Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software 22 (2007) 97–103
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations

S.I.V. Sousa, F.G. Martins*, M.C.M. Alvim-Ferraz, M.C. Pereira

LEPÆ – Laboratório de Engenharia de Processos, Ambiente e Energia, Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

Received 1 August 2005; received in revised form 29 November 2005; accepted 10 December 2005
Available online 27 January 2006

Abstract

The prediction of tropospheric ozone concentrations is very important due to the negative impacts of ozone on human health, climate and vegetation. The development of models to predict ozone concentrations is thus very useful because it can provide early warnings to the population and also reduce the number of measuring sites. The aim of this study was to predict next day hourly ozone concentrations through a new methodology based on feedforward artificial neural networks using principal components as inputs. The developed model was compared with multiple linear regression, feedforward artificial neural networks based on the original data and also with principal component regression. Results showed that the use of principal components as inputs improved both models prediction by reducing their complexity and eliminating data collinearity.

Keywords: Trophospheric ozone; Multiple linear regression; Artificial neural networks; Principal components

1. Introduction

In recent decades, global increase in tropospheric ozone concentrations has been attributed mainly to anthropogenic emissions (from industry and traffic). Photochemical interactions between emitted pollutants (nitrogen oxides and volatile organic compounds) and favourable meteorological conditions (high temperatures and strong solar radiation) can lead to high ozone concentrations. The residence time of ozone in the atmosphere is long, enabling long-range transport over hundreds to thousands of kilometres, thus allowing occurrence of high ozone levels at both, regional and urban scale (Sillman, 1999; San José et al., 2005).

The development of mathematical tools to predict ozone concentrations is very useful because it can provide early warnings to the population and reduce the number of measuring sites. Accordingly, the European directive concerning ozone in ambient air enhances the necessity of developing predicting models. Ozone concentrations are very difficult to model because of the different interactions between pollutants and meteorological variables (Borrego et al., 2003). One of the approaches to avoid this problem is the principal component analysis, which has been receiving increased attention as an accepted method in environmental pattern recognition. This multivariate statistical technique transforms the original data set into a set of linear combinations of the original variables. The uncorrelated new variables, designated by principal components, account for the majority of the original variance. In recent years, multiple linear regressions, feedforward artificial neural networks as well as principal component regressions (combining multiple linear regressions and principal component analysis) are being used to model ozone concentrations (Schlink et al., in press; Abdul-Wahab et al., 2005; Gonçalves et al., 2005; Lengyel et al., 2004; Abdul-Wahab and Al-Alawi, 2002). This work reports the use of a new methodology using feedforward artificial neural networks based on principal components, therefore combining statistical and artificial intelligence techniques.

* Corresponding author. Tel.: +351 22 508 1974; fax: +351 22 508 1449.
E-mail address: fgm@fe.up.pt (F.G. Martins).

1364-8152/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2005.12.002

The aim of this work was: (i) to evaluate the relative influence of precursor concentrations and meteorological variables on ozone formation, using principal component analysis; and (ii) to predict next day hourly ozone concentrations, through a new methodology based on feedforward artificial neural networks using principal components as inputs.

2. Methodology

2.1. Site characterization and data

Oporto is situated in Northern Portugal (41°10′N, 8°40′W). The mean annual temperature is around 15 °C (less than 10 °C of difference between warmer and colder months) and the total mean annual precipitation varies between 1000 and 1200 mm (40% occurring in the winter season). Annual air humidity is between 75 and 80% with prevailing winds from W and NW in the summer and from E and SE in winter (Pereira et al., 2005).

The air quality data were collected from an urban site with traffic influence, situated in Oporto and integrated in the Air Quality Monitoring Network of Oporto Metropolitan Area (Oporto-MA), managed by the Regional Commission of Coordination and Development of Northern Portugal (*Comissão de Coordenação e Desenvolvimento Regional do Norte*), under responsibility of the Ministry of Environment. The meteorological parameters were measured by the Geophysical Institute of Oporto University (*Instituto Geofísico da Faculdade de Ciências da Universidade do Porto*) on the left edge of Douro River, at an altitude of 90 m approximately.

This study considered as predictor variables the hourly concentrations of ozone (O_3), nitrogen monoxide (NO), nitrogen dioxide (NO_2) and hourly means of temperature (T), wind velocity (WV) and relative humidity (RH).

Ozone concentrations were monitored by UV-absorption photometry; NO and NO_2 were obtained through chemiluminescence method. Monitoring was continuous and hourly mean concentrations ($\mu g\, m^{-3}$) were recorded. All equipments were submitted to a rigid maintenance program, with periodical calibrations being preformed. The meteorological parameters were also continuously measured.

2.2. Models

Multiple linear regression (MLR) and feedforward artificial neural network (FANN) were used to predict the next day hourly ozone concentration using as predictors air pollutant concentrations (NO, NO_2 and O_3) and meteorological parameters (T, RH and WV). The same models, but based on principal component analysis (PCA), were also used, being referred to as principal component regression (PCR) and feedforward artificial neural network based on PC (PC-FANN), respectively.

PCA is a multivariate statistical method widely used in air pollution analysis. The objective of PCA, as previously referred, is to reduce the number of predictive variables and transform them into new variables, called principal components (PC); these new variables are independent linear combinations of the original data and retain the maximum possible variance of the original set. The eigenvalues of the standardized matrix are calculated from Eq. (1):

$$|C - \lambda I| = 0 \tag{1}$$

where C is the correlation matrix of the standardized data, λ is the eigenvalues and I is the identity matrix. The weights of the variables in the PC are then obtained by Eq. (2):

$$|C - \lambda I|W = 0 \tag{2}$$

where W is the matrix of the weights.

To evaluate the influence of each variable in the PC, varimax rotation was used to obtain values of rotated factor loadings. These loadings represent the contribution of each variable in a specific principal component.

The PC used for the prediction of O_3 concentrations were obtained through multiplication of the standardized data matrix by the previously calculated weights (W) (Çamdevýren et al., 2005; Slini et al., in press).

The applicability of the PCA to the data sets used in this study was verified through the application of modified Bartlett's sphericity test, expressed by the following equation:

$$\chi_k^2 = \left[n - k - \frac{2(p-k)+7+2/(p-k)}{6} + \sum_{j=1}^{k} \left(\frac{\overline{\lambda}}{\lambda_j - \overline{\lambda}} \right)^2 \right] \times \left[-\ln \prod_{j=k+1}^{p} \lambda_j + (p-k) \ln \overline{\lambda} \right] \tag{3}$$

where p is the number of components, λ_j represents the eigenvalue for the k_jth component, n is the number of observations in the sample and $\overline{\lambda}$ is obtained by the following equation:

$$\overline{\lambda} = \sum_{j=k+1}^{p} \frac{\lambda_j}{p-k} \tag{4}$$

The null hypothesis considered was that all variables were uncorrelated and when accepted, PCA could be applied (Peres-Neto et al., 2005).

Multiple linear regression models are often used in the prediction of ozone concentrations, being represented by the relationship between these concentrations and a set of predictor variables. The general equation is as follows:

$$\hat{Y} = P_0 + P_1 X_1 + \ldots + P_n X_n \tag{5}$$

where P_i ($i = 0,\ldots,n$) are the parameters generally estimated by least squares and X_i ($i = 1,\ldots,n$) are the explanatory variables (predictors).

Although these models are simply based on linear and additive associations of the explanatory variables, they have been extensively used with satisfactory results. Nevertheless, in regression equations, the collinearity between the independent variables can lead to incorrect identification of the most important predictors (Thompson et al., 2001; Heo and Kim, 2004).

Due to the non-linearities of ozone concentrations and the complex interactions between meteorological variables and ozone, the development of non-linear models, such as artificial neural networks, is currently being applied. These models perform a non-linear transformation of input data to approximate output data, learning from experimental data examples and exhibiting some ability for generalization beyond training data. The most common artificial neural network is the feedforward artificial neural network (FANN) where the nodes are grouped into three types of layers, i.e. input, hidden and output layers. Data are fed into the nodes in the input layer being after transferred to the subsequent layers. Cybenko (1989) has shown that a one hidden layer FANN is enough to approximate any function, if presenting enough hidden nodes. The topology of the network, along with the neuron processing function, determines the accuracy and degree of representation of the model developed to correctly represent the system behaviour.

To obtain the output value of the node, an activation function usually sigmoid, hyperbolic tangent or linear is applied. Each node in hidden and output layers has a bias value which is known as the activation threshold (Watanabe et al., 1989; Martins and Coelho, 2000; Aparício et al., 2002; Morabito and Versaci, 2003; Schlink et al., 2003; Heo and Kim, 2004; Mas et al., 2004).

In most cases, the FANN is obtained using two distinct data sets: training and validation. The training data set is used to determine the network topology and the associated weights by solving a non-linear optimization problem with the objective function being dictated by the mean squared error (MSE). The validation data set is used to compute the FANN performance.

Cross-validation is usually used to avoid the overfitting problem that often appears when applying FANN (Schenker and Agarwal, 1996; Warne et al., 2004). The best network topology corresponds to a FANN which presents a minimum value of MSE for the validation data set.

The non-linear features of FANN model and the possibility of incorporating pollutant concentrations and meteorological parameters as input variables, suggest good performance for the prediction of O_3 concentrations.

The application of PC in FANN models aims to reduce the collinearity of the data sets, which can lead to worst predictions and also to determine the relevant independent variables for the prediction of O_3 concentrations. The architecture of the PCA based neural network approach is shown in Fig. 1. The difference between this approach and the simple FANN model is that the input variables used are the principal components. Consequently, the network architecture will be less complex due to the decrease of input variables.

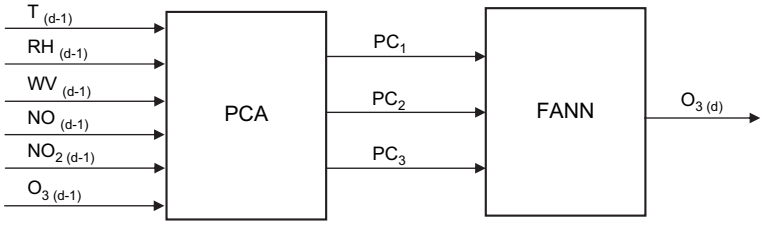

Fig. 1. Architecture of a PC-FANN model for the prediction of next day O_3 concentrations ($O_{3(d)}$).

According to Warne et al. (2004), the use of PC based neural networks eliminates the overfitting problem, i.e. both validation and training MSE continuously decrease.

2.3. Performance indexes

The models' behaviour in both, development and validation steps, was evaluated calculating the following statistical parameters: correlation coefficient (R), mean bias error (MBE), mean absolute error (MAE), root mean squared error (RMSE) and index of agreement (d_2), given by Eqs. (6–10), respectively:

$$R = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\overline{Y}_i)^2-\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\overline{Y}_i)^2}} \tag{6}$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i) \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2} \tag{9}$$

$$d_2 = 1 - \frac{\left[\sum_{i=1}^{n}|\hat{Y}_i - Y_i|^2\right]}{\left[\sum_{i=1}^{n}(|\hat{Y}_i-\overline{Y}_i|+|Y_i-\overline{Y}_i|)^2\right]} \tag{10}$$

R provides the variability measure of the data reproduced in the model. As this test does not give the accuracy of the model, other statistical parameters must be reported. MBE indicates if the observed concentrations are over- or under-estimated. MAE and RMSE measure residual errors, which give a global idea of the difference between the observed and modelled values. The values of d_2 compare the difference between the mean, the predicted and the observed concentrations, indicating the degree of error free for the predictions (Gardner and Dorling, 2000; Chaloulakou et al., 2003).

3. Results and discussion

Correlation coefficients between pollutants and meteorological variables were analysed to evaluate the influence of each variable on O_3 concentrations. These coefficients provide a measure of the linear relation between two variables and also indicate the existence of collinearity between the explanatory variables. The statistical significance of the regressions was analysed by calculating the critical correlation coefficient, R_{crit}, using a significance level of 0.05 (two-tailed test). R_{crit} was calculated by Eq. (11) using $\mathrm{DF} = n - k$ degrees of freedom.

$$R_{\mathrm{crit}} = \frac{t_{\mathrm{crit}}}{\sqrt{\mathrm{DF} + t_{\mathrm{crit}}^2}} \tag{11}$$

The regression is statistically valid if R_{crit} is lower than the correlation coefficient and results showed that all regressions were statistically valid.

The study performed for July 2003 considered the mean hourly concentrations of the above-mentioned variables. The training data set included the 26 first days of the month (616 data points) whereas the validation data set was constituted by the last five days (118 data points). O_3 concentrations varied between 0 and 95 μg m^{-3} (mean value of 36.6 μg m^{-3}) and 0 and 180 μg m^{-3} (mean value of 41.4 μg m^{-3}) during training and validation periods, respectively.

High correlation coefficients were found between O_3 and nitrogen monoxide (NO), nitrogen dioxide (NO_2), temperature (T), wind velocity (WV) and relative humidity (RH). Therefore, these variables were used to predict next day hourly O_3 concentrations. Also high correlation coefficients were achieved between some O_3 predictors, such as NO and NO_2 (0.56), RH and T (−0.77), demonstrating the existence of collinearity between the variables.

As previously mentioned, multiple linear regression and feedforward artificial neural networks were used to predict the next day hourly O_3 concentrations. These models were based on the original data (MLR and FANN) and on the PC (PCR and PC-FANN).

Table 1 shows the matrix of the weights for the PC, which demonstrates the relative importance of each standardized predictor in the PC calculations.

To apply the FANN models (based on original data and PC), several network structures were tested to find the most appropriate topology. Using original variables as inputs, the best architecture consisted of a three-layer network with six neurons in the input layer, eight neurons in the hidden layer and one neuron in the output layer. Considering PC as inputs,

Table 1
Matrix of the weights for the principal components

Variables	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
NO	−0.325	0.557	−0.555	−0.068	0.410	0.320
NO_2	−0.275	0.648	0.442	0.429	−0.328	−0.129
O_3	0.472	0.040	0.234	0.530	0.658	0.086
T	−0.461	0.286	0.281	−0.396	−0.215	0.651
RH	−0.433	−0.427	0.083	0.403	−0.105	0.670
WV	0.441	0.053	−0.596	0.459	−0.484	0.041

Table 2
Performance indexes achieved using MLR and FANN during training and validation periods

Performance indexes	Training		Validation	
	MLR	FANN	MLR	FANN
R	0.74	0.76	0.70	0.78
MBE	-1.0×10^{-5}	0.66	11.50	2.68
MAE	13.05	12.63	23.61	19.83
RMSE	16.00	15.37	29.50	25.64
d_2	0.83	0.85	0.81	0.84

the best architectures were achieved with different number of neurons in the input layer depending on the number of PC used. The hidden and the output layer consisted of eight and one neurons, respectively. Sigmoid and linear functions were used as activation functions in the neurons of the hidden layer and output neuron, respectively. The training was done for a maximum of 10 000 iterations. To avoid the overfitting problem, which generally appears with the application of FANN, cross-validation tests were used. The selection of the network was performed considering a minimum value of MSE for the validation data set.

Table 2 presents the values of the performance indexes using MLR and FANN for both, training and validation steps. A *t*-test (significance level of 0.05) was applied to calculate the statistically valid parameters. During the studied period, the coefficient of NO concentrations had a confidence interval [−0.02, 0.08] showing that it was statistically invalid; thus the NO concentrations were removed from the MLR model. The derived model is as follows:

$$[O_3]_{(d)} = -57.0 + 0.16[NO_2]_{(d-1)} + 2.86T_{(d-1)} + 0.14RH_{(d-1)} + 0.54WV_{(d-1)} + 0.40[O_3]_{(d-1)} \quad (12)$$

The results obtained during training and validation periods using MLR demonstrated that only the MBE value presented a significant difference, which means that although this model is only a simple linear additive association of the variables, it presented reasonable results.

Table 3
Values of the χ^2 distribution using Bartlett's sphericity test

Component	$(\chi_c)^2$	χ^2	DF
1	23.7	715	14
2	17.0	208	9
3	11.1	61	5
4	6.0	40	2

The prediction with FANN model was performed using 2000 iterations (cross-validation). The performance indexes, calculated for the training and validation periods, were quite similar, which indicates that the model performed good predictions. Better performance indexes were achieved with FANN model for both, training and validation steps, with the exception of the MBE value that was lower using the MLR in the training period. However, with FANN model, the values of MBE were positive, indicating a slight overprediction.

As previously mentioned, the next day hourly O_3 concentrations were also predicted based on the PC (PCR and PC-FANN models). Bartlett's sphericity test results (Table 3) showed that the PCA was applicable to this data set; the eigenvalues and respective variances were calculated through the PCA and are shown in Fig. 2.

For PCR and PC-FANN models, the forecasting was performed considering from two to six PC separately. Considering two PC the eigenvalues were higher than considering one (Kaiser Criterion), being responsible for 77% of the total variance (when considering six PC, all the variance was accounted).

A *t*-test was also performed for these regressions, to statistically evaluate the regression parameters. Considering the statistically valid parameters, new regressions were then performed.

Table 4 presents the values of performance indexes calculated for both models, using from two to six PC.

Using the PCR model, performance indexes for the validation step were generally slightly worse than for the training step. It is important to point out that for this model the MBE values were very low in the training step, independently of the number of PC used. The same model was achieved using either four or five PC. Considering all the models achieved

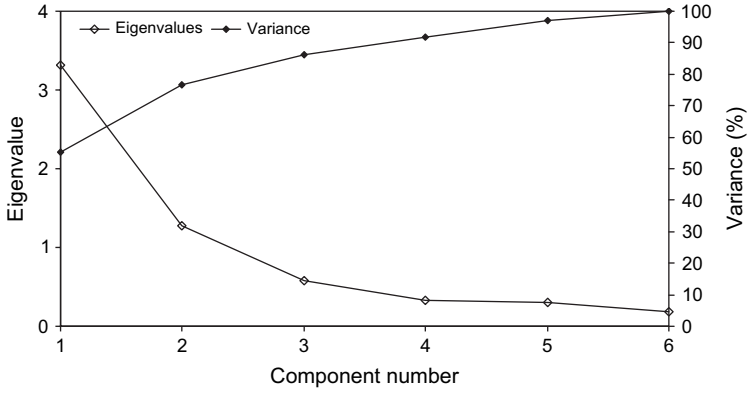

Fig. 2. Scree plot and respective cumulative variances (%).

Table 4
Performance indexes achieved with PCR and PC-FANN during training (Tra) and validation (Val) periods

Performance indexes	Two PC				Three PC				Four PC				Five PC[a]				Six PC			
	PCR		PC-FANN		PCR		PC-FANN		PCR		PC-FANN		PC-FANN		PCR		PCR		PC-FANN	
	Tra	Val	Tra	Val	Tra	Val	Tra	Val	Tra	Val	Tra	Val	Tra	Val	Tra	Val	Tra	Val	Tra	Val
R	0.70	0.68	0.70	0.63	0.70	0.67	0.73	0.65	0.72	0.73	0.73	0.73	0.73	0.78	0.74	0.68	0.77	0.76		
MBE	-3.1×10^{-7}	11.33	−0.04	16.18	-4.3×10^{-7}	12.78	0.10	14.43	-1.8×10^{-7}	9.13	0.62	6.49	0.74	8.49	-1.1×10^{-6}	12.12	−0.16	5.23		
MAE	13.99	23.29	14.00	25.16	13.82	24.04	13.18	24.27	13.51	22.06	13.24	21.84	13.28	20.38	13.02	24.04	12.18	19.79		
RMSE	16.87	29.93	16.84	31.96	16.72	30.44	16.12	31.13	16.45	28.13	16.21	21.78	16.07	25.72	15.98	30.03	14.94	26.69		
d_2	0.81	0.79	0.81	0.82	0.81	0.80	0.83	0.83	0.82	0.82	0.84	0.81	0.84	0.86	0.83	0.81	0.87	0.84		

[a] The performance indexes for PCR using five PC were the same as using four PC.

with different number of PC, best performance indexes were obtained when prediction was performed using four PC, which corresponds to a cumulative variance of 91.9%. These four PC were then used as predictor variables and the following model was obtained:

$$[O_3]_{(d)} = 38.8 + 9.57PC_1 + 5.24PC_2 + 2.93PC_3 + 5.59PC_4 \quad (13)$$

Table 5 shows, as an example, the rotated factor loadings and respective communalities using four and two PC, respectively. The bold marked loads indicate the variables that most influenced the correspondent component. Using four PC, those variables (associated with PC_1, PC_2, PC_3 and PC_4) were: (i) T and RH, (ii) NO and O_3 concentrations, (iii) WV and NO_2 concentration, and (iv) O_3 concentration. It was also observed that with lower number of PC, more variables were accounted for each one. In addition, communality values were higher using four PC.

The validation of PC-FANN models was performed according to the cross-validation, using different number of iterations depending on the number of PC used. The number of iterations, from two to six PC, was 500, 8000, 1000, 1000 and 3500, respectively. Also, the number of neurons in the input layer was different depending on the number of PC used.

The performance indexes, obtained using PC-FANN, were not very different in the training and validation periods, with exception of the MBE values being slightly different. For this model, the best performance was achieved when five PC were used.

The performance indexes calculated for PCR and PC-FANN showed that the approach using neural networks led to better predictions.

Fig. 3 shows, as an example, the predictions with all models and the measured data, corresponding to the validation period. It was shown that neural networks led to better predictions. Although the PC-FANN and the FANN models presented similar results, because PCA application led to the introduction of fewer variables and thus less complex networks, the first approach was considered to be better. Concluding, PC-FANN is a promising tool for the prediction of ozone concentration. The worst performance occurred with MLR. It is also important to refer that the neural networks achieved a significant power of generalization beyond the training data, i.e. in validation period they were tested in extrapolated regions being able to predict hourly O_3 concentrations. Although the MBE was generally positive for all the models, meaning that, in average, the predicted ozone concentrations were overestimated, it can be observed in Fig. 3 that the highest measured concentrations were underestimated. This problem occurred because the highest concentrations were not contemplated during the training step, which should be avoided. Also it was observed that the models were not able to predict lower concentrations. This occurs because, in the validation period, the percentage of data points lower than 20 $\mu g\, m^{-3}$ was 40% and in the training period it was of 25%.

Table 5
Rotated factor loadings using four and two PC and respective communalities

Variables	Four PC					Two PC		
	Rotated factor loadings				Communalities	Rotated factor loadings		Communalities
	PC_1	PC_2	PC_3	PC_4		PC_1	PC_2	
NO	−0.112	**0.879**	0.376	−0.061	0.93	−0.213	**0.838**	0.75
NO_2	−0.023	0.278	**0.923**	−0.019	0.96	−0.084	**0.884**	0.79
O_3	0.486	**−0.550**	0.011	**0.573**	0.87	**0.774**	−0.376	0.74
T	**0.885**	−0.287	−0.031	0.204	0.91	**0.892**	−0.123	0.81
RH	**−0.897**	−0.013	0.044	−0.329	0.91	**−0.924**	−0.042	0.86
WV	0.383	−0.060	−0.289	**0.833**	0.93	**0.733**	−0.336	0.65

Bold marked loads indicate the variables that most influence each parameter.

4. Conclusions

MLR and FANN were used to predict the next day hourly ozone concentrations using as predictors air pollutant concentrations (NO, NO_2 and O_3) and meteorological parameters (*T*, RH and WV). These predictors were selected through the calculation of correlation coefficients. Two different approaches were used, considering original data and PC as inputs.

During the studied period, the coefficient of NO concentrations was found to be statistically invalid and therefore NO concentrations were removed from MLR model.

Using four PC, the original variables (associated with PC_1, PC_2, PC_3 and PC_4) were: (i) *T* and RH, (ii) NO and O_3 concentrations, (iii) WV and NO_2 concentration, and (iv) O_3 concentration.

The results showed that the use of FANN led to more accurate results than linear models (MLR and PCR), due to the account of non-linearities. The application of PC in this model was considered better than using the original data, because it reduced the number of inputs and therefore decreased the model complexity. The performance indexes were similar using both approaches.

Considering MLR and PCR, the performance indexes were higher using PCR.

The use of PC based models was considered more efficient, due to elimination of collinearity problems and reduction of the number of predictor variables. It was also verified that the use of PC based neural networks improved the prediction of ozone concentrations, therefore proving to be a useful tool to public health protection because it can provide early warnings to the population. Although the predicted ozone concentrations were in average overestimated for all the models, the highest concentrations observed were underestimated, because the highest concentrations were not contemplated during the training step, which should be avoided. Also because the models were very sensitive to the training data set, they were not able to predict lower concentrations.

Acknowledgements

Authors are grateful to *Comissão de Coordenação da Direcção Regional-Norte* and to *Instituto Geofísico da Faculdade de Ciências da Universidade do Porto* for kindly providing the air quality and meteorological data. The authors also thank *Fundação Calouste Gulbenkian and Fundação para a Ciência e Tecnologia*.

References

Abdul-Wahab, S.A., Al-Alawi, S.M., 2002. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environmental Modelling & Software 17, 219–228.

Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone

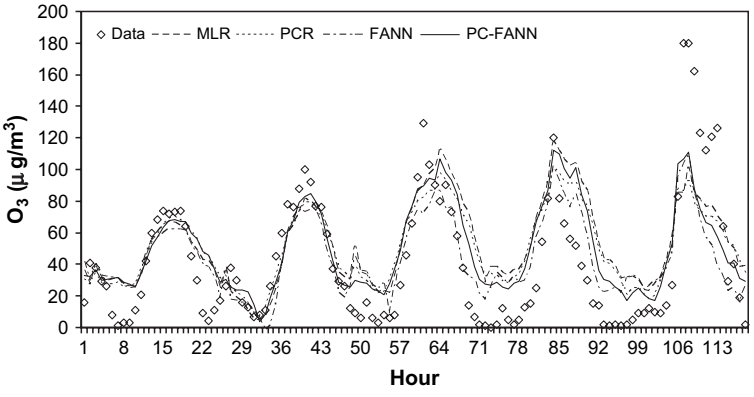

Fig. 3. Prediction of O_3 concentrations for the validation period.

and factors affecting its concentrations. Environmental Modelling & Software 20, 1263–1271.

Aparício, J.J.C., Jerónimo, M.A.S., Martins, F.G., Coelho, M.A.N., Martins, C., Braga, A.S., Costa, C.A.V., 2002. Two different approaches for RDC modelling when simulating a solvent deasphalting plant. Computers & Chemical Engineering 26, 1369–1377.

Borrego, C., Tchepel, O., Costa, A.M., Amorim, J.H., Miranda, A.I., 2003. Emission and dispersion modelling of Lisbon air quality at local scale. Atmospheric Environment 37, 5197–5205.

Çamdevýren, H., Demýr, N., Kanik, A., Keskýn, S., 2005. Use of principal component scores in multiple linear regression models for prediction of chlorophyll-*a* in reservoirs. Ecological Modelling 181, 581–589.

Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Science of the Total Environment 313, 1–13.

Cybenko, G., 1989. Approximations by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, 303–314.

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

Gonçalves, F.L.T., Carvalho, L.M.V., Conde, F.C., Latorre, P.H.N., Braga, A.L.F., 2005. The effects of air pollution and meteorological parameters on respiratory morbidity during the summer in São Paulo City. Environment International 31, 343–349.

Heo, J.-S., Kim, D.-S., 2004. A new method of ozone forecasting using fuzzy expert and neural network systems. Science of the Total Environment 325, 221–237.

Lengyel, A., Héberger, K., Paksy, L., Bánhidi, O., Rajkó, R., 2004. Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere 57, 889–896.

Martins, F.G., Coelho, M.A.N., 2000. Application of feedforward artificial neural networks to improve process control of PID-based control algorithms. Computers & Chemical Engineering 24, 853–858.

Mas, J.F., Puig, H., Palácio, J.L., Sosa-López, A., 2004. Modelling deforestation using GIS and artificial neural networks. Environmental Modelling & Software 19, 461–471.

Morabito, F.C., Versaci, M., 2003. Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data. Neural Networks 16, 493–506.

Pereira, M.C., Alvim-Ferraz, M.C.M., Santos, R.C., 2005. Relevant aspects of air quality in Oporto (Portugal): PM_{10} and O_3. Environmental Monitoring & Assessment 101, 203–221.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Computational Statistics & Data Analysis 49, 974–997.

San José, R., Stohl, A., Karatzas, K., Bohler, T., James, P., Pérez, J.L., 2005. A modelling study of an extraordinary night time ozone episode over Madrid domain. Environmental Modelling & Software 20, 587–593.

Schenker, B., Agarwal, M., 1996. Cross-validated structure selection for neural networks. Computers & Chemical Engineering 20, 175–186.

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxal, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., Doyle, M., 2003. A rigorous inter-comparison of ground-level ozone predictions. Atmospheric Environment 37, 3237–3253.

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E. Statistical models to assess the health effects and to forecast ground-level ozone. Environmental Modelling & Software, in press, doi:10.1016/j.envsoft.2004.12.002.

Sillman, S., 1999. The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. Atmospheric Environment 33, 1821–1845.

Slini, T., Kaprara, A., Karatzas, K., Moussiopoulos, N. PM10 forecasting for Thessaloniki, Greece. Environmental Modelling & Software, in press.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment 35, 617–630.

Warne, K., Prasad, G., Rezvani, S., Maguire, L., 2004. Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. Engineering Applications of Artificial Intelligence 17, 871–885.

Watanabe, K., Matsuura, I., Abe, M., Kubota, M., Himmelblau, D.M., 1989. Incipient fault diagnosis of chemical processes via artificial neural networks. AIChE Journal 35, 1803–1811.