Available online at www.sciencedirect.com
SCIENCE DIRECT®
Environmental Modelling & Software 20 (2005) 547–559
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua)

J.B. Ordieres[a,*], E.P. Vergara[a], R.S. Capuz[b], R.E. Salazar[c]

[a] *Universidad de La Rioja, c/ Luis de Ulloa 20, 26004, Logroño, La Rioja, Spain*
[b] *Universidad Politécnica de Valencia, Camino de Vera, s/n. 46022 Valencia, Spain*
[c] *Instituto Tecnológico de Mexicali, Av. Tecnológico, s/n Col. Elías Calles, 21396 Mexicali B.C., Mexico*

Received 5 June 2003; received in revised form 4 October 2003; accepted 8 March 2004

Abstract

The daily average $PM_{2.5}$ concentration forecast is a leading component nowadays in air quality research, which is necessary to perform in order to assess the impact of air on the health and welfare of every living being. The present work is aimed at analyzing and benchmarking a neural-network approach to the prediction of average $PM_{2.5}$ concentrations. The model thus obtained will be indispensable, as a control tool, for the purpose of preventing dangerous situations that may arise. To this end we have obtained data and measurements based on samples taken during the early hours of the day. Results from three different topologies of neural networks were compared so as to identify their potential uses, or rather, their strengths and weaknesses: Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Square Multilayer Perceptron (SMLP). Moreover, two classical models were built (a persistence model and a linear regression), so as to compare their results with the ones provided by the neural network models. The results clearly demonstrated that the neural approach not only outperformed the classical models but also showed fairly similar values among different topologies. Moreover, a differential behavior in terms of stability and length of the training phase emerged during testing as well. The RBF shows up to be the network with the shortest training times, combined with a greater stability during the prediction stage, thus characterizing this topology as an ideal solution for its use in environmental applications instead of the widely used and less effective MLP.

Keywords: US–Mexico border; Air quality; Particulate matter; $PM_{2.5}$; Neural Network Modeling; Multilayer Perceptron (MLP); Radial Basis Function (RBF); Square Multilayer Perceptron (SMLP)

1. Introduction

In the past few years, the U.S.–Mexican border has been an important environmental concern for both the U.S. and Mexico (Vega et al., 2002). Since 1983, when the La Paz Agreement[1] was signed, both countries combined efforts to improve the environmental conditions in the region.

The U.S.–Mexican border stretches over 100 km on both sides of the boundary line between both countries (Fig. 1); from the Gulf of Mexico to the Pacific Ocean, it is 3100 km long from end to end. The terrain therein includes large deserts, numerous mountain ranges, shared rivers, wet-lands, large estuaries, aquifers, national parks, and protected areas. Its weather dramatically varies from the Pacific Ocean to Arizona-Sonora, where harsh conditions rule day and night. Such differences stem from an extraordinary variety of wildlife which must be preserved.

At present, 11.8 million people live along the border, almost equally divided among the two countries. Projected population growth rates in the border region

* Corresponding author. Universidad de La Rioja, Edificio Departamental, Luis de Ulloa 20, 26004, Logroño, La Rioja, Spain. Fax: +34-941-299-478.
E-mail address: joaquin.ordieres@dim.unirioja.es (J.B. Ordieres).

[1] Acuerdo de Cooperación para la Protección y Mejoramiento del Medio Ambiente en la Región Fronteriza.

1364-8152/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2004.03.010

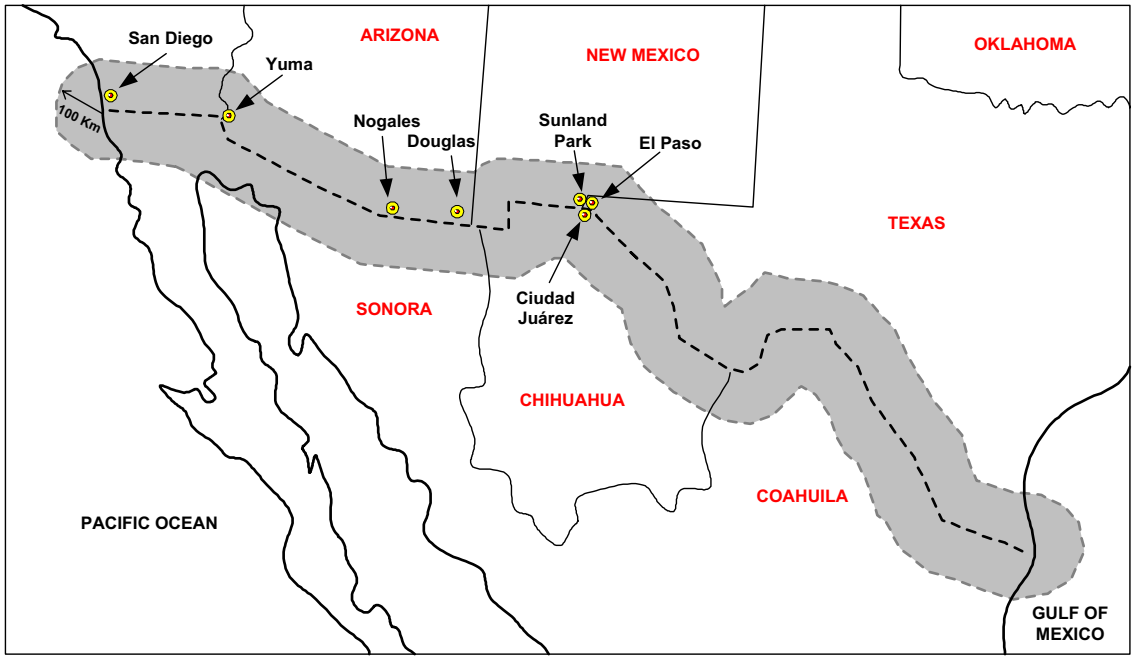

Fig. 1. U.S.–Mexican border region, as defined under the La Paz Agreement.

exceed anticipated national average growth rates for each country. If current growth rates remain consistent with the projected figures, the population in the border area will increase by 7.6 million people by 2020 (US-EPA and SEMARNAT, 2002).

In 1990, 1770 assembly plants were operating in Mexico. By 2001, this figure is projected to double, reaching a number of 3800 factories, 2700 of which will be based in borderline states. Compared with other regions in Mexico, the border area is characterized by very low unemployment rates and high wages. In spite of this economic growth, the region's infrastructure has yet to see much improvement. Thus, natural resources, environment and public health are adversely affected on both sides of the border.

The U.S. Environmental Protection Agency has designated cities including San Diego, CA; Yuma, AZ; Nogales, AZ; Douglas, AZ; Sunland Park, NM; and El Paso, TX, as non-attainment cities because they have failed to meet the National Ambient Air Quality Standards (US-EPA, 1998, 2000b; Mukerjee et al., 2001, Mukerjee, 2001; Watson and Chow, 2001).

The Border 2012 program proposes goals aimed to meet the most difficult challenges concerning air quality. Among these, goal #2 proposes three objectives to reduce air pollution (US-EPA, 2000a; US-EPA and SEMARNAT, 2002):

- By the year 2003, define baseline and alternative scenarios for emission reduction along the border area.
- By the year 2004, based on results from sub-objective #1, define specific emission reduction strategies.
- By 2012 or sooner, reduce air emissions, as much as possible in order to reach national ambient air quality standards.

Our work deals with pollution by particulate matter in the central-south border region of the U.S., particularly in the area of El Paso, Texas, and Ciudad Juárez in Chihuahua (Mexican central-north border). Air pollution in this area has been a major concern for years due to the environmental problems already described (US-EPA, 1996, 1998, 2000b and 2000c; US-EPA and SEMARNAT, 2002). According to the Texas Commission on Environmental Quality (TCEQ), El Paso is affected by the air emissions from Ciudad Juárez (Mexico), and in order to avoid federal penalties, must comply with the Clean Air Standard by 2007. Instead of the PM_{10} classical approach (Chow and Watson, 2001; Fuller et al., 2002), for this particular area the data must be based on $PM_{2.5}$, since natural emissions of minerals (i.e. sand storms) mask in that parameter the anthropogenic sources. Fortunately, the $PM_{2.5}$ measures allow us to identify anthropogenic sources for their later analysis (Magliano et al., 1999).

Several works have tried to isolate and identify the sources and composition of this type of air pollution, i.e. Querol et al., (2001a,b), Lenschow et al., (2001), Rodríguez et al., (2002), Lu (2002); in order to evaluate and avoid its health-related effects (McDonnell et al., 2000; Ostro et al., 1999a,b). Prediction models can be used to this end. Although there are few works on prediction models for $PM_{2.5}$, there have been several attempts to analyze its behavior and evaluate and/or predict the variations in fine particulate concentrations.

Mukerjee et al. (2001) and Kukkonen et al. (2001) developed a model based on the correlation between PM_{xx} and NO_x with a resulting fractional bias (FB) ranging from −0.05 to +0.09 and an index agreement (IA) from 0.85 to 0.96 to predict yearly average PM_{10} concentrations, but unfortunately with an index agreement (IA) ranging from 0.45 to 0.65 for hourly average concentrations. In Kuopio, Finland, Tiittaa et al. (2002) developed a semi-empirical model and applied Multiple Regression Analysis with a 0.67 square correlation coefficient, however, the concentrations predicted were much lower than the actual measurements. Jorquera et al. (2001) used box models to analyze air pollution in Santiago, Chile, and developed a linear model for $PM_{2.5}$, PM_{10} and coarse $PM_{2.5}$–PM_{10} fractions. They estimated a decrease of 50% and 22% for PM_{10} and $PM_{2.5}$, respectively, in the last 10 years.

Nevertheless, in spite of the various types of modeling methodologies available, the trend of using neural networks seems to be growing (Reich et al., 1999; Pérez and Reyes, 2001; Podnar et al., 2002; Chaloulakou et al., 2003), as they rest firmly upon the classical, statistical approaches. Some interesting cases in which these methodologies have been previously used for different purposes, in environmental sciences, and in particular, for atmospherical pollution modeling, are as follows. Pérez et al. (2000) developed a non-linear prediction model for $PM_{2.5}$ using neuronal networks, in Santiago de Chile, but the prediction errors of the model ranged from 30% in the early hours to 60% in the late hours. Later, Pérez and Reyes (2002) developed a model to predict daily average PM_{10} concentrations 30 h in advance, with a prediction error of about 20%. Pérez and Trier (2001) have also succeeded in modeling and forecasting other polluting agents by means of neural networks in the surrounding area of Santiago, Chile. Another approach based on neural networks as well is that of Kolehmainen et al. (2001), in which periodic components were employed in order to predict polluting agent concentrations.

However, it is important to keep on working along this line and to try to find the most accurate prediction model in order to prevent serious environmental damage, and health-related problems in susceptible groups such as children.

In this paper, we introduce three short term $PM_{2.5}$ non-linear prediction models by comparing the accuracy provided by different systems (Ho et al., 2002). The classical ARIMA linear modeling tool becomes unreliable inasmuch as air-pollution sources tend not to behave in a linear fashion. This is the reason why we used artificial neural networks (ANN): to find a prediction model with the lowest possible real prediction error according to the data. The first prediction model was developed using a Multilayer Perceptron neural network (MLP), in the second one, a Square Multilayer Perceptron (SMLP), and finally, a Radial Basis Function network (RBF).

In order to carefully appraise the need for complex models like the neural networks, their outcomes have been compared against the ones pertaining to the traditional models. Thus, a persistence model and a linear regression have been assessed.

These models predict $PM_{2.5}$ behavior using the data of the previous 24 h and the first 8 h of the day to determine $PM_{2.5}$ concentration in the remaining 16 h in the area of Paso del Norte. The demand for a prediction of the average level for the current day in the early morning hours led us to use only the first 8 h of the day, thus making available the estimation at 08:00 am, when typically, the labour day starts.

The tools used in the data pre-processing and conditioning steps were our own Linux-based software tools and R, (Ihaka, 1996), and the tools used in the neural network processing were some Linux-based tools we had built using the excellent NODELIB library (Flake, 1998). (Available from http://www.neci.nec.com/homepages/flake/nodelib/html).

2. Materials and methods

2.1. Data

Geographically, Ciudad Juárez and El Paso are located next to each other and have a common air shed. Air pollution sources on either side of the border have an impact on air quality on both cities. These cities share the Chihuahua Desert (3710 feet above sea level), are shielded by mountains on three sides, and enjoy more than 200 days of sunshine a year. The annual mean temperatures range from 27 °F (−3 °C) in winter to 100 °F (38 °C) in summer. We assumed that these data represent the urban environment on both sides of the border (Ciudad Juárez and El Paso) due to the reasons previously explained.

The data set was obtained from the urban air quality monitoring network of the Texas Commission of Environment Quality (TCEQ). The station used was El Paso UTEP-C12 (EPA number 48-141-0037). The station is located in downtown El Paso (31° 46′ 06″N, 106° 30′ 05″W), at 1158 m above the sea level. This station is constantly monitoring air and provides the latest hourly averaged data available in terms of carbon monoxide, sulfur dioxide, nitric-oxide, nitrogen dioxide, oxides of nitrogen, ozone, wind speed, resulting wind speed, resultant wind direction, maximum wind gust, standard deviation, wind direction, outdoor temperature, dew point temperature, relative humidity, solar radiation, ultraviolet radiation, PM_{10} standard conditions and $PM_{2.5}$ local conditions (Fig. 2).

The data were recorded on an hourly basis (24 h a day) from 2000 to 2002. According to the data, for the year 2000, the average $PM_{2.5}$ concentration was

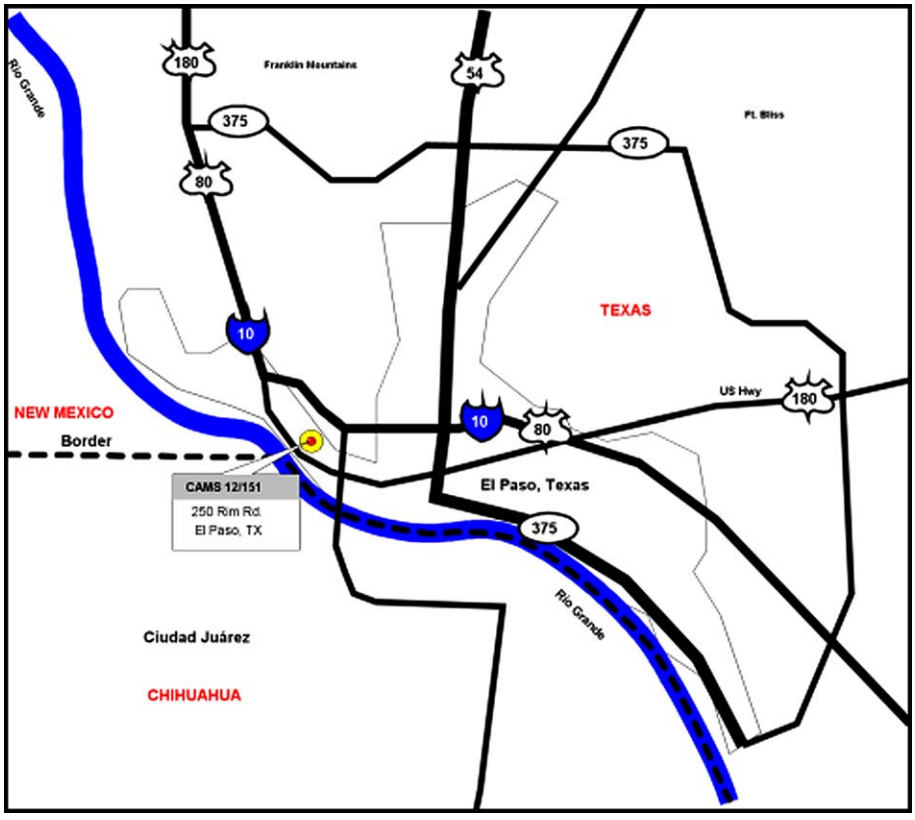

Fig. 2. Urban air quality monitoring network on the US–Mexico border in El Paso (Texas) and Ciudad Juárez, Chihuahua (Mexico).

8.27 μg/m^3, with a standard deviation of 4.71 μg/m^3, reaching a maximum value of 26.74 μg/m^3. For the year 2001, the average was 8.87 μg/m^3, with a standard deviation of 7.54 μg/m^3, and a maximum value of 87.44 μg/m^3. In turn, the average $PM_{2.5}$ concentration for the year 2002 was 9.61 μg/m^3, with a standard deviation of 7.42 μg/m^3, and a maximum value of 69.21 μg/m^3.

Because quality standards for $PM_{2.5}$ have not been defined in Mexico yet, U.S. NAAQS (National Ambient Air Quality Standards) guidelines and threshold values, shown in Table 1, were used in this work.

Other researchers, such as Jorquera et al. (2001), Rodríguez et al. (2001), and Yang (2002), to name a distinct few, found a seasonal behavior in different locations. Nevertheless, in El Paso–Ciudad Juárez, we could not identify such a component, but in this case we observed similarities between both years approximately on the same days. (Fig. 3).

On relatively similar dates in the three years, we observed that the highest peaks deviated from the 24-h average standard (NAAQS). Some of these peaks reached values including 87.4 μg/m^3 (April 10, 2001), 65.0 μg/m^3 (March 14, 2002), and 69.21 μg/m^3 (April 26, 2002). These peaks coincided with the so-called "spring dust storms" events reported by the Texas Commission Environmental Quality, although these events do not occur only in spring.

As these peaks were due to a meteorological phenomenon (dust storms), we assumed that it would be difficult to deduce proper rules and avoid these high $PM_{2.5}$ concentrations. However, it would be very useful to predict $PM_{2.5}$ concentrations before air pollution

Table 1
National Ambient Air Quality Standard (NAAQS) for the US

Particulate ($PM_{2.5}$)			
Annual arithmetic mean	15.1 μg/m^3	The three-year average of annual arithmetic mean concentrations from single or multiple community-oriented monitors is not to be at or above this level.	Primary and secondary NAAQS
24-h average	66 μg/m^3	The three-year average of the annual 98th percentile for each population-oriented monitor within an area is not to be at or above this level.	Primary and secondary NAAQS

Primary NAAQS: the levels of air quality that the EPA judges necessary, with an adequate margin of safety, to protect the public health.
Secondary NAAQS: the level of air quality that the EPA judges necessary to protect the public welfare from any known or anticipated adverse effects.

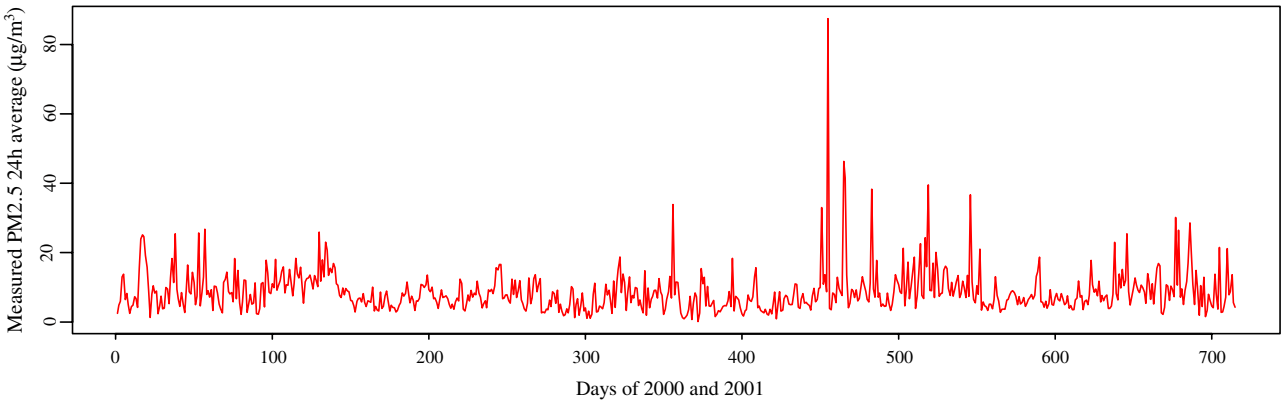

Fig. 3. Daily mean $PM_{2.5}$ value registered in 2000 and 2001.

events occur in order to take preventive actions such as alerting the population.

A new set of variables was designed (see Table 2) from the features actually available in the station, with the aim of complying with the requisite enforced, i.e. having a daily-average level prediction using only the samples from the first 8 h of the day. It is interesting to observe the variable corresponding to the Wind Direction Index (WDI), considered so as to avoid the discontinuity that it would cause the Wind Direction variable, if used instead. We define the WDI according to the following expression:

$$\mathrm{WDI} = 1 + \sin\left(\mathrm{WD} + \frac{\pi}{4}\right) \tag{1}$$

So as to select, among these features, those more appropriate to serve as explanatory variables in the different models, a linear correlation analysis was initially performed. The analysis revealed that the only notable relationships were those between the real average level of $PM_{2.5}$ and the average corresponding with the first 8 h, between the latter and the maximum level of $PM_{2.5}$, and between the wind direction and the Wind Direction Index. Fig. 4 shows the results of this analysis.

Table 2
Input variables used in the forecast models

Symbol	Description	Units
Pm25m8	Average levels of $PM_{2.5}$ during the first 8 h of the day.	$\mu g/m^3$
Pm25max8	Maximum level of $PM_{2.5}$ during the first 8 h of the day.	$\mu g/m^3$
Tam8	Average temperature during the first 8 h of the day.	°F
Hrm8	Average relative humidity during the first 8 h of the day.	%
Vvm8	Average wind speed during the first 8 h of the day.	m/s
rdvm8	Average wind bearing during the first 8 h of the day.	–
rdvsin8	Wind direction index.	–

Next, a stepwise regression analysis was performed, which measured only one explanatory variable, the Pm25m8. The residual analysis for such a model exhibited plainly its low quality. Thus, by virtue of the large number of samples, the whole set of the designed variables was considered in the input layer.

2.2. Classical models

The persistence model is an extremely simple model, with no adjustable parameter; consequently we can expect nothing but poor precision. Due to its simplicity, it represents the minimum acceptable quality out of any other model proposed. Basically, it accepts that the concentration levels in $PM_{2.5}$ at a particular time of day correspond to the value which occurred the day before at the same hour. That is to say:

$$y_t = x_t \tag{2}$$

On the other hand, linear regression models can be applied to both categorical and continuous explanatory variables in the prediction of continuous variables. The mathematical formulation is a model in which, for each observation i, $i = 1, \ldots, N$, the y_i value of the variable to be explained is linearly fitted according to the observed values of the samples. The error of the prediction is represented by ε. The complete model can be expressed as:

$$y_i = \beta_0 + \sum_{j=1}^{N} \beta_j x_{ij} + \varepsilon_i \tag{3}$$

The reader interested in applying linear models in an atmospheric context may find useful the work of Castejón et al. (2001).

2.3. Neural network models

Artificial neuronal networks (ANN) are powerful data modeling tools with a proven efficiency in dealing

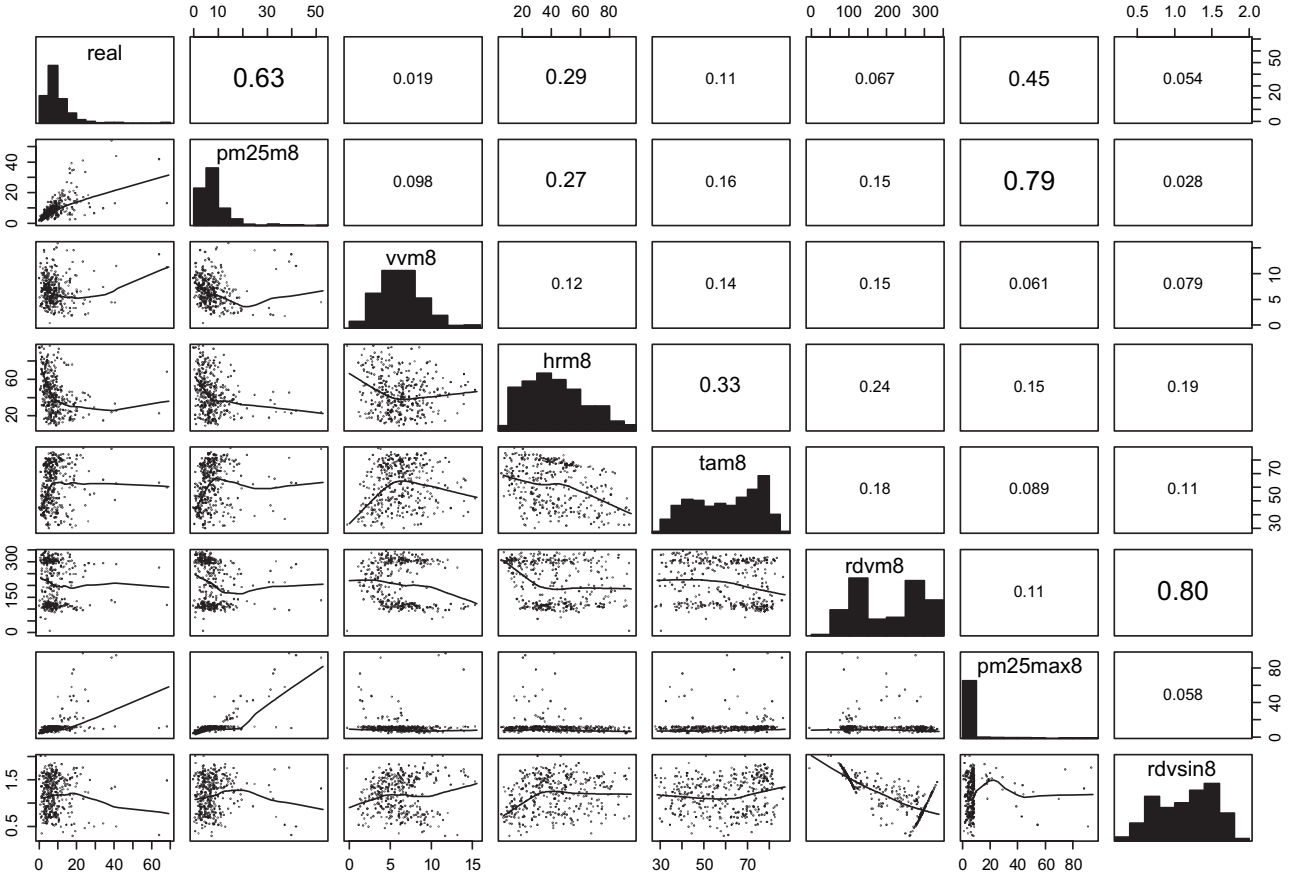

Fig. 4. Linear relations between variables.

with complex problems, particularly in the fields of association, classification and prediction. Many researchers have shown that neural networks can solve almost any problem more efficiently than the traditional modeling and statistical methods (Hornik, et al., 1989; Masters, 1993). In this paper, we compare three neuronal network architectures in order to obtain the best possible efficiency in the outcomes of Multilayer Perceptron (MLP), Square Multilayer Perceptron (SMLP), and Radial Basis Function (RBF) analyses.

Typically, a neural network is composed of a set of neurons laid out in layers. Commonly, those layers are classified as input layer, hidden layers and output layer.

Some neural networks do not have hidden layers and are used as more linear statistical techniques. These networks (with input and output layers only) are useful in many linear or semi-linear applications, but in general, it is difficult to get accurate results in non-linear problems (McCullagh and Nelder, 1989). We understand that particulate matter is clearly a non-linear

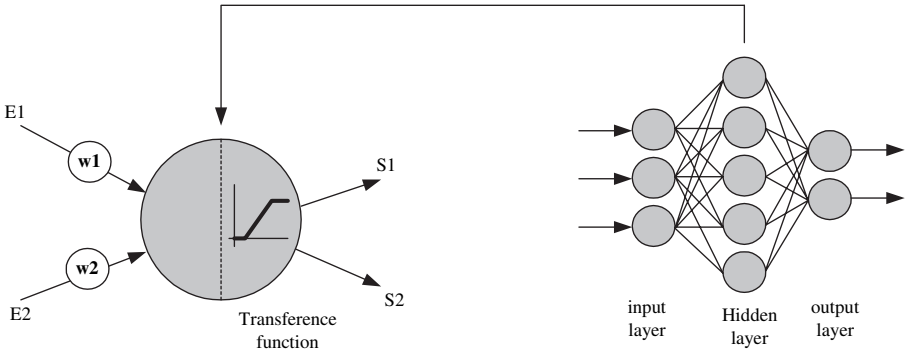

Fig. 5. Artificial neural network structure.

problem, at least in the current region of the study. However, there are no specific rules to define how many hidden layers a neural network must have (Fig. 5).

For MLP and RBF neural networks, one hidden layer with a large number of neurons usually yield good results (Hornik et al., 1989; Hornik, 1993; Bishop, 1995).

A similar situation occurs in terms of the quantity of data needed to obtain the best training results from the network. The neural network has the capacity to "learn" new skills and make predictions from new data; that is to say, it generalizes observed behavior, rather than simply memorizing a given training data set. As a "rule of the thumb", the quantity of data necessary in a neural network analysis would be, for a noise-free quantitative target variable, twice as many training cases as weights, while for a very noisy target variable, 30 times as many training cases as weights may not be enough. The high number of input variables frequently presented in these models implies an even higher number of weights to train—if the networks have a fully connected topology—thus turning the overwhelming size of the training data set into one of the main obstacles associated with this methodology.

2.3.1. Multilayer Perceptron (MLP) and Square Multilayer Perceptron (SMLP)

MLP is the most common and successful neural network architecture with feed-forward network (FFN) topologies (three layers of neurons: input layer, hidden layer and output layer).

Each layer uses a linear combination function. The inputs are fully connected to the hidden layer, which is fully connected to the outputs.

These networks are used to create a model and map the input to the output using historical data. Run-on in the model can be used to produce an output, even if the desired output is at that point unknown. These networks are called supervised networks because they need a desired output to learn (supervised training). For MLP applications in the atmospheric sciences, see Gardner and Dorling (1998).

The most common supervised training algorithm is the so-called backpropagation (Haykin, 1994). With backpropagation, the input data are repeatedly presented to the neural network. With each presentation, the output of the neural network is compared to the desired output and an error is computed. This error is then fed back (backpropagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to the desired output. This process is known as "training". This kind of training is relatively easy and offers good support for prediction applications.

It is generally accepted that the characteristics of a correctly designed MLP network are, though worthy of comparison, not better than the characteristics that can be obtained from classical statistical techniques. Nevertheless, MLP networks outperform classical statistical techniques in their much shorter time of development, and their adaptive capacity when faced with changes.

Generalized Additive Models (GAM), (Hastie and Tibshirani, 1990), are relevant statistical factors. Within this framework, Flake suggested an architecture similar to GAM, in which each hidden unit had a parametric activation function which could change from a projection-based to a radial function in a continuous way (Flake, 1998). He called this architecture SMLP.

2.3.2. Radial Basis Function (RBF)

The architecture of RBF neural networks is less well-known than that of the MLP, although it has been used in time series modeling predictions with good results. The input for this kind of architecture is a feed-forward network (i.e., an MLP neuron network), but every unit of the hidden layer has a "radial basis function" (statistical transformation based on Gaussian distribution function). Like MLP neural networks, RBF networks are suited for applications such as pattern discrimination and classification, interpolation, prediction, forecasting, and process modeling.

The "basis function" (often a Gaussian function) has the parameters "centre" and "width". Usually each unit of the network has a different central value. The center of the basis function is a vector of numbers C_i of the same size as the inputs to the unit. Normally, there is a different center for each unit in the neural network.

In the first computation, the "radial distance" is computed for every unit between the input vector and the center of the basis function using the Euclidean distance algorithm. In other words, the structure of the RBF has non-linear inputs (input vector) for every data (unit) and the radial distance is computed between the input vector and the center of the basis function. See Fig. 6.

The input of the RBF neural network is non-linear whereas the output is linear. Because of these properties, RBF neural networks can model complex maps more easily and quickly than MLP (Haykin, 1994). For further details and a more complete description, see Tao (1993).

2.4. Missing data

Missing data were carefully managed because there were relevant periods without information on one parameter, and therefore we could not apply the popular filling strategy described by Dixon (1979). In our particular case, such samples were discarded from the database, even when it was clear that this strategy

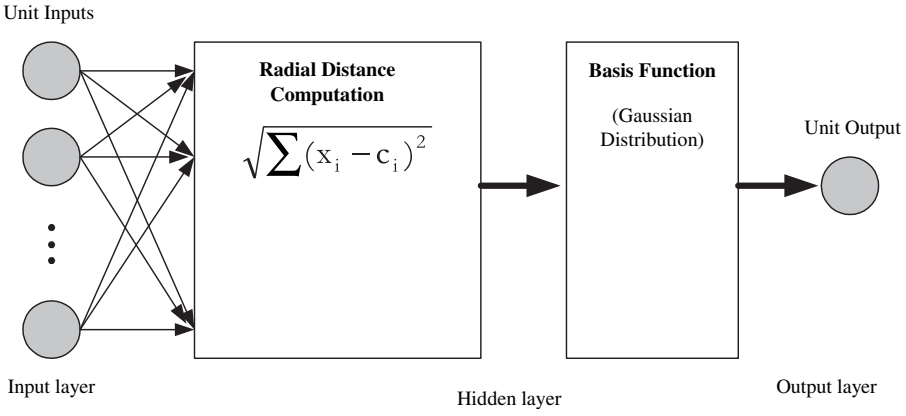

Fig. 6. Radial Basis Function structure.

could reduce the number of patterns available. We did not strive particularly to identify the outliers in the data that exceeded the first validation level since this level was considered to be sufficient in terms of quality.

2.5. Error measurement

In this kind of problem, there are many statistical indices to provide a numerical description of the accuracy of the estimates. One of those is the Root Mean Square Error (RMSE). This is calculated according to Eq. (4)

$$\mathrm{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{1/2} \tag{4}$$

where N is the number of observations, O_i is the observed value, and P_i is the predicted value.

Another is the *Mean Absolute Error* (MAE), defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|O_i - P_i| \tag{5}$$

Another interesting index is the *Correlation Factor* (R^2), defined as:

$$R^2 = \frac{\sum_{i=1}^{N}[P_i - \bar{O}]^2}{\sum_{i=1}^{N}[O_i - \bar{O}]^2} \tag{6}$$

where N is the number of observations, O_i is the observed value, P_i is the predicted value, $y\ \bar{O}$ is the average value of the explained variable on N observations.

So as to avoid distortions in the residuals due to the random initialization of the neuron weights during the training phase, each topology has been trained one hundred times, thus obtaining the distribution of the errors.

3. Results and discussion

The objective was to model the $PM_{2.5}$ daily average concentration using the mean $PM_{2.5}$, wind speed, maximum level of $PM_{2.5}$ during those first 8 h, wind direction, humidity, and temperature values registered in the first 8 h of the day, insofar as these were assumed to be the main parameters needed to predict it. The model was based on a neural network and we used

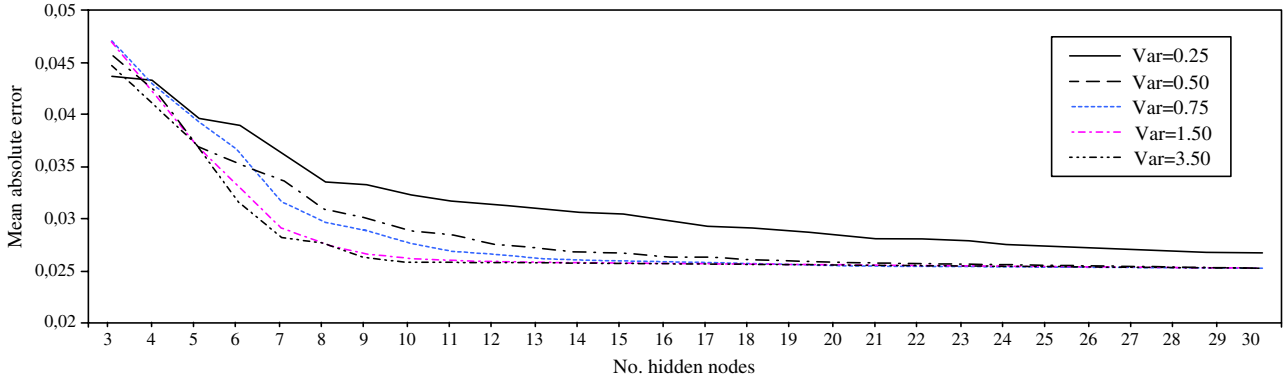

Fig. 7. MAE index evolution as a function of the number of hidden nodes for the different variances assessed.

MLP, SMLP, and RBF basic architectures for two reasons:

1. To evaluate the three kinds of neural computing topologies.
2. And particularly, to identify the best prediction model.

According to the assumption of the major parameters and the neural network architectures used, the neural network that was required had 7 nodes on the input layer and one node on the output layer. The quantity of neurons on the hidden layer was estimated by performing experimental analysis on the errors. So as to

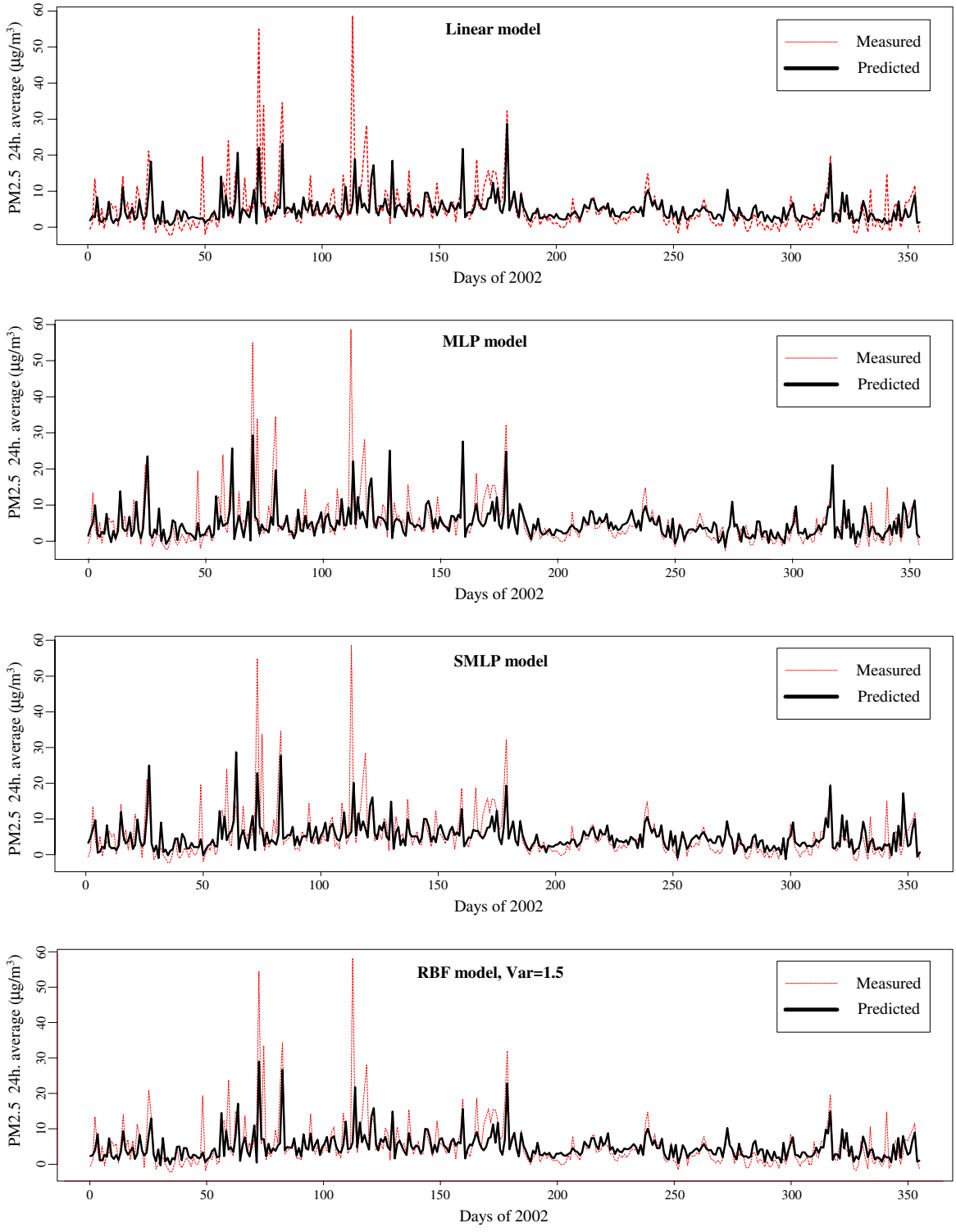

Fig. 8. Daily times series of the measured and predicted concentrations of $PM_{2.5}$.

determine the number of hidden nodes, the MAE index was used.

In all the training and test cycles, and as far as there were no observable data clusters, the 2000 and 2001 data sets were used as sources of the training data, and the 2002 data were used as the test data.

To handle the over-fitting problem, a regularization scheme is applied in neural networks.

In the case of the Multilayer Perceptron (MLP), a BPM (Back Propagation with Moment) was used as a learning rule. The number of nodes on the hidden layer was tested from 1 to 30. The stability reached, with the exception of some 'like resonant' topologies, was exceptional. In order to monitor the quality of the prediction over the time axis, Fig. 8 shows the real predicted values by MLP with 18 hidden nodes for the test data, i.e. 2002 data.

In the case of the SMLP network, some tests were carried out to see the Squared Multilayer Perceptron in action. Using the same chart structure, Fig. 9 shows the errors against the number of hidden units. It was apparent that a relatively low number of hidden neurons were giving a kind of weighted-mean estimator, and therefore, more members were necessary to represent the real model. To illustrate this particular aspect, an SMLP with 20 hidden neurons was used. As expected, the operation was slower with the test data.

As for Radial Basis Function (RBF) networks, they were trained with different variance values in the gaussian radial function: 025, 0.5, 0.75, 1.5, and 3.5. Clearly seen in Fig. 7 is the evolution of the MAE index as a function of the hidden nodes for the different variances assessed.

Fig. 8 shows the time series corresponding to the daily average values of the $PM_{2.5}$ observed during 2002 and the fitted values corresponding to different models. The behavior of the linear model lacks refinement, so to

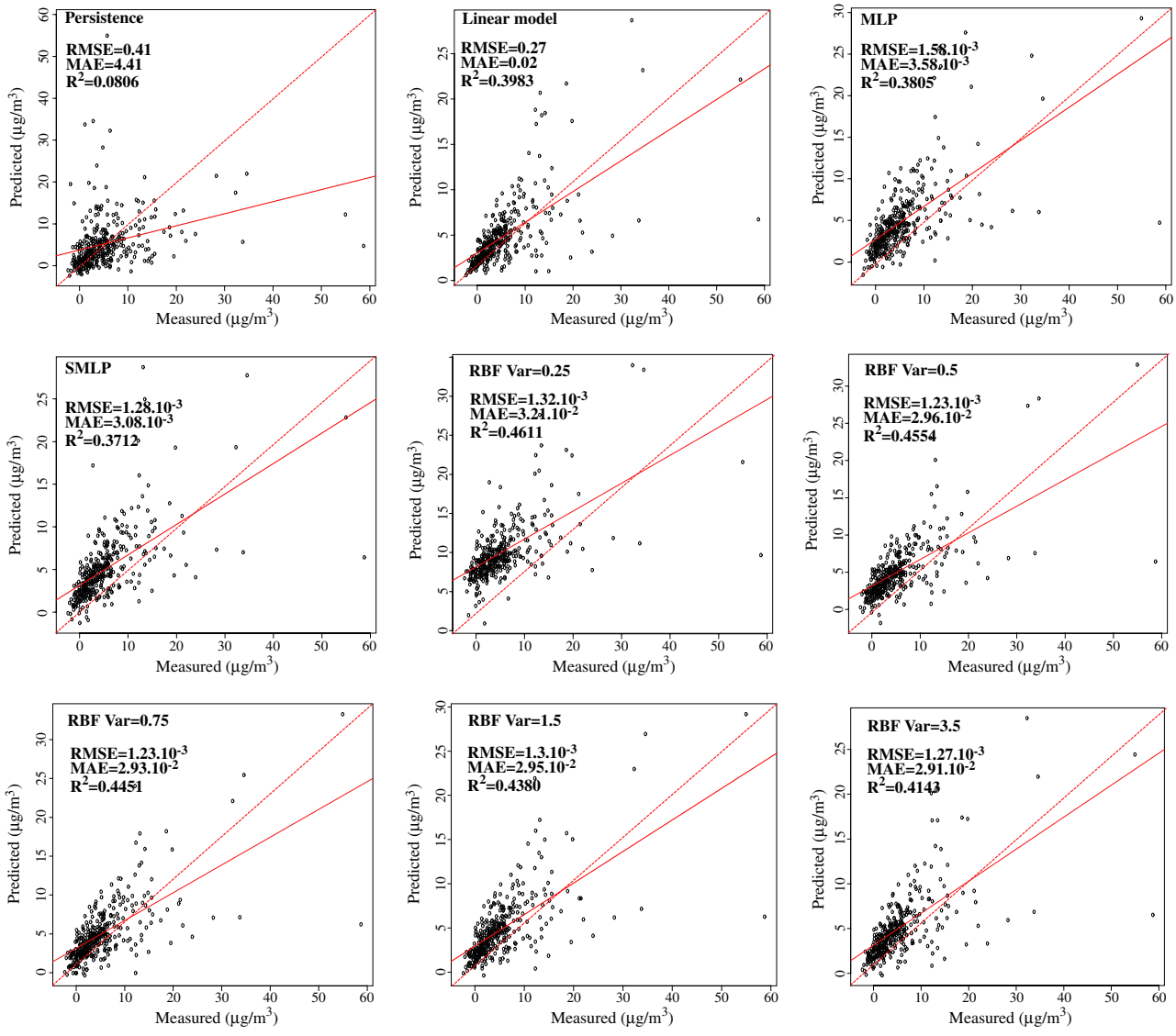

Fig. 9. Scatter plots of the measured and predicted concentrations of $PM_{2.5}$.

speak, insofar as it proves reluctant to defect from the low ranges. The MLP, SMLP and RBF models enforce the predictions to follow more precisely the observed reality. In Fig. 9, we have shown the dot plots of the observed measures vs. the fitted values, marking both the identity line and the trend obtained.

According to the results obtained, it is clear that the RBF neural network is particularly suited for our aims, with predictions as good as those obtained by the MLP, a lower training effort and more stability (probably due to the bounded, derivative property of RBF networks).

The results have also proven that SMLP does not have any particular advantage in our case compared to MLP in terms of prediction.

In Fig. 10, we have shown the histograms of the residuals for the different models assessed.

4. Conclusions

A short time $PM_{2.5}$ prediction model has been built taking into account a large number of samples from a non-linear data set with a high degree of internal noise. The model can be used as a tool for short time control and planning in difficult areas like the U.S.–Mexican border in El Paso–Ciudad Juárez. The comparative analysis of neural network architectures has provided very interesting results, comparing RBF networks with MLP and SMLP networks, which are the most commonly used.

The MLP network provides acceptable predictions, in spite of the difficult environmental conditions of the location (i.e. even though $PM_{2.5}$ data were considered, samples still show hard peaks of inmissions seasonally, mainly due to the common presence of dust storms in this area). The SMLP network shows a very similar behavior, although a few more neurons in the hidden layer are necessary to obtain the same error found in the former case. This fact has a direct influence on the sample size necessary for the right training, as previously mentioned. Finally, the RBF network shows the best behavior, with the shortest training times and best stability. These results suggest that the widely used MLP should be replaced by the more convenient RBF network.

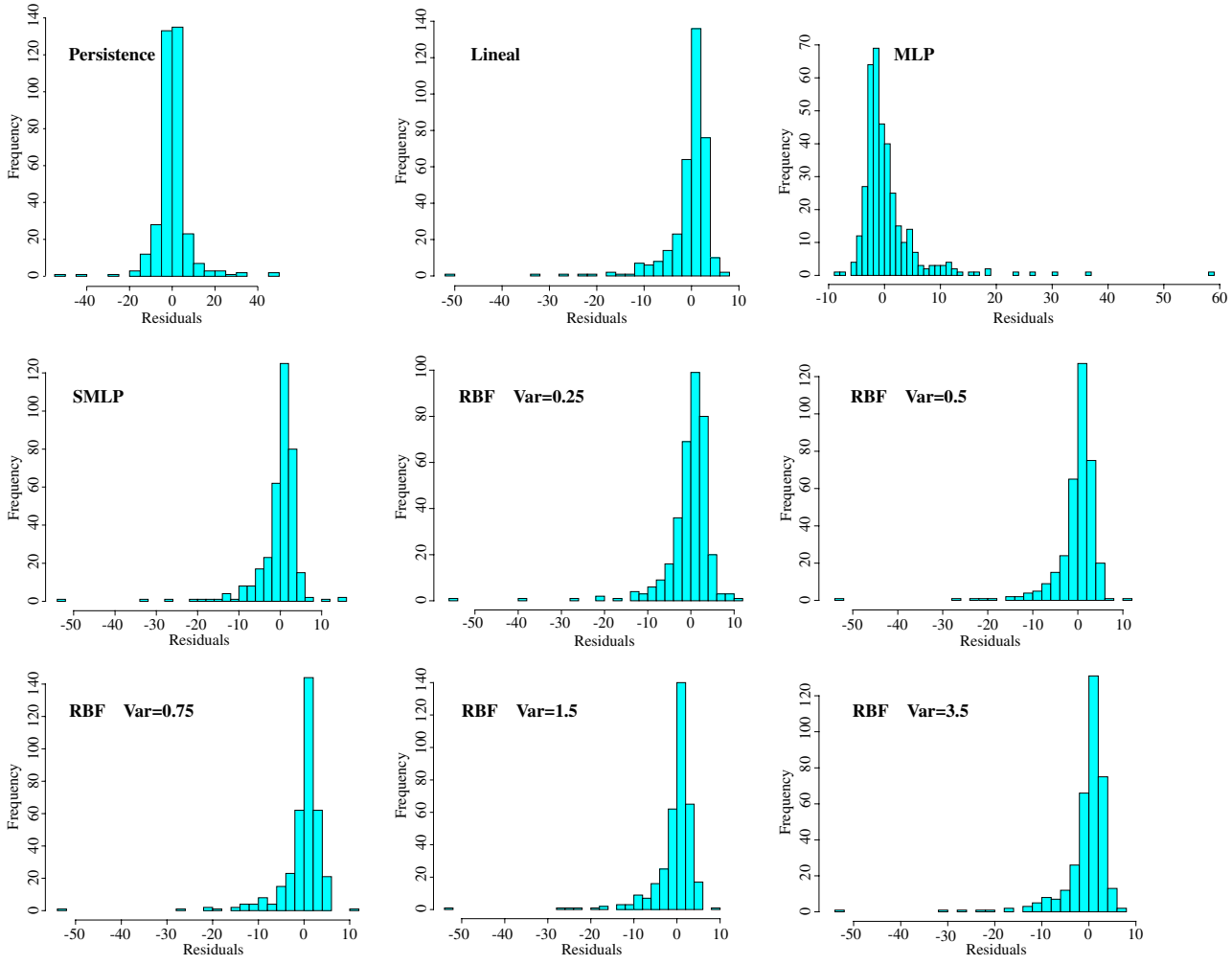

Fig. 10. Histograms of the residuals for the different models assessed.

Table 3
Performance statistics for the models

Performance measure	Model	Topology	Validation set
RMSE	Persistence	–	0.41
	Lineal regression	–	0.27
	NN_{MLP}	7-18-1	1.58×10^{-3}
	NN_{SMLP}	7-20-1	1.28×10^{-3}
	$NN_{RBF\ Var=0,25}$	7-21-1	1.32×10^{-3}
	$NN_{RBF\ Var=0.50}$	7-20-1	1.23×10^{-3}
	$NN_{RBF\ Var=0.75}$	7-16-1	1.23×10^{-3}
	$NN_{RBF\ Var=1.5}$	7-10-1	1.30×10^{-3}
	$NN_{RBF\ Var=3.5}$	7-10-1	1.27×10^{-3}
MAE	Persistence	–	4.41
	Lineal regression	–	0.02
	NN_{MLP}	7-18-1	3.58×10^{-3}
	NN_{SMLP}	7-20-1	3.08×10^{-3}
	$NN_{RBF\ Var=0,25}$	7-21-1	3.21×10^{-3}
	$NN_{RBF\ Var=0.50}$	7-20-1	2.96×10^{-3}
	$NN_{RBF\ Var=0.75}$	7-16-1	2.93×10^{-3}
	$NN_{RBF\ Var=1.5}$	7-10-1	2.95×10^{-3}
	$NN_{RBF\ Var=3.5}$	7-10-1	2.91×10^{-3}
R^2	Persistence	–	0.0806
	Linear regression	–	0.3983
	NN_{MLP}	7-18-1	0.3805
	NN_{SMLP}	7-20-1	0.3712
	$NN_{RBF\ Var=0,25}$	7-21-1	0.4611
	$NN_{RBF\ Var=0.50}$	7-20-1	0.4554
	$NN_{RBF\ Var=0.75}$	7-16-1	0.4451
	$NN_{RBF\ Var=1.5}$	7-10-1	0.4380
	$NN_{RBF\ Var=3.5}$	7-10-1	0.4143

Nevertheless, it seems necessary to point out that the ANN, in general, have limitations inherent to their own structure. The main handicap is the impossibility of generalizing what is trained for the 24-h range; the per-hour context, for instance. The location concept is also a key element; we must train the networks with data sets corresponding to the periods and locations in order to perform and analyse.

Table 3 summarizes the results of the analyses.

Acknowledgements

We gratefully acknowledge the funding support from the SUPERA program (MEXICO), and Spanish *Ministerio de Ciencia y Tecnología* grant DPI2001-1408, which made this work possible.

References

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Castejón Limas, M., Ordieres Meré, J.B., De Cos Juez, F.J., Martínez de Pisón, F.J., 2001. Control de Calidad, Metodología para el análisis previo a la modelización de datos en procesos industrials, Universidad de La Rioja.

Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. The Science of the Total Environment 313, 1–13.

Chow, J.C., Watson, J.G., 2001. Zones of representation for PM_{10} measurements along the US/Mexico border. The Science of the Total Environment 276, 49–68.

Dixon, J.K., 1979. Pattern recognition with partly missing missing data. IEEE Transactions on Systems Man and Cybernetics SMC-9 10, 617–621.

Flake, G.W., 1998. Square unit augmented, radially extended, multilayer perceptrons. In: Orr, G., Müller, K.-R., Caruana, R. (Eds.), Tricks of the Trade: How to Make Algorithms Really Work, LNCS State-of-the-Art-Surveys. Springer-Verlag.

Fuller, G.W., Carslaw, D.C., Lodge, H.W., 2002. An empirical approach for the prediction of daily mean PM10 concentrations. Atmospheric Environment 36, 1431–1441.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multi-layer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Prentice-Hall, Englewood Cliffs, NJ.

Ho, S.L., Xie, M., Goh, T.N., 2002. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. Computers & Industrial Engineering 42, 371–375.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Hornik, K., 1993. Some new results on neural network approximation. Neural Networks 6, 1069–1072.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 299–314.

Jorquera, H., Palma, W., Tapia, J., 2000. An intervention analysis of air quality data at Santiago, Chile. Atmospheric Environment 34, 4073–4084.

Kukkonen, J., Harkonen, J., Karppinen, A., Pohjola, M., Pietarila, H., Koskentalo, T., 2001. A semi-empirical model for urban PM10 concentrations, and its evaluation against data from an urban measurement network. Atmospheric Environment 35, 4433–4442.

Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815–825.

Lenschow, P., Abraham, H.-J., Kutzner, K., Lutz, M., Preuß, J.-D., Reichenbächer, W., 2001. Some ideas about the sources of PM10. Atmospheric Environment 35 (Supplement No. 1), S23–S33.

Lu, H.-C., 2002. The statistical characters of PM10 concentration in Taiwan area. Atmospheric Environment 36, 491–502.

Magliano, K.L., Hughes, V.M., Chinkin, L.R., Coe, D.L., Haste, T.L., Kumar, N., Lurmann, F.W., 1999. Spatial and temporal variations in PM_{10} and $PM_{2.5}$ source contributions and comparison to emissions during the 1995 integrated monitoring study. Atmospheric Environment 33, 4757–4773.

Masters, T., 1993. Practical Neural Network Recipes in C++. Academic Press, San Diego.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.

McDonnell, W.F., Nishino-Ishikawa, N., Peterson, F.F., Chen, L.H., Abbey, D.E., 2000. Relationship of mortality with the fine and coarse fraction of long term ambient PM10 concentrations in nonsmokers. Journal of Exposure Analysis and Environmental Epidemiology 10, 427–436.

Mukerjee, S., 2001. Selected air quality trends and recent air pollution investigations in the US–Mexico border region. The Science of the Total Environment 276, 1–18.

Mukerjee, S., Shadwick, D.S., Smith, L.A., Somerville, M.C., Dean, K.E., Bowser, J.J., 2001. Techniques to assess cross-border air pollution and application to a US–Mexico border region. The Science of the Total Environment 276, 205–224.

Ostro, B.D., Eskeland, G.S., Sánchez, J.M., Feyzioglu, T., 1999a. Air pollution and health effects: a study of medical visits among children in Santiago, Chile. Environmental Health Perspective 107, 69–73.

Ostro, B., Chesnut, L., Vichit-Vadakan, N., Laixuthai, A., 1999b. The impact of particulate matter an daily mortality in Bangkok, Thailand. Journal of Air and Waste Management Association 49, 100–107.

Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 34, 1189–1196.

Pérez, P., Reyes, J., 2001. Prediction of particulate air pollution using neural techniques. Neural Computing and Applications 10, 165–171.

Pérez, P., Trier, A., 2001. Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile. Atmospheric Environment 35, 1783–1789.

Pérez, P., Reyes, J., 2002. Prediction of maximum of 24-h average of PM10 concentrations 30 hours in advance in Santiago, Chile. Atmospheric Environment 36, 4555–4561.

Podnar, D., Koračin, D., Panorska, A., 2002. Application of artificial neural network to modeling the transport and dispersion of tracers in complex terrain. Atmospheric Environment 36, 561–570.

Querol, X., Alastuey, A., Rodríguez, S., Plana, F., Mantilla, E., Ruiz, C.R., 2001a. Monitoring of PM10 and PM2.5 around primary particulate anthropogenic emission sources. Atmospheric Environment 35, 845–858.

Querol, X., Alastuey, A., Rodríguez, S., Plana, F., Ruiz, C.R., Cost, N., Massagué, G., Puig, O., 2001b. PM10 and PM2.5 source apportionment in the Barcelona metropolitan area, Catalonia, Spain. Atmospheric Environment 35, 6407–6419.

Reich, S.L., Gómez, D.R., Dawidowski, L.E., 1999. Artificial neural network for the identification of unknown air pollution sources. Atmospheric Environment 33, 3045–3052.

Rodríguez, S., Querol, X., Alastuey, A., Kallos, G., Kakaliagou, O., 2001. Saharan dust contributions to PM10 and TSP levels in Southern and Eastern Spain. Atmospheric Environment 35, 2433–2447.

Rodríguez, S., Querol, X., Alastuey, A., Mantilla, E., 2002. Origin of high summer PM10 and TSP concentrations at rural sites in Eastern Spain. Atmospheric Environment 36, 3101–3212.

Tao, K.M., 1993. A Closer Look at the Radial Basis Function (RBF) Networks. Conference Record of the 27th Asilomar Conference on Signals, Systems and Computers, vol. 1. IEEE Comput. Soc. Press, Los Alamitos, CA, pp. 401–405.

Tiittaa, P., Raunemaa, T., Tissari, J., Yl-Tuomi, T., Leskinen, A., Kukkonen, J., Harkonen, J., Karppinen, A., 2002. Measurements and modelling of $PM_{2.5}$ concentrations near a major road in Kuopio, Finland. Atmospheric Environment 36, 4057–4068.

US-EPA, 1996. Air Quality Criteria for Particulate Matter. EPA/600/P-95/001F. US Environment Protection Agency, Washington, DC.

US-EPA, 1998. US–Mexico Border XXI Program. United States–Mexico Border Environmental Indicators EPA909-R-98-001. US Environment Protection Agency, Washington, DC.

US-EPA, 2000a. Summary of Selected Environmental Indicators from the U.S.–Mexico Border XXI Program: Progress Report 1996–2000. EPA 909-R-00-002. US Environment Protection Agency, Washington, DC.

US-EPA, 2000b. National Air Quality and Emissions Trends Report 1998. EPA 454-R-00-003. US Environment Protection Agency, Washington, DC.

US-EPA, 2000c. Latest Findings on National Air Quality: 1999 Status and Trends. EPA-454-F-00-002. US Environment Protection Agency, Washington, DC, 00–002.

US-EPA, SEMARNAT, 2002. FRONTERA 2012: Programa Ambiental Mexico-Estados Unidos, US Environment Protection Agency, Washintong, DC. Secretaría de Medio Ambiente y Recursos Naturales de México.

Watson, J.G., Chow, J.C., 2001. Source characterization of major emission sources in the Imperial and Mexicali Valleys along the US/Mexico border. The Science of the Total Environment 276, 33–47.

Yang, K.-L., 2002. Spatial and seasonal variation of PM10 mass concentrations in Taiwan. Atmospheric Environment 36, 3403–3411.

Further readings

Chen, L., Sandhu, H.S., Angle, R.P., McDonald, K.M., Myrick, R.H., 2000. Rural particulate matter in Alberta, Canada. Atmospheric Environment 34, 3365–3372.

Gauvin, S., Reungoat, P., Cassadou, S., Dechenaux, J., Momas, I., Just, J., Zmir, D., 2002. Contribution of indoor and outdoor environments to PM2.5 personal exposure of children VESTA study. The Science of the Total Environment 297, 175–181.

Hien, P.D., Binh, N.T., Truong, Y., Ngo, N.T., 1999. Temporal variations of source impacts at the receptor, as derived from air particulate monitoring data in Ho Chi Minh City, Vietnam. Atmospheric Environment 33, 3133–3142.

Hien, P.D., Bac, V.T., Tham, H.C., Nhan, D.D., Vinh, L.D., 2002. Influence of the meteorological conditions on PM2.5 and PM2.5-10 during the monsoon season in Hanoi, Vietnam. Atmospheric Environment 36, 3473–3484.

Hopke, P.K., 1985. Receptor Modeling in Environmental Chemistry. Wiley, New York, 319 pp.

McClellan, O.R., 2001. Setting ambient air quality standards for particulate matter, University of New Mexico, 13701 Quaking Aspen Place NE, Albuquerque, NM 87111, USA, Toxicology 00, 119 pp.

Morel, B., Yeh, S., Cifuentes, L., 1999. Statistical distributions for air pollution applied to the study of the particulate problem in Santiago. Atmospheric Environment 33, 2575–2585.

Salcedo, R.L.R., Alvim Ferraz, M.C.M., Alves, C.A., Martins, F.G., 1999. Time-series analysis of air pollution data. Atmospheric Environment 33, 2361–2372.

Vega, E., Reyes, E., Sanchez, G., Ortiz, E., Ruiz, M., Chow, J., Watson, J., Edgerton, S., 2002. Basic statistics of PM2.5 and PM10 in the atmosphere of Mexico City. The Science of the Total Environment 287, 167–176.