and SO_2 concentrations in the Campo de Gibraltar Region, Spain: a multiple comparison strategy

Ignacio J. Turias · Francisco J. González · Ma. Luz Martin · Pedro L. Galindo

Received: 30 April 2007 / Accepted: 27 August 2007 / Published online: 11 October 2007

Abstract The 'Campo de Gibraltar' region is a very industrialized area where very few air pollution studies have been carried out. Up to date, no model has been developed in order to predict air pollutant levels in the different towns spread in the region. Carbon monoxide (CO), Sulphur dioxide (SO_2) and suspended particulate matter (SPM) series have been investigated (years 1999–2000–2001). Multilayer perceptron models (MLPs) with backpropagation learning rule have been used. A resampling strategy with two-fold crossvalidation allowed the statistical comparison of the different models considered in this study. Artificial neural networks (ANN) models were compared with Persistence and ARIMA models and also with models based on standard Multiple Linear Regression (MLR) over test sets with data that had not been used in the training stage. The models based on ANNs showed better capability of generalization than those based on MLR. The designed procedure of random resampling permits an adequate and robust multiple comparison of the tested models. Principal component analysis (PCA) is used to reduce the dimensionality of data and to transform exogenous variables into significant and independent components. Short-term predictions were better than medium-term predictions in the case of CO and SO_2 series. Conversely, medium-term predictions were better in the case of SPM concentrations. The predictions are significantly promising (e.g., $d_{\text{SPM 24-ahead}}=0.906$, $d_{\text{CO 1-ahead}}=0.891$, $d_{\text{SO2 1-ahead}}=0.851$).

Keywords Air pollution · Artificial neural networks · Backpropagation · Multiple comparison · Forecasting models

I. J. Turias · P. L. Galindo
Department of Computer Science – Research Group "Intelligent Systems",
Polytechnic Engineering School of Algeciras,
University of Cádiz,
Algeciras, Spain

F. J. González
Department of Applied Physics, Polytechnic Engineering School of Algeciras, University of Cádiz,
Algeciras, Spain

M. L. Martin
Department of Chemical Engineering and Environmental Technologies, Polytechnic Engineering
School of Algeciras, University of Cádiz,
Algeciras, Spain

I. J. Turias (✉)
Department of Computer Science, Escuela Politécnica Superior de Algeciras,
Avda. Ramón Puyol s/n,
11202 Algeciras, Cádiz, Spain
e-mail: ignacio.turias@uca.es

Introduction

The forecasting of air pollutant trends has received much attention in recent years. It is an important and popular topic in environmental science, as concerns

have been raised about the health impacts caused by unacceptable ambient air pollutant levels. Hence, the study of the influence and the trends relating to these pollutants are extremely significant to the public health and the image of the cities.

Predicting atmospheric pollutant concentrations in both urban and industrial areas is of great significance for decision-making. The present study considers the possibility of using neural techniques to identify models for atmospheric pollutant prediction. Specifically, carbon monoxide (CO), sulphur dioxide (SO_2) and suspended particulate matter (SPM) have been considered in different points of the area under study (Campo de Gibraltar region, at Andalusia, in the South of Spain). This is a very industrialized area with the highest emissions of these pollutants in the Andalusian region.

In urban areas, anthropogenic sources, including fossil fuel combustion, industrial activities, biomass burning and anthropogenic hydrocarbons, contribute far more to the concentration of CO than the natural sources. SO_2 is a prominent anthropogenic pollutant and contributes to the formation of sulphuric acid, the formation of sulphate aerosols and the deposition of sulphate and SO_2 at the ground surface. SPM (or TSP, total suspended particulates) are the general term used for a mixture of solid particles and liquid droplets found in the air. These particles, which come in a wide range of sizes, originate from many different stationary and mobile sources as well as natural sources. The danger represented by these air pollutants has been largely demonstrated from toxicological studies both for short and long exposition times (Schwartz et al. 1996; Wilson and Suh 1997; Goldberg et al. 2001).

The study presented here is essentially focused on the power of neural techniques for the identification of short-term and medium-term prediction models based on recorded time-series data. The potential of ANNs as a predictive tool has also been tested in this work. ANNs require no priori assumptions about the model in terms of mathematical relationships or data distribution. Neural networks have found many applications on time series prediction in the literature. The most widespread ANN design is a multilayer perceptron (MLP) with a learning procedure based on the backpropagation algorithm (Bishop 1995; Rumelhart et al. 1986).

In general, ANNs are currently recognized as state-of-the-art approach for statistical prediction of air quality. Nunnari et al. (1998) compared MLP models with a neuro-fuzzy approach and a traditional ARMAX model. The results confirmed the superiority of MLP predictors. The results obtained showed that neural techniques have a good capacity for modelling air pollution when they are compared with the traditionally used autoregressive prediction models. Gardner and Dorling (1999) showed that MLP-based models give better results compared with linear regression methods. Perez et al. (2000) compared the forecasting produced by three different methods: MLP, multiple linear regression (MLR) and persistence methods. They concluded that the MLP models achieved more accurate regression results and better predictions. Different modelling approaches for the forecasting of CO concentrations can be found in the literature. Pelliccioni and Poli (2000) study MLP-based models in the forecasting of the CO and NO_2 concentration levels in Rome's urban city centre using standard R-correlation coefficient to measure the performance. Viotti et al. (2002) proposed an approach based on ANN to forecast 1 h-ahead CO concentrations with good results.

Chelani et al. (2002) also considered ANN approach for the forecasting of short-term (1-h and daily) SO_2 average concentrations. Nunnari et al. (2004) compared ANN, Fuzzy Logic and other statistical approaches for SO_2 predictions.

Regarding Particulate Matter (PM) forecasting, Kukkonen et al. (2003) showed improved performance for the MLP models compared with linear and deterministic models. Other studies of PM predictions are given by Corani (2005) and Grivas and Chaloulakou (2006).

The main objective of the present work has been the comparison of the ability of ANN models and classical models (Persistence, ARIMA and MLR) to forecast air pollution. The inputs of each model were the pollutant concentrations and some additional exogenous variables in an autoregressive arrangement, while the time in the past (the depth of the model) was another variable parameter.

Another objective was the development of a multiple comparison scheme. Many model selection algorithms have been proposed in the literature (Zucchini 2000). The existing procedures can roughly be categorized as analytical or resampling based methods. Analytical approaches require certain assumptions of the underlying statistical model. Resampling based methods involve much more

computation, but they remove the risk of making faulty statements due to unsatisfied assumptions (Feelders and Verkooijen 1996). With the computer power currently available, this does not seem to be an obstacle. Although there is an active debate within the research community regarding the best method for comparison, statistical model selection is a reasonable approach (Mitchell 1997, Pizarro et al. 2002). A procedure of random resampling simulation was designed to avoid variation coming from different sources, thus independence and randomness were guaranteed (Pizarro et al. 2002). The results obtained were statistically analyzed and compared through analysis of variance (ANOVA) (Jobson 1991) and Bonferroni method (Hochberg and Tambane 1987).

In the following section, the database collection procedure used in this work will be formulated. Principal component analysis (PCA) is described in "Principal component analysis". The well-known Persistence, ARIMA and multiple linear regression models are briefly outlined in "Persistence models", "Box-Jenkins ARIMA models" and "Multiple linear regression models", respectively, and the ANN approach is presented in "Backpropagation neural networks models". "Experimental procedure" discusses the experimental procedure developed to assure the best generalization performance of the models. These models will be analyzed and compared by the way of statistical methods in "Results and discussion". Finally, the conclusions are shown in "Conclusions".

Study area and methods

Area description

About 300,000 inhabitants live in the different towns spread in the 'Campo de Gibraltar' (Fig. 1), the southern most region of Andalusia (Spain). It is a complex industrial scenario, where many stationary sources are present: an oil-refinery and some petrochemical factories close to it, a coal-fired power plant, a fuel-oil power plant, a large steel factory and a paper factory. Traffic is especially concentrated in the urban areas and the main road of the region (N-340) which surrounds the Bay of Algeciras. The port of Algeciras, one of the most important ship-trading ports in Europe, is another possible source of particulate and gaseous air pollution in the area. It is known that air pollution has direct effects on human health through exposure to high concentration of ambient pollutants. Then, air pollution control and the associated prediction of pollutant levels are needed to take preventive and evasive actions during episodes of high air pollution. The monitoring stations (triangles on the map of Fig. 1) used for the hourly measurements of SO_2, CO, nitrogen monoxide (NO), nitrogen dioxide (NO_2), ozone (O_3) and SPM concentrations are controlled by the Enviromental Agency of the Andalusian Government. AL (within the town of Algeciras, with about 130,000 inhabitants) and LL (within the town of La Linea, with about 70,000 inhabitants) stations are located on a relatively flat terrain of urban areas, with many buildings around. The altitudes of AL and LL are about 20 and 2 m (a.s.l.), respectively. The nearest obstacles are about 5 m. CA (within Campamento urban area, with about 2,000 inhabitants) station is located on a flat terrain [about 8 m of altitude (a.s.l.)] of suburban areas which are close to the industrial sites.

SPM levels are measured by automatic beta radiation attenuation monitors. Gaseous pollutants are monitored by chemical analyzers. Specifically, SO_2 and CO concentrations are measured by ultraviolet (UV) fluorescence and infrared absorption, respectively. SO_2, NO, NO_2 and SPM are measured at all the stations, while CO and O_3 are only measured at AL and LL stations, respectively. Detailed information about the methodology followed for the analysis of SO2, CO and NOx concentrations can be found elsewhere (Hofzumahaus et al. 2006). The calibration process of all the sampling monitors is supervised by the Enviromental Agency of the Andalusian Government.

Data collection

During the period of analysis (1999–2001), the variables that showed the highest percentage of validated hourly data, and finally considered in this study, were: CO at AL site (95%), SO_2 at AL and LL (98 and 99%, respectively) and SPM at CA (93%). Therefore, artificial data were a very small part of the whole data. Missing values were replaced by linear interpolation (Junninen et al. 2004). Pollutant concentrations were measured in $\mu g/m^3$.

The choice of the exogenous variables was done considering a previous study of correlation between

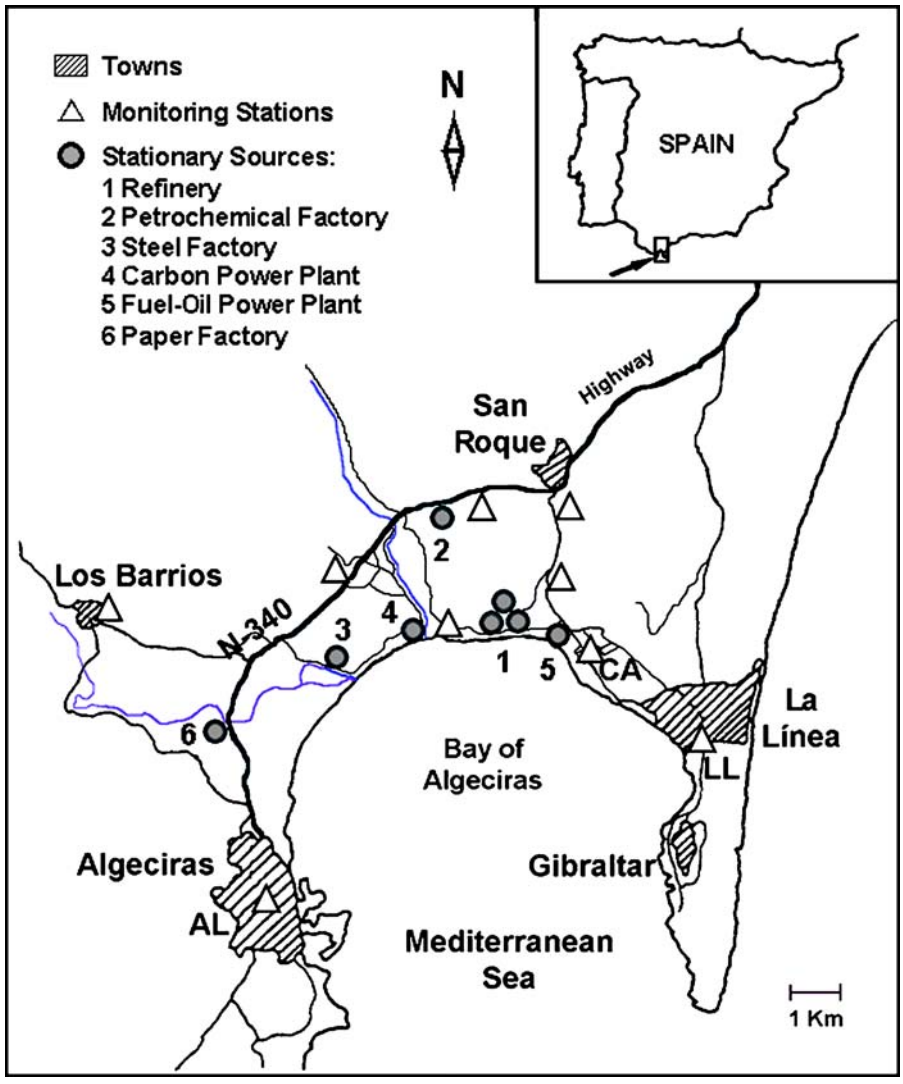

Fig. 1 Location of the towns, large factories and the monitoring stations in the 'Campo de Gibraltar' region (Spain)

all the air pollutant and meteorological variables measured at the monitoring stations (Gonzalez 2004). As an example, daily mean values of CO showed a statistically significant (at a 99% level) negative correlation with air temperature and wind speed, and a positive correlation with wind direction. Thus, for the prediction of CO concentrations at AL site, the following exogenous variables have been considered: three meteorological variables (wind direction, wind speed and air temperature) and NO concentrations measured at the AL station. It is interesting to note that CO and NO concentrations have a significant positive correlation (R^2=0.65). This indicates that both pollutants are emitted by the same source, that is, urban traffic. Therefore, the knowledge of NO concentrations could be helpful to the prediction process. The former three meteorological variables (wind direction, wind speed and air temperature) were also used as exogenous variables for the SO_2 and SPM predictions at AL and LL sites. NO, as a measure of traffic emissions, was also considered for SO_2 prediction at AL, while O_3 concentrations were used for SO_2 prediction at the LL station. The relationship between SO_2 and O_3 is based on atmospheric chemistry. Thus, in the presence of O_3 under wet conditions, the SO_2 dry-deposition velocity and oxidation from SO_2 to sulphate is enhanced (Sakamoto et al. 2004). Furthermore, in the presence of H_2O_2 and wet aerosols, SO_2 does participate in the chemistry of ozone by the absorp-

tion of light within the ultraviolet region (Gupta et al. 1986, Ruiz Suárez et al. 1995). Due to the high solar radiation and relative humidity in the area under investigation, the two former phenomena are likely to be strengthened.

It is interesting to observe the transformation considered to avoid the discontinuity wind direction variable could cause. The experiments presented here use the same expression as Ziomas et al. (1995).

The database of samples used to teach the forecasting models was arranged in the form of autoregressive data (see Fig. 2), where n was the width of the observation window in the past (lags). This information was used to make the prediction, as it is done in autoregressive models. nh was the time when pollutant concentration is predicted. That is, nh=24 means 24-ahead average prediction. The predictions made in this work were 1-ahead (short term) and 24-ahead (medium term), for which different prediction models (Persistence, ARIMA, MLR, ANN) and autoregressive inputs (with different lags and exogenous variables) have been combined (Table 1). The different numbers of lags used are in agreement with the autocorrelation coefficients computed onto the pollutant data series (Figs. 3 and 4) as in the work of Pelliccioni and Poli (2000). The randomised resampling procedure designed (see "Experimental procedure") permits the multiple comparison of models and the selection of the best prediction model (set of method-topology-lags) in each case.

In medium term predictions mean pollutant concentrations were used. Therefore, the daily averaged values were predicted from mean autoregressive information. Consequently, the number of samples of the database was 365 days/year×3 years=1,095 samples, in the case of daily mean forecasting, and

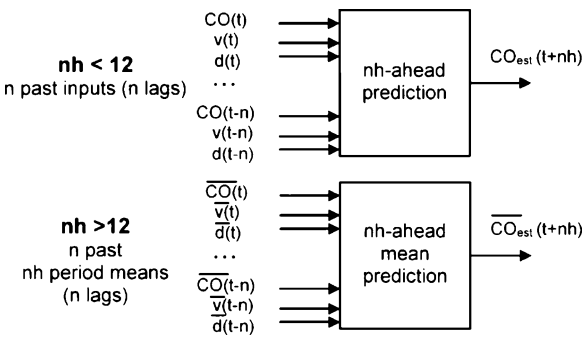

Fig. 2 Autoregressive scheme for CO forecasting

Table 1 Parameters of the different sets of models tested in the experiments

	Exogenous information	Number of lags	Number of hidden units
SET1	No	1–2–4–8–12–24	1–3–5–10–15–20–25–30–35–40–45–50
SET2	Yes	1–2–4–8–12–24	1–3–5–10–15–20–25–30–35–40–45–50
SET3	PCA	1–2–4–8–12–24	1–3–5–10–15–20–25–30–35–40–45–50

365 days/year×24 h/day×3 years=26,280 samples, in the case of hourly forecasting database.

Theoretical background

Principal component analysis

In order to avoid the tendency to sparseness, it was necessary to consider a feature selection procedure. In this paper, the PCA technique has been used. The objective of Principal Component Analysis (PCA) is to reduce the dimension, preserving as much of the relevant information as possible, finding out those directions which maximise the variance. The transformation maps vectors $\boldsymbol{x}^n$ in a d-dimensional space onto vectors $\boldsymbol{z}^n$ in another M-dimensional space, where $M<d$. The set of n patterns can be represented as a linear combination of the original d orthonormal vectors u_i (Jollife 1986). The minimum error can be obtained by choosing the $d-M$ smallest eigenvalues of covariance matrix of the set of data vectors. The new components are linear combination of the original features.

Persistence models

The persistence model is an extremely simple model, with no adjustable parameter. Due to its simplicity, it represents the minimum acceptable quality out of any other model proposed. Two different models have been considered: PER-1 h and PER-24 h.

- PER-1 h: It accepts that the concentration levels of a pollutant at a particular time of day correspond to the value which occurred the hour before (i.e. the

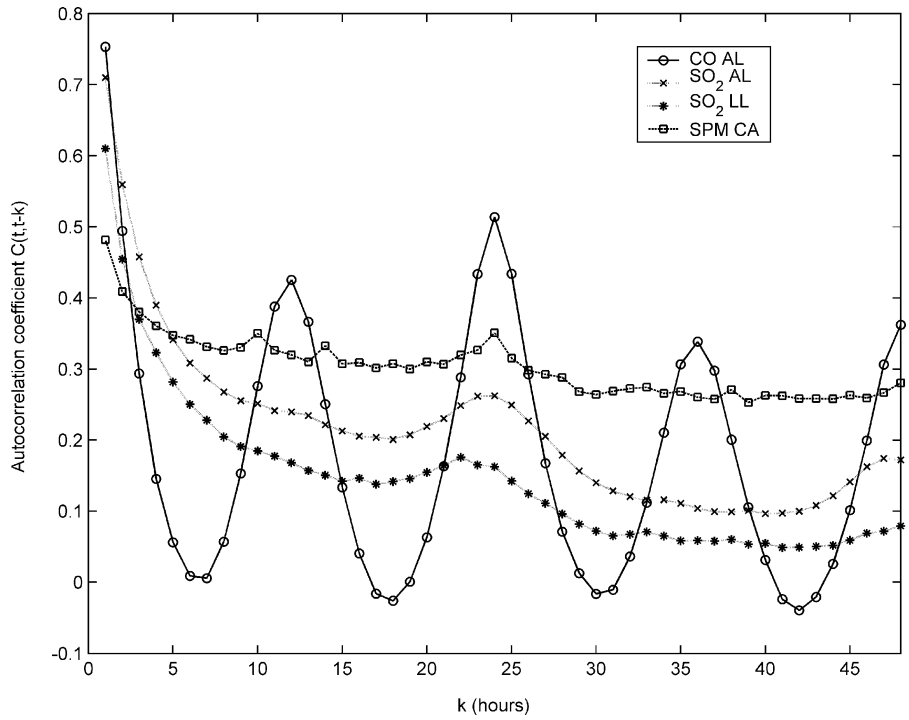

Fig. 3 Autocorrelation coefficients for the 1 h-sampled data series

predicted output at 7:00 A.M. correspond to the value of the pollutant at 6:00 A.M.).

- PER-24 h: It accepts that the concentration levels of a pollutant at a particular time of day correspond to the value which occurred the day before at the same hour. In the case of mean daily concentration prediction this model has no sense.

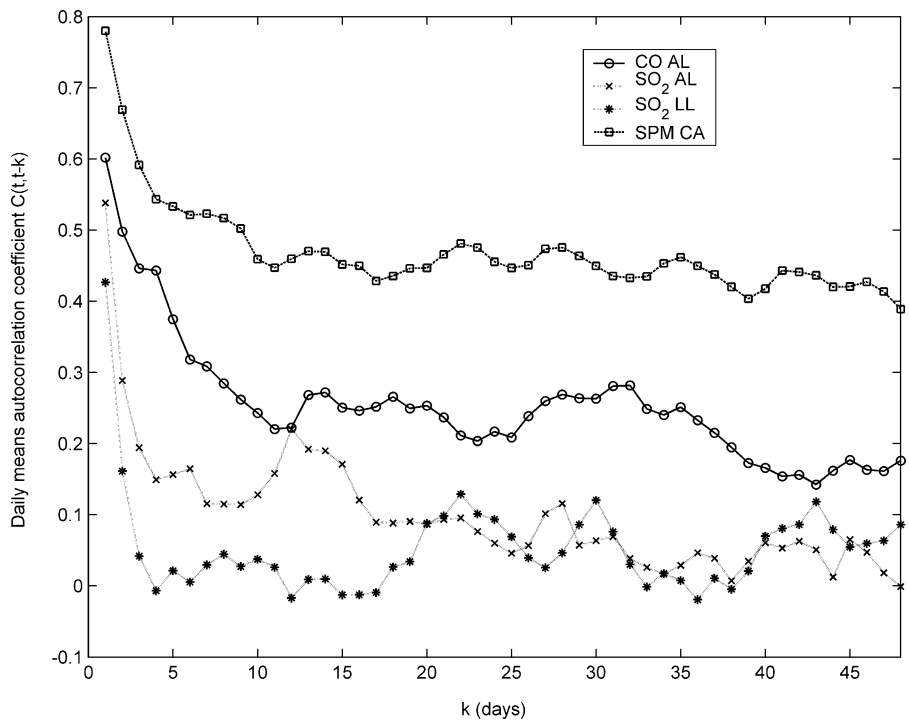

Fig. 4 Daily mean autocorrelation coefficients for the 24-h mean data series

Box-Jenkins ARIMA models

ARIMA linear models are described here briefly. In general, a nonseasonal time series, $x(t)$, $t=1\ldots n$, (n being the number of observations) of air pollutant concentrations measured at an equal time intervals, can be modelled as a combination of past values and past errors as:

$$x(t) = a_1 x(t-1) + a_2 x(t-2) + \ldots + a_p x(t-p) + e(t) - b_1 e(t-1) - b_2 e(t-2) - \ldots - b_q e(t-q) \quad (1)$$

where a and b are the vector of coefficients, p and q are the order of the autoregressive and moving average polynomials, respectively. The further details to estimate the parameters and order of the model are given in Box and Jenkins (1970).

Multiple linear regression models

The purpose of MLR is to establish a quantitative relationship between a group of predictor or independent variables, X, and a response, y (in this case, the pollutant concentration to be predicted). Using matrix notation, the linear model can be expressed as $y = X\beta + \varepsilon$(where ε is a vector of random disturbances). The solution to the regression problem is a vector, b, which estimates the unknown vector of parameters (β) and can be computed directly by inverting the matrix product $X'X$. In most of the real situations, this approach is really dangerous because X will be often either totally singular or ill-conditioned. There are different well-known methods to solve a MLR problem (Gauss–Jordan Elimination, LU decomposition, or QR decomposition). However, the most effective method is singular value decomposition (SVD), which is able to handle any problem that may arise, like singularities or ill-condition. The foundations of these algorithms are beyond the scope of this paper. Good references are Press et al. (1992) and Masters (1995). In the study presented here, SVD method has been used to solve the MLR problem.

Backpropagation neural networks models

Neural networks can extract the link between the input data and the corresponding output data. Thus, ANNs can be used to solve different problems of regression or classification (Bishop 1995), prediction and, more generally, black-box identification, in which 'a priori' knowledge of the model is not needed (Fu 1994). For such supervised networks, a prediction pattern is formed by inputs related to the past together with the pollutant concentration to be forecasted, named real or desired output. The ANN can be considered as a non-linear transformation which maps a set of input variables through several layers of processing elements or neurons (with activation functions) into a set of output variables. The form of the activation functions (typically sigmoid or hyperbolic functions) allows the approximation of complex non-linear functions and must be differentiable since backpropagation algorithm derives its name from the fact that error signals are propagated backwards through the network on a layer-by-layer basis. The learning process is stopped when a specified error goal or a number of epochs (a presentation of all the patterns is usually called epoch or cycle) is reached. There are many algorithms (Bishop 1995; Masters 1995) for training feedforward neural networks: conjugate gradients, quasi-Newton, Levenberg–Marquardt and others. Standard backpropagation training algorithms are often too slow for practical problems. The Levenberg–Marquardt algorithm (Hagan and Menhaj 1994) seems to be the fastest method for moderate-sized networks. It updates network weights following an iterative procedure to approximate the Hessian matrix with the Jacobian (Press et al. 1992). Levenberg–Marquardt algorithm operates in batch mode where the weights of the network are updated only after the entire training set has been applied to the network. The gradients calculated at each training example are added together to determine the change in the weights. Hagan et al. (1996) gives a complete discussion of the batch training with the backpropagation algorithm. In the study presented here, Levenberg–Marquardt method has been used to train the networks.

There is no way to determine the optimum topology of an ANN, although Kolmogorov's theorem (Fu 1994; Bishop 1995) and Vapnik–Chervonenkis (VCdim) dimension (Vapnik and Chervonenkis 1971) show the capabilities of backpropagation-based MLP feedforward networks. A simple choice would be to train many networks with a different number of

hidden units and layers, to estimate the generalization error for each one, and to select the network with the smallest error. However, it critically depends on the training/test sets and the initial weights. Therefore, it is necessary to compute the mean generalization error over a designed resampling experiment (Cobb 1998) as the one explained in the next section.

Experimental procedure

ANN models with different hidden units were compared to determine the impact of the addition of non-linear processing capabilities on model performance. A resampling procedure was found to reduce test set prediction error and to mitigate the effects of overfitting (Pizarro et al. 2002). The ANN models were also compared with those based on Persistence, ARIMA and MLR approaches.

Dietterich (1998) studied different statistical tests for comparing supervised learning algorithms and the sources of variation that a good statistical test should control. Ideally, the population is considered to have an infinite number of samples. However, in real situations, the amount of data available is only a subset of the overall population. For a finite set of data, these sources of variation should be controlled as follows (Pizarro et al. 2002):

- The learning algorithms should be executed multiple times over different training and test sets to control the variation due to the choice of training and test data sets.
- If any model is trained and tested on a given training and test data set, any other model should be trained and tested with the same set. This ensures that all models are compared under the same conditions. It also helps to control the variations due to the choice of training and test data sets, and allows us to apply statistical pairwise tests.
- Each unstable algorithm should be executed several times, taking different starting states for each training data set to reduce the variance due to internal randomness.

There are many methods to estimate the generalization performance of a model (hold-out, cross-validation, leave-one-out, penalization strategies...). The complete strategy used in this paper repeats 30 times a similar process: random splitting of data into a pair of equal sized portions (training and test sets) and two-fold crossvalidation for the estimation of generalization error using three indices: the standard correlation coefficient (R), the index of agreement (d) and the mean squared error (MSE), of each different prediction model (Pizarro et al. 2002). The parameters of each model were estimated using one of the groups (the training set) while its performance was tested using the remaining one (the test set). Therefore, the performance was measured over test data not used in the training or design of the model. This process was repeated twice each time, swapping the sets and averaging the results over these two runs. For a given training and test set, each algorithm is trained for 10 times (to avoid internal randomness). Therefore, the performance of each model is computed from a set of 600 ($30\times10\times2$) samples.

In the case of the neural approach, the tested models had only one hidden layer with different numbers of hidden units (Table 1), while the number of epochs were fixed at a maximum of 100. Levenberg–Marquardt method was used as a minimization strategy. The resampling procedure was designed in order to compare different models and to determine (by analysing the mean and the variance) if differences between them were statistically significant. Obviously, the main goal was to find out which models were better on average. While ANOVA is a good method to do an analysis of variance, it does not pinpoint where the significant differences lie. Bonferroni method is a well-known and easy to apply follow-up ANOVA f-test. It is a multiple comparison procedure for model selection that adjusts the observed significance level based on the number of pair comparisons. Two groups are not significantly different if the difference of their means follows Eq. 2 where M is the number of models, n_i is the number of data for model i, $\overline{y_i}, \overline{y_j}$ are the means for the models i and j, t is the Student pdf with n-M degrees of freedom, c is the Bonferroni correction, α is the statistical significance (0.1 has been used in the experiments), and S_{UV} is the unexplained variance due to the existence of different models.

$$\left|\overline{y_i}-\overline{y_j}\right| \le t_\alpha \Big/ 2\cdot C\cdot S_{UV}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} \qquad \mathrm{i,j}=1\ldots\mathrm{M} \tag{2}$$

Results and discussion

The simulations were run in MATLAB environment. Persistence models are considered as a starting point and baseline against which to compare forecasts generated from the other methods. For ARIMA modelling, the order of the model is selected by plotting the autocorrelation function (ACF) and partial autocorrelation functions (PACF). For all of the pollutants an autoregressive model of order 2, i.e. ARIMA(2,0,0) is found to be appropriate. With this model order, autoregressive model was fitted to the learning data. The model parameters were obtained by adopting the Box–Jenkins methodology. Table 2 shows the results compared with multiple linear regression (MLR) and persistence models.

In the case of CO predictions, PER-1 and ARIMA models perform better than MLR model. Therefore, it seems that exogenous variables can not explain very well the variability of predictions. A very high periodicity of the CO series even with the PER-24 model (d=0.71) is observed. Table 2 also shows the results obtained for SO_2 series in two different monitoring stations (AL y LL). The result of the PER-24 model is significantly smaller than that for CO series. The highest persistence is observed for the PER-1 model, but still smaller than the result for CO series. Considering the results of the different moni-

Table 2 Results obtained for 1 h-ahead and 24 h-ahead (daily) CO, SO_2 and SPM predictions with Persistence (PER), ARIMA and MLR models in the 30-times resampling procedure

Pollutant	nh-ahead	Method	R[a]	d[b]	MSE[c] (μg/m^3)
CO AL	1	PER-1	0.74	0.85	1.80E+05
CO AL	1	PER-24	0.51	0.71	3.54E+05
CO AL	1	ARIMA(2,0,0)	0.74	0.85	2.02E+05
CO AL	1	MLR	0.68	0.73	2.55E+05
CO AL	24	PER-1	0.61	0.80	5.58E+04
CO AL	24	ARIMA(2,0,0)	0.65	0.82	2.77E+04
CO AL	24	MLR	0.65	0.76	2.66E+04
SO_2 AL	1	PER-1	0.71	0.85	76.69
SO_2 AL	1	PER-24	0.26	0.48	217.56
SO_2 AL	1	ARIMA(2,0,0)	0.73	0.84	64.55
SO_2 AL	1	MLR	0.73	0.83	66.19
SO_2 LL	1	PER-1	0.61	0.74	269.25
SO_2 LL	1	PER-24	0.16	0.41	609.28
SO_2 LL	1	ARIMA(2,0,0)	0.63	0.74	210.86
SO_2 LL	1	MLR	0.64	0.75	213.49
SO_2 AL	24	PER-1	0.54	0.73	50.16
SO_2 AL	24	ARIMA(2,0,0)	0.52	0.71	15.45
SO_2 AL	24	MLR	0.65	0.75	16.03
SO_2 LL	24	PER-1	0.42	0.66	125.51
SO_2 LL	24	ARIMA(2,0,0)	0.39	0.63	42.01
SO_2 LL	24	MLR	0.49	0.63	35.54
SPM CA	1	PER-1	0.48	0.68	376.62
SPM CA	1	PER-24	0.33	0.57	471.85
SPM CA	1	ARIMA(2,0,0)	0.52	0.69	267.20
SPM CA	1	MLR	0.49	0.59	277.35
SPM CA	24	PER-1	0.80	0.89	54.19
SPM CA	24	ARIMA(2,0,0)	0.78	0.88	15.17
SPM CA	24	MLR	0.81	0.90	21.00

[a] Correlation coefficient

[b] Index of agreement

[c] Mean square error (a, b and c computed with test patterns)

toring stations it is worth mentioning that LL station exhibits significantly smaller persistence values than AL station. The results of SPM series show something unusual: the daily persistence is higher (d=0.89) than the hourly one. Daily CO and SO_2 time series have a higher variation than SPM series. Then, SPM sources of variation are likely to be more persistent than those of CO and SO_2. It seems that SPM has a longer-term background level probably due to soil dust resuspension, enhanced by the high solar radiation, African dust intrusions (Rodriguez et al. 2001) and construction activities in the area under study.

Results in Table 2 show that the performance of ARIMA and MLR models are quite similar. Obviously, the values of 24 h-ahead mean daily predictions are smaller than 1 h-ahead predictions (except in the case of SPM). ARIMA performs slightly better than MLR model for CO and SPM 1 h-ahead predictions. Results for SO_2 are very closed. Furthermore, MSE index seems to fluctuate more than R and d indices, specially in the case nh=24.

ARIMA models are quite flexible as they can represent several different types of time series, but their major limitation as well as MLR models is the pre-assumed linear form of the model. The approximation of linear models to real-world time series is not always satisfactory.

Table 3 shows the results of the best ANN models for each pollutant and time ahead. The selection of these best models was made by using the Bonferroni method within the three sets of models Table 1 shows. For each pollutant and time ahead, the results of three different models are shown: the best model without exogenous data (SET1), the best model with exogenous information (SET2) and the best model using PCA transformation of exogenous variables (SET3). The selection of the best model was done after training different ANN topologies and by using the Bonferroni criterion with the aid of three indices computed for test sets: the standard correlation coefficient (R), the index of agreement (d) and the mean squared error (MSE). The indices for each model were calculated as the average of the individual prediction coefficients in the experiments designed by using the above explained random resampling technique. This guarantees independency in the results and prevents the appearance of unexpected variation sources.

PCA is used in all cases to retain at least 92% of the total data variance. This way, dimensionality is reduced to two principal components which are linear combinations of the input exogenous variables. Nevertheless, the use of PCA does not guarantee the best prediction outcomes although prevents "the curse of dimensionality".

ANN models have in general (CO, SO_2 and SPM) better performance than the above classical (PER, ARIMA and MLR) models. The major advantage of ANN techniques over ARIMA or MLR models is their ability to take into account the nonlinear dynamics involved in the time series. The information about the linearity or nonlinearity of the time series is, however, not available in advance. Then, multiple comparison methods must be applied over the set of linear and nonlinear models and finally select the one which provides the most accurate results. In the experiments analyzed, the ANN-based models were able to make better predictions than those based on MLR approach. The indices between observed and predicted values for the samples in the test sets were higher for the neural models than for the standard regression ones. MLR models show values of the indices similar to those based on ANN approach in the training stage. However, when these MLR models were tested with data which were not included in the designing or training set, the results were worse. The error on the training set is driven to a very small value, but when new data is presented to the model the error is greater. This implies that the model has memorised the training examples but it has not learnt how to generalise new situations.

For some cases where ANNs perform worse than linear models, the reason may simply be that the data is linear without much disturbance. We can not expect ANNs to do better than linear models for linear relationships. Therefore, the neural networks models can behave as or worse than linear models do. In the present study, one of the best models selected (SPM at CA, see Table 4) is almost linear (with only one hidden unit).

Table 4 shows the global best models selected by Bonferroni method over the three sets of models (Table 1) and a majority voting scheme (MVS; Geok and Singh 1998). For example, in the case of 24 h-ahead prediction of SO_2 in Algeciras, the best model without using exogenous information has the performance indices 0.676, 0.771 and 15.830, respectively, while the best model using exogenous information and PCA leads to 0.680, 0.780 and 15.970. R and d

Table 3 Best models selected using the Bonferroni method in the 30-times resampling procedure (with/without using exogenous information and also using PCA) for the CO, SO_2 and SPM predictions

Pollutant	nh ahead	n lags	Ex. info.	ANN top[a]	R[b]	d[c]	MSE[d] ($\mu g/m^3$)
CO AL	1	24	No	50	0.809	0.891	1.30E+05
CO AL	1	2	Yes	20	0.786	0.869	1.39E+05
CO AL	1	24	PCA	20	0.816	0.891	1.22E+05
CO AL	24	1	No	10	0.662	0.820	2.52E+04
CO AL	24	1	Yes	10	0.570	0.729	3.11E+04
CO AL	24	1	PCA	10	0.600	0.730	2.88E+04
SO_2 AL	1	1	No	10	0.759	0.853	61.700
SO_2 AL	1	2	Yes	25	0.758	0.851	65.802
SO_2 AL	1	2	PCA	10	0.772	0.851	60.241
SO_2 LL	1	1	No	3	0.659	0.770	208.440
SO_2 LL	1	1	Yes	10	0.655	0.762	209.890
SO_2 LL	1	1	PCA	1	0.658	0.759	207.880
SO_2 AL	24	1	No	5	0.676	0.771	15.830
SO_2 AL	24	1	Yes	15	0.660	0.770	16.240
SO_2 AL	24	1	PCA	5	0.680	0.780	15.970
SO_2 LL	24	1	No	1	0.504	0.632	35.320
SO_2 LL	24	1	Yes	5	0.502	0.649	36.420
SO_2 LL	24	2	PCA	5	0.519	0.652	34.901
SPM CA	1	12	No	15	0.595	0.726	245.70
SPM CA	1	8	Yes	30	0.546	0.654	254.92
SPM CA	1	24	PCA	5	0.582	0.697	240.13
SPM CA	24	1	No	1	0.834	0.906	21.13
SPM CA	24	2	Yes	5	0.823	0.899	21.88
SPM CA	24	1	PCA	1	0.831	0.906	20.95

[a] ANN Topology (hidden units)

[b] Mean correlation coefficient

[c] Index of agreement

[d] Mean square error (b, c and d were computed with test patterns)

are better in the second model, while MSE is better in the first one. Thus, the model selected using MVS was the second.

The results show that most of the best models selected uses exogenous variables transformed via PCA. This is the case for CO and SO_2 at AL station (nh=1), SO_2 at AL and LL and SPM at CA (nh=24). The rest of the best models selected [SO_2 at LL and SPM at CA (nh=1) and CO at AL (nh=24)] do not use exogenous information.

The reason of these results can be explained considering the persistence measured at both stations (greater at AL than at LL). The greater persistence at AL station is probably due to a lower level and variability of SO_2 concentrations than at LL station. LL station is closer to stationary sources of SO_2 (a refinary and some petrochemical factories). Consequently, SO_2 concentration values exhibit a higher variability and a lower persistence which affects MLR and ANN models directly.

Results for test patterns are represented graphically in Fig. 5 (R-correlation coefficient observed-predicted data). Only SPM 24 h-ahead and CO 1 h-ahead have been shown in order to illustrate the predictions. Similar results were obtained for the rest of pollutants (Table 4). Figure 6a and b show the results for the predictions of CO 1 h-ahead and SPM 24 h-ahead forecasting results during two periods of high concentration values.

The values of CO forecasting presented in this study are better than those obtained elsewhere. Thus, for 1 h-ahead and 1 h average concentrations, Nagendra and Khare (2004) give a d value of 0.78, while Comrie and Diem (1999) give values ranging from 0.60 to 0.84. In the case of SO_2 at AL and LL stations the values of d are 0.780 and 0.652,

Table 4 Best global models selected using the Bonferroni method in the 30-times resampling procedure and a Mayority Voting Scheme (MVS) using the R, d and MSE indices

Pollutant	nh ahead	n lags	Ex. info[a]	ANN top[b]	R^c	d^d	MSE[e] ($\mu g/m^3$)
					2-CV	2-CV	2-CV
					10-CV	10-CV	10-CV
CO AL	1	24	PCA	20	0.816	0.891	1.22E+05
					0.779	0.894	1.24E+05
SO_2 AL	1	2	PCA	10	0.772	0.851	60.241
					0.783	0.850	60.030
SO_2 LL	1	1	No	3	0.659	0.770	208.440
					0.663	0.770	208.320
SPM CA	1	12	No	15	0.595	0.726	245.700
					0.599	0.727	246.920
CO AL	24	1	No	10	0.662	0.820	2.52E+04
					0.665	0.821	2.42E+04
SO_2 AL	24	1	PCA	5	0.680	0.780	15.970
					0.689	0.782	15.270
SO_2 LL	24	2	PCA	5	0.519	0.652	34.900
					0.515	0.654	38.250
SPM CA	24	1	PCA	1	0.834	0.906	20.950
					0.841	0.907	19.880

[a] Exogenous information

[b] ANN Topology (hidden units)

[c] Mean correlation coefficient

[d] Mean index of agreement

[e] Mean square error

respectively, which are in the range found in the literature. Nunnari et al. (2004) provide values of d ranging from 0.64 to 0.86, while Yildirim and Bayramoglu (2006) from 0.47 to 0.82. Particulate matter 1-ahead forecasting *d*-value is slightly smaller than those (*d* from 0.56 to 0.90) obtained by Grivas and Chaloulakou (2006). Regarding PM daily average forecasting at CA station (d=0.906), it can be pointed out that it seems to be better than those (0.53–0.78) reported by Yildirim and Bayramoglu (2006).

Table 4 also shows the values obtained for the three indices when using 2-CV or 10-CV. It can be seen that *d* is the most robust index with slight changes among the two cross validation schemes. R and MSE indices have higher changes, specially in the case of daily forecasting. The results from the 2-CV seem to be slightly more pessimistic than those from the 10-CV, as a higher number of training patterns are used.

The Andalusian Government has recently developed an environmental regional plan called 'CIUDAD 21', with different objectives on management, control and improvement of air quality at Andalusian urban areas. Then, the method proposed here could be an interesting tool for the environmental authorities to control air pollution and take the necessary measures in advance.

Conclusions

In the present work, the most convincing advantage of neural models is that the capability of generalization over the test data is higher than the one obtained by the other methods tested (Persistence, ARIMA and MLR). The use of autoregressive information with $n>1$ lag does not seem to improve significantly the capability of most of the models. The following concluding remarks can be made from the results discussed above. Neural networks are useful for modelling due to their ability to be trained using historical data and their capability for modelling non-linear relationships. Furthermore, PCA method prevents problems arising from "the curse of dimensionality" and although there is a loss of information, the overall performance of some of the selected models has increased. The random resam-

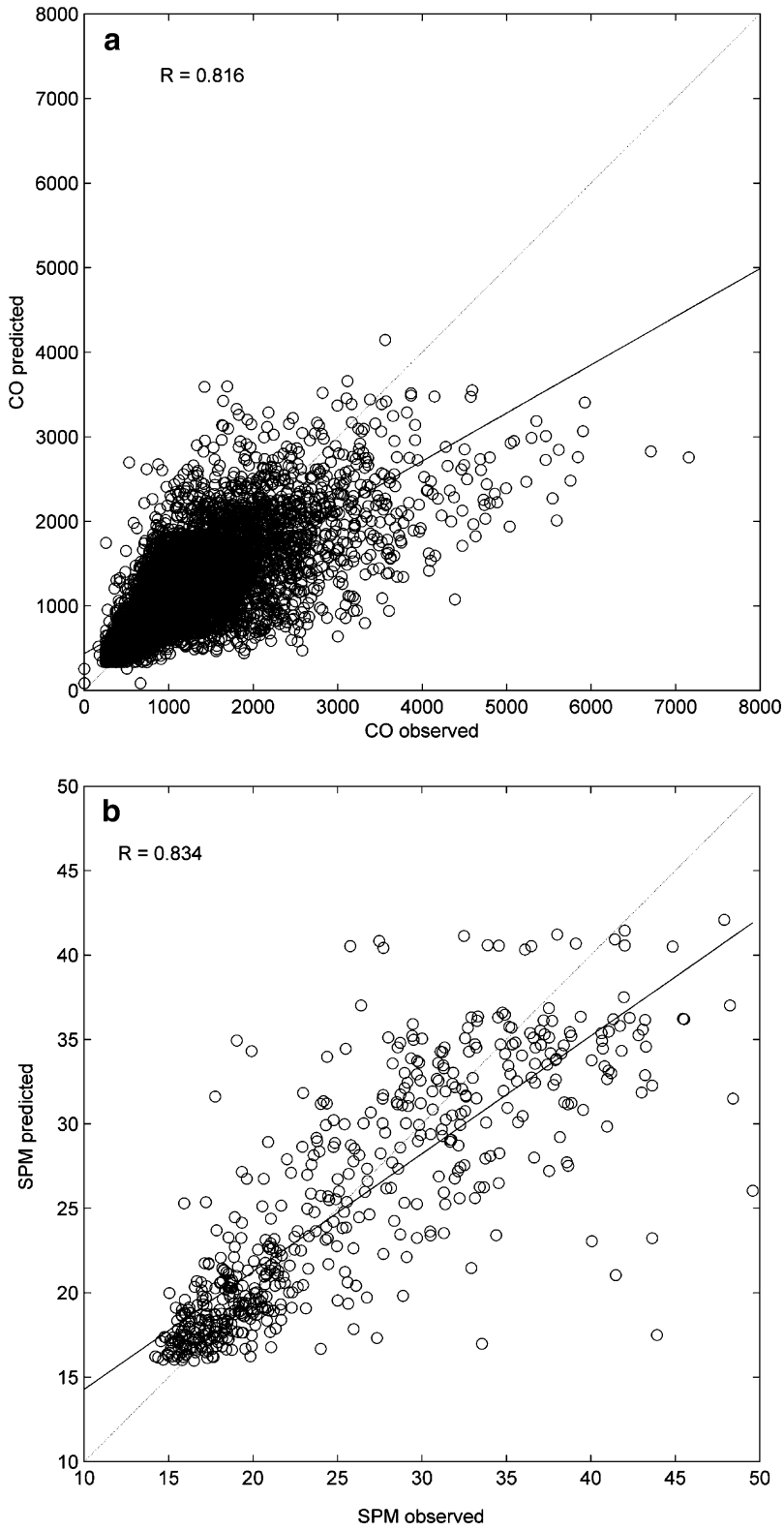

Fig. 5 Scatterplot of observed vs. predicted data and *R* correlation coefficients computed with one of the best neural models for the **a** CO 1 h-ahead forecasting and **b** SPM 24 h-ahead forecasting

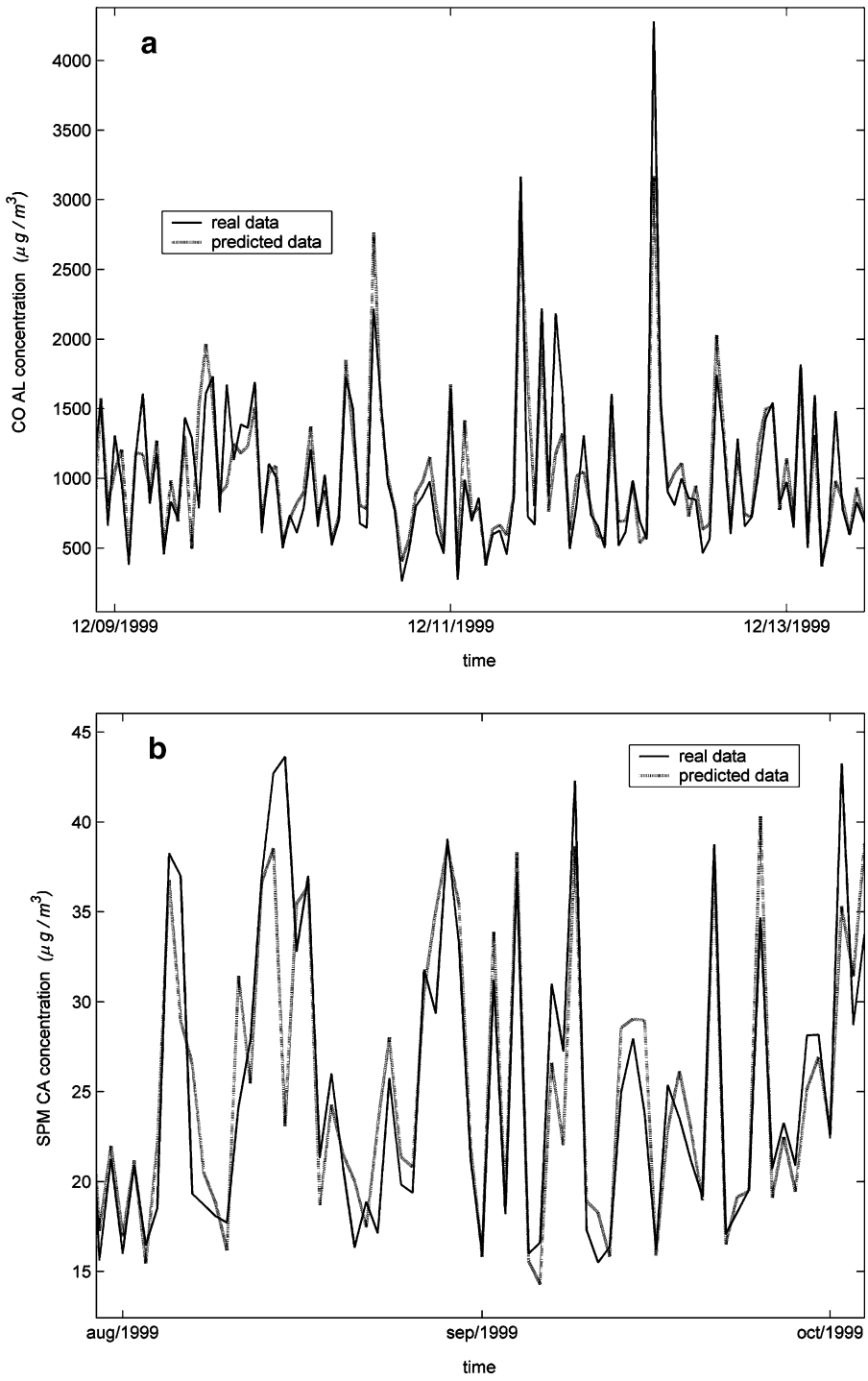

Fig. 6 Observed and predicted data for the **a** CO 1 h-ahead forecasting and **b** SPM 24 h-ahead forecasting during periods of high concentration values

pling procedure designed in this paper assures the accuracy of the results obtained and the certainty that the selected model in each case is the most suitable. Thus, this experimental procedure scheme can be used, together with ANOVA test and/or Bonferroni method, in order to perform a statistical comparison of tested models. Finally, the proposed multiple comparison methodology can be extended for other environmental applications.

Acknowledgements This research is supported in part by a grant from the Andalusian Government through P.A.I (Research Group TIC-145). Special thanks are due to Mrs. Ma Mar González who revised the English version of this paper. Finally, the authors are very grateful for the helpful comments and suggestions offered by reviewers.

References

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis, forecasting and control* pp. 131–164. San Francisco, CA: Holden-Day 137.

Chelani, A. B., Chalapati Rao, C. V., Phadke, K. M., & Hasan, M. Z. (2002). Prediction of sulphur dioxide concentration using artificial neural networks. *Environmental Modelling and Software*, *17*, 161–168.

Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer.

Comrie, A. C., & Diem, J. E. (1999). Climatology and forecast modelling of ambient carbon monoxide in Phoenix, Arizona. *Atmospheric Environment, 33,* 5023–5036.

Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecollogical Modelling*, *185*, 513–529.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, *7*(10), 1895–1923.

Feelders, A., & Verkooijen, W. (1996). On the statistical comparison of inductive learning methods. *Learning from data*, *Artificial intelligence and statistics V, lecture notes in statistics*, *vol. 112* (pp. 271–279). Berlin: Springer.

Fu, L. (1994). *Neural networks in computer intelligence*. New York: McGraw-Hill.

Gardner, M. W., & Dorling, S. R. (1999). Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment*, *33*, 709–719.

Geok, S. N., & Singh, H. (1998). Democracy in pattern classifications: combinations of votes from various pattern classifiers. *Artificial Intelligence in Engineering*, *12*, 189–204.

Goldberg, M. S., Burnett, R. T., Bailar, J. C., Brook, J., Bonvalot, Y., & Tamblyn, R., et al. (2001). The association between daily mortality and ambient air particle pollution in Montreal, Quebec. *Environmental Research*, *A86*, 12–25.

Gonzalez, F. J. (2004). Characterization and source apportionment to atmospheric particulate concentrations in the Campo de Gibraltar region. PhD. thesis. University of Cádiz.

Grivas, G., & 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.

Gupta, G., Sabaratnam, S., & Dadson, R. (1986). Linear regression analyses of ozone and sulphur dioxide in ambient air. *Science of the Total Environment*, *50*, 209–215.

Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). *Neural network design.* Boston, MA: PWS Publishing.

Hagan, M. T., & Menhaj, M. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, *5*(6), 989–993.

Hochberg, Y., & Tambane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hofzumahaus, A., Plane, J., Allan, J., Cohen, R., Fried, A., & Hamilton, J. (2006). *Analytical techniques for atmospheric measurement*. Oxford, UK: Blackwell.

Jobson, J. D. (1991). *Applied multivariate data analysis*.. *Springer texts in statistics*. New York: Springer.

Jollife, I. T. (1986). *Principal component analysis*. New York: Springer.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*, 2895–2907.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., & Kolehmainen, M., et al. (2003). Extensive evaluation of neural networks models for the prediction of NO_2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, *37*, 4539–4550.

Masters, T. (1995). *Advanced algorithms for neural networks*. New York: Wiley.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Nagendra, S. M. S., & 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.

Nunnari, G., Dorling, S., Schlink, U., Cawley, G., & Foxall, R. (2004). Modelling SO2 concentration at a point with statistical approaches. *Environmental Modelling & Software*, *19*, 887–905.

Nunnari, G., Nucifora, A. F. M., & Randieri, C. (1998). The application of neural techniques to the modelling of time-series of atmospheric pollution data. *Ecological Modelling*, *111*, 187–205.

Pelliccioni, A., & Poli, U. (2000). Use of neural net models to forecast atmospheric pollution. *Environmental Monitoting and Assessment*, *65*, 297–304.

Perez, P., Trier, A., & Reyes, J. (2000). Prediction of $PM_{2,5}$ concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment*, *34*, 1189–1196.

Pizarro, J., Guerrero, E., & Galindo, P. (2002). Multiple comparison procedures applied to model selection. *Neurocomputing*, *48*, 155–173.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). Numerical recipes in C, 2nd edition. Cambridge, UK: Cambridge University Press.

Rodriguez, S., Querol, X., Alastuey, A., Kallos, G., & Kakaliagou, O. (2001). Saharan dust contributions to PM10 and TSP levels in Southern and Eastern Spain. *Atmospheric Environment*, *35*, 2433–2447.

Ruiz Suárez, J. C., Mayorra-Ibarra, O. A., Torres-Jiménez, J., & Ruiz Suárez, L. G. (1995). Short-term ozone forecasting by artificial neural networks. *Advances in Engineering Software*, *23*, 143–149.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representation by error propagation.

Parallel distributed processing: explorations in the microstructures of cognition, Vol.1. Cambridge, MA: MIT Press.

Sakamoto, K., Takada, H., & Sekiguchi, K. (2004). Influence of ozone, relative humidity and flow rate on the deposition and oxidation of sulphur dioxide on yellow sand. *Atmospheric Environment*, *38*, 6961–6967.

Schwartz, J., Dockery, D. W., & Neas, L. M. (1996). Is daily mortality associated specifically with fine particles? *Journal of Air Waste Management*, *46*, 927–939.

Vapnik, V. N., & Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16*(2), 264–280.

Viotti, P., Liuti, G., & Di Genova, P. (2002). Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. *Ecological Modelling*, *148*, 27–46.

Wilson, W. E., & Suh, H. H. (1997). Fine and coarse particles: concentration relationships relevant to epidemiological studies. *Journal of Air Waste Management*, *47*, 1238–1249.

Yildirim, Y., & Bayramoglu, M. (2006). Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. *Chemosphere*, *63*, 1575–1582.

Ziomas, I. C., Melas, D., Zerefos, C. S., Bais, A. F., & Paliatsos, A. G. (1995). Forecasting peak pollutant levels from meteorological variables. *Atmospheric Environment*, *29*(24), 3703–3711.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*, 1.