ESTIMATION OF AMBIENT AIR LEVELS OF REGULATED HEAVY METALS BY MEANS OF PARTIAL LEAST SQUARES REGRESSION (PLSR)

Germán Santos and Ignacio Fernández-Olmo

Dept. Ingeniería Química y Química Inorgánica, Universidad de Cantabria, Santander, Spain

Abstract: The European Union through the Air Quality Framework Directive establishes that objective estimation techniques may be used for the assessment of the ambient air quality in all zones and agglomerations where the level of pollutants is below the lower assessment threshold as occurs with the levels of some regulated metals in some urban areas of Cantabria (Northern Spain). Multivariate regression techniques are widely used in the literature to estimate the concentration of air pollutants, e.g. Multiple Linear Regression (MLR) and Principal Component Regression (PCR). Nevertheless, Partial Least Squares Regression (PLSR) combines the advantages of both mentioned techniques. The aim of this work is to estimate the annual levels of the EU regulated metals i.e. arsenic, cadmium, nickel and lead, on airborne PM_{10} in 2008 at three urban sites in the Cantabria Region: Santander (SANT), Castro Urdiales (CAST) and Reinosa (REIN). For this purpose, statistical models based on PLSR have been developed. Furthermore, a comparison was conducted between the estimated metal levels using PLSR and those estimated using MLR and PCR techniques, employed in previous works. The results show that the estimations based on MLR and PLSR fulfill the EU uncertainty requirements for the objective estimations (lower than 100%), unlike PCR-based estimation models. Consequently, statistical estimation models based on MLR and PLSR provide valid approaches to estimate the concentration levels of the EU regulated heavy metals and could be employed to assess the air quality at the considered urban areas as an alternative to experimental measurements.

Key words: *Partial least square regression; PLS; PLSR; particulate matter; PM10; heavy metals.*

INTRODUCTION

The term *Particulate Matter* (PM) is used to describe a mixture of solid particles and liquid droplets suspended in the atmosphere. These particles originate from natural sources, such as volcanic eruptions, seismic activity, forest fires, winds of great intensity or natural particle transport from dry regions; and from anthropogenic sources, including all types of combustion (e.g. power plants, diesel engines, etc.) and some industrial processes (Pires, J.C.M. et al., 2008). The relationship between the presence of aerosol particles in the atmosphere and adverse health effects has been well recognized and reported in the literature (Pope III, C.A. et al., 2009 and Li, X. et al., 2011). Nevertheless, PM still remains one of the most important air pollutants responsible for causing damages to human health in Europe (Koelemeijer, R.B.A. et al., 2006). It has been demonstrated that the pernicious effects of PM are not only due to its physical attributes such as mass concentration and size distribution, but also to its chemical composition, which has been studied extensively during the last decades and reviewed in the literature (Querol, X. et al., 2004). It could include some acidic and toxic species such as heavy metals and aromatic compounds (Karar, K, and A.K. Gupta, 2006).

In this regard, the European Union (EU) has included limits for PM_{10} and $PM_{2.5}$ in the air quality directives; 1999/30/EC and 2008/50/EC. Furthermore, it has established a set of air quality targets for some trace metals in PM_{10}, namely: As, Cd, Ni (Directive 2004/107/EC) and Pb (Directive 2008/50/EC). Additionally, the EU regulation states that objective estimation techniques shall be sufficient for the assessment of the ambient air quality in all zones and agglomerations where the level of pollutants is below the lower assessment threshold. This situation occurs with the levels of certain metals in some of the studied urban areas of Cantabria (Arruti, A. et al., 2011). As shown in Figure 1, the annual mean concentration in 2008 of the four regulated metals was well below the lower assessment threshold in every study site. For this reason, it is concluded that objective estimation techniques are a proper alternative to experimental measurements to assess the air quality in these locations.

According to the Guidance on Assessment under the EU Air Quality Directives, “objective estimation techniques” is a fairly broad term which includes mathematical methods to calculate concentrations from values measured at other locations and/or times, based on scientific knowledge of the concentration distribution. Empirical data-based modelling falls within this definition and represents an attractive

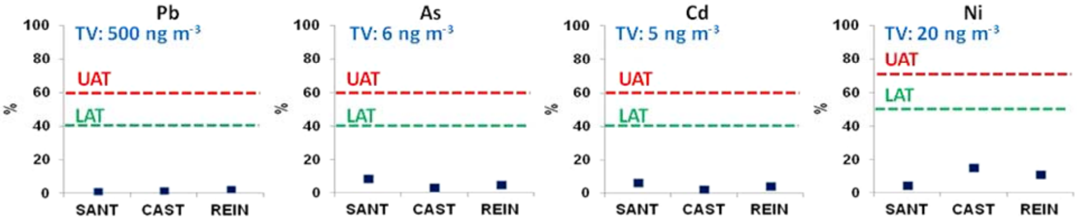

Figure 1. 2008 annual mean of regulated metals expressed as percent of their respective target values. TV: Target value; UAT: Upper assessment threshold; LAT: Lower assessment threshold; SANT: Santander; CAST: Castro Urdiales; REIN: Reinosa.

alternative to mechanistic modelling given that it requires less specific knowledge of the system under consideration. Empirical modelling techniques require data (measurements) of those variables believed to be representative of the process behaviour and of the properties of the system output (Kim, M. et al., 2009). Some studies have been published in the literature with the aim of estimating certain atmospheric pollutants by using techniques based on empirical models such as multiple linear regression (MLR), feedforward neural networks and principal component regression (PCR). Moreover, in a previous study carried out by Arruti, A. et al. (2011), estimations of regulated metal levels in ambient air by statistical MLR and PCR models have been conducted. Nonetheless, there are other promising techniques such as partial least squares regression (PLSR) that might lead to an improvement of the estimations, since it combines features from principal component analysis (PCA) and multiple linear regression (MLR): as MLR, PLSR creates a linear combination of the predictor variables (X-matrix, which is constituted by environmental observations in this study) that best correlates with the response variables (Y-matrix, composed of metal levels observations); and as PCA, it decomposes the X-matrix in order to obtain components that best explain X. Specifically, PLSR searches for a set of components (called *latent variables*) that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the covariance between X and Y. It is followed by a regression step where the latent variables obtained from X are used to predict Y (Abdi, H., 2010). This technique has been already used for the characterisation and the determination of profiles of polycyclic aromatic hydrocarbons in PM_{10} (Wingfors, H. et al., 2001) providing positive results.

The aim of this work is to estimate the annual levels of the EU regulated metals i.e. arsenic, cadmium, nickel and lead, on airborne PM_{10} for 2008 at three sites in the Cantabria Region (Northern Spain): Santander, Castro Urdiales and Reinosa.

METHODOLOGY

Area of study

This work was conducted at three different urban sites (Figure 2): Santander (SANT), capital of Cantabria extended over a bay, and with a broad industry presence; Castro Urdiales (CAST), a coastal urban site in the vicinity of a national highway and an industrial area; and Reinosa (REIN), an inland urban site close to a steel manufacturing plant. These sites are described in detail by Arruti, A. et al. (2011).

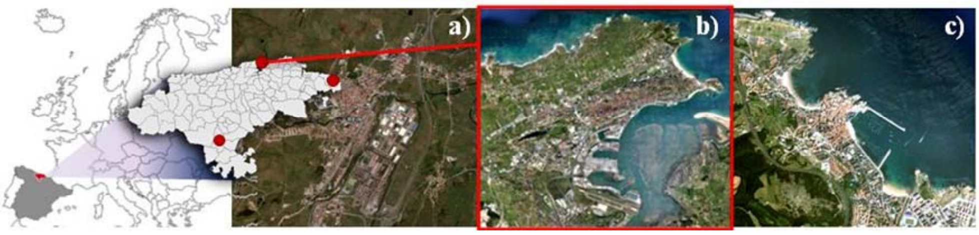

Figure 2. Areas of study and sampling sites in Cantabria: (a) Reinosa, (b) Santander and (c) Castro Urdiales.

Partial Least Squares Regression (PLSR) Model

Statistical models based on Partial Least Square Regression (PLSR) have been developed to estimate the ambient air concentrations of the EU regulated metals. For this purpose, the data set used in this study is divided into response variables and predictor variables. The former consist of regulated metal levels (ng m^{-3})

on airborne PM_{10} for 2008 at the three considered sites obtained from previous works (Arruti, A. et al., 2011). In turn, predictor variables are constituted by: qualitative or nominal variables (Table 1), taking into account the seasonal, the Saharan dust intrusion and the weekend effects; and quantitative or continuous variables, namely, meteorological data and major atmospheric pollutants concentration, which are detailed in Table 2. The continuous variables are measured automatically on real time at the monitoring stations of the Cantabria Regional Air Quality Monitoring Network and are available at the Regional Environment Ministry website.

Prior to analysis, predictor variables were auto-scaled. In addition, Cross-validation was used to estimate the number of significant components. PLS Toolbox (Eigenvector Research, Inc.) for MATLAB was used in the present study to develop the PLSR models.

Table 1. Nominal predictor variables

Notation	Description	Codification
SE	Season	1: Winter; 2: Spring; 3: Summer; 4: Fall
SD	Saharan dust intrusion	0: No intrusion; 1: Intrusion
WE	Weekend	0: No weekend; 1: Weekend

Table 2. Continuous predictor variables

Notation	Type	Description[a]	Units
PM_{10}	Major air pollutant	Average natural logarithm of PM_{10} concentration (µg m^{-3})	-
SO_2	Major air pollutant	Average concentration of sulphur dioxide	µg m^{-3}
O_3	Major air pollutant	Average concentration of ozone	µg m^{-3}
NO_x	Major air pollutant	Average concentration of nitrogen oxides	µg m^{-3}
T	Meteorological	Average temperature	ºC
RH	Meteorological	Average relative humidity	%
WD	Meteorological	Prevailing wind direction	º
PP	Meteorological	Cumulative precipitation	L m^{-2}

[a] According to the corresponding PM_{10} sampling periods, daily values of continuous variables were calculated at SANT site and 48-hour values were calculated at REIN and CAST sites

Evaluation of model performance

The statistical parameters used in the present work to evaluate the model performance are: the correlation coefficient (r), the fraction bias (FB), the root mean square error (RMSE), the normalised mean square error (NMSE) and the fractional variance (FV). Additionally, for the validation of the objective estimation and modelling techniques in the context of the Air Quality Directives, two indexes of uncertainty were used: the relative maximum error without timing (RME) and the relative directive error (RDE). The RME is the largest concentration difference of all percentile (*p*) differences normalized by the respective measures value. The RDE is the difference between the closest observed concentration to the limit/target value and the correspondingly ranked modelled concentration normalized by the limit/target value.

RESULTS AND DISCUSSION

Tables 3-5 show the performance indexes of different models developed for SANT, CAST and REIN sites, respectively. Due to lack of enough data in Reinosa and Castro Urdiales, PCR estimations are only available at SANT site.

The annual mean concentrations are estimated correctly, which results in low values of FB index. However, PLSR improves the FB index when is possible. PLSR provides greater or equal correlation coefficients than MLR, but higher than those provided by PCR. The r values are within the range of 0.4-0.8 at the SANT site, and 0.6-0.9 at the CAST and REIN sites. Even though the three techniques provide acceptable estimations, their performance is heavily dependent on the place being studied. In this respect, some difficulties were found in estimating the daily concentrations, which is reflected on NMSE and FV indexes. To illustrate this, daily Pb concentrations at SANT and CAST site are presented in Figure 3(a) and Figure 3(b), respectively. The highest observed Pb concentrations at SANT site are underestimated, unlike than those at CAST site. As a result, greater values of NMSE and FV indexes are obtained for Pb at SANT site with respect to CAST site.

Table 3. Performance indexes for the estimations at SANT site

Metal	Technique	Annual mean (ng m^{-3})		Performance indexes					EU uncertainty	
		Observed	Estimated	r	FB	RMSE	FV	NMSE	RME (%)	RDE (%)
Pb	MLR	6.4	6.3	0.6	0.0	5.6	1.07	0.8	48	3.2
	PCR	6.4	6.4	0.5	0.0	6.0	1.27	0.9	59	2.1
	PLSR	6.4	6.4	0.6	0.0	5.7	1.06	0.8	60	3.8
As	MLR	0.8	0.9	0.8	0.11	1.1	0.45	1.8	33	69
	PCR	0.8	0.9	0.6	-0.2	1.7	1.24	3.6	67	42
	PLSR	0.8	0.8	0.8	0.0	1.2	0.38	2.1	32	79
Ni	MLR	0.9	0.9	0.5	0.0	0.7	1.24	0.6	59	12
	PCR	0.9	0.9	0.4	0.0	0.7	1.50	0.7	55	12
	PLSR	0.9	0.9	0.5	0.0	0.7	1.16	0.6	68	14
Cd	MLR	0.3	0.2	0.4	0.0	0.4	1.32	2.9	72	42
	PCR	0.3	0.3	0.4	-1.8	4.0	1.82	181	259	151
	PLSR	0.3	0.3	0.5	0.0	0.4	1.32	2.8	70	41

Table 4. Performance indexes for the estimations at CAST site

Metal	Technique	Annual mean (ng m^{-3})		Performance indexes					EU uncertainty	
		Observed	Estimated	r	FB	RMSE	FV	NMSE	RME (%)	RDE (%)
Pb	MLR	8.0	8.4	0.9	0.0	17.4	0.26	0.1	20	1.5
	PLSR	8.0	8.0	0.9	0.0	3.2	0.15	0.2	19	1.5
As	MLR	0.2	0.2	0.6	0.0	0.0	0.85	0.8	48	4.0
	PLSR	0.2	0.2	0.7	0.0	0.1	0.77	0.8	49	3.9
Ni	MLR	3.0	3.0	0.8	0.0	2.2	0.48	0.3	36	22
	PLSR	3.0	3.0	0.8	0.0	1.7	0.48	0.3	34	22
Cd	MLR	0.1	0.1	0.8	-0.1	0.1	0.63	0.7	29	3.2
	PLSR	0.1	0.1	0.9	0.0	0.1	0.64	0.7	34	3.9

Table 5. Performance indexes for the estimations at REIN site

Metal	Technique	Annual mean (ng m^{-3})		Performance indexes					EU uncertainty	
		Observed	Estimated	r	FB	RMSE	FV	NMSE	RME (%)	RDE (%)
Pb	MLR	11.2	11.2	0.9	0.0	4.4	0.11	0.2	16	0.3
	PLSR	11.2	11.2	0.9	0.0	4.2	010	0.1	18	1.0
As	MLR	0.3	0.3	0.8	0.0	0.2	0.42	0.3	35	1.8
	PLSR	0.3	0.3	0.8	0.0	0.2	0.36	0.2	55	3.8
Ni	MLR	2.0	2.0	0.8	0.0	0.8	0.42	0.2	28	9.0
	PLSR	2.0	2.0	0.9	0.0	0.7	0.28	0.1	95	6.9
Cd	MLR	0.2	0.2	0.9	-0.2	0.2	0.37	0.8	22	10.0
	PLSR	0.2	0.2	0.9	0.0	0.2	0.22	1.2	22	10.0

Finally, as shown in Figure 3, PLSR and MLR estimations overlap throughout the period of study. This trend is observed in the estimations of each metal at all sites. In turn, PCR provides worse estimations, what is reflected in lower r values and greater RMSE, NMSE and FV values.

EU Uncertainty

Regarding the EU uncertainty requirements, the RME and RDE indexes for PLSR estimations as well as MLR estimations are below 100%. As a consequence, it is concluded that PLSR and MLR statistical models fulfill the EU uncertainty requirements for objective estimations. This cannot be extended to PCR models, as seen in the cadmium RME and RDE at SANT site (Table 3).

CONCLUSIONS

Partial least squares regression (PLSR) statistical models have been developed to estimate the ambient air

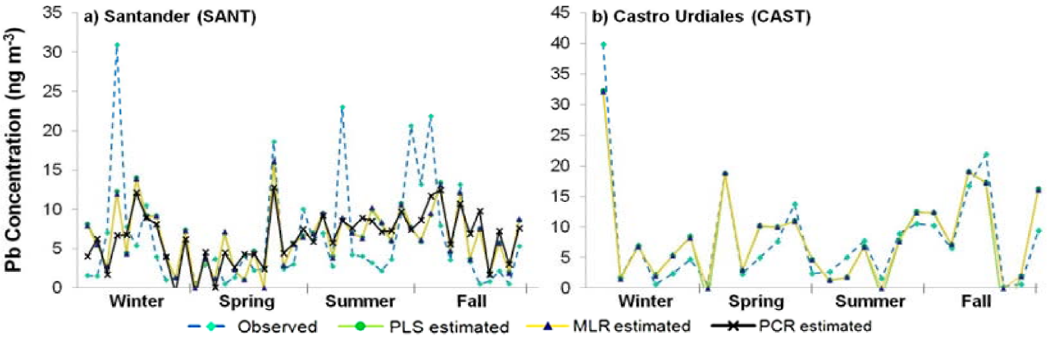

Figure 3. Comparison between observed and estimated levels of Pb at: a) SANT site; b) CAST site.

levels of the EU regulated metals i.e. arsenic, cadmium, nickel and lead, from the airborne PM_{10} in 2008 at three sites in Cantabria: Santander, Castro Urdiales and Reinosa. Furthermore, a comparison was conducted between the estimated metal levels using PLSR and those estimated using MLR and PCR techniques, employed in previous works. Results show that PLSR and MLR provide valid approaches to estimate the concentration levels of the regulated metals fulfilling the uncertainty requirements for objective estimations (RME and RDE lower than 100%). Therefore, statistical estimation models based on MLR and PLSR could be employed to assess the levels of metals in air at the considered urban areas as an alternative to experimental measurements, which would lead to save time, effort and resources. Further work will imply the application of more powerful estimation tools (e.g. neural networks) and the development of estimations of non-regulated metals with higher concentration levels on ambient air (e.g. Mn or Zn in the studied areas), which will demand more strict uncertainty requirements.

AKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the Spanish Ministry of Economy and Competitiveness through the Project CMT2010-16068 and also the granting of a FPI research fellowship to Germán Santos.

REFERENCES

Abdi, H, 2010: Partial least squares regression and projection on latent structure regression (PLSR Regression). *Wiley Interdiscip. Rev. Comput. Stat.*, **2(1)**, 97-106.

Arruti, A, I. Fernández-Olmo and A. Irabien, 2011: Assessment of regional metal levels in ambient air by statistical regression models. *J. Environ. Monit.*, **13(7)**, 1991-2000.

EC Working group for ambient air quality directives, 2000: Guidance on assessment under the EU air quality directives - final draft. (http://europa.eu.int/comm/environment/air/pdf/guidanceunderairquality.pdf)

Karar, K, and A.K. Gupta, 2006: Seasonal variations and chemicals characterization of ambient PM_{10} at residential and industrial sites of an urban region of Kolkata (Calcutta). India. *Atmos. Res.*, **81**, 36-53.

Kim, M, Y. Kim, S. Sung and C. Yoo, 2009: Data-Driven Prediction Model of Indoor Air Quality by the Preprocessed Recurrent Neural Networks. *Korean J. Chem. Eng.*, **27(6)**, 1675-1680.

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.

Li, X, T. Hede, Y. Tu, C. Leck and H. Ågren, 2011: Glycine in aerosol water droplets: a critical assessment of Köhler theory by predicting surface tension from molecular dynamics simulations. *Atmos. Chem. Phys.*, **11**, 519-527.

Pires, J.C.M, F.G. Martins, S.I.V. Sousa, M.C.M. Alvim-Ferraz and M.C. Pereira, 2008: Prediction of the Daily Mean PM_{10} Concentrations Using Linear Models. *Am. J. Environ. Sci.*, **4(5)**, 445-453.

Pope III, C.A, M. Ezzati and D.W. Dockery, 2009: Fine-particulate air pollution and life expectancy in the United States. *N. Eng. J. Med.*, **360**, 376-386.

Querol, X, A. Alastuey, M.M. Viana, S. Rodríguez, B. Artiñano, P. Salvador, S. Garcia do Santos, R. Fernández Patier, C.R. Ruiz, J. de la Rosa, A. Sánchez de la Campa, M. Menéndez and J.I. Gil, 2004: Speciation and origin of PM_{10} and $PM_{2.5}$ in Spain. *J. Aerosol Sci.*, **35**, 1151-1172.

Wingfors, H, Å. Sjödin, P. Haglund and E. Brorström-Lundén, 2001: Characterisation and determination of profiles of polycyclic aromatic hydrocarbons in a traffic tunnel in Gothenburg, Sweden. *Atmos. Environ.*, **35**, 6361-6369.