American Journal of Environmental Sciences 4 (5): 445-453, 2008
ISSN 1553-345X
© 2008 Science Publications

Prediction of the Daily Mean PM_{10} Concentrations Using Linear Models

J. C. M. Pires, F. G. Martins, S. I. V. Sousa, M. C. M. Alvim-Ferraz, M. C. Pereira
LEPAE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal.

Abstract: The performance of five linear models to predict the daily mean PM_{10} concentrations was compared. The linear models proposed were: (i) multiple linear regression; (ii) principal component regression; (iii) independent component regression; (iv) quantile regression; and (v) partial least squares regression. The study was based on data from an urban site in Oporto Metropolitan Area and the analysed period was from January 2003 to December 2005. The linear models were evaluated with two datasets of different sizes belonging to the analysed period. Environmental data (SO_2, CO, NO, NO_2 and PM_{10} concentrations) and meteorological data (temperature, relative humidity and wind speed) were used as PM_{10} predictors.During the training step, quantile regression presented the lowest residual errors for the two datasets. Independent component regression was the worst model using the larger dataset. Multiple linear regression, principal component regression and partial least squares regression presented similar results for both datasets. During the test step, independent component regression and quantile regression showed bad performance, while multiple linear regression, principal component regression and partial least squares regression presented similar results using the larger dataset. For the smaller dataset, the models that remove the correlation of the variables (principal component regression, independent component regression and partial least squares regression) presented better results than multiple linear regression and quantile regression. Independent component regression was the linear model with the lowest value of residual error. Concluding, the dataset size is also an important parameter for the evaluation of the models concerning the prediction of variables. The prediction of the daily mean PM_{10} concentrations was more efficient when using independent component regression for the smaller dataset and partial least squares regression for the larger datasets.

Key words: PM_{10} concentrations, multiple linear regression, principal component regression, independent component regression, quantile regression, partial least squares regression

INTRODUCTION

Atmospheric particulate matter is made up of solid and liquid particles suspended in the atmosphere. They are emitted by: (i) natural (volcanic eruptions, seismic activity, forest fires, winds of great intensity or natural particle transport from the dry regions); and (ii) anthropogenic sources (all types of combustion and some industrial processes). In Europe, particulate matter is one of the most important air pollutants responsible for loss of human health[1]. In the last decade, several studies about health effects of particulate matter were published[2-5]. Long exposure to PM_{10} (particles with diameter smaller than 10 μm) and to $PM_{2.5}$ (particles with diameter smaller than 2.5 μm) has been associated with respiratory and cardiovascular diseases. Recent research seems to indicate that particles with smaller sizes are associated with childhood morbidity and mortality[5].

The selection of the modelling techniques must consider some features, such as, complexity, flexibility, accuracy and speed of computation. The interpretability is also a very important characteristic of a model[6]. Without interpretability, the model is only used for prediction. In many situations, this type of model is enough. But in some cases, it is relevant to know the correlations between input variables and predictive variables. Furthermore, an interpretable model provides a sense of confidence. Comparing to the nonlinear models, linear models are simple and interpretable, taking less computation time.

Recently, several attempts have been made to model PM_{10} concentrations using different models. G. Corani[7] tried to predict this pollutant using feed-

Corresponding Author: F. G. Martins, LEPAE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

forward neural networks, pruned neural networks (nonlinear approaches) and lazy learning (local linear modelling approach). Comparing these three methodologies, lazy learning presents slightly better results than the other methods. Perez et al.[8] developed a neural network (nonlinear approach) to predict the maximum of the 24-h moving average of PM_{10} concentration on the next day. This method was compared with linear perceptron (linear approach) and presented slightly better performance. Fuller et al.[9] used an empirical model to predict concentrations of PM_{10} at background and roadside locations. The method was based on the regression analysis between PM_{10} and NO_x. The model accurately predicted daily mean PM_{10} concentrations but presented some limitations. For example, it depends of the existence of a consistent relationship between PM_{10} and NO_x emissions. Thus, the studies that compared linear and nonlinear models[7,8] did not find a significant difference in the results obtained by the different methodologies.

This paper aims to analyse the performance of linear models, obtained by different methodologies, to predict the next day daily mean PM_{10} concentrations. The considered models were: (i) multiple linear regression; (ii) principal component regression; (iii) independent component regression; (iv) quantile regression; and (v) partial least squares regression. The explanatory variables were meteorological data (daily means of temperature, relative humidity and wind speed) and environmental data (daily means of CO, SO_2, NO, NO_2 and PM_{10} concentrations of the previous day).

MODELS

Multiple Linear Regression: Multiple linear regression (MLR) attempts to model the relationship between two or more explanatory variables and a response variable, by fitting a linear equation to the observed data. The dependent variable (y) is calculated by:

$$y = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \varepsilon \quad (1)$$

where x_i (i=1,…,k) are the explanatory variables, $\hat{\beta}_i$ (i=0,…,k) are the regression coefficients, and ε is the error associated with the regression and assumed to be normally distributed with both expectation value zero and constant variance[10].

The predicted value given by the regression model ($\hat{y}$) is calculated by:

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i \quad (2)$$

The most common method to estimate the regression parameters $\hat{\beta}_i$ is the minimization of the sum of square errors (SSE). The equation is as follows:

$$\hat{\beta}_i = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

Principal Component Regression: Principal component regression (PCR) is a method that combines linear regression and principal component analysis (PCA). PCA creates new variables, the principal components (PC), that are orthogonal and uncorrelated. These variables are linear combinations of the original variables. The PC are ordered in such a way that the first component has the largest fraction of the original data variability[11-13]. To evaluate the influence of each variable in the PC, varimax rotation is generally used to obtain the rotated factor loadings that represent the contribution of each variable in a specific PC. PCR establishes a relationship between the output variable (y) and the selected PC obtained from the explanatory variables (x_i).

Independent Component Regression: Independent component regression (ICR) is a method that combines linear regression and independent component analysis (ICA). In ICA, the input variables are considered linear combinations of latent variables. These latent variables are considered nongaussian and independent[14]. Therefore, they are called independent components (IC). PCA and ICA are considered linear representation models. While PCA determines the orthogonal variables (PC), ICA tries to find independent variables (IC). ICR establishes a relationship between the output variable (y) and the selected IC obtained from the explanatory variables (x_i)

Quantile Regression: Quantile regression (QR) was introduced by Koenker and Bassett[15] and can be seen as a natural extension of the least squares estimation of conditional mean models. This method presents some advantages when compared with ordinary least squares regression. For example, it allows the examination of the entire distribution of the variable of interest rather than a single measure of the central tendency of its distribution. It can also provide information about any linear or nonlinear relationships between the dependent variable and the explanatory variables without an a priori knowledge of the type of (potential) non-linearities. Thus, it is more flexible to model data with

heterogeneous conditional distribution. To describe the quantile function, a random variable Y with the distribution function $F(y) = Pr(Y \le y)$ is considered. The quantile function $Q(\tau)$ with $\tau \in [0, 1]$ is defined as follows:

$$Q(\tau) = \inf\{y : F(y) \ge \tau\} \quad (4)$$

The median is Q(1/2), the first quartile is Q(1/4) and the first decile is Q(1/10). The median regression minimizes a sum of absolute errors. The remaining conditional quantile functions are estimated by minimizing an asymmetrically weighted sum of absolute errors:

$$\hat{Q}(\tau) = \arg\min_{a}\left\{\sum_{i:yi\ge a}\tau|y_i - a| + \sum_{i:yi<a}(1-\tau)|y_i - a|\right\} \quad (5)$$

Equation 6 presents a way to calculate the model parameters, considering quantile approach and the regression equation given by Equation 2:

$$\hat{\beta}(\tau) = \arg\min_{\beta(\tau)}\left\{\sum_{i:yi\ge \hat{y}_i}\tau|y_i - \hat{y}_i| + \sum_{i:yi<\hat{y}_i}(1-\tau)|y_i - \hat{y}_i|\right\} \quad (6)$$

Partial Least Squares Regression: Partial least squares regression (PLSR) is probably the least restrictive of the various multivariate extensions of the MLR model[16]. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. In these situations, the MLR approach is not feasible due to the multicollinearity between the explanatory variables.

PLSR is based on linear transition from a large number of original descriptors to a new variable space based on small number of orthogonal factors (latent variables - LV)[16]. In other words, factors are orthogonal and linear combinations of original descriptors. Unlike some similar approaches (e.g. PCR), LV are chosen in such a way that provides maximum correlation with dependent variable. Thus, PLSR model contains the smallest necessary number of factors.

PLSR decomposes both explanatory variables (X) and the output variables (Y) as a product of a common set of orthogonal factors and a set of specific loadings. The complete procedure is as follows[16]:

Step 1: Normalization of X and Y: $X_0 = X/\|X\|$ and $Y_0 = Y/\|Y\|$;

Step 2: Definition of the vector u with random values;

Step 3: Estimation of the X weights: $w = \frac{X_k{}^T u}{\|X_k{}^T u\|}$;

Step 4: Estimation of the X factor scores (LV): $t = \frac{X_k w}{\|X_k w\|}$;

Step 5: Estimation of the Y weights: $c = \frac{Y_k{}^T t}{\|Y_k{}^T t\|}$;

Step 6: Estimation of the Y factor scores: $u = Y_k c$;

Step 7: Repetition of the steps 3 to 6 until the convergence of t;

Step 8: Determination of the value of b used to predict Y from t: $b = t^T u$;

Step 9: Determination of the X factor loadings: $p = X_k{}^T t$;

Step 10: Elimination of the effect of t from X and Y: $X_{k+1} = X_k - tp^T$ and $Y_{k+1} = Y_k - btc^T$;

Step 11: Repetition of the steps 2 to 10 until the determination of a selected number of LV.

Each vector t, u, w, c and p is stored in the columns of the correspondent matrices (T, U, W, C and P) and the scalar b is stored in a diagonal matrix (B). If X_k is a null matrix, all latent variables were determined.

The prediction of the dependent variable is done by $\hat{Y} = TBC^T = XB_{PLS}$. If all latent variables are used, the results of PLSR are similar to that obtained by PCR.

Regression parameters validation: It is important to know which explanatory variables are relevant to predict the dependent variable. For the studied models, PLSR is the only one that includes this step in its procedure. For MLR, PCR and ICR, the significance of each regression parameter in the models was evaluated through the calculation of their confidence interval. The parameter $\hat{\beta}_i$ is statistically significant if[17]:

$$\left|\hat{\beta}_i\right| > \frac{t_{n-k-1}^{\alpha/2}\hat{\sigma}}{\sqrt{Sxx_i}} \quad (7)$$

where t is the Student t distribution, n is the number of points, k is the number of parameters, α is the significance level, $\hat{\sigma}$ is the standard deviation given by $\sqrt{\frac{SSE}{n-k-1}}$ and Sxx_i is the sum of squares related to x_i given by $\sum_{j=1}^{n}\left(x_{i,j} - \bar{x}_i\right)^2$.

For QR, bootstrap estimates of standard error (at the 95% confidence level) were calculated by randomly sampling each dataset with replacement (1000 times).

Performance indexes: The linear models were compared through the calculation of the following statistical parameters: mean bias error (MBE), mean absolute error (MAE), root mean squared error (RMSE), index of agreement (d_2), that are commonly referred in literature[18, 19].

MBE indicates if the observed values are over or under estimated. MAE and RMSE measure residual errors, which give a global idea of the difference between the observed and the modelled values. The values of d_2 compare the difference between the mean, the predicted and the observed variables, indicating the degree of error free for the predictions[18, 19].

DATA

This study aims to predict the daily mean PM_{10} concentration of the next day. The concentrations of pollutants (SO_2, CO, NO, NO_2 and PM_{10}) were recorded in an urban site (Matosinhos) with traffic influences situated in Oporto Metropolitan Area, Northern Portugal. SO_2 concentrations were obtained by the ultraviolet fluorescence method; CO concentrations were measured through nondispersive infrared spectrometric; NO_2 was obtained through the chemiluminescence method; PM_{10} concentrations were obtained through the beta radiation attenuation method. These equipments were submitted to a rigid maintenance program and calibrated periodically. Measurements were continuously made and hourly average concentrations (in micrograms per cubic meter) were registered. Meteorological variables were measured on the left edge of Douro River at an altitude of 90 m approximately. These variables are hourly means of: air temperature (T), relative humidity (RH) and wind speed (WS). Daily average values for these variables were calculated and used if more that 75% of hourly values were available. The period of measurement was from January 2003 to December 2005. To evaluate the influence of the dataset size in the performance of the linear models, two datasets with different sizes were considered: (i) dataset 1 (DS1) considered 2003, 2004 and the first three trimesters of 2005 as training period and the last trimester of 2005 as test period; and (ii) dataset 2 (DS2), January 2003 to May 2003 was considered as training period and June 2003 was considered as test period. Additionally, QR and PLSR models required a validation period. For DS1, the last trimester of 2004 was considered as a validation period. For DS2, May 2003 was considered as a validation period. The explanatory variables were standardized (zero mean and unit standard deviation).

RESULTS AND DISCUSSION

Table 1 presents the correlation coefficients between pollutants and meteorological variables for the two analysed datasets (DS1 – upper triangular matrix; DS2 – lower triangular matrix). These coefficients provide a measure of linear relationship between two variables. Values in bold correspond to statistically significant coefficients [20]. For the evaluation of their statistical significance, the critical correlation coefficient was calculated (with a significance level of 0.05). A correlation coefficient is significant if its absolute value is greater than the critical value. In this preliminary analysis, concerning the correlations with the next day PM_{10} concentration ($PM_{10|d+1}$), only temperature was not considered statistically significant for DS2. The explanatory variables that presented highest correlation with $PM_{10|d+1}$ were CO, NO, NO_2 and PM_{10} concentrations (for both datasets). In both datasets, as it was expected, $PM_{10|d+1}$ concentration had positive correlations with SO_2, CO, NO and NO_2. Some combustion and industrial processes (important PM_{10} sources) also increase the emission of these pollutants. $PM_{10|d+1}$ concentration had negative correlation with RH and WS. In wet weather, the particles in suspension can stick on the surface and can be whirled up into the air in dry weather. For high values of WS, there is an efficient dispersion of pollutants. Thus, high values of WS correspond to low pollutant concentrations. $PM_{10|d+1}$ concentration had also positive correlation with its concentration of the previous day.

For the statistical evaluation of the regression parameters, a t-test (significance level of 0.05) was performed for MLR, PCR and ICR. Considering the statistically significant parameters, new regressions were performed. Table 2 shows the statistically significant regression parameters for MLR, PCR and ICR. For MLR, the parameters β_1 to β_8 correspond to SO_2, CO, NO, NO_2, T, RH, WS and PM_{10}, respectively. The variables considered important for the prediction of the next day PM_{10} concentration were: (i) CO, NO_2, T, RH and PM_{10} for DS1; and (ii) CO, NO, NO_2 and T for DS2. For PCR, the parameters β_i (i=1, 8) correspond to each PC. All PC were used as input variables in PCR. However, the important predictors of PM_{10} concentration considered by this method were: (i) PC1 to PC6 for DS1; and (ii) PC1, PC2, PC4 and PC8 for DS2. Table 3 shows the rotated factor loadings for DS1 and DS2 that represent the contribution of each variable

Table 1: Correlation coefficients between pollutants and meteorological variables for the two analysed datasets (DS1 – upper triangular matrix; DS2 – lower triangular matrix).

	SO_2	CO	NO	NO_2	T	RH	WS	PM_{10}	$PM_{10}\|_{d+1}$
SO_2	1	**0.171**	**0.329**	**0.430**	**0.297**	**-0.163**	**-0.155**	**0.371**	**0.276**
CO	**0.510**	1	**0.860**	**0.786**	**-0.374**	**-0.086**	**-0.343**	**0.638**	**0.486**
NO	**0.575**	**0.917**	1	**0.798**	**-0.211**	-0.062	**-0.328**	**0.565**	**0.439**
NO_2	**0.619**	**0.882**	**0.832**	1	-0.043	**-0.318**	**-0.283**	**0.734**	**0.594**
T	-0.089	**-0.433**	**-0.366**	**-0.295**	1	**-0.251**	-0.022	**0.168**	**0.153**
RH	**-0.333**	-0.138	-0.167	**-0.338**	**-0.270**	1	**-0.278**	**-0.275**	**-0.323**
WS	**-0.221**	**-0.532**	**-0.447**	**-0.482**	**0.188**	-0.065	1	**-0.234**	**-0.150**
PM_{10}	**0.436**	**0.773**	**0.684**	**0.778**	-0.012	**-0.289**	**-0.401**	1	**0.685**
$PM_{10}\|_{d+1}$	**0.352**	**0.640**	**0.551**	**0.675**	-0.052	**-0.337**	**-0.289**	**0.669**	1

Values in bold correspond to statistically significant correlation coefficients

Table 2. Statistically significant regression parameters for MLR, PCR and ICR using both datasets (DS1 and DS2).

		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
MLR			SO_2	CO	NO	NO_2	T	RH	WS	PM_{10}
	DS1	42.12		2.36		2.88	1.98	-2.74		10.82
	DS2	40.89		16.17	-8.17	10.54	5.87			
PCR			PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
	DS1	42.12	7.35	4.35	1.41	4.82	4.87	5.19		
	DS2	40.89	7.62	3.90		7.23				-13.49
ICR			IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8
	DS1	26.39	7.57		3.14		-2.95	1.65	2.49	3.19
	DS2	6.11	7.50	8.15	6.42		-10.83	-3.55		4.22

Table 3. Varimax rotated loadings using both datasets (DS1 and DS2).

		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
DS1	SO_2	-0.171	0.159	-0.061	**0.960**	0.065	-0.121	-0.035	0.002
	CO	**-0.842**	-0.277	-0.176	0.020	0.050	-0.318	-0.008	-0.280
	NO	**-0.952**	-0.090	-0.132	0.162	-0.003	-0.144	0.034	0.139
	NO_2	**-0.742**	-0.014	-0.129	0.234	0.223	-0.355	-0.449	-0.004
	T	0.177	**0.957**	-0.033	0.160	0.132	-0.088	0.000	0.010
	RH	0.076	-0.127	-0.160	-0.065	**-0.967**	0.106	0.033	0.004
	WS	0.194	-0.031	**0.962**	-0.062	0.161	0.077	0.021	0.008
	PM_{10}	-0.440	0.129	-0.094	0.161	0.141	**-0.855**	-0.053	-0.011
DS2	SO_2	-0.281	-0.028	-0.076	**0.929**	0.167	-0.146	0.045	0.006
	CO	**-0.753**	-0.276	-0.275	0.209	0.053	-0.429	0.081	0.211
	NO	**-0.878**	-0.186	-0.190	0.280	0.065	-0.258	0.000	-0.096
	NO_2	**-0.607**	-0.180	-0.250	0.327	0.225	-0.439	0.432	0.013
	T	0.204	**0.964**	0.068	-0.029	0.150	-0.016	-0.027	-0.005
	RH	0.082	-0.147	-0.050	-0.150	**-0.967**	0.107	-0.031	-0.001
	WS	0.225	0.070	**0.955**	-0.074	0.055	0.151	-0.037	-0.006
	PM_{10}	-0.442	0.064	-0.181	0.162	0.141	**-0.848**	0.046	-0.004

Values in bold indicate the variables that most influence each principal component.

in a specific PC. In both datasets, PC1 is heavily loaded by CO, NO and NO_2 and PC2 to PC6 had greater contributions of T, WS, SO_2, RH and PM_{10}, respectively. Thus, the most important original variables selected by PCR were: (i) all variables for DS1; and (ii) SO_2, CO, NO, NO_2 and T for DS2. Similarly to PCR, with ICR the parameters β_i (i=1, 8) correspond to each IC. All IC were used as input variables in ICR. However, the important predictors of PM_{10} concentration considered by this method were: (i) IC1, IC3, IC5, IC6, IC7 and IC8 for DS1; and (ii) IC1, IC2, IC3, IC5, IC6 and IC8 for DS2. Table 4 presents the correlation matrix between the original variables and IC. These correlation values showed the importance of each original variable on the prediction of PM_{10} concentration. Thus, the original variables considered relevant for prediction of PM_{10} concentration were: (i) SO_2, CO, NO, NO_2, WS and PM_{10} for DS1; and (ii) CO, NO, NO_2, T, RH, WS and PM_{10} for DS2.

Table 4. Correlation matrix between the original variables and the IC using both datasets (DS1 and DS2).

		IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8
DS1	SO_2	**0.329**	0.038	0.012	**-0.129**	0.014	**0.923**	**0.107**	**0.096**
	CO	**0.410**	**0.203**	0.005	**0.486**	**-0.244**	0.003	**0.288**	**0.642**
	NO	**0.428**	**0.146**	0.041	**0.195**	**0.138**	**0.110**	**0.235**	**0.818**
	NO_2	**0.807**	**0.170**	**0.167**	**0.228**	0.027	**0.128**	**0.323**	**0.346**
	T	**0.291**	0.040	-0.053	**-0.901**	**-0.235**	**0.120**	**-0.081**	**-0.152**
	RH	**-0.521**	**0.776**	-0.011	**0.102**	**0.212**	0.010	**-0.256**	**0.071**
	WS	**0.133**	**-0.543**	-0.018	**0.127**	0.037	**-0.065**	**-0.798**	**-0.172**
	PM_{10}	**0.526**	**0.141**	**0.582**	0.010	**-0.426**	**0.152**	**0.163**	**0.364**
DS2	SO_2	-0.137	**0.300**	**0.389**	0.118	**-0.470**	-0.111	**0.700**	-0.013
	CO	**0.369**	**0.287**	**0.600**	-0.102	**-0.636**	0.065	-0.055	-0.019
	NO	0.018	**0.290**	**0.775**	-0.064	**-0.550**	0.012	-0.088	-0.031
	NO_2	0.124	**0.498**	**0.331**	-0.011	**-0.782**	-0.043	-0.017	0.110
	T	-0.053	**0.201**	-0.078	**0.475**	**0.550**	-0.143	0.097	**0.621**
	RH	-0.060	**-0.616**	0.000	**-0.465**	0.035	**0.591**	-0.023	**0.223**
	WS	**-0.232**	**-0.661**	**-0.241**	**0.551**	0.058	**-0.364**	-0.057	-0.082
	PM_{10}	**0.385**	**0.306**	**0.431**	-0.153	**-0.474**	**-0.384**	-0.017	**0.421**

Values in bold correspond to statistically significant correlation coefficients

Table 5. Statistically significant regression parameters for QR using both datasets (DS1 and DS2).

	τ	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
			SO_2	CO	NO	NO_2	T	RH	WS	PM_{10}
DS1	0.10	23.7		5.39			3.72	-3.50		4.6
	0.30	33.6					0.64	-2.73		12.3
	0.50	40.8					0.47	-2.41		15.7
	0.70	49.0				2.25		-2.32		16.3
	0.90	62.3				7.33				16.7
			SO_2	CO	NO	NO_2	T	RH	WS	PM_{10}
DS2	0.10	22.7		13.80			2.99	-2.45		
	0.30	22.9								
	0.50	33.3								
	0.70	47.2								21.0
	0.90	62.3	-5.85			26.86				

Table 5 shows the statistically significant regression parameters for quantile regression. Five percentiles (τ) were selected. For the evaluation of the statistical significance of the regression parameters, their confidence intervals (significance level of 0.05) were determined by the application of bootstrap (calculated with 1000 times replacements). For DS1, the results showed that RH and PM_{10} concentration were the most important explanatory variables. CO concentration and T were important only in low values of τ, while NO_2 concentration was relevant in high values of τ. For DS2, SO_2, NO_2 and PM_{10} concentrations were important for high values of τ, while CO concentration, T and RH were relevant for low values of τ.

For PLSR, it is important to determine the number of the LV needed to obtain the best generalisation for the prediction of PM_{10} concentrations. The validation period was used to determine the number of LV that presents the lowest value of error and in both datasets, two variables were achieved. Table 6 shows the regression parameters for PLSR (B_{PLS}) and the rotated factor loadings that represent the contribution of each variable in a specific LV. For DS1, LV1 was heavily loaded by the CO, NO, NO_2 and PM_{10} concentrations, while LV2 had a great contribution of T. For DS2, LV1 was heavily loaded by the SO_2, CO, NO, NO_2 and PM_{10} concentrations, while LV2 had a great contribution of T.

Table 7 shows the performance indexes for the different linear models for the training period. ICR presented bad performance in the larger dataset (DS1). QR was the linear method that presented better performance using the two datasets.

In the test step and for MLR, PCR and ICR, PM_{10} concentrations were determined by the application of the regression equations achieved in the training step. For this step, the application of QR needs to predict the PM_{10} percentile for each validation point, following the application of the correspondent regression equation. The PM_{10} percentile was determined applying the k-nearest neighbour (k-NN) algorithm. This algorithm was used for classifying objects based on closest examples in the training data. It was based on the

Table 6. Regression parameters for PLSR (B_{PLS}) and the rotated factor loadings using both datasets (DS1 and DS2).

	DS1			DS2		
	B_{PLS}	LV1	LV2	B_{PLS}	LV1	LV2
SO_2	0.025	0.504	-0.060	-0.010	**0.722**	-0.273
CO	0.115	**0.826**	-0.371	0.196	**0.960**	-0.041
NO	0.041	**0.822**	-0.431	0.081	**0.918**	-0.112
NO_2	0.172	**0.936**	-0.097	0.189	**0.955**	-0.003
T	0.149	0.087	**0.685**	0.096	-0.438	**0.700**
RH	-0.180	-0.395	-0.602	-0.138	-0.371	-0.459
WS	0.008	-0.309	0.374	0.017	-0.545	0.286
PM_{10}	0.327	**0.868**	0.266	0.287	**0.847**	0.393

Values in bold indicate the variables that most influence each latent variable.

Table 7: Performance indexes of the different linear models for the training period using both datasets (DS1 and DS2).

		R	MBE	MAE	RMSE	d_2
DS1	MLR	0.71	0.00	11.99	15.92	0.81
	PCR	0.71	0.00	11.96	15.91	0.81
	ICR	0.43	0.00	15.58	20.45	0.55
	QR	0.79	2.60	9.93	14.21	0.92
	PLSR	0.70	0.00	12.45	16.38	0.81
DS2	MLR	0.72	0.00	12.78	16.55	0.82
	PCR	0.73	0.00	12.51	16.42	0.83
	ICR	0.74	0.00	12.18	16.27	0.83
	QR	0.84	1.24	7.99	13.07	0.94
	PLSR	0.72	0.00	12.78	16.80	0.82

Table 8: Performance indexes of the different linear models for the test period using both datasets (DS1 and DS2).

		R	MBE	MAE	RMSE	d_2
DS1	MLR	0.74	-1.12	12.66	18.43	0.83
	PCR	0.74	-0.90	12.68	18.37	0.84
	ICR	0.68	-1.90	13.97	20.22	0.71
	QR	0.60	2.16	15.20	21.95	0.86
	PLSR	0.75	-2.07	12.24	18.13	0.83
DS2	MLR	0.70	5.92	11.76	13.84	0.84
	PCR	0.76	-0.25	10.87	12.64	0.86
	ICR	0.79	0.64	10.02	11.83	0.88
	QR	0.72	-3.44	11.14	13.53	0.86
	PLSR	0.77	4.34	10.50	12.48	0.87

Euclidean distance between the correspondent validation point and the training points. The evaluation of the optimal value of *k* nearest training samples depends of the dataset. A good value of *k* can be achieved using cross-validation. The k-NN algorithm is as follows:

Step 1: Selection of the *k* value;

Step 2: Determination of *k* nearest training points from the validation point;

Step 3: Determination of the percentile of PM_{10} concentration values correspondent to these training points;

Step 4: Application of the QR equations correspondent to these percentiles using the validation point:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i ;$$

Step 5: Determination of the average of the *k* values of $\hat{y}_i$;

Step 6: Repetition of the steps 2 to 5 for all validation points;

Step 7: Determination of the error associated to the value of *k*, based on the difference of the average values calculated above and the true values;

Step 8: Repetition of the steps 1 to 7 for different values of *k*;

Step 9: Determination of the lowest value of error associated to the optimal value of *k*.

For the test step, it was necessary 23 and 5 nearest points for prediction of the percentile of the test points using DS1 and DS2, respectively. For PLSR, PM_{10} concentrations were determined through the B_{PLS} (equation $\hat{Y} = XB_{PLS}$) achieved with the number of LV obtained in the validation step. Table 8 presents the performance indexes achieved in the test step using DS1 and DS2. The results showed that for the larger dataset (DS1), QR and ICR presented the highest values of residual error and the remained models had similar results. For DS2, ICR presented the best performance, achieving the lowest value of RMSE. The other models that remove the correlation between the input variables (PCR and PLS) also presented good results. MLR and QR achieved similar performance indexes.

CONCLUSIONS

Five linear models were used to predict the daily mean PM_{10} concentrations using as predictors air pollutant (SO_2, CO, NO, NO_2 and PM_{10}) concentrations and meteorological parameters (temperature, relative humidity and wind speed) for two datasets with different sizes. PM_{10} concentration presented positive correlation with SO_2, CO, NO and NO_2 concentrations due to the similar emission sources. It was also positively correlated with temperature and negatively with relative humidity and wind speed. In wet weather, the particles in suspension can stick on the surface and can be whirled up into the air in dry weather. For high values of wind speed, there is an efficient dispersion of pollutants.

As the selected models are all linear, they are interpretable and their results were used to determine which variables were important in the prediction of daily mean PM_{10} concentrations. All models considered different group of important variables, however CO and NO_2 were always selected.

During the training step, quantile regression presented the lowest residual errors using the two datasets. Multiple linear regression, principal component regression and partial least squares regression presented similar results. Independent component regression presented bad performance in the dataset 1, which may be the result of the dataset size.

During the test step and for the dataset 1, independent component regression and quantile regression showed bad performances, while multiple linear regression, principal component regression and partial least squares regression presented similar results. Using the smaller dataset (dataset 2), the models that remove the correlation of the variables (principal component regression, independent component regression and partial least squares regression) presented better results than multiple linear regression and quantile regression. Independent component regression was the linear model with the lowest value of residual error.

Concluding, the dataset size is also an important parameter for the evaluation of the models concerning the prediction of variables. The prediction of the daily mean PM_{10} concentrations was more efficient when using independent component regression for the smaller dataset and partial least squares regression for the larger dataset.

ACKNOWLEDGMENTS

Authors are grateful to Comissão de Coordenação da Direcção Regional-Norte and to Instituto Geofísico da Universidade do Porto, for kindly providing the air quality and meteorological data. This work was supported by Fundação para a Ciência e Tecnologia (FCT). J. C. M. Pires also thanks the FCT for the fellowship SFRD/BD/23302/2005.

REFERENCES

1. Koelemeijer, R.B.A., C.D. Homan and J. Matthijsen, 2006. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. Atmos. Environ., 40 (27): 5304-5315.
2. Alvim-Ferraz, M.C., M.C. Pereira, J.M. Ferraz, A.M.C. Almeida e Mello and F.G. Martins, 2005. European Directives for Air Quality: Analysis of the New Limits in Comparison with Asthmatic Symptoms in Children Living in the Oporto Metropolitan Area, Portugal. Hum. Ecol. Risk Assess., 11(3): 607-616.
3. Brunekreef, B. and S. T. Holgate, 2002. Air pollution and health. The Lancet, 360 (9341): 1233-1242.
4. Hoek, G., B. Brunekreef, S. Goldbohm, P. Fischer and P.A. van der Brand, 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. The Lancet, 360 (9341): 1203-1209.
5. Kappos, A.D., P. Bruckmann, T. Eikmann, N. Englert, U. Heinrich, P. Hoppe, E. Koch, G. H. M. Krause, W.G. Kreyling, K. Rauchfuss, P. Rombout, V. Schulz-Klemp, W.R. Thiel and H.E. Wichmann, 2004. Health effects of particles in ambient air. Int. J. Hyg. Environ. Health, 207 (4): 399-407.
6. Guha, R., D.T. Staton, P.C. Jurs, 2005. Interpreting Computational Neural Network Quantitative Structure-Activity Relationship Models: A Detailed Interpretation of the Weights and Biases. J. Chem. Inf. Model., 45: 1109-1121.
7. Corani, G., 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model., 185 (2-4): 513-529.
8. Perez, P. and J. Reyes, 2002. Prediction of maximum of 24-h average of PM_{10} concentrations 30 h in advance in Santiago, Chile. Atmos. Environ., 36 (28): 4555-4561.
9. Fuller, G.W., D.C. Carslaw and H.W. Lodge, 2002. An empirical approach for the prediction of daily mean PM_{10} concentrations. Atmos. Environ., 36 (9): 1431-1441.
10. Agirre-Basurko, E., G. Ibarra-Berastegi and I. Madariaga, 2006. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environ. Modell. Softw., 21 (4): 430-446.
11. Abdul-Wahab, S.A., C.S. Bakheit, S.M. Al-Alawi, 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Modell. Softw., 20 (10): 1263-1271.
12. Wang, S. and F. Xiao, 2004. AHU sensor fault diagnosis using principal component analysis method. Energy Build., 36(2): 147-160.
13. Sousa, S.I.V., F.G. Martins, M.C.M. Alvim-Ferraz and M.C. Pereira, 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ. Modell. Softw., 22 (1): 97-103.

14. Shao, X., W. Wang, Z. Hou and W. Cai, 2006. A new regression method based on independent component analysis. Talanta, 69 (3): 676-680.
15. Koenker, R. and G. Basset, 1978. Regression quantiles. Econometrica, 44: 33-50.
16. Abdi, H., 2003. Partial Least Squares (PLS) Regression. In: Encyclopedia of Social Sciences Research Methods (eds M. Lewis-Beck, A. Bryman and T. Futing) pp. 1-7. Thousand Oaks: Sage.
17. Hayter, A.J., H.P. Wynn and W. Liu, 2005. Slope modified confidence bands for a simple linear regression model. Stat. Methodol., 3 (2): 186-192.
18. Chaloulakou, A., M. Saisana and N. Spyrellis, 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Sci. Total Environ., 313 (1-3): 1-13.
19. Gardner M.W. and S.R. Dorling, 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmos. Environ., 34 (1): 21-34.
20. Sousa, S.I.V., F.G. Martins, M.C. Pereira and M.C.M. Alvim-Ferraz, 2006. Prediction of ozone concentrations in Oporto city with statistical approaches, Chemosphere, 64 (7): 1141-1149.