Asian Journal of Information Technology 6 (11): 1148-1153, 2007
ISSN: 1682-3915

Prediction of SO_2 Ground Level Concentrations by Means of RBF Neural Networks

[1]A. Boumerah, [2]A. Abidi, [3]S. Chenikher, [2]T. Bouchami and [3]M. Ramdani
[1]Department of Chemistry,
[2]Department of Process Engineering,
[3]Department of Electronics, Badji-Mokhtar University, Algeria
BP. 12, 23000, Annaba, Algeria

Abstract: In this study, a nonlinear model for forecasting the SO_2 ground level concentrations is build by using a Radial Basis Function Network (RBFN) based on hybrid learning algorithm. Ground level concentrations of pollutants were analysed in the area under study, in particular the high levels of SO_2 occuring during relatively rare episodes. These events are influenced by many factors, such as local meteorology aspects, topography and industrial emissions. The model structure is identified by using a fuzzy C-means clustering algorithm. The proposed RBFN is trained by hybrid learning algorithm to obtain the centre and width of each radial basis function and the least squares method to obtain the output weights. An improved learning scheme is used to avoid the local minima. The developed model concerns an urban area in the Annaba City (North-East Algeria), but it can be adapted to other locations.

Key words: Air pollution, ground level SO_2, multiple regression, analysis, RBF network, hybrid learning

INTRODUCTION

Air pollution is one of the most serious problems confronting our modern word. Air pollution depends on the quantity and quality of fuels used, on technology used by industrial and transportation units, on the high concentration of population and factories and on the prevalent meteorological conditions. The main source of SO_2 emissions is the combustion of fossil fuels and it is among the most prevalent air pollutants in many industrialized areas. In particular, the combustion of fuels for power generation is believed to be responsible for most of the SO_2 to which the population is exposed. Exposure of humans to high levels of SO_2 has been related to increase in hospital admission for chronic bronchitis and to low birth weights. The World Health Organization has determined that the safety limit for SO_2 concentrations is $\mu g\ m^{-3}$ for 24 h averages (ECH, 1979).

In recent environmental studies, it has been reported that the level of sulphur emissions, mainly as SO_2, have over the last two decades been reduced in western and northern Europe (Holland *et al.*, 1999). However, localised SO_2 pollution still exist related to local emission, meteorological and topographical factors. By contrast, sulphur emissions are increasing in many emerging industrialised and developing countries around the world. Hence, environmental problems associated with sulphur emissions are still far from being fully solved. It seems then very useful to have at hand reliable methods to forecast sulphur dioxide concentrations several hours in advance, in order to control the phenomenon, to diagnose the sensors of air quality network, or even more simply to improve the knowledge about the pollution phenomenon.

Existing models for SO_2 forecasting are deterministic and empirical type. A difficulty encountered with deterministic models is that emissions from natural and anthropogenic sources, such as industry and traffic, are often uncertain and sometimes unavailable. By contrast, empirical models which include statistical techniques and neural networks have some advantages over the deterministic ones. Firstly, they do not need data about emissions since they are based on the use of air quality and measurements only. Secondly, the structure of empirical models is often simpler than deterministic models and they can more easily be implemented and used by non-experts, although an obvious drawback is that they are not portable from site to site since they are developed and calibrated on local data (Dorling *et al.*, 2003; Gardner and Dorling, 1998, 2000) .

Concerning empirical approaches, linear, nonlinear, neural and fuzzy models are proposed to predict pollution levels. The results obtained by linear models were with no

Corresponding Author: S. Chenikher, Department of Electronics, Badji-Mokhtar University, Algeria BP. 12, 23000, Annaba, Algeria

doubt encouraging, but due to the complexity of the underlying environmental processes and the nonlinear interactions between meteorological variables and pollution, the development of nonlinear models, such as artificial neural networks, is currently being applied. The nonlinear models have been considered for air pollution time series modelling by several authors such as (Dorling *et al*., 2003; Gardner and Dorling, 2000, 1998). The use of neural networks has been reported by Boznar *et al*. (1993), Dorling *et al*. (2003), Nunnari *et al*. (1998) among others. Comparisons between neural networks and linear models were also done in Chaloulakou *et al*. (2003) and the results demonstrated the improvements of neural networks over linear models. Fuzzy techniques were used by Mintz *et al*. (2005) with good performances.

In this study, we report a study on the possibility to forecast hourly averages of SO_2 concentrations based on data obtained in a station located at fixed point in the city of Annaba (North-East of Algeria) where the expected emissions in its neighbourhood come from industries, heating and vehicle traffic. The selected station is one among the several stations that form the air quality monitoring network.

MATERIALS AND METHODS

Data description: In the city of Annaba (North-East of Algeria), the pollution control is of great importance, as it affects the life quality of about one million inhabitants as well as the ecosystem. The high rate of emissions and consequently, of contaminants is particularly notorious in the area of Annaba City. The inhabitants are thus critically exposed to those contaminants as it is the case of carbon monoxide, nitrogen oxides, ozone, sulphur dioxide (SO_2) and particulate matter. With about one million inhabitants and relatively high population density, Annaba is limited on the North by the Mediterranean Sea, with about 80 km of coastline and intersected by the Seybouse river. The average temperatures in warmer and cold periods are 22,8 and 12,1°C, respectively. The annual air humidity is 75% and the total annual mean precipitation varies between 675 and 700 mm. Prevailing winds are from North-West in summer and North-South in winter. The main pollution sources are one petrochemical plant, one steel plant, one thermoelectric power plant working with natural gas.

The monitoring facilities provided hourly concentrations of pollutants such nitrogen monoxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO) and particulate matter with an equivalent aerodynamic diameter smaller than 10 mm (PM_{10}), sulphur dioxide (SO_2) as well as meteorological variables such as Temperature (T), Wind Velocity (WV) and Relative Humidity (RH). Since the SO_2 dynamics is strictly related to the dynamics of other pollutants with which it is likely to combine and recombine, only one station has been considered because it provides the richest number of meteorological and pollutant information, together with, of course, the measures of SO_2. We consider data corresponding to hourly averages for the period that goes from 30/08/2004 to 20/09/2004.

Models: Let us indicate by y (t) a pollutant time series and by u_i (t), (i = 1,..., q) to the time series correlated to y (t) (e.g., meteorological variables as well as other pollutant concentrations). A d-step ahead prediction model for y (t) can be represented in Nonlinear AutoRegressive with eXogeneous inputs (NARX) form as follows:

$$\hat{y}(t) = F(x(t)) + \varepsilon(t) \quad (1)$$

F being an unknown nonlinear function, n_y, n_i integer numbers related to the model order, x (t) is a regression vector expressed as follows:

$$\begin{aligned} x(t) = [&y(t-d),.., y(t-d-n_y),\\ &u_1(t-d),.., u_1(t-d-n_i),\\ &u_q(t-d),.., u_q(t-d-n_q)] \in R^n \end{aligned} \quad (2)$$

The variables u_i (t) in expression (2) are usually referred to as the exogeneous model inputs while y (t) is the model output. When F is linear in its arguments, the NARX model becomes the well-known ARX model. In this study, the hourly SO_2 concentrations are predicted using a neural network NARX model and a linear ARX model. The predictors are Particulate Solids concentrations (PS) and meteorological parameters (Wind Velocity (WV), Temperature (T), Relative Humidity (RH)).

Multiple linear regression: Multiple Regression Analysis (MRA) has been widely used to model the cause-effect relationship between inputs and outputs and can generally expressed as:

$$y = f(x_1,..., x_n; \theta_1,..., \theta_p) + \varepsilon \quad (3)$$

Where, y is a dependent variable (i.e, output variable), $x_1,..., x_n$ are independent or explanatory variables (i.e., input variables), $\theta_1,..., \theta_p$ are regression parameters, ε is a random error, which is assumed to be normally distributed with zero mean and constant variance σ^2 and f is a known function, which may be linear or nonlinear. If f is linear, then (3) becomes a Multiple Linear Regression (MLR) and can be expressed as:

$$y = b_0 + b_1 x_1 + b_n x_n \quad (4)$$

The learning problem is a straightforward application of linear regression techniques to find parameters $\theta = [b_0\ b_1 \ldots b_n]^T$ which best fit the data.

RBF neural network: A RBF consists of an input layer, a nonlinear hidden layer and a linear output layer. The nodes of each layer are fully connected to the previous layer nodes. The input variables are each assigned to nodes in the input layer and connected directly to the hidden layer without weights. The hidden layer nodes are RBF units. The nodes calculate the Euclidean distances between the centres and the network input vector and pass the results through a nonlinear function. The output layer nodes are weighted linear combination of the RBF in hidden layer (Haykin, 1989; Moody and Darken, 1989). The structure of a RBF neural network with n inputs, C hidden nodes and m output nodes is given in Fig. 1.

Where, input $x = [x_1, x_2, \ldots, x_n]^T$ and $w\ [w_1, w_2, \ldots, w_c]^T$ is the neural network weight. u_i is a nonlinear function and here is chosen as a Gaussian activation function:

$$u_i = \exp\left[-\frac{(x - v_i)^T (x - v_i)}{2b_i^2}\right],\ i = 1,2,\ldots,c \quad (5)$$

Where $v_i = [v_{i1}, v_{i2}, \ldots, v_{in}]^T$ is the centre of the ith RBF hidden unit and b_i is the width of the *i*th RBF hidden unit. Then the jth RBF network output can be represented as a linearly weighted sum of c basis functions:

$$\hat{y}(k) = \sum_{i=1}^{c} w_i u_i \quad (6)$$

Let y (k) represent the target vector of the network at time k. The error of the network at time k is defined as:

$$e(k) = \hat{y}(k) - y(k) \quad (7)$$

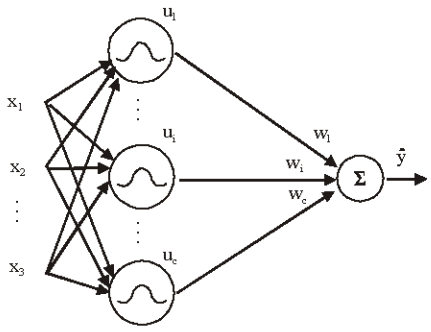

Fig. 1: The structure of RBF neural networks

The cost function of the network is the squared error between the target and the predicted values, which is given by the following equation:

$$J(k) = \frac{1}{2} e(k)^2 \quad (8)$$

The learning algorithm aims to minimize the squared error using a gradient descent or a hybrid learning algorithm. The later category is based on the idea of sequentially using nonlinear and linear optimisation techniques to train the nonlinear and linear parameters. Here, we chose to apply the least squares method for the linear weights parameters and conjugate-gradient iterative optimisation algorithm for nonlinear ones, i.e., centres and widths of radial basis functions.

Determination of initial centres and widths: When designing an RBF network, the most critical task is certainly the determination of the parameters of the hidden layer. Hence, in this research, the fuzzy c-means algorithm is applied to determine the initial parameters. Given a dataset of input-output pairs, the design matrix Z is formed by concatenating the regression data matrix X and the output vector y:

$$X = \begin{bmatrix} x_1^T \\ . \\ . \\ . \\ x_N^T \end{bmatrix},\ y = \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{bmatrix},\ Z = [Xy]$$

Each observation thus is an (n + 1) dimensional column vector:

$$z_k = [x_k\ y_k]^T = [x_{1,k}, x_{2,k}, \ldots, x_{n,k}, y_k]^T$$

The fuzzy partitioning space for Z has to satisfy the following conditions:

$$\begin{aligned} &U \in R^{c \times N} \ \ |\mu_{i,k} \in [0,1],\ \forall i,k; \\ &\sum_{i=1}^{c} \mu_{i,k} = 1,\ \forall k;\ \ 0 < \sum_{k=1}^{N} \mu_{i,k} < N,\ \forall i \end{aligned} \quad (9)$$

The clustering of the data is based on the minimization of of the c-means functional defined as follows:

$$J(Z,U,V) = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{i,k}^m D_{i,k}^2 \quad (10)$$

where, U contains the cluster and m > 1 is a weighting exponent that determines the fuzziness of the resulting partition and it is often chosen as m = 2. The minimization of J, which represents a nonlinear optimization problem that can be solved by a simple Picard iteration through the first-order conditions for stationary points of (the objective function, where U contains the cluster and m>1 is a weighting exponent that determines the fuzziness of the resulting partition and it is often chosen as m = 2. The minimization of J, which represents a nonlinear optimization problem that can be solved by a simple Picard iteration through the first-order conditions for stationary points of (the objective function, known as the Fuzzy C-Means (FCM) algorithm.

The stationary points of the objective function can be found by adjoining the constraint $\sum_{i=1}^{c} \mu_{i,k} = 1, 1 \le k \le N$ to J by means of Lagrange multipliers:

$$\bar{J}(Z,U,V,?) = \sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{i,k}^{m}D_{i,k}^{2} + \sum_{k=1}^{N}?_{k}\left[\sum_{i=1}^{c}\mu_{i,k} - 1\right] \quad (11)$$

and by setting the gradients of J with respect to U,V and λ to zero. It can be proven that $D_{i,k}^{2} > 0, \forall i,k$ if and m > 1, then (11) is minimized only if:

$$\mu_{i,k} = \frac{1}{\sum_{j=1}^{c}\left(D_{i,k}/D_{j,k}\right)^{2/(m-1)}}, \quad 1 \le i \le c; 1 \le k \le N; \quad (12)$$

$$v_i = \frac{\sum_{k=1}^{N}\left(\mu_{i,k}\right)^{m} z_k}{\sum_{k=1}^{N}\left(\mu_{i,k}\right)^{m}}; 1 \le i \le c \quad (13)$$

The fuzzy c-means scheme is summarized as follows: Given the data set X and the number of clusters 1<c<N, the cluster centres are chosen randomly from X.
Iterate for t = 1, 2, . . .,
Step 1: Compute the cluster means:

$$v_i^{(t)} = \frac{\sum_{k=1}^{N}\left(\mu_{i,k}^{(t-1)}\right)^{m} z_k}{\sum_{k=1}^{N}\left(\mu_{i,k}^{(t-1)}\right)^{m}}; 1 \le i \le c \quad (14)$$

It is worth noticing that $v_i = \left[v_i^x\ v_i^y\right]^T$ because the clustering is done in the input/output space.
Step 2: Compute the distance measure:

$$D_{i,k}^{2}\left(z_k, v_i^{(t)}\right) = \left(z_k - v_i^{(t)}\right)F_i^{-1}\left(z_k - v_i^{(t)}\right)^{T} \quad (15)$$

With F_i is the weighted covariance matrix:

$$F_i^{(t)} = \frac{\sum_{k=1}^{N}\left(\mu_{i,k}^{(t)}\right)^{m}\left(z_k - v_i^{(t)}\right)\left(z_k - v_i^{(t)}\right)^{T}}{\sum_{k=1}^{N}\left(\mu_{i,k}^{(t)}\right)^{m}}$$

Step 3: Update the partition matrix:

$$\mu_{i,k}^{(t)} = \frac{1}{\sum_{j=1}^{c}\left(D_{i,k}\left(z_k, v_i^{(t)}\right)/D_{j,k}\left(z_k, v_i^{(t)}\right)\right)^{2/(m-1)}} \quad (16)$$

$$\text{Until } \left\|U^{(t)} - U^{(t-1)}\right\| < e$$

RESULTS

The quantities that have been taken into account for the pollution prediction are hourly sampled measurements values of SO_2 and several pollutants, as well as meteorological variables (Table 1). The aim is to perform the pollution prediction and therefore to build-up a d-step ahead forecasting model (d = 6 h) for the hourly SO_2 concentrations. The data set, which we used to build the database for the neural network, is constituted by the hourly values related to the period that goes from 30/08/2004 to 20/09/2004. Time series of 15 days (360 points) was used to build the training data set, while the remaining the data of the rest 7 days were used to build the testing data set (168 points). The data were pre-processed in order to eliminate instrument errors, replacing the missing data with the linear interpolative function. In addition, each value in the neural network was normalized in the range [-1,1] using the following linear transformation:

$$x' = \left(x - x_m\right)/\left(x_{max} - x_{min}\right) \quad (17)$$

Where, x' is the new normalized value, x is the old value, x_{max}, x_{min} and x_m are the maximum, minimum and mean values, respectively. The set of normalized values was used as neural network input.

Table 1: Characteristic variables of the phenomena

Input variable	Output variable
Particulate matter PM_{10} μg m^{-3}	SO_2 μg m^{-3}
Temperature T, °C	
Wind velocity WV m sec^{-1}	
Relative humidity RH, %	

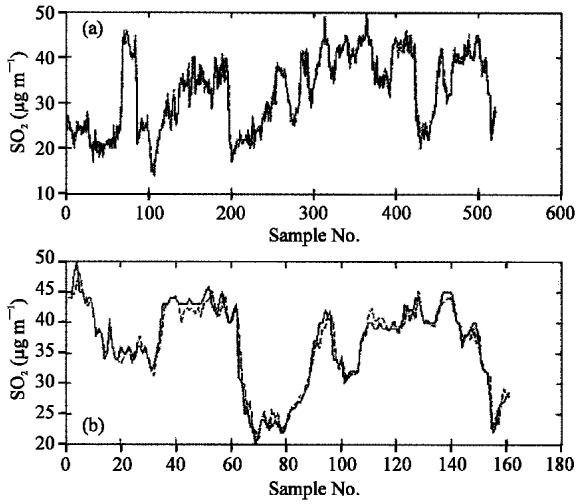

Fig. 2: SO_2 concentrations forecasting using the RBF network with the measured data (solid line) and forecasted (dashed line) for the training data set (a) and testing data set (b)

To evaluate the validity of the suggested modelling strategy in the prediction of pollution, the traditional Multiple Linear Regression (MLR) has been applied to the same task. The prediction results are evaluated by using as performance index the root mean square error, which is defined as:

$$RMSE = \sqrt{\frac{1}{N}\left(\hat{y}(t) - y(t)\right)^2} \quad (18)$$

with t is the sample number, $\hat{y}$ (t) the predicted value and y (t) the measured value. The missing data are reconstructed by using the linear interpolation function. The MLR model is given by:

$$\begin{aligned} SO_2(t) = {} & -0.3908 + 0.9649PM_{10}(t-d) \\ & + 0.0012T(t-d) + 0.0282RH(t-d) \\ & + 0.0037WV(t-d) + 0.5250SO_2(t-d) \end{aligned} \quad (19)$$

and the selected RBF network has 6 inputs, 12 neurons in the hidden layer and one output neuron. The fuzzy c-means algorithm was applied with c = 12. The learning process based on the hybrid learning algorithm has taken 100 epochs.

As shown in Table 2, the RBF network presents the most accurate prediction capability. This indicates that the linear modelling is inappropriate to describe the functional relationship between sulphur dioxide concentration and its precursors. Figure 2 indicates clearly that the RBF network is able to predict the SO_2 concentrations with a good accuracy.

Table 2: Comparison of RMSE between two different methods

	RMSE (Training data)	RMSE (Test data)
MLR	2.1448	2.1767
RBF	1.5768	1.6759

CONCLUSION

The aim of this preliminary study is to develop a predictive nonlinear model to forecast sulphur dioxide concentrations several hours in advance, in order to have the opportunity to take emergency actions when conditions that favour high levels are foreseen or to diagnose the operation of the air quality network. The results have shown that the proposed model can produce good performances. The knowledge of actual meteorological conditions improves pollutant forecasting. Nonlinear effects are important when combining pollutant and meteorological information as input. Our future research should address the issue of input variable selection and the use of other learning strategies.

REFERENCES

Boznar, M., M. Lesjak and P. Mlakar, 1993. A neural network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terain. Atmospheric Environ., 27B: 221-230.

Chaloulakou, A., M. Saisana and N. Spyrellis, 2003. Comparative assessment of neural networks and regression models for forecasting summer time ozone in Athens. Sci. Total Environ., 313: 1-13.

Dorling, S.R., R.J. Foxall, P.D. Mandic and G. C. Cawley, 2003. Maximum likelihood cost function for neural networks models of air quality data. Atmospheric Environ., 37: 3435-3443.

Gardner, M.W. and S.R. Dorling, 2000. Statistical surface ozone models: An improved methodology to account for nonlinear behaviour. Atmospheric Environ., 34, 21-34.

Gardner, M.W. and S.R. Dorling, 1998. Artificial neural networks (the multilayer perceptron). A review of applications in the atmospheric sciences. Atmospheric Environ., 32: 2627-2636.

Haykin, S., 1989. Neural networks: A Comprehensive Foundation. 2nd Edn. Englewood Cliffs, NJ: Prentice-Hall.

Holland, D.M., P.P. Principe and J.E. Sickles, 1999. Trends in atmospheric sulfur and nitrogen species in the eastern United States for 1989-1995. Atmospheric Environ., 33: 37-49.

Mintz, R., B.R. Young and W.Y. Svrcek, 2005. Fuzzy logic modelling of surface ozone concentrations. Comput. Chem. Eng., 29: 2049-2050.

Moody, J. and C.J. Darken, 1989. Fast learning in networks of locally tuned processing units. Neural Comput., 1: 281-294.

Nunnari, G., A. Nucifora and C. Randieri, 1998. The application of neural techniques to the modelling of time series of atmospheric pollution data. Ecol. Modelling, pp: 187-205.

Environmental Health Criteria, 8 1979, World Health Organisation, Geneva.