Regulated metal levels in particulate matter in the Cantabria region (Northern Spain) using multivariate linear regression (MLR)

A. Arruti, I. Fernández & A. Irabien
Departamento de Ingeniería Química y Química Inorgánica, ETSIIyT, Universidad de Cantabria, Spain

Abstract

The levels and chemical composition of the particulate matter (PM) are linked to the effects of this atmospheric pollutant on human health. An assessment of the PM levels and its constituents present in the atmosphere is an important requirement of the air quality management and air pollution abatement. Taking into account that (i) EC Directives allow the Regional Government to assess the air quality by objective estimation and modelling techniques and (ii) the experimental effort required in the analysis of heavy metals in air, the present work aims to estimate the annual levels of the heavy metals regulated by the EC Directives (Pb in 1999/30/EC, and As, Ni and Cd in 2004/107/EC) in PM10 by means of multivariate linear regression (MLR). The main results show that although important deviations are found for individual measurements, the 2008 annual average metal concentrations are well estimated by the MLR technique at the studied areas. So, these estimations may be used by Regional Governments for the level assessment of regulated metals when their concentrations are below the lower assessment threshold.

Keywords: multivariate linear regression, objective estimation, particulate matter, metal levels.

1 Introduction

The European Union (EU) Framework Directive on ambient air quality and the daughter directives designate the use of modelling and estimation techniques for the assessment of air quality, especially for zones where concentrations of

WIT Transactions on Ecology and the Environment, Vol 136, © 2010 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR100051

pollutants in ambient air do not exceed the lower assessment threshold [1]. According to these directives the objective estimations can be applied by Regional Governments to estimate the regulated metal levels (Pb, As, Ni and Cd) in particulate matter less than 10 µm (PM10). In contrast to measurements, the EU directive does not define a reference method for objective estimations; nevertheless, the estimated values must fulfil some data quality objectives.

The concentration and composition of the particulate matter is very difficult to estimate due to the complexity of the processes, which control the formation, transportation and removal of aerosol in the atmosphere [2]; the composition of the PM also depends on the emission sources [3]. However, the development of estimation tools of the metal levels in PM10 is very useful because it reduces the required analysis of PM10.

In recent years, multiple linear regression (MLR), feedforward artificial neural networks as well as principal component regressions (PCR) are being used to estimate some air pollutants such as sulphur dioxide, ozone or particulate matter, PM10 and PM2.5 [2, 4, 5].

This work aims to estimate the annual levels of the regulated heavy metals (Pb in 1999/30/EC and As, Ni and Cd in 2004/107/EC) in PM10 by means of MLR and PCR; the regressions were applied with the software package SPSS 17.0. The statistical model is developed as a proper alternative to the experimental determination at some selected urban sites in the region of Cantabria (Northern Spain). A comparative study between the estimated and measured values is also carried out to test the fulfilment of EU requirements for data quality; the comparison is performed by statistical parameters such as the correlation coefficient, the fractional bias, the root mean square error or the normalised mean square error.

2 Methodology

2.1 The area of study

Cantabria Region (Northern Spain) covers an area of 5321 km^2 between the mountains of the Cordillera Cantábrica and the Cantabrian Sea. Santander is the most important urban area of the region; this city and some surrounding towns make up an agglomeration area (106 km^2) with about 250000 inhabitants (almost the half of the population of Cantabria).

The present study was performed at three different urban sites at the Cantabria Region: Santander, Castro Urdiales and Reinosa (figure 1).

Santander, SANT, (182700 inhabitants, 2009) extends over a wide bay. An industrial area mostly related to steel and ferroalloys manufacturing plants is located at Santander suburbs, 5-10 Km SW.

Castro Urdiales, CU, (31670 inhabitants, 2009) is a coastal urban site at the oriental zone of Cantabria. This town is in close vicinity to a national highway; furthermore, it is also located at 10-30 Km NW to a highly industrial area, which is close to Bilbao agglomeration (refinery, steel manufacturing plants).

Reinosa, REIN, (10307 inhabitants, 2009) is located in the interior of the Cantabria Region; this town is near to an industrial area (steel manufacturing plant).

2.2 The data

The present study is based on data measured during the year 2008. In particular, the data set consists of one year long daily values with a time coverage greater than 14%. The data were selected on the basis of homogeneity and completeness; hence, some quality criteria were established such as 2-3 data per month or 30% of the data must be associated to weekend. The data are divided into two groups; predictor/input variables and predictands/output variables [6].

Table 1: Predictor variables.

Variables	*Type*	*Maximum value*	*Minimum value*
PM10	C	62 μg/m^3	8 μg/m^3
SO_2	C	15 μg/m^3	0.4 μg/m^3
O_3	C	95 μg/m^3	10 μg/m^3
NO_x	C	88 μg/m^3	2 μg/m^3
T	C	25 ºC	1 ºC
RH	C	98 %	42 %
WD	C	360º	0º
PP	C	45 L/m^2	0 L/m^2
SE	N	1 Winter;2 Spring;3 Summer;4 Fall	
SD	N	0 No intrusion; 1 Intrusion	
WE	N	0 No weekend; 1 Weekend	

C; Continuous variable.
N; Nominal variable.

The study considered as predictors two types of variables, table 1: continuous variables and nominal variables. The continuous variables are the daily concentrations of particulate matter (PM10, μg/m^3), sulphur dioxide (SO_2, μg/m^3), ozone (O_3, μg/m^3), nitrogen oxides (NO_x, μg/m^3) and the daily meteorological variables of temperature (T, ºC), relative humidity (RH, %), wind direction (WD, º) and precipitation (PP, L/m^2). The seasonal (SE), the Saharan dust intrusion (SD) and the weekend (WE) effects are taken into account using nominal variables. All the predictor variables are measured using automatic equipment; these data are available at the Regional Environment Ministry of the Cantabria Government website.

The output variables are the daily regulated metal levels in PM10, which were determined using UNE-EN 12341 PM10 samplers. The filter digestion was carried out in a microwave oven; during the first hour, the temperature increases slowly to the top temperature, 185ºC. As, Ni, Cd and Pb were determined by inductively coupled quadrupole mass spectroscopy (ICP-MS). In order to validate the method of analysis, the procedure was tested with Standard Reference Material (SRM) 1649a "urban dust".

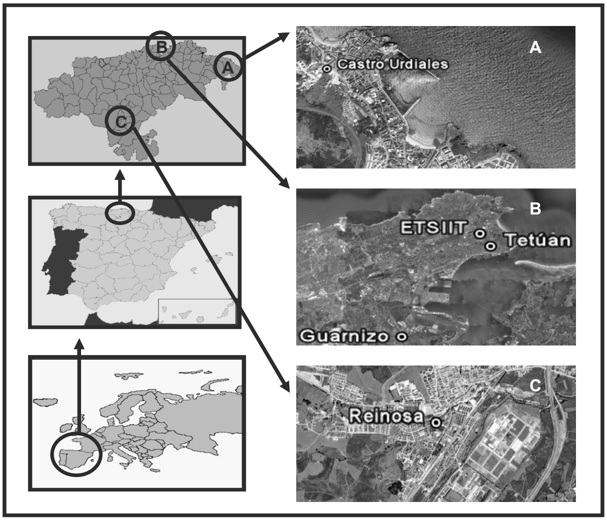

Figure 1: Selected urban sites in the Cantabria region: (A) Castro Urdiales, (B) Santander and (C) Reinosa.

The predictor variables at SANT are monitored at different sites; (i) the daily concentration data at Tetúan station and (ii) the meteorological data at Guarnizo station, which places to 10 Km SW of Tetúan station. The predictands were measured at the rooftop of the building ETSIIyT; this urban sampling site is in very close vicinity of Tetúan site, figure 1. The Tetúan and Guarnizo monitoring stations are integrated in the Air Quality Monitoring Network of the Cantabria Regional Government. Pollutant and meteorological data for CU and REIN are monitored at the same stations in both cases, which are also integrated in the Regional Air Quality Monitoring Network.

At SANT site, additional data were measured from January to September 2009 that were used as validation data set for the developed MLR models.

2.3 Performance indexes

The air quality models and estimations need to be properly evaluated before their predictions can be used with confidence. The information about uncertainties should be correctly calculated and interpreted since it is as important as modelling data [7]. Uncertainties can be characterized and objective estimation evaluation can be determined by statistical analysis, where estimation is examined to see how well they match the observations [7].

Table 2 presents the main statistical parameters used as quality indicators. Some studies conclude that every statistical parameters play a role in the evaluation of the objective estimation, but some of the these statistical parameters could be considered more important and useful: the correlation coefficient (r), the fractional bias (FB), the root mean square error (RMSE) and the normalised mean square error (NMSE) [4, 8]. The correlation coefficient can be interpreted as the proportion of the predictand variation that is described by the regression, the fractional bias gives information about the difference between average observed and estimated results and the normalised mean square error is based on the errors obtained within the observed and estimated pairs of results. A normalized form of the parameter, normalised mean square error, could be more adequate because the normalized parameter ignores the range of the variable in consideration.

Table 2: Performance indexes.

Quality indicators	***Formula***	***Ideal value***
Correlation coefficient	$r = \left[\frac{\sum_{i=1}^{n} \left(C_{O,i} - \overline{C_O}\right)\left(C_{E,i} - \overline{C_E}\right)}{\sqrt{\rho_O \rho_E}} \right]$	1
Fractional Bias	$FB = \frac{\overline{C_O} - \overline{C_E}}{0.5\left(\overline{C_O} + \overline{C_E}\right)}$	0
Root Mean Square Error	$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_{O,i} - C_{E,i}\right)^2}$	0
Normalized Mean Square Error	$NMSE = \frac{\overline{\left(C_O - C_E\right)^2}}{\overline{C_O}\,\overline{C_E}}$	0

O index is observed
E index is estimated

2.3.1 Uncertainty according to EU directives

The directive 96/62/EC and the daughter directives establish the requirements for air quality modelling and objective estimation. In this context, the uncertainty is defined as the maximum deviation of the measured and calculated concentration levels, over a full year, without taking into account the timing of the events. The quality objective for the estimation of Pb, As, Ni and Cd is that the uncertainty shall not exceed 100%.

The uncertainty described in the air quality directives has been interpreted as the relative maximum error without timing (RME), eqn. (1), which is the largest concentration difference of all percentile (p) differences normalized by the respective measured value [1, 7, 8].

$$RME(\%) = \frac{\max\left(\left|C_{O,p} - C_{E,p}\right|\right)}{C_{O,p}}.100 \tag{1}$$

3 Results and discussion

At the selected sites, the observed annual concentration values of regulated metals in PM10 during 2008 were well bellow the EC proposed limit/target values; so the objective estimation techniques are an alternative to the experimental determination, which is allowed by the EU directives on ambient air.

All the predictor variables are physically relevant but a good set of predictors must be selected. The predictor variables were subjected to a filtering procedure using the backward elimination approach; the wind variable was rejected because in all the studied cases the performance indexes were not improved. Predictor variables were also tested for collinearity using the variance inflation factor (VIF) statistic indicator; at REIN, the daily concentration of NOx and O_3 were collinear. Since various studies have reported that the use of some transformations of the input variables can result to more powerful regression models [2], the PM10 concentrations were transformed by neperian logarithmics reporting better results.

$$y = a + b_1 \ln(PM.10) + b_2 C_1 + b_3 C_2 + ... + b_7 C_6 + b_8 N_1 + ... + b_{10} N_3 \qquad (2)$$

The multivariate linear regression models used in the present study are based on the eqn (2), where a is the intercept, b_1, b_2, b_3,…,b_{10} are the partial slope coefficients to be determined by the regression model, C_2, C_3,…, C_6 are the predictor continuous variables, except the PM10, and the N_1, N_2 and N_3 are the

Table 3: Performance indexes for the developed estimations.

		Pb	***As***	***Ni***	***Cd***
SANT site					
Annual mean	Observed	6.2	0.8	0.9	0.3
(ng/m^3)	Estimated	6.3	0.9	0.9	0.2
r		0.6	0.8	0.5	0.4
FB		0.0	0.11	0.0	0.0
RMS		5.6	1.1	0.7	0.4
NMSE		0.8	1.8	0.6	2.9
RME		48	33	59	72
CU site					
Annual mean	Observed	8.0	0.2	3.0	0.1
(ng/m^3)	Estimated	8.4	0.2	3.0	0.1
r		0.9	0.6	0.8	0.8
FB		0.0	0.0	0.0	-0.1
RMS		17.4	0.0	2.2	0.1
NMSE		0.1	0.8	0.3	0.7
RME		20	48	36	29
REIN site					
Annual mean	Observed	11.2	0.3	2.0	0.2
(ng/m^3)	Estimated	11.2	0.3	2.0	0.2
r		0.9	0.8	0.8	0.9
FB		0.0	0.0	0.0	-0.2
RMS		4.4	0.2	0.8	0.2
NMSE		0.2	0.3	0.2	0.8
RME		16	35	28	22

predictor nominal variables; at REIN the variable O_3 was rejected. Statistical Package of Social Sciences (SPSS) version 17.0 for Windows software has been used in the present work.

The results for the developed estimations and the performance index values at the three sites are summarised in table 3. Table 3 shows that all the developed estimations fulfil the quality objective fixed by the EU directives for the objective estimations; RME is less than 100%. Furthermore, since RME at REIN and CU sites is always less than 50% the quality objective for modelling is also fulfilled. The performance of estimation models was better at CU and REIN sites than at SANT site; the fact that at REIN and CU all the predictor variables are monitored at the same station could be a possible explanation.

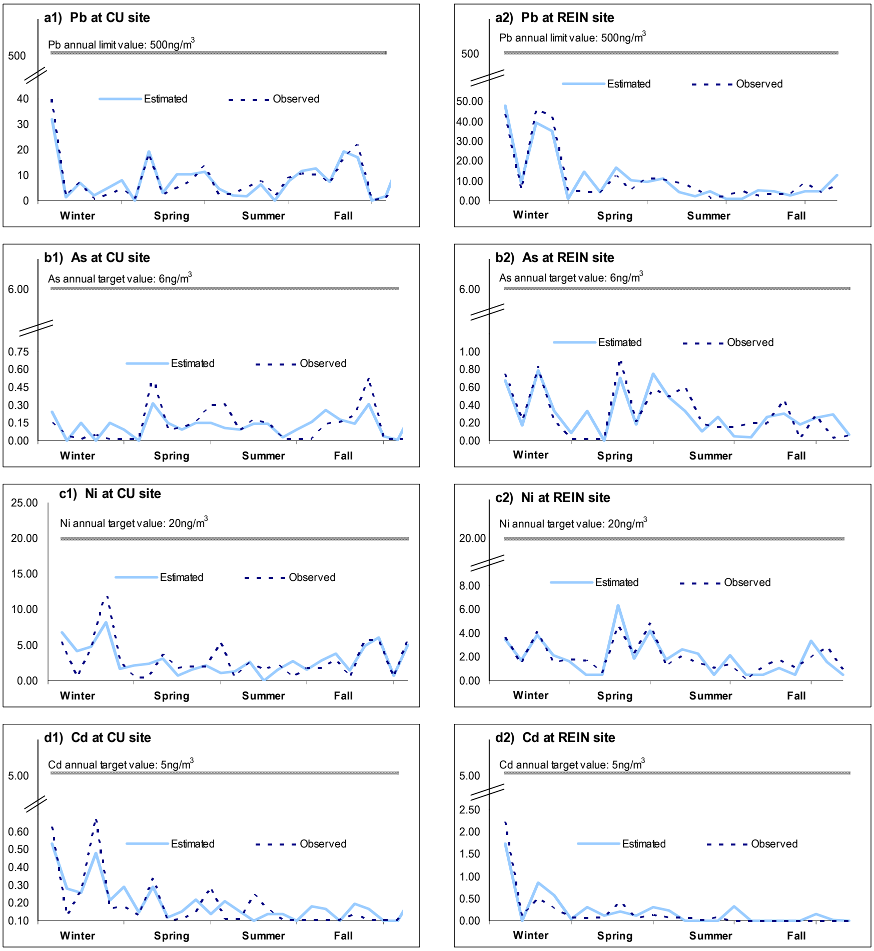

Figure 2: Comparison between estimated and observed daily concentration values (ng/m^3) at CU (1) and REIN (2) sites, a) Pb, b) As, c) Ni and d) Cd. Year 2008.

The fit of the regression could be tested by some statistical parameters such as the mean squared error, the F-Snedecor or the coefficient of correlation [6]; in the present work, this study is carried out by the correlation coefficient. The values for the r coefficient are presented in table 3. At CU and REIN sites the r value varies between 0.8-0.9, showing a proper regression strength. At SANT site, since the r coefficients are lower, an additional validation step is carried out in order to check the estimation models.

At all the urban sites, the annual concentration values for the regulated metals are estimated correctly; the FB indexes are low and the errors between the observed and estimated annual values are not greater than 18%. However, the developed estimations can not completely describe the daily variation of the data sets; RMS and NMSE indexes show this variability. The NMSE values for cadmium are the worst at the three selected sites. Figure 2 shows the estimated and observed daily concentration values of the regulated metals at CU and REIN sites; a quite good agreement between both values is presented.

The same procedure used for the estimations based on MLR was carried out in order to develop estimations based on principal component analysis (PCR). The objective of PCR is to reduce the number of predictive variables and transform them into new variables; these new variables are independent linear combinations of the original data, so the collinearity study is not a necessary step [4]. The PCR was performed considering only the principal components (PC) with an eigenvalue greater than one; in the case of SANT four PC are considered with a total explained variance of 70%. The PCR estimations report slightly worse performance indexes; furthermore, the quality objective for objective estimation fixed by the EU is not fulfilled for some pollutants, such as cadmium at SANT site. On the basis of these results, it can be concluded that MLR is a better alternative than PCR in order to estimate the metals in PM10.

Table 4: Estimated and observed annual concentration values at SANT site. Year 2009.

	Annual concentration values (ng/m^3)		*RME*	*Error between observed and estimated values (%)*
	Observed	*Estimated*		
Pb	6.8	7.1	66	5
As	<d.l[a]	0.9		
Ni	1.4	1.0	54	-24
Cd	0.2	0.3	71	38

[a] Detection limit (d.l) for As: 0.5ng/m^3

At SANT urban site, the validation of the developed MLR estimations was performed using the metal levels in PM10 for the year 2009; this validation is not carried out at the other urban sites because the correlation strength is well tested by the coefficient of correlation. The results obtained from the validation of the MLR estimations are presented in table 4. The observed and the estimated annual values are quite similar, the error between the two values being not greater than 40%. In addition, RME is always less than 100%, so the quality objective for estimation is fulfilled. Figure 3 shows the Pb estimated and observed concentration values during 2009; some higher values are not well estimated. In the case of the arsenic, the performance indexes were not calculated because the observed average value during 2009 is lower than the experimental detection limit; since the MLR estimations are developed for values greater than the detection limit, it is usual that the observed and estimated values are not completely agree.

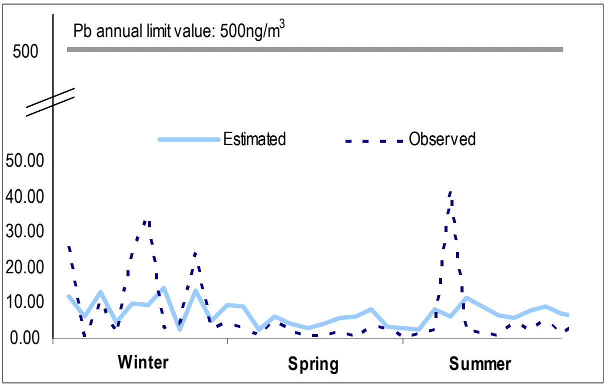

Figure 3: Comparison between estimated and observed Pb daily concentration values (ng/m^3) at SANT. Year 2009.

Therefore, the developed MLR estimations could be used by Regional Governments as a proper alternative to the experimental determination of regulated metals in order to assess the air quality.

4 Conclusions

In the present study, MLR and PCR are used for the estimation of regulated metals levels in PM10 at Santander, Reinosa and Castro Urdiales (Cantabria Region, Northern Spain). The predictor variables are divided into two groups: continuous variables (air pollutant concentration and meteorological parameters) and nominal variables to take into account the influence of the weekend, the seasonal and the Saharian dust intrusions effects. The estimated and observed values are compared during the 2008 year showing that MLR gives a more accurate results than PCR, so MLR estimations are selected.

At the studied urban sites, the annual concentration values are well estimated. The daily concentrations are also correctly estimated at REIN and CU sites. However, at SANT site, the highest observed values of some metals such as cadmium are underestimated; NMSE index shows this variation. A validation, which is carried out only at SANT site using 2009 data, agrees well with these conclusions.

The quality objective fixed by the EU directives for the estimations, RME index lower than 100%, is fulfilled for all the regulated metals at the three studied areas. Hence, the developed estimations could be used by Regional Governments as an alternative to the experimental determination in order to assess the air quality at the studied urban areas; nevertheless, a greater number of training data is necessary in order to improve the estimation models.

Acknowledgements

The authors would like to thank the Spanish Ministry of Education and Science (CTM2006-00317) and the Government of Cantabria ("Actions for improving the air quality and its diagnosis in Cantabria") for their funding. The authors also thank the Regional Environment Ministry of the Cantabria Government for providing the PM10 samples at CU and REIN site.

References

[1] Flemming J., Stern R. Testing model accuracy measures according to the EU directives-examples using the chemical transport model REM-CALGRID. Atmospheric Environment, 41, 9206-9216, 2007.

[2] Grivas G., Chaloulakou A. Artificial neuronal network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment, 40, 1216-1229, 2006.

[3] Viana M., Querol, X., Alastuey, A., Ballester, F., Llop, S., Esplugues, A., Fernández-Patier, R., García dos Santos, S., Herce, M.D. Characterising exposure to PM aerosols for an epidemiological study. Atmospheric Environment, 42, 1552-1568, 2008.

[4] Sousa S.I.V, Martins F.G., Alvim-Ferraz M.C.M, Pereira M.C. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental, Modelling & Software, 27, 97-103, 2007.

[5] Pérez P., Reyes J. An integrated neural network model for PM10 forecasting. Atmospheric Environment 40, 2854-2851, 2006.

[6] Wilks D.S. Statistical methods in the atmospheric sciences. International Geophysics Series. Second edition, 197-212, 2006.

[7] Borrego C., Monteiro A., Ferreira J., Miranda A.I., Costa A.M., Carvalho A.C., Lopes M. Procedures for estimation of modelling uncertainty in air quality assessment. Environment International, 34, 613-620, 2008.

[8] Heinke K. and Sokhi R.S. Overview of tools and methods for meteorological and air pollution mesoscale model evaluation and user training. Joint Report of COST Action 728, 25-26, 2008.