Prediction of PM_{10} Concentrations for Bahía Blanca, Argentina

Lucila L. Chiarvetto Peralta[1], Mónica F. Díaz[1,2], Nélida B. Brignole[1,2]

[1]Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC).
Universidad Nacional del Sur (DCIC-UNS) Av. Alem 1253 - Bahía Blanca (B8000CPB) - Argentina
[2] Planta Piloto de Ingeniería Química (UNS - CONICET) Cno. La Carringanda Km.7 Bahía Blanca (8000) - Argentina

mdiaz@plapiqui.edu.ar

Abstract. PM_{10} non-traditional modelling for e-government development is described in detail. Ambient PM_{10} concentrations were predicted using meteorological variables as inputs, whose relevance for a generated Artificial Neural Network was analyzed by a feature selection method. The work is specially focused on the surroundings of Bahía Blanca city, its petrochemical pole and Ing. White grain port. Its accuracy was tested with time windows ranging from 2004 to 2006. A trustworthy simulation of the physical phenomena was built. As a result, this predictive model will contribute to the local observatory in order to trigger early-alert warnings.

Keywords: artificial neural network, particulate matter, PM_{10}, forecast model.

1 Introduction

Bahía Blanca city is located at the South East of Buenos Aires province, Argentina, and its population is around 284,000 inhabitants. It is surrounded by a big petrochemical pole and a busy deep-water port (Ing. White), where ships load and unload plenty of cereal. As a consequence of various accidents that were originated by those industrial activities, environmental damages have been recorded and the population has urged the government to exert a severe control policy on the area. In 2000 Buenos Aires Province delegated and decentralized the environmental control to Bahía Blanca Town Hall. Thus, the Executive Technical Committee (CTE) was created. It was stated [1] that an environmental observatory would be useful for an efficient control and identification of dangerous phenomena. The CTE has embarked on this challenging enterprise with the help of local researchers. Besides, an epidemiologic study on asthma was made in the area [10], recognizing it as a major public health problem frequently associated with air pollution. Higher prevalence of asthma was found associated with areas where the particulate matter (PM) pollutant presents higher values. This elevated PM concentration in our monitoring area of interest was attributed mainly to grain activity in the harbour.

PM is a mixture of tiny pieces of solid or liquid material suspended in a gas. PM sources may be man-made or natural. Averaged over the globe, anthropogenic aerosols -those made by human activities- currently account for about 10% of the total amount of aerosols in our atmosphere. Increased levels of fine particles in the air are linked to health hazards, such as heart disease, altered lung function and lung cancer. PM_{10} is a name that refers to a measure of atmospheric particles whose aerodynamic diameters are smaller or equal to 10 µm. PM_{10} has been the pollutant particulate level standard EPA has been measuring against. EPA revised the air quality standards for particle pollution in 2006, deciding to retain the existing 24-hour PM_{10} standard of 150 µg/m^3 [9].

In 2006 an innovative proposal for environmental control in Bahía Blanca and its surroundings was defined [1]. It basically consisted in mounting a governmental observatory that would profit from novel communication aids and grown-up computing technology so as to control the industrial emission sources and to prevent environmental risks. Afterwards, the implementation of this project was devised [2]. A critical part of this implementation was grounded in the proper design of a data warehouse environment [3]. While planning how to build the data warehouse, the need for predictive models that served to trigger contaminant alarms was identified. This environmental observatory would play a strict control role aiming at an e-government policy. E-government essentially refers to centering the focus on the use of technologies so as to facilitate the governmental operation and services. Then, our interest in environmental protection demands the integration of technology in order to be able to incorporate data, store them, display and report them.

As to contaminant prediction, it is impossible to turn into a model based on phenomenon description because there are many factors that jointly affect contaminant distribution at any time. It is important to remark that none of the classical approaches succeeds in addressing nonlinear relationships whose complex behaviour either lacks interpretation or is difficult to represent using a straightforward function. Then, in the chemical engineering context the growing number of data available on the Internet or in specialized databases makes the establishment of innovative modelling techniques a significant priority in order to profit from the content of this valuable information. In this context, this paper focuses on defining a method that efficiently allows the modelling of contaminant presence at a key site by means of predictions based on neural networks.

There are some interesting environmental applications that are worthwhile mentioning [4-8]. In Helsinki, Artificial Neural Networks (ANNs) were employed to predict particulate matter (PM_{10}) and NO_2 [6]. In great Athens, during the 2004 Olympic Games, models based on ANNs were applied to predict hourly PM_{10} concentrations, proving to be more efficient than the simple regression models [7]. By comparing several models for PM_{10} in Santiago, which is the capital city of Chile, Perez and Reyes [8] also reported the best performance for the ANNs, emphasizing the importance of a careful choice of input variables. Then, on these bases the ANN models are recommended as the most promising candidates for operative use.

In this work we have aimed at developing computational tools for the forecast of contaminants that affect Bahía Blanca city and its surroundings. A neural network was trained using data about the daily averages of PM_{10} and weather variables. The results obtained with this particulate-matter predictor, together with its description, are reported.

This paper presents a PM_{10} predictor and describes in detail the steps for its design. Besides, this is a valuable contribution for the Bahía Blanca community, whose members are particularly sensitive to the industrial impact on the neighborhood.

The data set and the characterization of its variables are detailed in Section 2. Secondly, the modeling approach is outlined in Section 3. Then, some relevant details about the modeling steps together with the key results are considered. Finally, the last section refers to the main conclusions.

2 The Data

The data about PM_{10} and the meteorological variables employed in this work were provided by the CTE, the authorization coming from a collaboration agreement between the Town Hall and the university (UNS-MBB: File 713/82). These data were gathered by the EMCABB (Estación de Monitoreo Continuo del Aire de Bahía Blanca) for the period 2004-2006. This station was settled at Villa Delfina, urban zone located 2 km away from Ing. White Port, which is a sector called Zone B by Carignano et al. [10].

The weather variables that could be chosen as input variables were the following: average temperature (T), average pressure (P), average humidity (H), prevailing wind direction (WD), and average dominating wind speed (WS). The daily averages for T, P and H were calculated based on hourly averages. In particular, the prevailing WD was determined as the daily direction where the wind blew more frequently from. In turn, the WS was calculated as an average restricted to the most frequent direction. Then, WS was based on the average wind speed measured for each hour, taking into account the prevailing wind directions for the period.

Some seasonal indicators- i.e. the month of the year and the day of the week- were also included in order to capture the influence of seasonal changes and vehicular traffic on PM_{10}. These variables were discretely represented through integer numbers.

On Table 1 the climatic variables used in this work are shown with their maximum and minimum values in the data set captured during the period between 2004 and 2006. In the case of prevailing wind direction, Table 1 shows the most and the least frequent wind direction.

Table 1. Weather variables for a three-year period, ranging from 2004 to 2006.

Daily weather variables	Maximum value	Minimum value
Prevailing wind direction	North-northwest	calm
Average dominating wind speed (km/hour)	36.0	1.5
Average temperature (°C)	32.0	1.5
Average pressure (hPa)	1041.8	998.7
Average humidity (%)	96.0	28.8

Another relevant aspect to be taken into account is the climatic description of the region under study. Bahía Blanca is located in an area where climate behaviour lies near the warm sub-humid (600 mm of yearly rainfall to the northeast) zone and also close to the semiarid area (400 mm of yearly rainfall to the southeast). Its wetter periods take place at the end of spring and the beginning of summertime. Its winds blow more frequently from the northwest [11].

3 Modelling Approach

Artificial Neural Networks are typically used when a large number of observations are available and a nonlinear relationship is expected, or when the problem is not understood well enough to apply other methods [12]. In this paper the neural-network approach was adopted as the forecast method since the ANN models have frequently been shown to possess better predictive characteristics, compared to the models that employ standard multilinear regressions, and their flexibility also exhibits a better ability to represent predictive models [12].

ANNs belong to the research area called artificial intelligence. A type of feed-forward neural network called the multilayer perceptron was adopted in this work. In particular the multilayer perceptron has been applied to atmospheric science in prediction, function approximation and pattern classification [13]. It approximates highly non-linear functions with no prior knowledge about the nature of the relationships among inputs and outputs.

Gardner and Dorling [14] made a detailed review of the application of multilayer perceptrons in atmospheric science and concluded that the Multilayer Perceptron (MLP) offers an attractive alternative to the development of numerical models, even among statistical approaches. Nevertheless, various architectures- including Radial Basis Network and MLP- were evaluated and only the best one was reported below.

The Development Process Model (DPM) that allowed this predictor's construction focuses on the development of multilayer feed-forward ANNs. Accordingly, the DPM proposed in this paper is the result of an improvement from the process described in Chiarvetto et al. [15], where too many stages of development had been considered. The DPM that allowed our ANN creation can be applied in order to develop novel ANN models over new time conditions (other time series), a different space (other geographic sites), and over even new weather variables. Our process consists of the following six stages: I. Choice of the input variables; II. Data normalization and selection of the activation function; III. Architecture selection; IV.Selection of the learning algorithm; V.ANN training, testing and validation; VI. Scientific assessment

The aims of stages I to IV are to define ANN design aspects in particular, making use of other authors' previous experience. Different sets of heuristics and techniques that have been productive in the ANN development for the air-quality prediction were found in the literature [15, 16]. For each of these heuristics, a prototype was built to find the most effective archetype in our context. The construction of the prototype for stage *i*+1 is based on the results obtained in phase *i*. Then, at stage V the ANN was built, while at stage VI it was evaluated from the scientific viewpoint.

4 Results and Discussions

4.1 Choice of input variables

The selection of input variables is limited to measurement availability. The data on hand proved to be enough to model PM_{10} behaviour. The input variables are divided in two groups: seasonal and weather variables. In Section 2 and Table 1 these variables are reported in detail. Besides, the day of the week was added. It was considered as the most useful variable for the prediction of hourly PM_{10} in Greater Athens [7]. Although in this paper there are no predictions for hourly intervals, the result obtained by Grivas and Chaloulakou [7] was considered as a heuristics for the selection of variables because it is related to the vehicular traffic.

4.2 Data normalization and selection of the activation function

In order to be able to put together an assortment of variables of different nature, normalization becomes necessary. In this work the normalization function (Eq. 1), which was also used by Gardner and Dorling [4], was employed.

$$N(x) = 2 * \left(\frac{(x - x_{\min})}{(x_{\max} - x_{\min})} \right) - 1.0 \tag{1}$$

where $x_{\min}$ is the lowest value of the variable, and $x_{\max}$ is the highest value of the variable.

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

As to the activation function, the following common alternatives were considered: 1. Hyperbolic tangent (Eq. 2) 2. Sigmoid function (Eq. 3) The hyperbolic tangent was employed by Shiva Nagendra and Khare [5] and Gardner and Dorling [4], while Grivas and Chaloulakou [7] made use of both. Besides, it is also uncertain whether to take an angular or scalar measurement of the wind direction. Both alternatives were evaluated. For the angles, the trigonometric function cos θ was employed, like Grivas and Chaloulakou [7] and Shiva Nagendra and Khare [5] did. For the scalars, both the 16-point compass rose and the calm were considered. The combinations of these possibilities were evaluated on a multilayer net with 10 units in the hidden layer. It was trained during 2000 epochs with the back-propagation algorithm (learning rate= 0.2 and momentum= 0.3).

For the evaluation measurement, the descriptive index of agreement called ***d*** was employed. This measure has been defined by Willmott [17], and it was used by other authors [5,7,18,19]. The index of agreement was calculated through Equation (4):

$$d = 1 - \left[\frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} \left(|P_i'| + |O_i'| \right)^2} \right] \tag{4}$$

where *N* is the number of cases, $P_i' = P_i - \overline{O}$ and $O_i' = O_i - \overline{O}$, *P* is the model-predicted variable, *O* is the observed variable, and $\overline{O}$ is the mean of *O*.

The index of agreement may take values in the interval [0,1]. ***d***=1 implies a perfect match between the observed and the predicted value, while ***d***=0 represents a total disagreement between them. According to Willmott [17], the index of agreement ***d*** is recommended as the best parameter for the performance evaluation of prediction methods in the atmospheric sciences. Nevertheless, the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) were also calculated to complete the information. The combination of a Hyperbolic Tangent function and a scalar wind direction was finally chosen because it proved to be more accurate.

4.3 Architecture Selection

A multilayer architecture was chosen. The hidden units were free to construct their own representations of the input. The choice of the number of units in the hidden layer (cardinality) seems to be uncertain when a neural network design is carried out. Then, a statistical analysis became necessary. Too many units might introduce unnecessary blocks for a proper modelling. Conversely, few units might guide towards a frail problem-solution capability.

Shiva Nagendra and Khare [5] proposed different heuristics for cardinality selection; these heuristics pointed at a minimum of eight units, and a maximum of fourteen. Therefore, prototypes were built by varying the amount of units in the hidden layer within the interval (8-14). Then, the interval yielded seven alternative prototypes that were individually trained during 2000 epochs by means of the back-propagation algorithm (learning rate= 0.15 and momentum= 0.20) as the learning tool. According to the ***d*** values together with the MAE and the RMSE, the most convenient architecture had thirteen units in its hidden layer.

4.4 Selection of the learning algorithm

The learning process applied for this ANN training is the method known as back-propagation, which has classically been adopted in the first place by several scientists [20]. In our approach the following kinds of back-propagation algorithms [21] were employed: basic on-line and resilient.

The tests were carried out with 11 hidden units during 5000 epochs. Unlike the resilient algorithm, the basic on-line one employs two parameters: the momentum and the learning rate, which were also evaluated. The performance was assessed by means of the index of agreement ***d***, together with the MAE and the RMSE. The basic resilient back-propagation algorithm was chosen because it seems to be the most effective in terms of its prediction capability.

4.5 ANN training, testing and validation

The complete data set was randomly divided in three subgroups: training set (70%), testing set (15%) and validation set (15%). For the training, cross validation and early stopping techniques were combined in order to avoid overfitting.

The ANN was trained during 140 000 epochs; the results for the index of agreement ***d***, the MAE and the RMSE are shown in Table 2 where the global values were calculated by using the complete data set. All values exhibited good performance for each stage. In comparison with the ***d*** values reported in the literature for environmental contaminants [5,7,16], the model performance was satisfactory.

Table 2. Performance analysis for each stage of ANN building.

Stages	***d***	**MAE**	**RMSE**
Training	0,840	18.177	26.736
Testing	0,785	21.586	28.943
Validation	0,778	21.608	23.820
Global	**0.899**	**19.206**	**26.662**

In Figure 1 the predicted PM_{10} amount can be compared with the observed data as a function of time. Some lower values could not be predicted accurately by the ANN model. In contrast, the prediction worked well for the medium and higher values. In accordance with our target, these results are satisfactory because the model tends to predict accurately the maximum values.

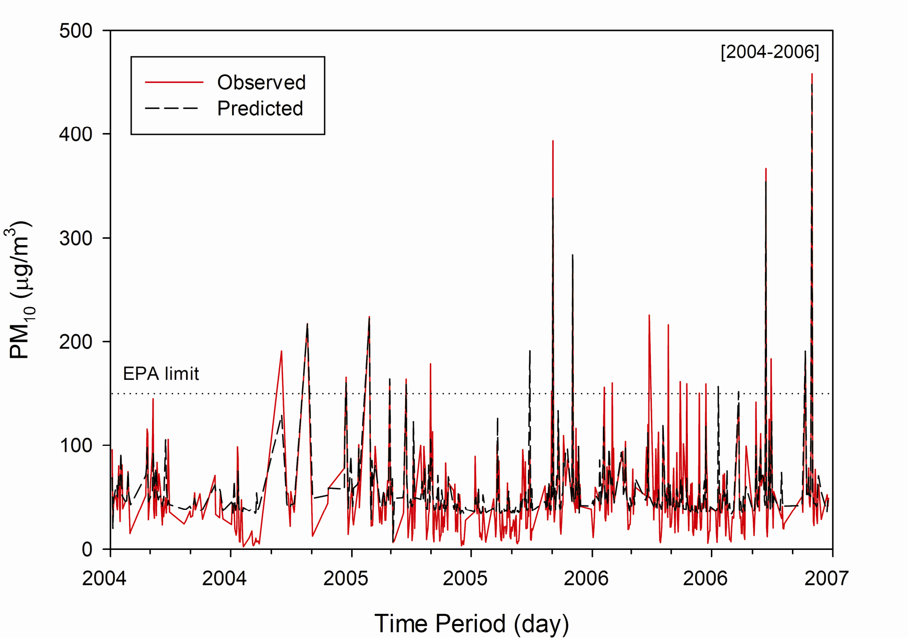

Fig. 1. Predicted and observed PM_{10} data as a function of time between 2004 and 2006.

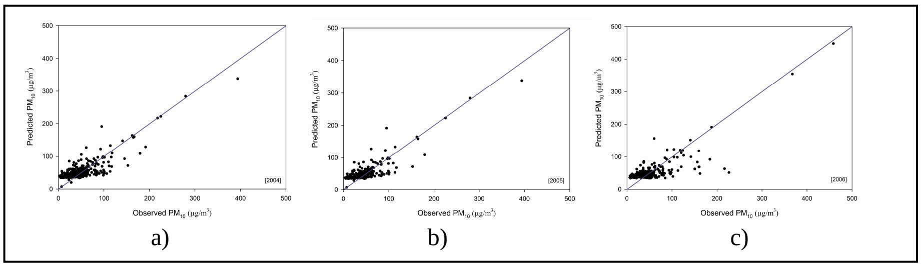

Fig. 2. Predicted vs observed PM_{10} data for: a) 2004; b) 2005; c) 2006. Ideal fit: solid line.

It is illustrative to make a detailed analysis as follows. It should be taken into account that according to EPA norms, the upper limit for PM_{10} is 150 µg/m^3 per day. It was considered the maximum amount allowable (see Fig. 1). Since our goal is to forecast when the concentrations will be far past the limit of what is acceptable, we can consider a wrong prediction whenever a value that is higher than 150 µg/m^3 is predicted, while the real value is lower. Then, this case will be called false-positive. Conversely, the case will be called false-negative whenever a value that is lower than the upper limit is estimated, while the real one overcomes this roof. For the designed ANN, 15 false-positive cases (2.7% of the occasions) were detected, while 44 false-negative cases (7.94% of the occasions) aroused. Then, the prediction was correct for 89.35% of the occasions. We have considered a forecast correct when the predicted value accurately indicates whether the standard upper limit has been overcome. In short, when this model is implemented in the local observatory, contamination warnings will be trustworthy.

4.6 Scientific Assessment

Apart from the statistical measures, a scientific evaluation was regarded as imperative since the eighties [22]. In this work the assessment involved both testing and interpreting model variations with input parameters. For this purpose, we adopted a tailor-made feature selection method that works as follows: an *i*-eth input variable was removed, the network was trained without it and the resulting model was appraised by evaluating the index of agreement $\boldsymbol{d}_\mathrm{i}$. The indices were compared with the reference value obtained globally for the complete model ($\boldsymbol{d}$=0.899, see Table 2) so as to rank all the input variables in order of importance. When an index $\boldsymbol{d}_\mathrm{i}$ denounced a significant decrease, it could be concluded that the presence of the associated *i*-eth input variable was mandatory for the model. When $\boldsymbol{d}_\mathrm{i}$ increases or remains similar to $\boldsymbol{d}$, the *i*-eth input variable should be taken out. If the value is higher, the variable is affecting the model negatively. In turn, when it is similar, the variable seems irrelevant. The results are shown on Table 3, where the rows have been hierarchically arranged.

Table 3. Agreement indexes, MAEs and RMSEs for input variables after scientific assessment.

Variables	**Global dataset**		
	d	**MAE**	**RMSE**
Humidity	0.315	57.663	115.295
Month	0.409	57.596	95.880
Temperature	0.467	0.46.188	77.465
Pressure	0.528	42.885	67.664
Day of the week	0.623	36.730	53.966
Wind Speed	0.670	28.272	45.683
Wind direction	0.782	23.388	33.562

The ranking was made by means of the global ***d***. It is evident that all the input variables play an important role in the model since none of them neither reach nor overcome ***d***=0.899. The humidity, the month and the temperature are the most relevant ones because their absence in the model affects ***d*** remarkably. On the other hand, the least significant ones are the wind direction and the wind speed, followed by the day of the week and the pressure.

5 Conclusions

Due to their usefulness and versatility, which were ratified through this application, ANNs should remain an active field of chemical engineering research. Vital guidelines on ANN-based modelling are provided by describing the design steps in detail. The proposed DPM can be applied in order to develop novel ANN models over new time conditions (other time series) and space (different geographic sites), even over new weather variables. The daily PM_{10} concentrations, together with the corresponding meteorological variables, which were collected for the years from 2004 to 2006, were used for model training, testing and evaluation. Besides, the implementation was also judged scientifically by means of a feature selection process in order to determine the relevance of the input variables in the ANN model. All the chosen input variables proved to be necessary, being the most relevant ones the humidity, temperature and month and the least relevant ones the wind speed and direction. The performance of each stage was evaluated by means of an agreement index. The resulting values were satisfactory and comparable with literature reports.

The data taken to model the impact of PM_{10} on the environment were enough to predict the phenomena accurately. This estimator will contribute to the local observatory in order to trigger alert warnings related to health risk. It is interesting to remark that although the methodology developed in this study was applied to the surroundings of Bahía Blanca city, its petrochemical pole and Ing. White grain port, it can be generalized for other locations or other contaminants.

6 Acknowledgements

We would like to thank Ing. Fernando Rey Saravia and Lic. Marcelo Pereyra for their assistance and data provision during our research work. FRS is the coordinator

of the CTE-MBB (Comité Técnico Ejecutivo, Municipalidad de Bahía Blanca, Argentina); MP is the subcoordinator and chief monitor of the CTE-MBB.

7 References

[1] F.A. Rey Saravia, E. Puliafitto, N.B. Brignole, Mounting an Environmental Observatory in an Urban-Industrial Area, XXII-CIIQ-2006: XXII Congreso Interamericano de Ingeniería Química y V Congreso Argentino de Ingeniería Química, Buenos Aires, Argentina, 2006.

[2] L.L. Chiarvetto Peralta, F.A. Rey Saravia, N.B. Brignole, Diseño de Sistema de un Almacén de Datos para la Gestión Ambiental, 9° Congreso Interamericano De Computación Aplicada a la Industria de Procesos CAIP 2009, Montevideo, Uruguay, 2009.

[3] R. Kimball, M. Ross, The Data Warehouse Toolkit, 2nd Ed., John Wiley, New York, 2002.

[4] M.W. Gardner, S.R. Dorling, Neuronal Network Modelling and Prediction of Hourly NOx and NO_2 Concentrations in Urban Air in London, Atm. Environment 33 (1999) 709-719.

[5] S.M. Shiva Nagendra, M. Khare, ANN Approach for Modelling Nitrogen Dioxide Dispersion from Vehicular Exhaust Emissions, Ecol Mod 190 (2006) 99-115.

[6] J. Kukkonen, L. Partanen, A. Karppinen, J. Ruuskanen, H. Junninen, M. Kolehmainen, H. Niska, S. Dorling, T. Chatterton, R. Foxall, G. Cawley, Extensive Evaluation of Neuronal Networks Models for the Prediction of NO_2 and PM_{10} Conc, Compared with a Deterministic Modelling System and Measurements in Central Helsinki, Atm Env 37 (2003) 4539-4550.

[7] G. Grivas, A. Chaloulakou, Artificial Neuronal Network Model for Prediction of PM_{10} Hourly Concentrations, in the Greater Area of Athens, Greece, Atmospheric Environment 40 (2006) 1216-1229.

[8] P. Perez, J. Reyes, An Integrated Neural Network Model for PM_{10} Forecasting, Atmospheric Environment 40 (2006) 2845–2851.

[9] EPA., 2010, U.S. Environmental Protection Agency. "PM Standards" Last updated on Sept 30, 2010, from http://www.epa.gov/PM/standards.html.

[10] C.O. Carignano, L. Elosegui, M.P. Abrego, S. Spagnolo, M.E. Esandi, R. Frapichini, O.E. Reissing, Asthma and Indicators Symptoms Prevalence in Three Urban Areas in a Multiple Purpose Survey Archivos De Alergia E Inmunología Clínica 34 4, (2003) 119-128.

[11] Atlas Total de la República Argentina, La Bahía Blanca, Area de Contacto entre Ambientes Diferenciados, Centro Editor de América Latina, 7, 1981.

[12] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Dennis Hall, M. Karelson, I. Kahn, D.A. Dobchev, Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction, Chem. Rev. 110 (2010) 5714–5789.

[13] S. Lek, J.F. Guégan, Artificial Neuronal Networks as a Tool in Ecological Modelling: an Introduction, Ecological Modelling 120 (1999) 65-73.

[14] M.W. Gardner, S.R. Dorling, ANNs (The Multilayer Perceptron)—A Review of Applications in the Atmospheric Sciences, Atmospheric Environment 32 (1998) 2627- 2636.

[15] L.L. Chiarvetto Peralta, F.A. Rey Saravia, N.B. Brignole, Aplicación de RNAs para la Predicción de Calidad de Aire, Mec Comput. XXVII (2008) 3607-3625.

[16] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 2nd edition, Prentice Hall, New York, 2002.

[17] C.J. Willmott, Some Comments on the Evaluation of Model Performance, Bulletin American Meteorological Society 63 (1982) 1309–1313.

[18] G. Corani, Air Quality Prediction in Milan: Feed-Forward Neuronal Networks, Pruned Neuronal Networks and Lazy Learning, Ecological Modelling 185 (2005) 513-529.

[19] S.A. Abdul-Wahab, S.M. Al-Alawi, Assessment and Prediction of Tropospheric Ozone Concentration Levels using ANNs, Env. Modelling & Software 17 (2002) 219-228.

[20] A.J. Soto, "Técnicas de Aprendizaje Automático y Computación Científica Aplicadas a la Predicción de Parámetros ADME-Tox". Doctoral Thesis, UNS, B. Blanca, Argentina, 2010.

[21] P. Marrone, "Joone (Java Object Oriented Neural Engine) Complete Guide", 2007, from http://sourceforge.net/projects/joone/files/Documentation/engine/JooneCompleteGuide.pdf.

[22] D.G. Fox, Judging Air Quality Model Performance, Bulletin American Meteorological Society 62 (1980) 599–561.