Prediction of the next day maximum ozone concentration using multiple linear and principal component regressions

S. I. V. Sousa, F. G. Martins, M. C. Pereira
& M. C. M. Alvim-Ferraz
Departamento de Engenharia Química, LEPAE,
Faculdade de Engenharia, Universidade do Porto, Portugal

Abstract

Prediction of ground-level ozone concentrations is very important due to the negative impacts of this pollutant on human health and environment. Multiple linear regression (MLR) and principal component (PCR) regression were used as statistical models for the forecasting of ozone concentrations. The aim of this study was to predict the next day maximum ozone concentration. The studies were performed considering separately the year 2002 and the respective four trimesters. A subset of the last 10 and 30 days was used, respectively, for each trimester and for the year to validate the models. The predictor variables were inferred by the analysis of the linear correlation with ozone. For that, maxima hourly values for ozone, ratio of nitrogen dioxide and nitrogen monoxide, temperature, wind velocity and the minima hourly values for carbon monoxide and relative humidity were used. The main results achieved were: i) the performance indexes obtained for validation datasets were usually higher with PCR; ii) the number of principal components considered to develop the PCR was dependent of the dataset considered; iii) the PCR is more robust than MLR because collinearity effects are accounted with the first approach; iv) PCR model is shown to be a useful tool to provide protection for the public, through the use of early warnings for the population.
Keywords: ground-level ozone forecasting, multiple linear regression, principal component analysis, principal component regression.

WIT Transactions on Ecology and the Environment, Vol 86, © 2006 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR06011

1 Introduction

Tropospheric ozone is a secondary photochemical air pollutant which has negative impacts on human health, climate, vegetation, materials and atmospheric composition. The most significant effects on human health are related with a decrease in pulmonary function, attacks of asthma and chronic bronchitis, and globe ocular damages White *et al.* [1]. In terms of climate it is considered an important greenhouse gas so the increasing of ozone concentrations is related with an increasing of the surface temperature. In vegetation it has harmful effects especially during the growing season WHO [2].

Surface ozone formation results from a chain mechanism involving photochemical reactions of NO_x, VOCs, CO and CH_4 Seinfeld and Pandis [3]. Ozone concentrations have been increasing significantly since pre-industrial times due to the increased photochemical production associated with the increase of anthropogenic emissions Alvim-Ferraz *et al.* [4]. Thus, it is very important to develop tools able to predict ozone concentrations and to provide early warnings to the population. Generally, ozone concentrations are very difficult to model, because of the different interactions between pollutants and between pollutants and meteorological variables. Principal component analysis is one of the approaches, which has been receiving attention as an accepted method in environmental pattern recognition. This multivariate statistical technique transforms the original data set into an equal set of linear combinations of the original variables. The new variables named principal components (PC) are uncorrelated and account for the majority of the original variance Gonçalves *et al.* [5]; Abdul-Wahab *et al.* [6]; Lengyel *et al.* [7].

The aim of this work was to analyse the relative importance of the concentration of precursors and meteorological variables in ozone formation, using principal component analysis, and to predict ozone concentrations. The performance of these models was compared through the evaluation of the mean bias error, the mean absolute error, the root mean squared error and the index of agreement.

2 Methodology

2.1 Data

The air quality data used was collected from an urban site with traffic influences, situated in Oporto, integrated in the measuring conducted by 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 the responsibility of the Ministry of Environment. The meteorological parameters used were measured in the left edge of Douro River, at an approximate altitude of 90 m, by the Geophysical Institute of Oporto University (*Instituto Geofísico da Faculdade de Ciências da Universidade do Porto*).

Oporto is situated in the North of Portugal and has a latitude and longitude of approximately 41º10' N and 8º40'W, respectively. The annual average temperature is around 15ºC and the difference between warmer and colder monthly averages is less than 10ºC. Annual air humidity is between 75% and 80%, and the total annual mean precipitation varies between 1000 mm and 1200 mm, with about 40% in the winter season. Prevailing winds are from W and NW in summer and from E and SE in winter Pereira *et al.* [8].

This study considered as variables the values of ozone (O_3), ratio of nitrogen dioxide and nitrogen monoxide (NO_2/NO), carbon monoxide (CO), particulate matter with an equivalent aerodynamic diameter smaller than 10 μm (PM_{10}), sulphur dioxide (SO_2), temperature (T), wind velocity (WV) and relative humidity (RH).

Ozone concentrations were monitored by UV-absorption photometry; PM_{10} concentrations were obtained through the beta radiation attenuation method, SO_2 concentrations were obtained through UV Fluorescence method; CO concentrations were measured through IV spectroscopy without dispersion; NO and NO_2 were obtained through chemiluminescence method. The monitoring is continuous and hourly averages are recorded expressing the concentration in $\mu g.m^{-3}$. All the equipments were submitted to a rigid maintenance program being periodically calibrated.

The meteorological parameters were continuously measured, the hourly averages being considered in this study.

This study considered as predictor variables the maxima hourly values of ozone (O_3), ratio of nitrogen dioxide and nitrogen monoxide (NO_2/NO), temperature (T), wind velocity (WV) and the minima hourly values of carbon monoxide (CO) and relative humidity (RH). The concentrations of particulate matter with an equivalent aerodynamic diameter smaller than 10 μm (PM_{10}) and of sulfur dioxide (SO_2) were considered in the correlation study but no correlation was found between this variables and O_3.

2.2 Models

Multiple linear regression (MLR) and principal component regression (PCR) were used to predict the next day maximum ozone concentration, with other air pollutant concentrations and meteorological parameters as predictors.

MLR is an extension of a simple linear regression model incorporating several explanatory variables in a prediction equation, for a response variable. The general equation is as follows:

$$\hat{Y} = P_0 + P_1 X_1 + \ldots + P_n X_n \tag{1}$$

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

MLR models have been extensively used for O_3 prediction, although these predictions are simply based on linear and additive associations of the

explanatory variables and are highly sensitive to colinearities of the data Heo and Kim [9]; Thompson *et al.* [10].

The PCR combines the principal component analysis (PCA) and the MLR to determine the relevant independent variables for the prediction of O_3 concentrations. PCA is a multivariate statistical method widely used in air pollution analysis. The objective of PCA is to reduce the number of predictive variables and transform them into new variables or principal components (PC) that are independent linear combinations of the original data, retaining the maximum possible variance of the original set. The eigenvalues of the standardized matrix are calculated through eqn (2):

$$\left|C - \lambda I\right| = 0\,, \tag{2}$$

where C is the correlation matrix of the standardized data, λ are the eigenvalues and I is the identity matrix. The weights of the variables in the PC are then obtained by eqn (3):

$$\left|C - \lambda I\right| W = 0\,, \tag{3}$$

where W is the matrix containing the weights. To analyse the influence of the variables in PC, values of rotated factor loadings were calculated through varimax rotation. These loadings represent the contribution of corresponding variables to each principal component.

The PC used for the prediction of O_3 concentrations were obtained through the multiplication of the standardized data matrix by the weights (W) previously calculated Çamdevýren *et al.* [11]; Slini *et al.* [12].

The applicability of the PCA to the datasets used in the study was verified through the application of Bartlett's sphericity test expressed by the following equation Peres-Neto *et al.* [13]:

$$\chi^2 = -\left[n - \frac{1}{6}(2p + 11)\right] \ln\left|R\right|\,, \tag{4}$$

where |R| is the determinant of the correlation matrix, n is the sample size and p the number of variables. The distribution is χ^2 with p(p-1)/2 degrees of freedom. The null hypothesis here considered was that all variables are uncorrelated, and if accepted the PCA can be applied.

2.3 Performance indexes

The statistical parameters taken into account to evaluate the behaviour of the MLR and the PCR in the two steps of the models implementation (development

and validation) were mean bias error (MBE), mean absolute error (MAE), root mean squared error (RMSE) and index of agreement (IA) given by eqns (5), (6), (7) and (8), respectively:

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right) \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{7}$$

$$\mathrm{IA} = 1 - \frac{\left[\sum_{i=1}^{N}\left|\hat{Y}_i - Y_i\right|^2\right]}{\left[\sum_{i=1}^{N}\left(\left|\hat{Y}_i - \overline{Y}_i\right| + \left|Y_i - \overline{Y}_i\right|\right)^2\right]}. \tag{8}$$

The MBE indicates if the observed concentrations are over or under estimated. The MAE and the RMSE measure residual errors which give a global idea of the difference between the observed and modelled values. The values of IA indicate the degree of which the predictions are error free, because it compares the difference between the mean, the predicted and the observed concentrations Chaloulakou *et al*. [14]; Gardner and Dorling [15].

3 Results and discussion

An analysis of the correlation coefficients between pollutants and meteorological parameters available was performed to evaluate the influence of each variable on the O_3 concentrations. These coefficients provide a measure of the linear relation between the two considered variables.

The results showed positive correlations between ozone and the ratio nitrogen dioxide and nitrogen monoxide (NO_2/NO), the temperature (T) and the wind velocity (WV); and negative correlations with relative humidity (RH) and carbon monoxide (CO). The concentrations of particulate matter with an equivalent aerodynamic diameter smaller than 10 μm (PM_{10}) and of sulphur dioxide (SO_2) were considered in the correlation study but no correlation was found between this variables and O_3. Thus, the variables used to predict the next day maxima O_3 concentrations were the maxima hourly values for O_3, ratio of NO_2/NO, T, WV and the minima hourly values for CO and RH.

Five datasets were considered corresponding to the complete year (1 dataset) and the trimesters (4 datasets) of 2002.

The results of Bartlett's sphericity test showed that the principal component analysis is applicable to all five datasets.

The eigenvalues and respective variances were calculated through the PCA. Table 1 shows an example for the 4th trimester. Two different approaches were considered in the PCA. The first using the PC with eigenvalues higher than one (the Kaiser criterion) responsible for 54% to 72% of the total variance; the second using six PC responsible for all the variance. Considering the first approach, only the first two PC were selected in all datasets with exception for the 4th trimester, where the first three PC were used.

Table 1: Eigenvalues and respective cumulative variances (%) for each principal component for the 4th trimester.

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
Eigenvalue	1.87	1.35	1.08	0.82	0.58	0.29
Cumulative variance (%)	31.2	53.7	71.8	85.5	95.1	100.0

Table 2 shows, as an example, the rotated factor loadings using three PC and six PC, for the 4th trimester. The bold marked loads indicate the variables that most influence the correspondent component. It was observed that using two PC (or three in the 4th trimester) each component accounts for a higher number of variables than with six PC.

Table 2: Rotated factor loadings using two and six PC, for the 4th trimester.

Variables	Rotated factor loadings								
	Two PC			Six PC					
	PC_1	PC_2	PC_3	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
NO_2/NO	0.076	0.033	**-0.952**	0.074	-0.054	**-0.983**	-0.134	-0.078	0.033
CO	**-0.676**	0.202	0.338	-0.048	-0.013	0.148	**0.956**	-0.054	0.241
T	0.115	**0.766**	0.218	0.045	**0.983**	0.055	-0.013	-0.170	-0.025
RH	0.035	**-0.828**	0.248	-0.056	-0.172	0.078	-0.049	**0.979**	0.013
WV	**0.906**	0.018	0.108	0.286	0.029	0.043	-0.274	-0.016	**-0.916**
O_3	**0.702**	0.238	-0.049	**0.962**	0.047	-0.081	-0.047	-0.059	-0.246

In the example shown, the variables that had most significant loads when using three PC were the concentration of CO and O_3 and the WV for PC_1, the T and RH for PC_2 and the ratio NO_2/NO for PC_3. When using six PC, each principal component had one variable with most significant load.

As mentioned before, the next day maximum ozone concentration was predicted through MLR and PCR models. The PC used in PCR were obtained multiplying standardized values of the original independent variables by the weights. The PCR was performed considering two (or three for the 4th trimester) and six PC separately.

In both model approaches, a t-Test was used to statistically evaluate the regression coefficients. The following procedure was the development of new regressions using only the predictor variables with statistically valid coefficients.

Considering three PCR, the statistically valid coefficients were the same using two or six PC, for all datasets, except for the 1st trimester and for the annual datasets with one more valid parameter when using six PC. As an example, for the 4th trimester, the variable used in the MLR was the wind velocity (WV) and in PCR, the valid PC were the PC_1 and the PC_3, which account for the CO and O_3 concentrations and for the ratio NO_2/NO and WV, respectively.

In the development step, the performance indexes computed were very similar for both models with the exception of MBE, whose values were always lower with PCR (values between -7.5×10^{-7} and -3.2×10^{-6}) than with MLR (values between -0.39 and -3.39).

For the models validation the periods used for each trimester and for the complete year were, respectively, the last 10 and 30 days.

The performance indexes calculated for the PCR model were better than those calculated for the MLR model. Table 3 presents the values of the performance indexes calculated during the validation step with both models, for the 4th trimester and annual datasets.

Table 3: Performance indexes for both models, for the 4th trimester and annual datasets, in the validation step.

Performance indexes	4th trimester		Annual	
	MLR	PCR	MLR	PCR
MBE	-37.47	-3.30	-2.32	0.23
MAE	37.47	7.88	9.47	9.48
RMSE	39.51	9.90	11.72	11.43
IA	0.140	0.89	0.84	0.82

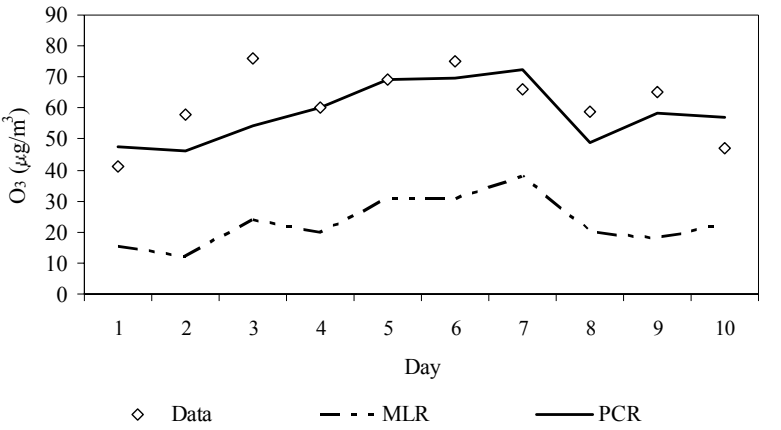
As an example, figs. 1a and 1b show the predictions with both models and the measured data corresponding to the 4th trimester and the annual validation periods. As can be seen the PCR model performance was superior throughout the validation period. It was also verified that the PCR model improved significantly the prediction of the ozone concentrations, showing to be a useful tool to provide the protection of the public health, through the early warnings of the population.

4 Conclusions

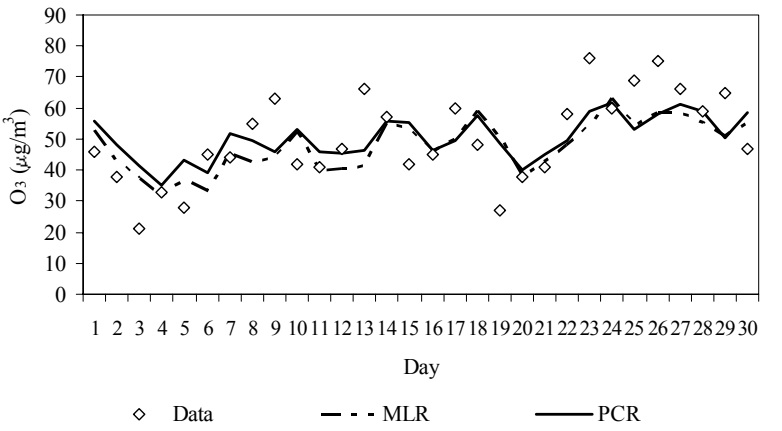
MLR and PCR were applied to predict the next day maximum of O_3 concentration. The variables initially used in the prediction models were the maxima hourly values for O_3, ratio of NO_2/NO, T, WV and the minima hourly values for CO and RH. Several variables were removed during the procedure to obtain significant statistical models.

In the development step, the performance indexes computed were very similar for both models with the exception of MBE, whose values were always lower with PCR than with MLR. In the validation step, the performance indexes calculated for the PCR model were better than those calculated for the MLR model.

Concluding, multiple linear regression based on principal components was more robust because it eliminated collinearity problems and reduced the number of variables presented in multiple regression models. It was also verified that PCR model improved significantly the prediction of the ozone concentrations, showing to be a useful tool to provide the protection of the public health, through the early warnings of the population.



(a)



(b)

Figure 1: Prediction of O_3 concentrations for: a) 4th trimester dataset, b) annual dataset.

References

[1] White, M. C., Etzel, R.A., Wilcox, W. D. & Lloyd C., Exacerbation of childhood asthma and ozone pollution in Atlanta. *Environmental Research*, **35**, pp. 56-68, 1994.

[2] WHO, Air Quality Guidelines, World Health Organization Regional Office, Copenhagen, 2000.
[3] Seinfeld, J. H. & Pandis, S. N., *Atmospheric Chemistry and Physics – from Air Pollution to Climate Changes*, John Wiley Sons, USA, 1998.
[4] Alvim-Ferraz, M. C. M., Sousa, S. I. V., Pereira, M. C. & Martins, F. G., Contribution of anthropogenic pollutants to the increase of tropospheric ozone levels in Oporto Metropolitan Area, Portugal since the 19th century. *Environmental Pollution*, in press, 2005.
[5] Gonçalves, F. L. T., Carvalho, L. M. V., Conde, F. C., Latorre, P. H. N. & Braga, A. L. F., The effects of air pollution and meteorological parameters on respiratory morbidity during the summer in São Paulo City. *Environment International*, **31**, pp. 343-349, 2005.
[6] Abdul-Wahab, S. A., Bakheit, C. S. & Al-Alawi, S. M., Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, **20**, pp. 1263-1271, 2005.
[7] Lengyel, A., Héberger, K, Paksy, L., Bánhidi, O. & Rajkó, R., Prediction of ozone concentration in ambient air using multivariate methods. *Chemosphere*, **57**, pp. 889-896, 2004.
[8] Pereira, M. C., Alvim-Ferraz, M. C. M. & Santos, R. C., Relevant aspects of air quality in Oporto Portugal): PM_{10} and O_3. *Environmental Monitoring and Assessment*, **101**, pp. 203-221, 2005.
[9] Heo, J-S. & Kim, D-S. A new method of ozone forecasting using fuzzy expert and neural network systems. *Science of the Total Environment*, **325**, pp. 221-237, 2004.
[10] Thompson, M. L., Reynolds, J., Cox, L. H., Guttorp, P. & Sampson P. D., A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, **35**, pp. 617-630, 2001.
[11] Çamdevýren, H., Demýr, N., Kanik, A. & Keskýn, S., Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. *Ecological Modelling*, **181**, pp. 581-589, 2005.
[12] Slini, T., Kaprara, A., Karatzas, K. & Moussiopoulos N., PM_{10} forecasting for Thessaloniki, Greece. *Environmental Modelling & Software*, in press.
[13] Peres-Neto, P. R., Jackson, D., A. & Somers, K., M., How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, in press.
[14] Chaloulakou, A., Saisana, M. & Spyrellis, N., Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment*, **313**, pp. 1-13, 2003.
[15] Gardner M., W. & Dorling, S. R., Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, **34**, pp. 21-34, 2000.