Available at www.scientevents.com/proscience/

ScientEvents

ProScience 1 (2014) 100-105

ProScience

Conference Proceedings

1st International Conference on Atmospheric Dust - DUST2014

PLSR and ANN estimation models for PM_{10}-bound heavy metals in Dunkerque (Northern France)

Germán Santos[1*], Ignacio Fernández-Olmo[1], Ángel Irabien[1], Frédéric Ledoux[2], Dominique Courcot[2]

[1]*Departamento de Ingenierías Química y Biomolecular, Universidad de Cantabria, Santander, 39005, Spain*

[2]*Unité de Chimie Environnementale et Interactions sur le Vivant (UCEIV) EA 4492, Université du Littoral Côte d'Opale, Dunkerque, 59140, France*

Abstract

The aim of this work is to develop statistical estimation models of some EU regulated heavy metal levels (Pb, Ni) and some non-regulated heavy metal levels (Mn, V and Cr) in the ambient air of the city of Dunkerque (Northern France) so that they might be used for air quality assessment as an alternative to experimental measurements, since these levels are relatively low compared to the EU limit/target values and other air quality guidelines. Three different approaches were considered: Partial Least Squares Regression (PLSR), Artificial Neural Networks (ANN) and Principal Component Analysis (PCA) coupled with ANN. External validation results evidence that PLSR and ANN-based statistical models for regulated metals and for Mn and V provide adequate mean values estimations while fulfill the EU uncertainty requirements.

Keywords: Partial least squares regression; PLSR, artificial neural networks; ANN, statistical models; particulate matter; PM10, heavy metals.

1. Introduction

Particulate Matter (PM) still remains a concerning environmental problem in urban areas, which is due not only to its physical properties such as mass distribution, particle size and shape, but also to its chemical composition. With respect to the metal content in PM, the European Union (EU) throughout the Air Quality Framework Directive (EC, 2008) and the 4th Daughter Directive (EC, 2004) has established limit/target values for some metals, as shown on Table 1. The air quality assessment criteria depend on the pollutant

*Corresponding Author: santosg@unican.es
ISSN: 2283-5954 © 2014 The Authors. Published by Digilabs
Selection and peer-review under responsibility of DUST2014 Scientific Committee
DOI:10.14644/dust.2014.016

concentration. Thus, when the levels of pollutants are bellow the Lower Assessment Threshold (LAT), modelling and objective estimation techniques are allowed to be used as air quality assessment methods (EC, 2008). Since the analytical determination of the levels of these metals is rather expensive and time consuming, it might be interesting to try to find new alternatives for air quality assessment in relation to heavy metals so that less experimental measurements are required. In this respect, the main objective of this work is to estimate the levels of some EU regulated and non-regulated metals in airborne PM_{10} in an urban area. For this purpose, statistical models based on Partial Least Squares Regression (PLSR) and Artificial Neural Networks (ANN) have been developed as objective estimation techniques. It should be mentioned that since this work is conceived as an air quality assessment tool at a later stage it is not about forecasting but estimation.

Table 1. EU quality objectives and evaluation thresholds for regulated metals.

Pollutant	TV/LV[a] (ng m^{-3})	UAT[b]	LAT[b]	Directive
Pb	500	60	40	2008/50/EC
As	6	60	40	
Cd	5	60	40	2004/107/EC
Ni	20	70	50	

TV: Target Value; LV: Limit Value; UAT: Upper Assessment Threshold; LAT: Lower Assessment Threshold
[a] For the total content in the PM_{10} fraction averaged over a calendar year
[b] Percent of the target value

2. Methodology

2.1 Area of study and input database

The database consists of dependent (response) variables, metal concentration in PM_{10} at Les Darses site in Dunkerque (Northern France) corresponding to an intensive sampling campaign performed from February to May 2008 (Hleis, 2010), and independent (predictor) variables (listed on Table 2). Meteorological data were obtained at the Meteorological station.

Table 2. Input variables.

Notation	Description[a]	Type	Units
SE	Season (1: Winter; 2: Spring; 3: Summer; 4: Fall)	Nominal	-
WE	Weekend (0: No weekend; 1: Weekend)	Nominal	-
$LnPM_{10}$	Average natural logarithm of PM_{10} concentration (µg m^{-3})	Continuous (Major air pollutant)	-
SO_2	Average concentration of sulphur dioxide	Continuous (Major air pollutant)	µg m^{-3}
O_3	Average concentration of ozone	Continuous (Major air pollutant)	µg m^{-3}
NO_x	Average concentration of nitrogen oxides	Continuous (Major air pollutant)	µg m^{-3}
T	Average temperature	Continuous (Meteorological)	ºC
RH	Average relative humidity	Continuous (Meteorological)	%
WD	Prevailing wind direction	Continuous (Meteorological)	º
WS	Prevailing wind speed	Continuous (Meteorological)	ms^{-1}
P	Average pressure	Continuous (Meteorological)	mbar
PP	Cumulative precipitation	Continuous (Meteorological)	L m^{-2}

[a] Average values were calculated according to the corresponding duration of the PM_{10} sampling periods

in Dunkerque Harbor, and major pollutant data were measured at St. Pol sur Mer monitoring station.

A total number of 78 samples were analyzed for different trace metals. Among the EU regulated metals, Pb and Ni were determined. Additionally, three non-regulated metals were also considered: Mn, V and Cr. These metals were tracers of some industrial activities found in Dunkerque and Santander Bay (Spain), where previous studies on trace metal levels have been developed: Mn, for ferromanganese alloys manufacturing; V, for marine traffic and liquid fuel combustion; and Cr, for non-integral steel manufacturing and coal combustion. Since these metals are not regulated by the EU, they do not have a limit/target or LAT value. So, in order to normalize the metal concentration and to calculate the EU uncertainty indexes, the following values were considered as LAT for non-regulated metals: the annual air quality guideline for Mn (150 ng/m^3) proposed by the World Health Organization (WHO); and the maximum observed concentration for V and Cr, in absence of a standard value for a period of duration comparable to that of the period of study.

As shown in Table 3 Pb and Ni mean values are below their respective LAT. Therefore, according to the EU Air Quality Directives, objective estimation techniques can be applied for the air quality assessment in relation to Pb and Ni.

Table 3. Levels of Pb, Ni, Mn, V and Cr for the period of study. Table adapted from Hleis (2010).

Pollutant	**LAT[a] (ng m^{-3})**	**Mean value (ng m^{-3})**	**Max. value (ng m^{-3})**	**Min. value (ng m^{-3})**	**Stand. deviation (ng m^{-3})**
Pb	200	14.3	79.1	0.35	15.4
Ni	10	9.2	104.0	0.17	14.7
Mn	150	74.4	872.8	0.03	141.0
V	49.0	7.8	49.0	0.17	7.9
Cr	190.7	9.6	190.7	0.17	28.5

[a] LAT-equivalent value for Mn, V and Cr

Having been removed five identified outliers, the resulting complete database was divided into three subsets: 60% for models development, 20% for verification to avoid overfitting and 20% for external validation. In order to avoid scale effects, the dependent variables were normalized dividing the metal concentration by their respective LAT.

2.2 Partial Least Squares Regression and Artificial Neural Network statistical models

Partial Least Squares Regression and Artificial Neural Networks have been proposed in this study to estimate PM_{10}-bound heavy metals due to the fact that both of them have been used in the literature as mathematical techniques to forecast the air concentration of a number of pollutants. Pires et al. (2008), Polat & Durduran (2012), Singh et al. (2012) applied PLSR to predict PM concentrations; and numerous authors over the years have investigated on developing ANN models to predict PM concentrations as well as gaseous pollutants (Gardner & Dorling 1999; Kukkonen et al., 2003; Perez & Reyes 2002), to cite but a few. Chelani et al. (2002) even included prediction of ambient air metal levels. Since the number of independent input variables is relatively high with respect to the number of samples, an alternative approach based on applying Principal Component Analysis (PCA) prior to ANN was considered as reported in the literature to be an effective strategy to improve the models.

2.3 Model performance criteria

In this study the evaluation criteria to determine whether a model is suitable for air quality assessment purposes is principally based on: (i) the fulfilment of the European Union uncertainty requirements for objective estimation techniques, and (ii) the accuracy of estimated mean values, since the metal limit/target values correspond to annual mean concentrations. Two indexes of uncertainty were calculated: the relative maximum error without timing (RME) and the relative directive error (RDE). The former is the largest concentration difference of all percentile (*p*) differences normalized by the respective measured value. The latter 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. Additionally, a number of statistical parameters were considered to evaluate the model performance and are shown in Table 4:

Table 4. Statistical parameters to evaluate the model performance.

Statistic	**Equation**[a]
Relative maximum error without timing	$RME = \max(\lvert C_{O,p}-C_{E,p}\rvert)/C_{O,p}$
Relative directive error	$RDE = \lvert C_{O,LV}-C_{E,LV}\rvert/LV$
Fractional bias	$FB = \dfrac{\overline{C_O}-\overline{C_E}}{0.5\,(\overline{C_O}+\overline{C_E})}$
Correlation coefficient	$r = \left[\dfrac{\sum_{i=1}^{n}(C_{O,i}-\overline{C_O})(C_{E,i}-\overline{C_E})}{\sqrt{\sigma_O\sigma_E}}\right]$
Root mean square error	$RMSE = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(C_{O,i}-C_{E,i})^2}$
Fractional variance	$FV = 2\,\dfrac{\sigma_O-\sigma_E}{\sigma_O+\sigma_E}$

[a] O: Observed; E: Estimated

3. Results and discussion

3.1. Regulated metals: Pb and Ni

Table 5 presents the results obtained for the best developed models for the two considered EU regulated metals (Pb and Ni) using the three different contemplated approaches: PLSR, ANN and PCA-ANN. Results relating to training (T) and external validation (V) subsets are displayed for each model. As shown, both RME and RDE indexes are well below the uncertainty requirements for objective estimation techniques, which are 100% for both of them. PLSR shows more accuracy for the training stage, however the best pair of training and external validation FB is found when using ANN, which indicates a better generalization ability for this technique. Table 5 also shows that using PCA before ANN increases uncertainty and produces a decrease in r.

With regard to the performance indexes, the models show difficulties in estimating accurately the individual sample concentrations leading to an underestimation of the highest concentrations. However, PLSR and ANN capture the underlying trend during training although there is a better fitting when using ANN with lower values of RMSE, and

Table 5. Uncertainty, mean concentration and performance statistics for the best models developed for Pb and Ni.

Metal	Model	Subset[a]	EU Uncertainty RME (%)	RDE (%)	Mean Concentration[b] C_O 10^2	C_E 10^2	FB 10^2	r	Performance RMSE 10^2	FV 10
Pb	PLSR	T	28.1	1.44	6.52	6.52	3.7 10^{-08}	0.823	3.94	1.95
		V	31.9	0.31	7.46	8.88	-17.4	0.837	4.48	-2.78
	ANN	T	18.3	2.10	6.38	6.84	-7.0	0.932	2.72	-0.85
		V	54.0	0.54	7.46	8.31	-10.8	0.861	4.90	-4.12
	PCA-ANN	T	40.1	2.22	6.57	6.78	-3.2	0.663	3.90	5.82
		V	90.4	1.38	3.69	7.95	-73.1	0.266	5.64	-2.56
Ni	PLSR	T	65.9	12.87	68.5	68.5	-7.510^{-11}	0.560	80.5	5.64
		V	83.6	11.70	156.6	98.2	45.8	0.556	241.3	14.62
	ANN	T	29.2	18.55	73.4	73.8	-0.5	0.873	54.2	1.44
		V	50.0	17.60	156.6	115.8	30.0	0.702	186.6	4.77
	PCA-ANN	T	64.9	24.86	95.9	95.8	0.1	0.470	161.1	7.20
		V	42.6	2.50	68.6	94.9	-32.2	0.443	63.2	-3.04

[a] T: Training; V: Validation
[b] O: Observed; E: Estimated

FV. With respect to the external validation, there is a RMSE and FV increase as a result of an accuracy decrease. Still, they can be considered satisfactory estimations.

3.2. Non-regulated metals: Mn, V and Cr

The results of the best developed models for these three metals are shown on Table 6. As can be observed, the uncertainty requirements are fulfilled with a RME and RED lower than 100% in all cases, except for the RME index of the Cr PCA-ANN external validation model. With respect to the mean concentration, low values of FB for Mn and V indicate acceptable training and external validation estimations. Nevertheless, correlation coefficient values lower than 0.66 for external validation evidence an unsatisfactory fitting of the individual sample concentrations. ANN models correlation coefficients are greater than those obtained

Table 6. Uncertainty, mean value and performance statistics for the best models developed for Mn, V and Cr.

Metal	Model	Subset[a]	EU Uncertainty RME (%)	RDE_{eq} (%)	Mean Concentration[b] C_O 10^2	C_E 10^2	FB 10^2	r	Performance RMSE 10^2	FV 10
Mn	PLSR	T	53.9	4.71	32.55	33.43	-2.7	0.580	41.49	3.57
		V	53.1	50.23	64.95	58.24	10.9	0.184	92.80	6.99
	ANN	T	52.6	60.51	33.61	21.38	44.5	0.704	39.20	3.42
		V	48.2	68.86	64.95	63.77	1.8	0.457	81.56	4.64
	PCA-ANN	T	66.4	78.51	46.05	46.05	0.0	0.463	64.77	7.34
		V	35.3	11.62	29.49	30.33	-2.8	0.431	42.72	1.74
V	PLSR	T	42.7	1.20	13.12	13.42	-2.3	0.694	8.17	2.33
		V	31.5	4.74	18.07	18.48	-2.2	0.590	11.21	3.16
	ANN	T	41.6	4.28	14.27	14.00	1.9	0.806	7.19	1.78
		V	30.7	5.45	18.07	18.45	-2.1	0.663	10.43	2.11
	PCA-ANN	T	42.9	1.60	13.14	13.20	-0.5	0.747	5.91	2.93
		V	12.5	15.60	12.79	16.50	-25.4	0.366	13.97	0.56
Cr	PLSR	T	88.8	1.01	4.58	3.96	14.6	-0.031	15.28	11.34
		V	78.5	25.74	6.97	5.78	18.7	0.077	18.09	12.78
	ANN	T	50.0	39.17	5.59	3.25	53.1	-0.040	19.24	6.53
		V	83.6	27.89	6.97	3.55	65.0	-0.240	19.06	14.76
	PCA-ANN	T	79.2	0.43	7.15	7.90	-10.0	0.331	18.33	9.70
		V	489.9	0.43	1.45	9.31	-146.3	0.275	10.04	-13.57

[a] T: Training; V: Validation
[b] O: Observed; E: Estimated

for PLSR and PCA-ANN models. This, together with the fact that ANN provide the lowest FB and adequate RME and RDE, places ANN as the most suitable approach out of the three studied.

4. Conclusions

Models based on PLSR and ANN to estimate the levels of some EU regulated metals (Pb and Ni) and some non-regulated metals (Mn, V, Cr) were developed for an intensive PM_{10} sampling campaign carried out at Les Darses site in Dunkerque (Northern France) from February to May 2008. An alternative approach consisting in performing PCA prior to ANN was also considered.

PLSR and ANN techniques provide acceptable mean concentration estimations. Moreover, uncertainty requirements for objective estimations (RME and RDE lower than 100%) are fulfilled when estimating regulated metals. Consequently, according to EU Air Quality Directives, PLSR and ANN represent valid approaches as tools for air quality assessment in relation to regulated heavy metals in the studied area. The results obtained for the studied non-regulated metals are slightly worse, mainly for Cr.

With respect to the applied techniques, ANN showed a slight better performance than PLSR. However, PCA did not improve the ANN model performance.

5. Aknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Project CTM2010-16068 and the FPI short stay EEBB-I-13-07691.

References

Chelani A.B., Gajghate D.G., Hasan M.Z. (2002). Prediction of ambient PM_{10} and toxic metals using artificial neural networks. Journal of the Air and Waste Management Association 52, 805-810.

EC (European Commission). (2004). Council Directive 2004/107/EC of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air. Official Journal of the European Communities L23, 3-16.

EC (European Commission). (2008). Council Directive 2008/50/EC, of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Communities L 152, 1-44.

Gardner M.W., Dorling S.R. (1999). Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 33, 709-719.

Hleis D. (2010). Evaluation de la contribution d'émissions sidérurgiques à la teneur en particules en suspension dans l'atmosphère à une échelle locale. PhD thesis, Université du Littoral Côte d'Opale.

Kukkonen J., Partanen L., Karppinen A., Ruuskanen J., Junninen H., Kolehmainen M., Niska H., Dorling S., Chatterton T., Foxall R., Cawley G. (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki, Atmospheric Environment 37, 4539-4550.

Perez P., Reyes J. (2002). Prediction of maximum of 24-h average of PM_{10} concentrations 30h in advance in Santiago, Chile. Atmospheric Environment 36, 4555-4561.

Pires J.C.M., Martins F.G., Sousa S.I.V., Alvim-Ferraz M.C.M., Pereira M.C. (2008). Prediction of the daily mean PM_{10} concentrations using linear models. American Journal of Environmental Sciences 4, 445-453.

Polat K., Durduran S.S. (2012). Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values (PM_{10}) in the city of Konya, Neural Computing and Applications 21, 2153-2162.

Singh K.P., Gupta S., Kumar A., Shukla S.P. (2012). Linear and nonlinear modeling approaches for urban air quality prediction, Science of the Total Environment 426, 244-255.