Estimation of PM_{10}-bound As, Cd, Ni and Pb levels by means of statistical modelling: PLSR and ANN approaches

Authors: Germán Santos*, Ignacio Fernández-Olmo, Ángel Irabien

Universidad de Cantabria, Dep. Ingenierías Química y Biomolecular, Avda. Los Castros s/n, 39005 Santander (Cantabria), Spain

*Corresponding author:

Tel.:+34942201579; fax: +34942201591

E-mail address: santosg@unican.es

1. Introduction

Mathematical modelling for air quality assessment purposes has become increasingly important in recent years. These models consist of a set of analytical/numerical algorithms that describe the physical and chemical aspects of a problem and can be divided into two main groups: (i) deterministic models based on fundamental mathematical descriptions of atmospheric processes, where emissions (causes) generate air pollution (effects); (ii) statistical (empirical) models based on semiempirical statistical relations between available data of input variables that are believed to be representative of the process behaviour and measurements of the target parameters/properties of the system output. Moreover, the European Union Air Quality Framework Directive establishes that in all zones and agglomerations where the level of pollutants is below the lower assessment threshold (LAT), which is expressed as a percentage of the corresponding target/limit value, modelling techniques or objective estimation techniques (or both) shall be sufficient for the assessment of the ambient air quality (European Council Directive 2008/50/EC). Both statistical and deterministic methods are currently used in regulatory air pollution forecasting by environmental authorities.

Although deterministic models have some advantages over statistical models, such as a full-coverage 3D domain, in some particular situations they may have some drawbacks in terms of accuracy and input data uncertainty. According to Hanna (1989), generally, a larger number of input parameters corresponds to a lower model uncertainty and smaller prediction errors, but unfortunately, by extending the number of input parameters, the error and uncertainty attached to the input data also increase. Therefore, complex deterministic models work well when their extensive input data requirements are satisfied, which rarely occurs with some pollutants, such as As, Ni, Cd and Pb. This is due to the fact that the presence of these pollutants in the atmosphere normally originates in a variety of pollution sources, not exclusively bound to specific industrial activities at a certain location. As a consequence, the emission rates of these pollutants from all the point or area sources are difficult to estimate. A solution to address this problem consists in performing a spatial disaggregation of emission inventories (Maes et al. 2009). Notwithstanding, there is an underlying uncertainty associated with the method of disaggregation together with the inherent uncertainty of the emission inventories themselves. For that reason, for pollutants like the ones under study in this work, the performance of complex models is often equal to that of simpler methodologies. This fact highlights the interest of statistical models (e.g., linear regression techniques and non-linear modelling techniques) to estimate the ambient air concentration of atmospheric pollutants even though a wide range of deterministic models, as reviewed by El-Harbawi (2013), have been already developed and studied in the literature. Nevertheless, techniques such as partial least squares regression (PLSR), which presents advantages over other statistical linear regression techniques because it combines features from factor analysis statistical methods, such as principal component analysis (PCA) and linear regression techniques, as multiple linear regression (MLR), may potentially lead to more accurate estimations than those provided by MLR or principal component regression (PCR). Furthermore, according to Wold et al. (2001), although regression techniques such as MLR works reasonably well with problems involving fairly few uncorrelated independent variables, PLSR is preferable when analysing more intricate problems because it is able to manage simultaneously numerous and collinear predictor variables and responses. Despite the fact that it has been widely applied in other disciplines, chemometrics in particular, and used in some works related to atmospheric pollution (Ogulei et al. 2006; Wingfors et al. 2001), there are few studies on the application of PLSR to predict atmospheric pollutant concentrations. Pires et al. (2008) tested the ability of different linear models, including PLSR, to predict daily mean concentrations of particles with an aerodynamic diameter of less than 10 µm (PM_{10}) in Oporto (Portugal). It was obtained that even though every model fitted the data similarly well, PLSR shows higher generalization ability than other linear techniques. Polat and Durduran (2012) used regression models such as least squares regression (LSR), PLSR and MLR to predict daily particulate matter concentration values in the city of Konya (Turkey). PLSR performance, slightly better than those of the other regression models, was remarkably improved by considering data pre-processing methods such as output-dependent data scaling (ODDS). Singh et al. (2012) compared PLSR with non-linear modelling approaches to predict respirable suspended particulate matter (RSPM), SO_2 and NO_2 in Lucknow city (India). Both linear and non-linear approaches provided adequate estimations, especially for the RSPM, with values of correlation coefficient up to 0.9. Nonetheless, non-linear models performed relatively better than the linear PLSR models.

With respect to non-linear modelling approaches, artificial neural networks (ANNs) have been suggested as fair alternatives to statistical linear regression methods because they usually provide equal or superior results, especially when there is non-linear behaviour involved in the problem under analysis, i.e., cases in the atmospheric sciences (Gardner and Dorling 1998). For this reason, ANNs are particularly expected to produce good predictive results when modelling PM mass concentrations compared with common gaseous pollutants based on their ability to capture the highly non-linear character of the complex processes that control the formation, transportation and removal of aerosols in the atmosphere (Grivas and Chaloulakou 2006). Furthermore, ANNs have been extensively applied in the past in the atmospheric literature with successful results regarding forecasting major gaseous air pollutant concentrations, such as nitrogen oxides (Gardner and Dorling 1999; Kolehmainen et al. 2001; Lu et al. 2003), sulphur dioxide (Chelani et al. 2002a), and (commonly) ozone (Abdul-Wahab and Al-Alawi 2002; Chaloulakou et al. 2003; Comrie 1997; Inal 2010; Sousa et al. 2007; Wang et al. 2003; Yi and Prybutok 1996). Moreover, a number of studies have been conducted using ANN approaches to forecast airborne PM mass concentrations (Caselli et al. 2009; Chelani 2005; Grivas and Chaloulakou 2006; Hoi et al. 2009; Kim et al. 2009; Papanastasiou et al. 2007; Paschalidou et al. 2011; Perez and Reyes 2002; Perez and Reyes 2006; Pérez et al. 2000; Voukantsis et al. 2011), predict PM mass concentrations, and predict other gaseous pollutant concentrations (Brunelli et al. 2007; Cai et al. 2009; Hrust et al. 2009; Jiang et al. 2004; Kukkonen et al. 2003; Kurt et al. 2008; Lu et al. 2004; Lu et al. 2003; Niska et al. 2005; Turias et al. 2008). Nevertheless, regarding the PM composition and estimation of PM constituents, few studies have been conducted. In particular, with respect to the metal content in PM, Chelani et al. (2002b) used ANNs to predict ambient PM_{10} and metals, such as Cd, Cr, Fe, Ni, Pb and Zn, in the air of Jaipur, India, in 1999. It was observed that the ANN models were able to predict all the pollutant concentrations with low values of root means square error (RMSE). Nonetheless, more studies related to atmospheric metal concentration estimations by means of ANNs have been conducted, such as the study performed by Li et al. (2009) in which statistical models based on back-propagation ANNs and MLR are applied to reconstruct occupational manganese exposure. Apart from ANNs, some research has been conducted to model metal concentrations in ambient air using other statistical approaches. Hernández et al. (1992) applied state-space modelling, Box-Jenkins modelling and time series autoregressive integrated moving average (ARIMA) models to estimate the daily concentrations of air-particulate Fe and Pb in Madrid (Spain). Predictions of daily Fe were better than those of Pb. No difference being found between State-space and Box-Jenkins models, their outcomes were better than those of ARIMA models in terms of root mean squared error (RMSE), correlation coefficient and efficiency. Chelani et al. (2001) used a state-space model coupled with Kalman filter and an autoregressive model with external input (ARX model) to forecast Pb, Fe and Zn along with RSPM in Delhi (India). The state space model performed better than the ARX model. On the other hand, Vicente et al. (2012) developed predictive models based on multiple regression analysis together with time series (ARIMA) models to predict the concentration of total suspended particles (TSP), PM_{10}, As, Cd, Ni and Pb in the ambient air of Castellón (Spain). Furthermore, in a previous study conducted by Arruti et al. (2011), estimations of As, Cd, Ni and Pb levels in Cantabria (Spain) by means of statistical MLR and PCR models have been conducted. It is concluded that both represent valid approaches as objective estimation techniques.

This paper is focused on the development of PLSR and ANN statistical models to estimate the levels of As, Cd, Ni and Pb in the ambient air of two urban areas: Castro Urdiales and Reinosa in the Cantabria region (northern Spain). These models are evaluated according to the uncertainty requirements established by the EU for objective estimation techniques as well as for their ability to estimate the mean concentration. Additionally, an external validation of the models developed is performed.

2. Materials and methods

2.1. Statistical model fundamentals

2.1.1. Partial least squares regression (PLSR)

Partial least squares regression is a multivariate calibration technique whose aim is to investigate the relationship between a set of dependent variables or responses and a set of independent variables known as predictors. Firstly, in a similar manner to PCA, PLSR performs a decomposition of the original predictor variables (X-matrix, which consists of environmental observations in this study) by projecting them to a new space and extracts a set of orthogonal factors, called latent variables, which have the best predictive ability. Simultaneously, a decomposition of the response variables (Y-matrix, composed of metal level observations) is also performed. This decomposition step is made in a manner that the projections (scores) of X have maximum covariance with the projections of Y. This procedure is followed by a regression stage, where PLSR (just as MLR) creates a linear combination of the predictor variables in order to predict Y (Abdi 2010).

In this work, cross-validation techniques were used to select the more suitable number of significant components. PLS Toolbox (Eigenvector Research, Inc.) for MATLAB was used in the present study to develop the PLSR models.

2.1.2. Artificial neural networks (ANNs)

Artificial neural networks are computational systems inspired by the biological central nervous system. They consist of a number of simple process elements, commonly referred to as artificial neurons, which are logically arranged into layers, highly interconnected, and interact with each other via weighted connections. Through a supervised training process, in which they are successively presented with a series of input and associated output data, ANNs are capable to learn to model highly non-linear relationships and, as a result, to accurately generalise when previously unseen data are presented afterwards. The reader is referred the handbooks of Bishop (1995) and Hassoun (1995) for a comprehensive description of the ANN technique.

Plenty of neural network architectures exist. In this work, based on the different ANN approaches found in the air quality related literature, a multilayer perceptron (MLP) neural network architecture was selected; details of the architecture are provided in Gardner and Dorling (1998).

Because the ratio of input variables/number of samples is relatively high in this work due to the number of samples that were collected by the Regional Environmental Ministry, applying a dimension reduction technique prior to the ANN models was expected to produce an improvement in the estimations as reported in some studies (Lu et al. 2003; Sousa et al. 2007). Therefore, an alternative approach in which the PCA is performed before the development of the ANN models (hereafter known as PCA-ANNs) is considered.

The ANN models in this study were developed using the Neural Network Toolbox for MATLAB (MathWorks, Inc.).

2.2. Study area

Two urban areas in the Cantabria region (northern Spain) whose air quality may be influenced by the presence of metallurgical and other industrial activities in their vicinity were selected: Castro Urdiales and Reinosa (Fig. 1). The former area is a coastal urban site at the NE zone of Cantabria which has 32258 (2010) inhabitants and encompasses an area of approximately 97 km^2. Pollution in this area has a marked anthropogenic origin which is caused by traffic, not in vain Castro Urdiales is surrounded by the main national highway in the northern part of Spain. Pollution also proceeds from industrial activities, such as chemical and metallurgical plants and an oil refinery, located 10-30 km SE (near the city of Bilbao). The monitoring station is located at 43º22'53"N, 3º13'22" W and 20 m above sea level, in the core of the urban area. In contrast, Reinosa, covering nearly 4 km^2 with approximately 10277 inhabitants (2010), is located inland, at about 50 km off the shore, in the southern part of the region. The sampling station is located at 43º00'01"N, 4º08'13"W and 850 m above sea level. It is in close proximity to a steel manufacturing plant and also to a national highway, main exit route from Cantabria, which establishes connection with the central Iberian Peninsula.

2.3. Input dataset

The dataset used in this study is divided into response variables and predictor variables. The former data consist of As, Cd, Ni and Pb concentrations (ng m^{-3}) in airborne PM$_{10}$ for the period from 2008 to 2010 at the two study sites. The PM$_{10}$ sampling was performed by the Cantabrian Regional Environmental Ministry according to the reference method for the determination of the PM$_{10}$ fraction of suspended particulate matter detailed in standard UNE-EN 12341:1999. 48h averaged samples of PM$_{10}$ were taken once every two weeks for the period from 2008 to 2009 and 24h averaged samples of PM$_{10}$ were collected for 2010 with a weekly sampling frequency. The content of a number of metals and metalloids in the PM$_{10}$ samples was determined by our research group based on the standard method for the measurement of Pb, Cd, As and Ni in the PM$_{10}$ fraction of the suspended particulate matter described in standard UNE-EN 14902:2006. According to this, after gravimetric determination of the particle concentration levels, the PM$_{10}$ filters were treated with microwave-assisted acid digestion to extract the analytes into an aqueous solution prior to the analytical determination of their concentration by inductively coupled plasma mass spectroscopy (ICP-MS). Further details of this analytical method can be found in Arruti et al. (2010).

As a consequence of the high cost associated with the analytical determination of the content of this sort of pollutants in particulate matter, a considerably low number of samples was selected for the analysis. However, this number was sufficient to guarantee the minimum time coverage (14%) for indicative measurements as European Council Directive 2004/107/EC requires.

The predictor variables are qualitative or nominal variables (Table 1) that take into account seasonal effects, Saharan dust intrusion and weekend effects or quantitative or continuous variables, namely, meteorological data and major atmospheric pollutant concentration, which are detailed in Table 2. With respect to the nominal variables, the information regarding the occurrence of Saharan dust intrusion events has been obtained from annual reports on African dust episodes over Spain (MAGRAMA 2015), which are developed by the Spanish National Research Council (CSIC) in collaboration with the Spanish Ministry of Agriculture, Food and Environment. In contrast, the continuous variables are measured automatically in real time (maximum time resolution of fifteen minutes) at the monitoring stations of the Cantabrian Regional Air Quality Monitoring Network located in the study sites and are available at the Regional Environment Ministry website. Average values of continuous variables were calculated according to the corresponding duration of the PM_{10} sampling periods (48 hours for 2008-2009 samples and 24 hours for 2010 samples). Moreover, as regards to PM_{10} concentration, it has been included as input variable in the form of natural logarithm because of this transformation being reported to improve the performance of regression models (Arruti et al. 2011).

Prior to model development it is always rather convenient to take account of the application of a data pre-processing method, especially if there is lack of knowledge regarding the relative importance of the variables. In this study, the following data pre-treatment procedure was applied:

1. Dependent variable normalisation by the respective LAT in order to minimise scale effects.
2. Input variable auto-scaling, subtracting the mean and dividing by the standard deviation, in an attempt to make each variable a priori equally important.
3. Multivariate outlier identification and removal method based on Mahalanobis distance. It is a well-known classical approach that computes the Mahalanobis distance (MD) of each observation as an indicative measure of the distance of each data point from the centre of the multivariate data cloud. By

convention, this method identifies as outliers those observations with a large MD (exceeding the 99% quantile of a chi-square distribution).

Apart from the data pre-processing treatment, over-fitting is another decisive matter that must be taken into consideration beforehand so that it could be prevented. This term refers to the circumstance that occurs when a model fit the data in such a manner that not only captures the underlying trend in the data but also the unexplained variation or statistical noise and therefore it is unable to generalize properly - that is, to correctly perform when new observations are presented. In order to overcome this phenomenon it is highly recommended the consideration of an additional verification or cross-validation data subset, besides the training or fitting dataset, to check the models performance during the model development stage (usually known as calibration or fitting for PLSR and training for ANNs). Additionally, if the generalisation ability of a model is to be tested, a subset of samples has to be kept in reserve to perform an external validation with previously unused observations once the models have been developed. For that reason, the complete dataset was divided into three different subsets: 60% for training/fitting, 20% for verification and 20% for external validation. Data partition of the available data, often randomly conducted, was carried out in this work by means of the Kennard-Stone algorithm (Kennard and Stone 1969) with the purpose that the resulting subsets are statistically representative. This data division method, originally developed for design of experiments, has been traditionally applied to select calibration samples extracting subsets, as much diverse as possible, from a large set of candidate samples based on the Euclidean distance, which is employed as a measure of similarity between samples (the lower the Euclidean distance, the higher the similarity). Initially, the pair of samples with the largest Euclidean distance are selected. Subsequently, by means of an iterative process that concludes when the number of required objects is reached, more samples are selected, maximizing the minimal Euclidean distances between those already selected and the remaining samples.

2.4. Model evaluation

The main criteria employed in this work to determine whether a model is suitable for air quality assessment purposes is principally based on two aspects: (i) the fulfilment of the European Union uncertainty requirements for objective estimation techniques, which are shown in Table 3 and (ii) the accuracy of estimated mean values. Additionally, a number of statistical parameters has been considered to evaluate the modelling performance and are also shown in Table 3.

3. Results and discussion

3.1. As, Cd, Ni and Pb levels in Castro Urdiales and Reinosa

Fig. 2 summarises the levels of As, Cd, Ni and Pb in PM_{10} at Castro Urdiales and Reinosa for the period from 2008 to 2010. According to the European Council Directive 2008/50/EC, because these levels did not exceed their lower assessment threshold and did not present significant variations throughout the period of study, modelling and objective estimation techniques are permitted as an alternative method to experimental measurements for air quality assessment.

3.2. Statistical estimation models for Castro Urdiales

Table 4 shows the results relating to the best-developed models at the Castro Urdiales site for the four pollutants under study using the three approaches: PLSR, ANNs and PCA coupled with ANNs. The results obtained for both the training and the external validation subsets are presented.

Limit/target values for As, Cd, Ni and Pb in ambient air in the European regulations are given in annual mean concentration values. Therefore, attention should be paid to the estimated mean concentrations in the study period. The normalised mean concentrations are presented in Table 4. The accuracy in the estimation of the mean concentration is evaluated by means of the fractional bias (FB) index. In this respect, the estimations are more accurate for the training step. At this step, PLSR provides a FB index lower than those obtained for ANNs and PCA-ANNs because the mean metal concentration estimated by the PLSR models are equal —up to two significant figures— to the corresponding observed values and that, according to the corresponding equation (Table 3), yields lower FB index values. However, the differences between estimated and observed mean concentrations using the three considered techniques are not remarkably significant. As for external validation, the precision is inferior to that of the training phase.

In a more illustrative way, Fig. 3 represents the mean metal concentration estimation expressed as a percentage of the corresponding limit/target value. The vertical axis is presented in logarithmic scale. The green area represents the zone below the LAT, the yellow area represents the zone between the UAT and the LAT, and the red area is the zone between the limit/target value and the UAT. Fig. 3 shows that, even though there are some differences between the estimated and the observed mean levels, they are similar. Moreover, because the observed metal(loid) levels are within the green area, well below the LAT, even higher discrepancy could be allowed. Therefore, the developed models provide satisfactory mean concentration estimations.

It is necessary to validate objective estimation techniques in the context of the EU Directives in terms of uncertainty. In this sense, according to Arruti et al. (2011), two indices have been considered: on the one hand, the RME, which is defined as the largest concentration difference of all percentile differences normalized by the respective observed value (Fleming and Stern 2007); on the other hand, the RDE, which evaluates the accuracy in the estimation of the observation closest to the limit/target value (Denby 2009). As observed in Table 4, the values of these indices for the four pollutants in question for the training and the external validation are well below 100%, which is the maximum permissible uncertainty limit for using objective estimation techniques as air quality assessment tools according to the European Council Directive 2008/50/EC. For As, Ni and Cd, ANNs provide higher RME values than PLSR and PCA-ANN. In the majority of cases, except for the As and Cd ANN models, the RME values are below 50%, which is the uncertainty requirement for modelling techniques. In all cases, the RDE values are below 10%. However, these indices have some limitations: it has been discussed that RME is sensitive to the presence of outliers resulting in an increase of the uncertainty values (Fleming and Stern 2007); RDE only evaluates the uncertainty of just one sample, the closest to the limit/target value.

From a scientific point of view, apart from a precise estimation of mean values to comply with the policy framework, a model should be able to correctly describe the temporal variations of dependent variables. For this purpose, a set of statistics has been used in this work. In the first place, the correlation coefficient is employed to measure the goodness of fit between the observed and the estimated values. The results show that PLSR correlation coefficients, which are within the range of 0.6-0.7, are less variable than those of ANNs and PCA-ANNs: whereas the highest correlation coefficient, an r value of 0.82, is found when using ANNs for Pb training, the correlation coefficients for As and Cd ANN models are significantly low and therefore unacceptable. This could be explained because in the area of study As and Cd tend to be in lower concentration than Ni and Pb and consequently in the period of study a number of samples have levels of As and Cd below their detection limits. As a result, models are trained to produce the same output from different inputs, a detrimental contradiction that may negatively affect the estimation of the rest of the samples. Moreover, as expected, the r values for external validation are often lower than those for training. Nonetheless, the PCA-ANN external validation correlation coefficients are systematically below 0.5.

In addition to the correlation coefficient, the precision of the individual sample concentration estimation is quantified by the RMSE, the NMSE and the FV, (see equations in Table 3). The RMSE values, which provide information regarding the differences between the observed and estimated concentrations, are shown in Table 4. However, to compare these differences for different approaches and pollutants, a normalised version of this parameter (NMSE) is more preferable because it does not take into account the range of the independent variable. In general, the three considered approaches provide low values of NMSE in the order of 10^{-1}.

With respect to the FV index, positive values can be observed in Table 4; this indicates that the estimated variance is lower than the observed variance. Therefore, estimated values are less dispersed than observed values, which tend to be more distanced from the mean value. This fact, together with a positive FB corresponding to a slight mean value underestimation, indicates that there are some shortcomings in the model capacity to perfectly describe all the concentration variations, especially regarding peak values. Nevertheless, despite no substantial differences being found when comparing PLSR and ANNs, in general both models are able to capture the underlying trend and provide temporal variations with similar shape to that of the observed values as depicted in Fig. 4 for Pb and Ni in the training stage.

Based on the results obtained, there is no improvement associated with considering a dimension reduction technique such as PCA before the development of the ANNs. This could be accounted for the fact that most ANNs suffer less from the curse of dimensionality than some other techniques, as they can concentrate on a lower dimensional section of the high-dimensional space, which may be done, for instance, by disregarding completely an input, setting the corresponding weights to zero. Hence, for this specific application dimensionality reduction has been proven not to be effective because removing input variables from the analysis entails a loss on the predictive ability of the model.

Furthermore, because these models are devised to be used when the pollutant levels are sufficiently lower at a certain location, in principle the moderated inaccuracy to estimate peak values should not represent an unacceptable drawback to acknowledge these models as proper approaches complying with regulatory requirements: the uncertainty values obtained with the developed models and the accuracy in the estimation of the mean values would be favourable enough from a regulatory perspective. Nonetheless, some refinement is possible because, as mentioned, there are some difficulties in estimating the highest observed concentrations, which are underestimated. In this regard, further work involving new additional input variables and the enlargement of the database with additional samples from different periods of time would be recommendable.

3.3. Statistical estimation models for Reinosa

Analogously to the results at the Castro Urdiales site, the statistical parameters corresponding to the best-developed models at the Reinosa site are presented in Table 5.

Regarding the uncertainty indices, it is observed that, as in Castro Urdiales, the RME and RDE values at the Reinosa site are below 100% for the estimations obtained with the three different models developed for the four pollutants. Hence, the quality objectives for ambient air quality assessment by means of objective estimation techniques are met. However, there is a general increase in the obtained RDE values, especially for As and Ni, which are significantly greater than those obtained at the Castro Urdiales site.

In relation to the mean values, again, PLSR provides the lowest FB training values, but the FB external validation values are greater than the training values. Although there are still evident differences between the observed and estimated mean concentrations, 90% of the estimations do not differ by more than 50%. Therefore, as shown in Fig. 5, the three developed models provide satisfactory estimations. Nonetheless, a substantial increase in FB values is found in Reinosa compared with Castro Urdiales.

Results at the Reinosa site present more variability between the training and external validation correlation coefficient values for each pollutant than the results at the Castro Urdiales site, which may be partially accounted for the higher inherent variance of metal levels in Reinosa compared to those obtained in Castro Urdiales. However, the ANN correlation coefficient values are generally equal or superior to those of PLSR and PCA-ANNs. As for the errors in the individual sample concentration estimations, the NMSE values for Reinosa and Castro Urdiales are within the same range. Nevertheless, the FV values are slightly greater in Reinosa than in Castro Urdiales but still lower than 1.0, which represents 50% of the observed variance.

Results prove that these models provide an acceptable performance in varied areas of a region, even when there is a complex pollution framework with diverse emission sources, as is the case of Castro Urdiales. Nevertheless, because the models were trained on data for particular sites and having been demonstrated that the precision in the estimation is dependent on the specific location, these models can therefore only be used with confidence at those sites. This dependence is especially pronounced in the ANN models, which produced a higher variability in the results than the PLSR or PCA-ANN models. This may be influenced by the fact that a limited number of samples are used for developing the models due to the unavailability of additional observations stemming from their costliness and time consumption. Thus, it could be inferred that for small datasets, linear regression techniques can work as well as non-linear modelling approaches in terms of the estimation of metal(loid) levels in ambient air.

4. Conclusions

Statistical models are developed as objective estimation techniques to estimate the As, Cd, Ni and Pb in ambient air at a local scale in two urban areas in the Cantabria region (northern Spain): Castro Urdiales and Reinosa. These models were built based on linear regression techniques, partial least squares regression (PLSR), and the non-linear modelling technique of artificial neural networks (ANNs). Additionally, an alternative approach is considered that performs principal component analysis (PCA) prior to the ANN analysis (PCA-ANNs). Furthermore, these models were externally validated using previously unseen data.

The models are evaluated by means of a number of statistical parameters, including uncertainty indices, to determine if they comply with the EU quality requirements for objective estimation techniques.

Additionally, the model performance in estimating the individual sample concentrations is evaluated by means of a number of statistical parameters, including a correlation coefficient, RMSE, NMSE and FV.

Based on the results obtained, PLSR and ANN techniques are acceptable alternatives to estimate the mean concentration of As, Cd, Ni and Pb for the period of study in the two considered sites while fulfilling the uncertainty requirements for objective estimation techniques established in the EU Directives. Consequently, PLSR and ANN-based statistical models represent a proper alternative to experimental measurements for air quality assessment purposes in the area of study. However, ANNs have not demonstrated to offer a clear superior performance over the linear regression technique, what may be attributed to the modest size of the available database. Furthermore, the three considered approaches had some difficulties providing accurate estimations of the levels of individual samples, particularly for the external validation subset. Moreover, the application of PCA before the ANN model development did not yield an improvement of the models.

Acknowledgements

The authors gratefully acknowledge the financial support from the Spanish Ministry of Economy and Competitiveness through the Project CMT2010-16068. The authors also thank the Regional Environment Ministry of the Cantabria Government for providing the PM_{10} samples at the Castro Urdiales and Reinosa sites.

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 97-106.

Abdul-Wahab, S. A., & Al-Alawi, S. M. (2002). Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling and Software*, 17(3), 219-228.

Arruti, A., Fernández-Olmo, I., & Irabien, A. (2010). Evaluation of the contribution of local sources to trace metals levels in urban $PM_{2.5}$ and PM_{10} in the Cantabria Region (Northern Spain). *Journal of Environmental Monitoring*, 12, 1451-1458.

Arruti, A., Fernández-Olmo, I., & Irabien, A. (2011). Assessment of regional metal levels in ambient air by statistical regression models. *Journal of Environmental Monitoring*, 13(7), 1991-2000.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition and Machine Learning*. Oxford: Clarendon Press.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007). Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy. *Atmospheric Environment*, 41(14), 2967-2995.

Cai, M., Yin, Y., & Xie, M. (2009). Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transportation Research Part D: Transport and Environment*, 14(1), 32-41.

Caselli, M., Trizio, L., De Gennaro, G., & Ielpo, P. (2009). A simple feedforward neural network for the PM_{10} forecasting: Comparison with a radial basis function network and a multivariate linear regression model. *Water, air, and soil pollution*, 201(1-4), 365-377.

Chaloulakou, A., Saisana, M., & Spyrellis, N. (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment*, 313(1-3), 1-13.

Chelani, A. B., Gajghate, D. G., Tamhane, S. M., & Hasan, M. Z. (2001). Statistical modeling of ambient air pollutants in Delhi. *Water, air, and soil pollution*, 132(3-4), 315-331.

Chelani, A. B., Chalapati Rao, C. V., Phadke, K. M., & Hasan, M. Z. (2002a). Prediction of sulphur dioxide concentration using artificial neural networks. *Environmental Modelling and Software*, 17(2), 161-168.

Chelani, A. B., Gajghate, D. G., & Hasan, M. Z. (2002b). Prediction of ambient PM_{10} and toxic metals using artificial neural networks. *Journal of the Air and Waste Management Association*, 52(7), 805-810.

Chelani, A. B. (2005). Predicting chaotic time series of PM_{10} concentration using artificial neural network. *International Journal of Environmental Studies*, 62(2), 181-191.

Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air and Waste Management Association*, 47(6), 653-663.

Denby, B. (2009). Guidance on the use of the models for the European Air Quality Directive. A Working document of the Forum for Air Qualtiy Modelling in Europe, FAIRMODE. ETC/ACC Report.

El-Harbawi, M. (2013). Air quality modelling, simulation, and computational methods: A review. *Environmental Reviews*, 21(3), 149-179.

Fleming, J., & Stern, R. (2007). Testing model accuracy measures according to the EU directives-examples using the chemical transport model REM-CALGRID. *Atmospheric Environment*, 41, 9206-9216.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14-15), 2627-2636.

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(5), 709-719.

Grivas, G., & Chaloulakou, A. (2006). Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment*, 40(7), 1216-1229.

Hanna, S. R. (1989). Plume dispersion and concentration fluctuations in the atmosphere. In P.N. Cheremisinoff (Ed.), *Encyclopedia of Environmental Control Technology, Vol. 2, Air Pollution Control*. Houston, Texas: Gulf Publishing Co.

Hassoun, M. H. (1995). *Fundamentals of Artificial Neural Networks*. London, England: MIT Press.

Hernández, E., Martín, F., & Valero, F. (1992). Statistical forecast models for daily air particulate iron and lead concentrations for Madrid, Spain. *Atmospheric Environment*, 26B, 107-116.

Hoi, K. I., Yuen, K. V., & Mok, K. M. (2009). Prediction of daily averaged PM_{10} concentrations by statistical time-varying model. *Atmospheric Environment*, 43(16), 2579-2581.

Hrust, L., Klaic, Z. B., Križan, J., Antonic, O., & Hercog, P. (2009). Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment*, 43(35), 5588-5596.

Inal, F. (2010). Artificial Neural Network Prediction of Tropospheric Ozone Concentrations in Istanbul, Turkey. *Clean - Soil, Air, Water*, 38(10), 981.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., & Shao, D. (2004). Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment*, 38(40 SPEC.ISS.), 7055-7064.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137-148.

Kim, M., Kim, Y., Sung, S., & Yoo, C. (2009). Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks. *ICCAS-SICE 2009 - ICROS-SICE International Joint Conference 2009, Proceedings*, 1688-1692.

Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment*, 35(5), 815-825.

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(32), 4539-4550.

Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment international*, 34(5), 592-598.

Li, Y., Luo, F., Jiang, Y., Lu, Y., Huang, J., & Zhang, Z. (2009). A prediction model of occupational manganese exposure based on artificial neural network. *Toxicology Mechanisms and Methods*, 19(5), 337-345.

Lu, W. Z., Wang, W. J., Wang, X. K., Xu, Z. B., & Leung, A. Y. T. (2003). Using improved neural network model to analyze RSP, NO_X and NO_2 levels in urban air in Mong Kok, Hong Kong. *Environmental monitoring and assessment*, 87(3), 235-254.

Lu, W., Wang, W., Wang, X., Yan, S., & Lam, J. C. (2004). Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. *Environmental research*, 96(1), 79-87.

Maes, J., Vliegen, J., Van de Vel, K., Janssen, S., Deutsch, F., De Ridder, K., Mensink, C. (2009). Spatial surrogates for the disaggregation of CORINAIR emission inventories. *Atmospheric Environment*, 43, 1246-1254.

Ministerio de Agricultura, Alimentación y Medio Ambiente (MAGRAMA), 2015. Histórico de informes de episodios naturales. http://www.magrama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/gestion/anuales.aspx

Niska, H., Rantamäki, M., Hiltunen, T., Karppinen, A., Kukkonen, J., Ruuskanen, J., & Kolehmainen, M. (2005). Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. *Atmospheric Environment*, 39(35), 6524-6536.

Ogulei, D., Hopke, P. K., Zhou, L., Patrick Pancras, J., Nair, N., & Ondov, J. M. (2006). Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data. *Atmospheric Environment*, 40(SUPPL. 2), 396-410.

Papanastasiou, D. K., Melas, D., & Kioutsioukis, I. (2007). Development and assessment of neural network and multiple regression models in order to predict PM_{10} levels in a medium-sized Mediterranean city. *Water, air, and soil pollution*, 182(1-4), 325-334.

Paschalidou, A. K., Karakitsios, S., Kleanthous, S., & Kassomenos, P. A. (2011). Forecasting hourly PM_{10} concentration in Cyprus through artificial neural networks and multiple regression models: Implications to local environmental management. *Environmental Science and Pollution Research*, 18(2), 316-327.

Pérez, P., Trier, A., & Reyes, J. (2000). Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment*, 34(8), 1189-1196.

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(28), 4555-4561.

Perez, P., & Reyes, J. (2006). An integrated neural network model for PM_{10} forecasting. *Atmospheric Environment*, 40(16), 2845-2851.

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(5), 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(8), 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.

Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M., & Pereira, M. C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling and Software*, 22(1), 97-103.

Turias, I. J., González, F. J., Martin, M. L., & Galindo, P. L. (2008). Prediction models of CO, SPM and SO_2 concentrations in the Campo de Gibraltar Region, Spain: A multiple comparison strategy. *Environmental monitoring and assessment*, 143(1-3), 131-146.

Vicente, A. B., Jordán, M. M., Sanfeliu, T., Sánchez, A., & Esteban, M. D. (2012). Air pollution prediction models of particles, As, Cd, Ni and Pb in a highly industrialized area in Castellón (NE, Spain). *Environmental Earth Sciences*, 66(3), 879-888.

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., & Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Science of the Total Environment*, 409(7), 1266-1276.

Wang, W., Lu, W., Wang, X., & Leung, A. Y. T. (2003). Prediction of maximum daily ozone level using combined neural network and statistical characteristics. *Environment international*, 29(5), 555-562.

Wingfors, H., Sjödin, A., Haglund, P., & Brorström-Lundén, E. (2001). Characterisation and determination of profiles of polycyclic aromatic hydrocarbons in a traffic tunnel in Gothenburg, Sweden. *Atmospheric Environment*, 35(36), 6361-6369.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.

Yi, J., & Prybutok, V. R. (1996). A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3), 349-357.

Figure captions

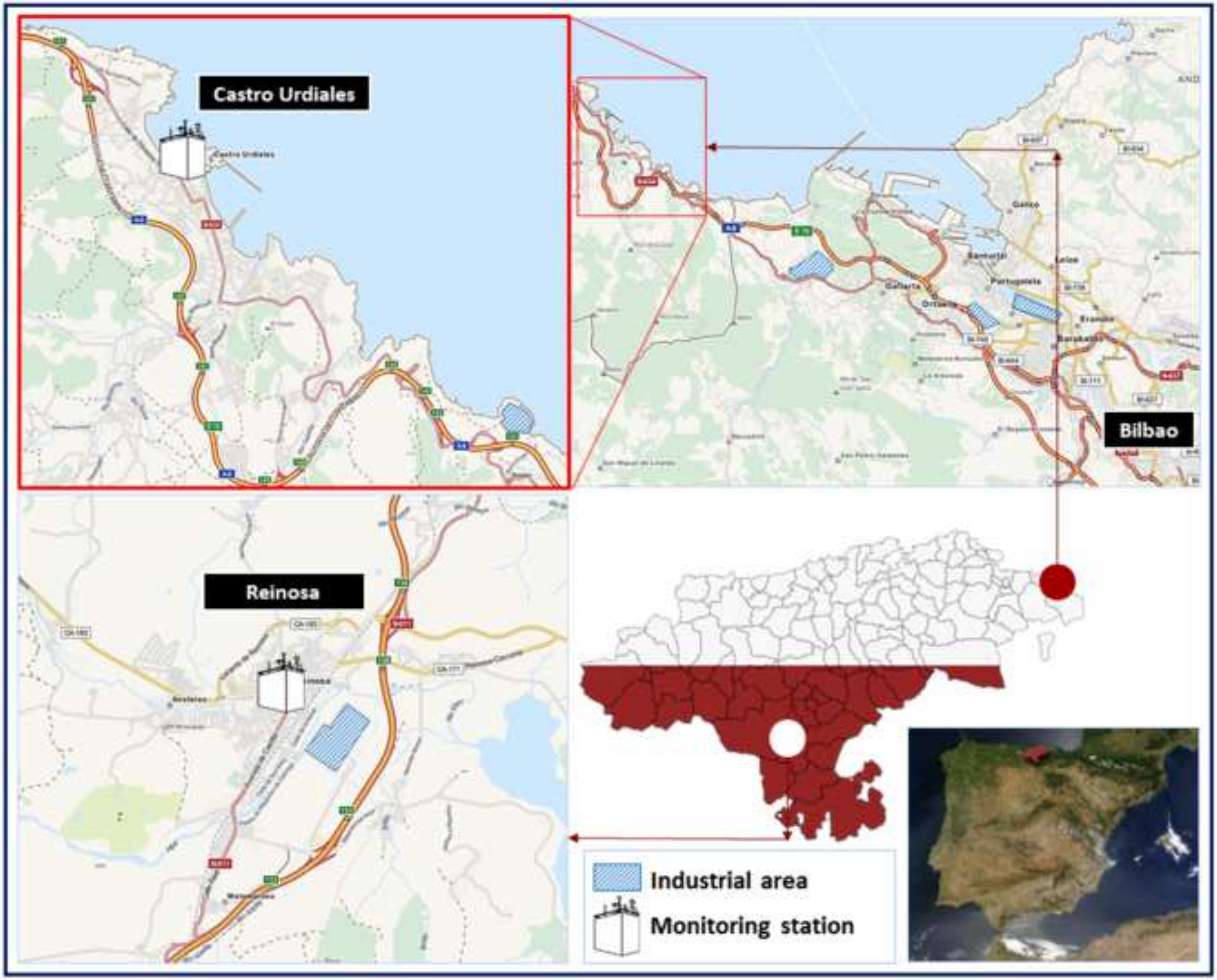
Fig. 1 Location of the monitoring stations

Fig. 2 As, Cd, Ni and Pb levels in PM_{10}, normalized with respect to their corresponding LAT, for the period of study at (a) the Castro Urdiales site and (b) the Reinosa site. The 2008 mean values are obtained from Arruti et al. (2011). LAT: 250 ng m^{-3} (Pb), 2.4 ng m^{-3} (As), 10 ng m^{-3} (Ni), 2 ng m^{-3} (Cd)

Fig. 3 Comparison between the observed and estimated mean concentrations at the Castro Urdiales site and their respective assessment thresholds and limit/target values. (a) Pb; (b) As; (c) Ni and (d) Cd. TV: 500 ng m^{-3} (Pb), 6 ng m^{-3} (As), 20 ng m^{-3} (Ni), 5 ng m^{-3} (Cd); UAT: 70% (Pb and Ni), 60% (As and Cd); LAT: 50% (Pb and Ni), 40% (As and Cd)

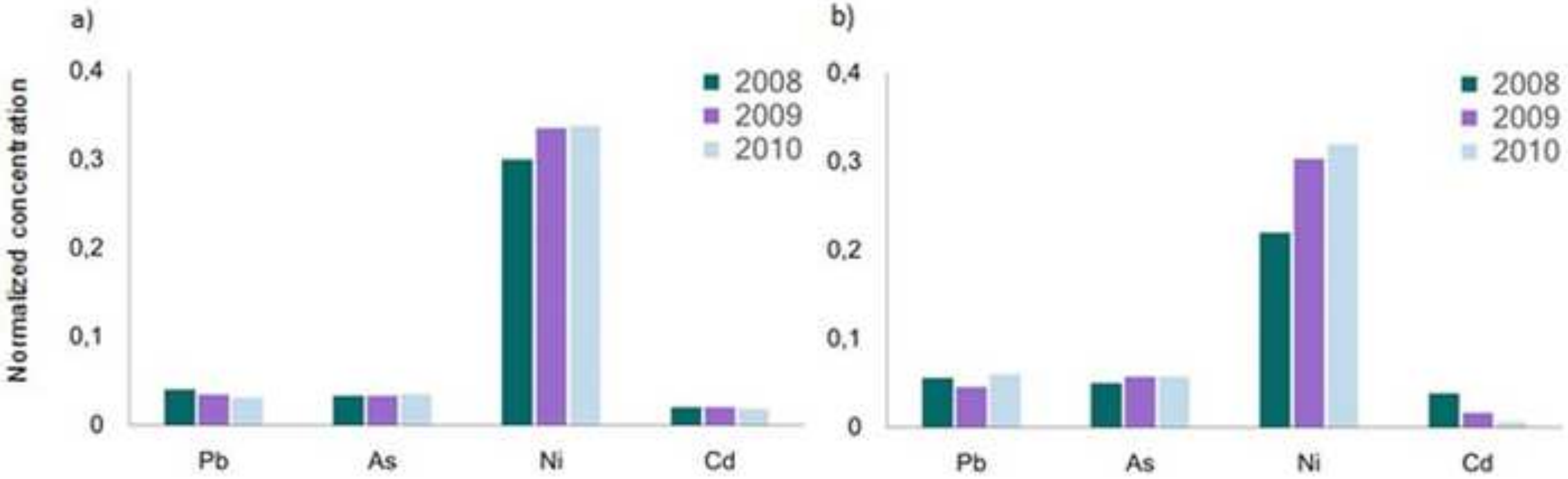
Fig. 4 Fitting of the Pb and Ni models for the training subset at the Castro Urdiales site

Fig. 5 Comparison between the observed and estimated mean concentrations at the Reinosa site and their respective assessment thresholds and limit/target values. (a) Pb; (b) As; (c) Ni and (d) Cd. TV: 500 ng m^{-3} (Pb), 6 ng m^{-3} (As), 20 ng m^{-3} (Ni), 5 ng m^{-3} (Cd); UAT: 70% (Pb and Ni), 60% (As and Cd); LAT: 50% (Pb and Ni), 40% (As and Cd)


Castro Urdiales
Castro Urdiales
Getxo
Leioa
Santurtzi
Portugalete
Gallarta
Ortuella
Erandio
Barakaldo
Bilbao
Reinosa
Industrial area
Monitoring station

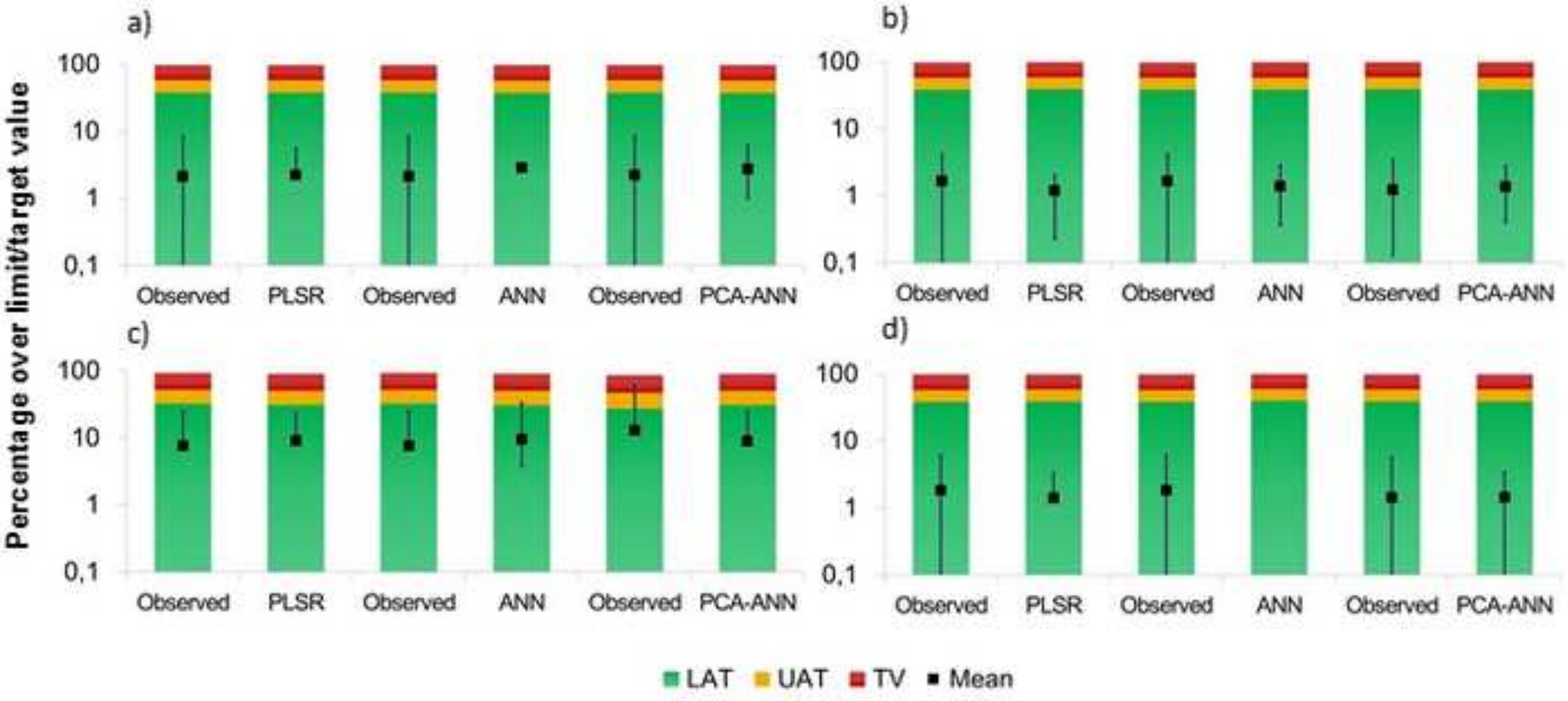
colour figure

Click here to download colour figure: Fig 2.tif



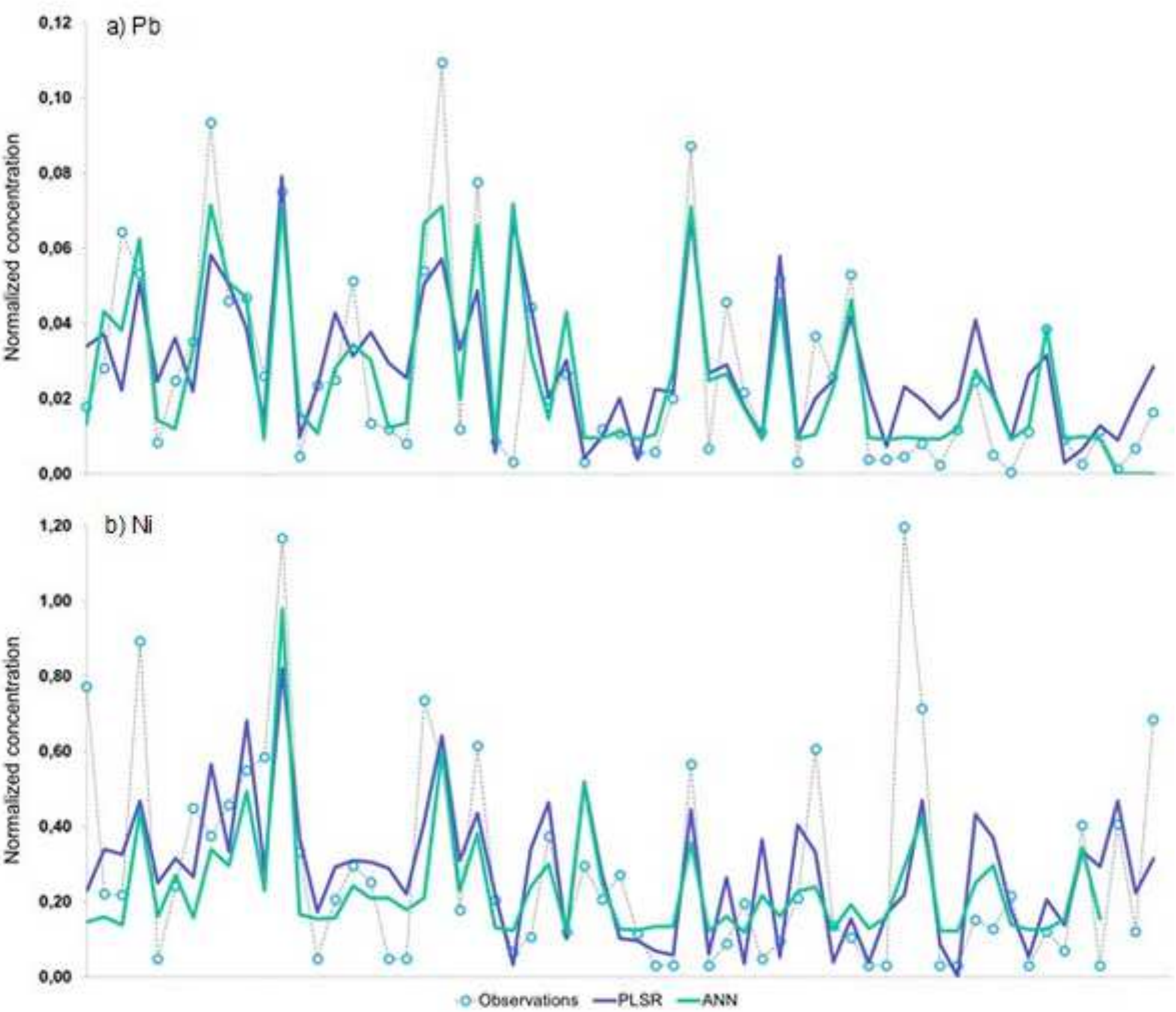
colour figure

Click here to download colour figure: Fig 3.tif



colour figure

Click here to download colour figure: Fig 4.tif



colour figure

Click here to download colour figure: Fig 5.tif

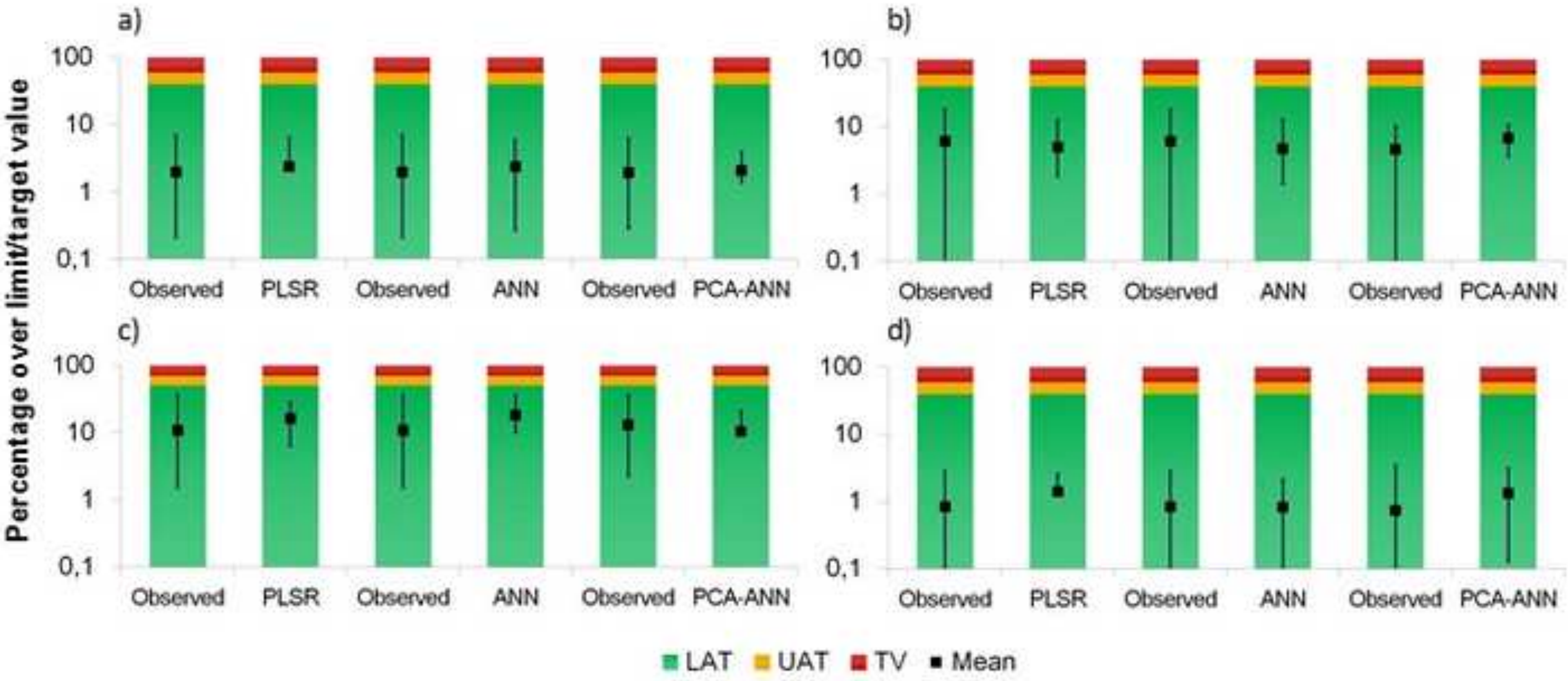


Table 1. List of nominal variables used as input for the models

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: Working day; 1: Weekend

Table 2. List of continuous variables used as input for the models.

Notation	Description[a]	Type	Units
$LnPM_{10}$	Average natural logarithm of PM_{10} concentration (µg m^{-3})	Major air pollutant	-
SO_2	Average concentration of sulphur dioxide	Major air pollutant	µg m^{-3}
O_3	Average concentration of ozone	Major air pollutant	µg m^{-3}
NO_x	Average concentration of nitrogen oxides	Major air pollutant	µg m^{-3}
T	Average temperature	Meteorological	ºC
RH	Average relative humidity	Meteorological	%
WD	Prevailing wind direction	Meteorological	º
WS	Prevailing wind speed	Meteorological	ms^{-1}
P	Average pressure	Meteorological	mbar
PP	Cumulative precipitation	Meteorological	L m^{-2}

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

Table 3. Statistical parameters used for evaluating the model performance

Evaluation	Statistic	Equation
EU Uncertainty	Relative maximum error without timing	$\mathrm{RME} = \max\left(\left\|C_{O,p}\text{-}C_{E,p}\right\|\right)/C_{O,p}$
	Relative directive error	$\mathrm{RDE} = \left\|C_{O,LV}\text{-}C_{E,LV}\right\|/LV$
Mean concentration	Fractional bias	$\mathrm{FB} = \frac{\overline{C_O}\text{-}\ \overline{C_E}}{0.5\ (\overline{C_O} + \overline{C_E})}$
Performance	Correlation coefficient	$r = \left[\frac{\sum_{i=1}^{n}\left(C_{O,i} - \overline{C_O}\right)\left(C_{E,i} - \overline{C_E}\right)}{\sqrt{\sigma_O \sigma_E}}\right]$
	Root mean square error	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_{O,i}\text{-}C_{E,i}\right)^2}$
	Normalised mean square error	$\mathrm{NMSE} = \frac{\overline{(C_O\text{-}\ C_E)^2}}{\overline{C_O}\ \overline{C_E}}$
	Fractional variance	$\mathrm{FV} = 2\ \frac{\sigma_O\text{-}\sigma_E}{\sigma_O + \sigma_E}$

Table 4. Training and external validation performance indices of the various developed models at the Castro Urdiales site

			EU Uncertainty		**Mean Concentration**[b]			**Performance**			
Pollutant	**Model**	**Subset**[a]	**RME (%)**	**RDE (%)**	**C_O 10^2**	**C_E 10^2**	**FB 10^2**	**r**	**RMSE 10^2**	**NMSE 10**	**FV 10**
Pb	PLSR	T	26.7	0.09	2.79	2.79	-9.8 10^{-15}	0.704	1.77	4.04	3.48
		V	49.6	0.71	4.13	3.01	31.4	0.620	2.71	5.90	7.44
	ANN	T	34.5	0.66	2.67	2.70	-0.9	0.820	1.44	2.88	1.98
		V	32.0	1.76	4.34	3.48	22.0	0.676	2.41	3.85	2.47
	PCA-ANN	T	36.9	0.41	3.14	3.11	1.0	0.681	2.03	4.25	4.41
		V	30.6	0.07	3.08	3.43	-10.8	0.269	2.60	6.41	3.60
As	PLSR	T	42.8	0.30	6.66	6.66	-4.2 10^{-12}	0.656	5.17	6.02	4.16
		V	34.8	1.74	5.32	5.50	-3.4	0.629	4.63	7.34	0.88
	ANN	T	77.0	0.17	6.96	6.81	2.1	0.130	6.96	10.22	12.74
		V	66.3	0.25	5.32	6.81	-24.6	0.193	5.70	8.99	11.26
	PCA-ANN	T	54.6	0.86	6.96	6.35	9.1	0.536	5.87	7.80	7.36
		V	33.5	0.96	5.55	6.33	-13.3	0.190	5.72	9.30	2.77
Ni	PLSR	T	34.7	10.83	27.61	27.61	3.1 10^{-13}	0.642	21.52	6.07	4.36
		V	22.1	2.64	18.91	22.67	-18.1	0.663	12.27	3.51	-0.97
	ANN	T	45.0	6.19	28.27	23.30	19.3	0.676	21.59	7.08	6.18
		V	32.6	1.24	19.36	23.71	-20.0	0.387	16.15	5.67	-0.58
	PCA-ANN	T	33.9	1.77	24.54	26.01	-5.8	0.643	17.60	4.85	4.92
		V	25.6	2.06	23.64	23.42	0.9	0.216	21.19	8.11	1.64
Cd	PLSR	T	40.9	0.14	3.75	3.75	-1.1 10^{-12}	0.672	3.36	8.03	3.92
		V	46.2	0.86	4.55	3.52	25.6	0.628	3.42	7.30	4.69
	ANN	T	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]
		V	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]	n.c.[c]
	PCA-ANN	T	33.5	0.47	3.84	3.88	-1.1	0.613	3.11	6.49	5.36
		V	41.5	0.42	3.56	3.62	-1.7	0.534	3.11	7.48	5.05

[a] T: Training; V: External validation

[b] O: Observed; E: Estimated

[c] Not calculated (n.c.)

Table 5. Training and external validation performance indices of the various developed models at the Reinosa site

Pollutant	**Model**	**Subset**[a]	**EU Uncertainty**		**Mean Concentration**[b]			**Performance**			
			RME (%)	**RDE (%)**	**C_O 10^2**	**C_E 10^2**	**FB 10^2**	**r**	**RMSE 10^2**	**NMSE 10**	**FV 10**
Pb	PLSR	T	31.1	0.61	5.60	5.60	$-6.2\ 10^{-14}$	0.723	3.48	3.86	3.21
		V	14.0	0.21	5.11	6.01	-16.2	0.553	3.76	4.60	0.17
	ANN	T	30.3	0.83	6.01	5.71	5.2	0.829	2.86	2.38	3.38
		V	35.9	1.07	5.11	5.95	-15.2	0.563	3.58	4.22	1.05
	PCA-ANN	T	42.3	1.48	5.50	6.02	-8.9	0.679	3.77	4.29	7.17
		V	42.2	1.38	4.96	5.68	-13.4	0.374	3.80	5.13	9.70
As	PLSR	T	28.2	1.25	13.61	13.61	$4.1\ 10^{-14}$	0.446	8.68	4.07	7.67
		V	31.7	5.66	15.12	12.46	19.3	0.441	12.66	8.51	7.23
	ANN	T	23.6	3.56	14.47	13.39	7.7	0.765	6.75	2.35	4.89
		V	35.4	6.01	15.12	11.82	24.5	0.393	13.17	9.71	6.78
	PCA-ANN	T	25.0	1.41	16.03	15.61	2.7	0.572	10.03	4.02	6.69
		V	37.4	8.26	11.52	16.96	-38.3	0.132	10.09	5.21	2.75
Ni	PLSR	T	53.4	5.45	30.61	30.61	$-3.9\ 10^{-6}$	0.386	20.49	4.48	8.86
		V	25.2	2.51	21.54	32.26	-39.9	0.549	19.23	5.32	6.13
	ANN	T	38.9	22.21	33.60	34.60	-2.9	0.460	20.92	3.78	8.54
		V	28.2	2.51	21.54	36.09	-50.5	0.455	22.71	6.63	4.10
	PCA-ANN	T	26.9	9.01	30.81	28.36	8.3	0.677	17.43	3.48	3.09
		V	42.5	20.85	25.95	20.94	21.4	0.304	22.50	9.32	5.53
Cd	PLSR	T	48.5	0.46	3.06	3.06	$-3.8\ 10^{-13}$	0.644	3.30	1.16	4.32
		V	46.2	0.59	2.09	3.56	-51.8	0.338	2.74	1.01	3.26
	ANN	T	59.7	0.19	3.40	2.47	31.5	0.641	3.74	1.67	7.03
		V	36.7	0.26	2.09	2.07	1.2	0.521	1.94	8.70	4.30
	PCA-ANN	T	67.5	7.02	3.21	3.56	-10.3	0.518	4.10	1.47	8.42
		V	34.3	0.40	1.83	3.32	-57.7	0.579	2.48	1.01	2.74

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