Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software 23 (2008) 1056–1069
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US)

E. Salazar-Ruiz [a], J.B. Ordieres [b,*], E.P. Vergara [b], S.F. Capuz-Rizo [c]

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

Received 13 August 2007; received in revised form 25 November 2007; accepted 28 November 2007
Available online 21 February 2008

Abstract

This study developed 12 prediction models using two types of data matrix (daily means and a selection of the mean for the first 6 h of the day). The Persistence parametric prediction technique was applied separately to these matrices, as well as semiparametric Ridge Regression and three non-parametric or artificial intelligence techniques: Support Vector Machine, Multilayer Perceptron and ELMAN networks. The target was the prediction of maximum tropospheric ozone concentrations for the next day in the Mexicali–Calexico border area. The main ozone precursors and meteorological parameters were used for the different models. The proposals were evaluated using specific performance measurements for the air quality models established in the Model Validation Kit and recommended by the US Environmental Protection Agency.

Results with similar margins of error were obtained in various models developed in this study, and some of them have provided smaller margins of error than similar prediction models existing in the literature developed in other regions. For this reason, we consider it feasible to apply the prediction models developed and they could be useful for supporting decisions in the matter of ozone pollution in the region under study, as well as for use in daily forecasting in this area.

Keywords: US–Mexico border; Ozone neural network modeling; Multilayer Perceptron (MLP); ELMAN neural network; Support Vector Machine; Ridge Regression; Model Validation Kit (MvK); Transboundary air quality

1. Introduction

One of the main problems of atmospheric pollution in urban areas is contamination caused by photochemical oxidants such as ozone (O_3) and nitrogen dioxide (NO_2) (Lee et al., 1996; Kongtip et al., 2006). Ozone is considered to be one of the main greenhouse gases and a component of photochemical smog with potentially harmful effects on human health, mainly in high-risk populations (Sousa et al., 2007; Filleul et al., 2006; Weschler, 2006), and on habitats and their vegetation (Davis and Orendovici, 2006; Scebba et al., 2006).

Due to the nature of ozone, its photolysis in the troposphere has been shown to be directly related to ultraviolet solar radiation at a wavelength of around 300 nm, followed by reaction with water molecules, sources of OH radicals, which take part in reactions responsible for the oxidation of other gases present in the atmosphere (Guicherit and Roemer, 2000). The nitrogen dioxide (NO_2) photo-dissociates to form nitrogen oxide (NO) and atomic oxygen (O), which immediately combines with oxygen (O_2) to form ozone (O_3). The nitrogen oxides (NO_x) act as a catalyst in the ozone formation process (Frost et al., 1998). Studies by Monks (2000), Kleinman (2000) and Trainer et al. (2000), all based on observations in rural environments, showed that ozone production was limited by NO_x availability. In the presence of a sufficient amount of NO_x, the main source of

* Corresponding author. Fax: +34 941 299 277.
E-mail address: joaquin.ordieres@unirioja.es (J.B. Ordieres).

1364-8152/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2007.11.009

ozone production is the oxidation of carbon monoxide (CO) and volatile organic compounds (VOCs).

Of particular interest is the chemical coupling between ozone and nitrogen oxides, prompting different authors to study the atmospheric relationship between O_3 and NO_x, in order to obtain further knowledge of this phenomenon (Clapp and Jenkin, 2001). Sillman (1999) affirms that two photochemical regimes may be differentiated in ozone production: an initial regime, before NO_x saturation takes place, when photochemical ozone production in urban areas increases with emissions of NO_x, but is less sensitive to emissions of VOCs; and a second regime, after NO_x saturation has occurred, when ozone levels rise as VOC levels increase, and fall when NO_x levels decrease.

Furthermore, it should not be forgotten that daily variations in ozone are controlled not only by variations in precursor gases and VOCs but also by local weather conditions (Fischer et al., 2003). Factors such as temperature, wind behavior and relative humidity have an important influence on daytime ozone levels. The specific meteorological characteristics of the area studied here make the development of ozone prediction models all the more interesting. The study was performed in two neighboring cities: Calexico, in California (US), and Mexicali, in Baja California (Mexico). These cities are situated in semi-desert regions, with high summertime temperatures being a natural characteristic of the area. Moreover, air quality in both cities has binational features due to their geographical location and the natural, two-way flow of contaminants.

Legislation in both countries has laid down regulations to govern acceptable national emission levels in the first instance, and regional levels in agreement with local governments. A National Ambient Air Quality Standard (NAAQS) is in force in the US, but California has established its own standards: California Ambient Air Quality Standards (CAAQS). Californian standards differ substantially as regards ozone. The national standard for ozone is 0.080 ppm average concentration over 8 h and 0.12 ppm over 1 h as maximum values (US-EPA, 2006). In contrast, the maximum average concentrations in California are 0.070 ppm over 8 h and 0.090 ppm over 1 h (CARB, 2005, California).

Using the various mechanisms, techniques and methodologies available, numerous authors have proposed different strategies for resolving air contaminant prediction problems of highest peaks, both short-term and mid-term prediction accuracy, etc (McCullagh and Nelder, 1989; Hastie and Tibshirani, 1990; Salcedo et al., 1999; Ho et al., 2002; Gardner and Dorling, 2000a; Podnar et al., 2002; Sousa et al., 2007). Sokhi et al. (2006) and Han (2007) used Eulerian grid models with good results. Thompson et al. (2001) and Gardner and Dorling (2000b) made an interesting review of statistical methods for the meteorological adjustment of ozone. Along the same lines, Schlink et al. (2006) confirm the efficient performance of non-linear multivariate tools, such as generalized additive and neural network models for application in warning systems of high ozone concentrations. A detailed review of prediction techniques can be found in Gardner and Dorling (1998). The first Position Paper of this journal, Jakeman et al. (2006), contains a comprehensive set of guidelines for evaluating environmental models including quantitative and qualitative measures of performance.

Among the different strategies reported in the literature for the development of models, neuronal networks and, specifically, the MultiLayer Perceptron (MLP) are being increasingly used in applications for predicting contamination levels or for estimating meteorological adjustments in ozone trends (Hornik, 1993; Bishop, 1997; Gardner and Dorling, 1998; Flake, 1998; Haykin, 1999; Kolehmainen et al., 2001; Ordieres et al., 2005; Dudot et al., 2007). In particular, nitrogen oxide (NO_x and NO_2) concentrations were predicted by Gardner and Dorling (1999) applying a Multilayer Perceptron-based model and other statistical models, and the comparison of the different results revealed the benefit of using a Multilayer Perceptron. Additionally, Artificial Neural Networks have recently been used to predict SO_2 levels and have proven to be of greater efficiency than linear methods (Chelani et al., 2002). Multilayer Perceptrons, in particular, have provided better results than statistical linear methods. Artificial Neural Networks (ANNs) are mathematical models capable of determining a non-linear relationship between two data sets (Haykin, 1999). ANNs are universal functions of approximation that can be applied to problems, in which there is *a priori*, no knowledge of the relevance of the input variables (Hornik et al., 1989; Hornik, 1993; Pernía-Espinoza et al., 2005; Martínez-De-Pisón et al., 2006). Since the mapping carried out by ANNs is non-linear, it is complex to understand; nevertheless, certain simple methods can be used to explore input relevance. Recently, Pires et al. (2008) used Multiple Linear Regression (MLR) and Principal Components Regression (PCA) for meteorological and environmental parameter validation for tropospheric ozone forecasting models.

This study has two main objectives: firstly, to provide an advanced model for predicting maximum ozone levels 1 day ahead in order to establish a strategic decision-making process; and secondly, to explore the capability of recurrent neural networks, such as ELMAN, in order to test their capabilities. As regards the first objective, an in-depth analysis was performed for the types of models, and an exhaustive search was performed for each model to identify relevant variables, since the aim was to build a model with the lowest possible margins of error. Although there were no such models in the region studied, there are those that have been developed for predicting maximum ozone levels in other parts of the world; hence, we also sought to improve on models described in the prior literature or, in their absence, on strategies that were not hitherto useful. As regards the second objective, in the case of the MLP and ELMAN networks, the aim with the recurrent networks was to try to improve the slow learning times displayed by the neuronal networks with MLP, without any noticeable loss of quality in the solution.

To begin this paper, Section 1.1 describes the geographical area where the construction of the models was validated, as well as the data available and the data management strategy. This is followed by a description of the models used, from the most traditional models to the non-parametric models, as well as the different error criteria used to measure the quality

of each model, with special attention paid to their significance. Sections 5 and 6 present the results, discussions and conclusions obtained. Although not all the results were positive, the use of a data matrix with 24 h daily means yielded a real improvement.

1.1. Site characterization

The geographical area of study is part of the border strip established since 1983 according to the La Paz Agreement (1983) between the US and Mexico on joint cooperation to protect and improve the environment in the border area.

This border region extends for 100 km (62.5 miles) on either side of the international boundary and corresponds to the shaded strip in Fig. 1. This delimitation has been maintained in subsequent binational agreements (US-EPA, 1996; US-EPA and SEDUE, 1991), including the most recent "Border 2012/Frontera 2012" US–Mexico Environmental Program (US-EPA and SEMARNAT, 2003).

Geographically, Mexicali is situated in Baja California in the northwestern corner of Mexico (32°40′ north 115°28′ west), near the city of Calexico (32°40′42″ north, 115°29′53″ west), which is located in southwestern California, US. Both cities are densely populated and have numerous factories known as "maquiladoras" (essentially assembly plants) established mainly on the Mexican side. These industries are supplied with raw materials from the US, and normally the finished products return to the US for assembly in end products or for global marketing. Mexicali–Calexico is therefore a very important border crossing.

Since the cities are next to each other, they share a common air shed. Air pollution sources on either side of the border have an impact on air quality in both cities. The semi-desert climate in this area means that both cities have predominantly dry climates, with summer temperatures reaching 50 °C. This prevailing climate compounded by economic and political activities favors air quality problems.

Mexicali is a non-attainment area for ozone and carbon monoxide, and the concentration level of PM_{10} is considered serious. Most emissions of ozone precursors (81% of NO_x and 61% of VOCs) come from motor vehicles, 91.1% of carbon monoxide comes from the same source and 94% of PM_{10} comes from "fugitive dust", largely from the use of unpaved roads and, to a lesser degree, from wind erosion.

Calexico is considered a non-attainment area for ozone and particulate matter. In the case of ozone, 78% of NO_x and 63% of VOCs come from motor vehicles. The biggest air pollution problem is PM_{10}, which has been linked to asthma and other health problems. PM_{10} (54%) comes from "fugitive dust", and another 30% from agricultural tilling and animal feedlots (SCERP, 2003; US-EPA, 2000a).

2. Prediction models

2.1. Parametric models: Persistence and Linear Regression

Persistence is the easier method. It basically assumes that O_3 maximum concentration levels on a specific day correspond to the value occurring the day before. As a result, it is extremely simple to develop a Persistence model, with no adjustable

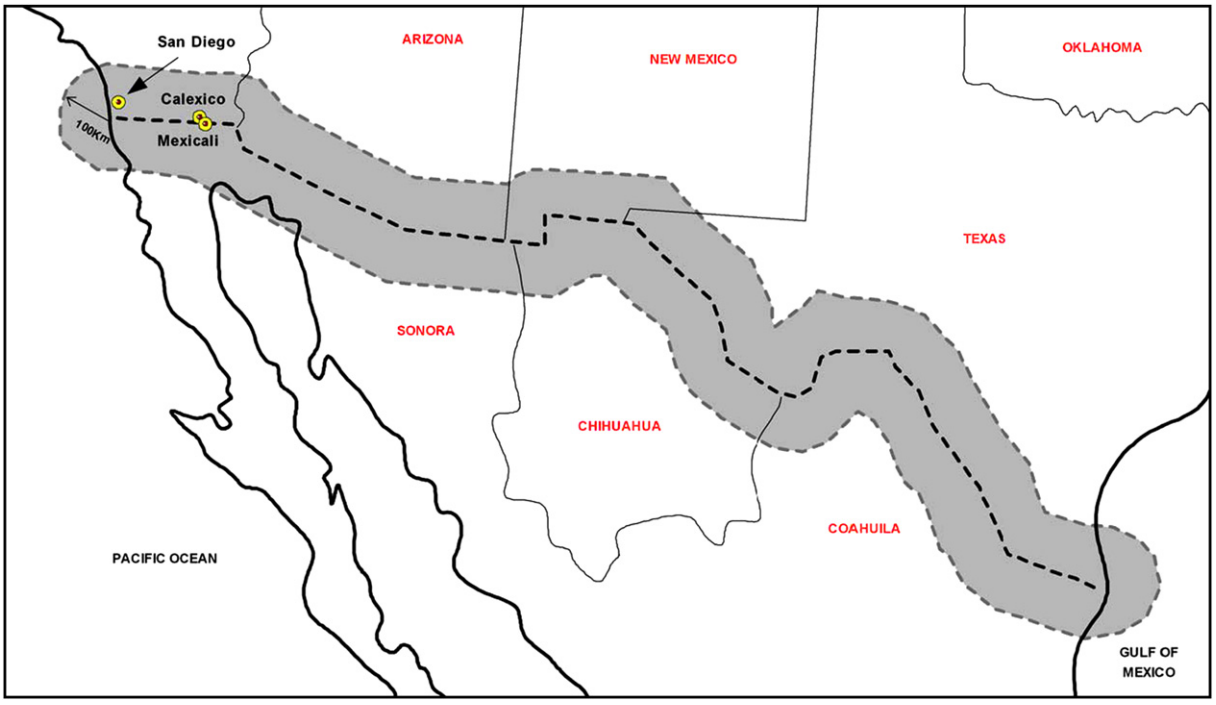

Fig. 1. Calexico, California (US) and Mexicali, Baja California (Mexico) are situated in the northwest. Both cities belong to the shaded border strip established under the La Paz Agreement (1983).

parameter. In addition, ozone and other atmospheric variables show a positive statistical association with their own past or future values (Wilks, 1995). Nevertheless, Persistence models frequently represent baseline precision against the prediction using other models proposed. The representative mathematical model is:

$$y_i = y_{(i-1)} \tag{1}$$

$y_i = \mathrm{O}_{3\max}$ at day $i \quad i = 1, 2, 3, \ldots, n$ days

Linear Regression models can be applied to both categorical and continuous explanatory variables for the prediction of continuous variables. The mathematical formula is a model in which, for each observation i, the y_i value of the variable to be explained is linearly fitted according to the observed values of the samples. The prediction error is represented by ε. The complete model can be expressed as:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \varepsilon_i \tag{2}$$

Readers interested in applying Linear models within an atmospheric context may find the work of Castejón-Limas et al. (2001) useful.

2.2. Semiparametric model: Ridge Regression

Semiparametric regression models play an interesting role, as they use regression models that contain at least one function, being modeled non-parametrically. Accordingly, they can be of substantial value in the solution of complex scientific problems.

Semiparametric regression models reduce complex data sets to summaries that we can understand, and properly applied they retain essential features of the data while discarding unimportant details, and hence they help to understand and characterize different kinds of phenomena. In particular, air quality problems have pronounced nonlinearity, which suggests that better predictions and managerial decisions can be made through the use of semiparametric regression (Ruppert et al., 2003).

2.2.1. Ridge Regression

Ridge Regression is an extension of a simple Linear Regression to the case of multiple predictor variables. The main reason for solving by Ridge Regression is the multicollinearity relationship with the different predictor variables. Multi-collinearity is a supposed characteristic of air quality prediction models, and a simple Linear Regression model does not take this important situation into account. Accordingly, the Ridge Regression model equation used in this paper is:

$$\widehat{y}_{\mathrm{new}} = \widehat{\beta}^{\mathrm{T}} x_{\mathrm{new}} = \sum_i \alpha_i x_i^{\mathrm{T}} x_{\mathrm{new}} \tag{3}$$

2.3. Non-parametric models: Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are powerful data modeling tools with proven efficiency for dealing with complex problems, particularly in the fields of association, classification and prediction. A neural network typically comprises a set of neurons distributed in layers. These layers are often classified as input layer, hidden layer 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 obtain accurate results in non-linear problems (McCullagh and Nelder, 1989). As noted in Section 1, ozone behavior is clearly a non-linear phenomenon.

A similar situation occurs in terms of the quantity of data needed to obtain the best training results from the network. The neural network aims to achieve the necessary skills to make predictions from new data, that is, to generalize observed behavior rather than simply memorizing the training data set. As a rule of the thumb, the quantity of data required in a neural network analysis would be, for a noise-free quantitative target variable, twice as many training cases as weights; this would be enough. However, for an extremely noisy target variable, 30 times as many training cases as weights may not be enough. The high number of input variables frequently present in the models implies an even higher number of weights to train, if the networks have fully connected topologies; hence, the large size of the training data set is one of the main obstacles associated with this methodology.

2.3.1. Multilayer Perceptron (MLP)

MLP is the most common and successful neural network architecture, with feedforward 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, and this hidden layer is fully connected to the outputs, see Fig. 2.

These networks are used for creating models and for mapping the input to the output using historical data so that the model can then be used to produce an output, even if the desired output is unknown.

Some networks are called supervised networks because they need a desired output to learn (supervised training). The most common supervised training algorithm is the so-called ‘‘backpropagation’’ rule. With backpropagation, the input data are repeatedly presented to the neural network. With each presentation, the output of the neural network is compared with 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. As a result, the error decreases with each iteration and the neural model gets closer and closer to the desired output. This process is known as ‘‘training’’ (Haykin, 1994; Flake, 1998). This kind of training is relatively easy and offers good support for prediction applications.

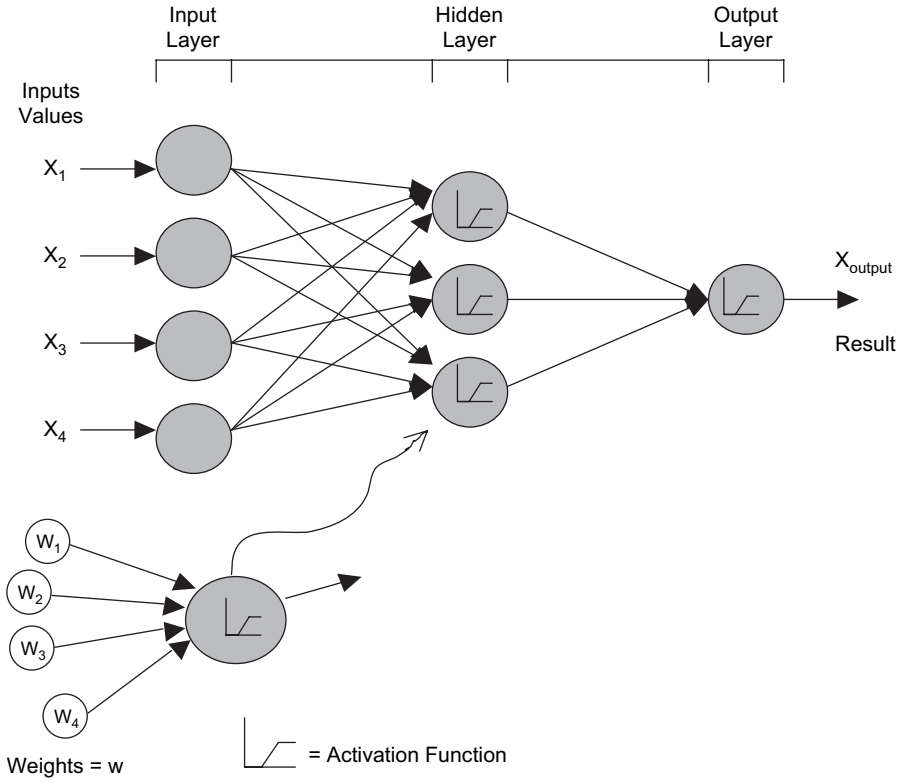

Fig. 2. Typical MLP Artificial Neural Networks.

It is generally accepted that the characteristics of a correctly designed MLP network are worthy of comparison with the characteristics obtained using classical statistical techniques.

2.3.2. Support Vector Machine (SVM)

SVMs are sets of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers; a special case of Tikhonov regularization can also be considered. Support Vector Machines non-linearly map their *n*-dimensional input space into a high-dimensional feature space. A linear classifier is constructed in this high-dimensional feature space. A special property of this family of classifiers is their capacity for simultaneously minimizing empirical classification errors and

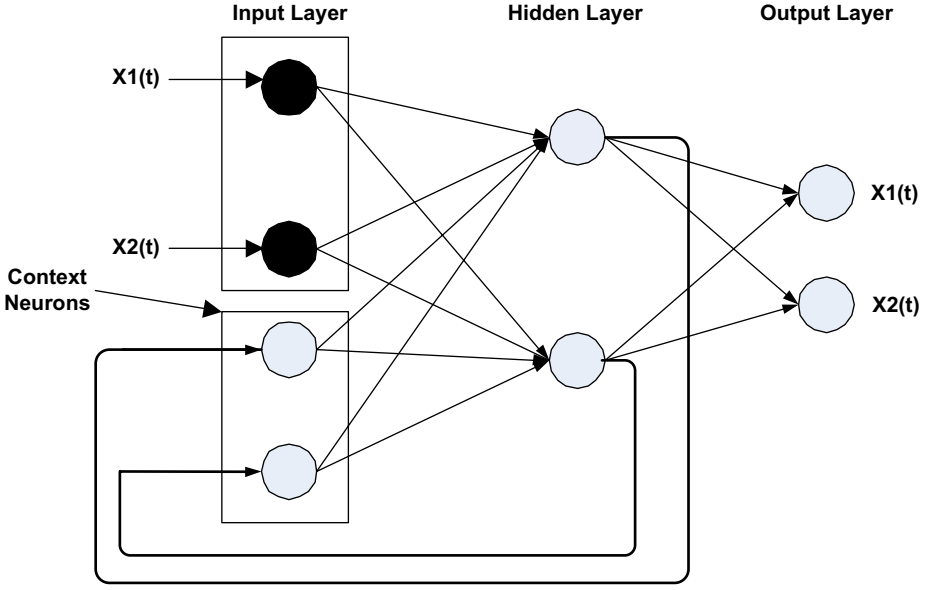

Fig. 3. General model for an ELMAN recurrent neural network.

maximizing the geometric margin of the hyper-plane making the classification. Hence, they are also known as maximum margin classifiers.

2.3.3. ELMAN neural network (ELMAN)

The ELMAN network is recognized as a two-layer network with feedback from the first-layer output to the first-layer input. This recurrent connection allows the ELMAN network to both detect and generate time-varying patterns, see Fig. 3.

The ELMAN network has tansig neurons in its hidden (recurrent) layer and purelin neurons in its output layer. This combination is special in that two-layer networks with these transfer functions can approximate any function (with a finite number of discontinuities) with arbitrary accuracy. The only requirement is that the hidden layer must have enough neurons. More hidden neurons are needed as the function being fitted increases in complexity.

The ELMAN network differs from conventional two-layer networks in that the first-layer has a recurrent connection. The delay in this connection stores values from the previous time step, which can be used in the current time step. This means that even when two ELMAN networks, with the same weights and biases, are given identical inputs at a given time step, their outputs may differ because of different feedback states.

Considering that the network can store information for future reference, it is able to learn both temporal and spatial patterns. The ELMAN network can be trained to both respond to and generate both kinds of patterns. This study aims to evaluate its capacity for modeling the concept of precursor variables for ozone, and whether the method is robust enough to adapt to scenarios with a high noise component.

3. Measuring model performance

In order to compare the performance of the different models developed in this study, we used the statistical standardized evaluation tools included in the Model Validation Kit (MvK), released within the framework of the workshops "*Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*" held by www.harmo.org since

Table 1
Model Validation Kit error measurements

	Description	Formula
i.	Root Mean Square Error. Provides a global idea of the difference between the observed and modeled values.	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(C_\mathrm{p} - C_\mathrm{o}\right)^2}$
ii.	Correlation coefficient between C_o and C_p. Provides a global description of the model.	$R = \frac{\mathrm{Mean}\left(\left(C_\mathrm{o} - \mathrm{Mean}(C_\mathrm{o})\right)\left(C_\mathrm{p} - \mathrm{Mean}\left(C_\mathrm{p}\right)\right)\right)}{\sigma_{C_\mathrm{o}}\sigma_{C_\mathrm{p}}}$
iii.	Geometric Mean Bias. Quantifies the geometric mean deviation.	$\mathrm{MG} = \mathrm{e}^{(\mathrm{Mean}(\ln(C_\mathrm{o})) - \mathrm{Mean}(\ln(C_\mathrm{p})))}$
iv	Normalized Mean Square Error. A version of the MSE, but normalized in order to establish comparisons among different models.	$\mathrm{NMSE} = \frac{\mathrm{Mean}\left(C_\mathrm{o} - C_\mathrm{p}\right)^2}{\mathrm{Mean}(C_\mathrm{o})\mathrm{Mean}\left(C_\mathrm{p}\right)}$
v.	Fractional Bias. Normalized measure that enables the mean observed values and the mean predicted values to be compared.	$\mathrm{FB} = 2\frac{\mathrm{Mean}(C_\mathrm{o}) - \mathrm{Mean}\left(C_\mathrm{p}\right)}{\mathrm{Mean}(C_\mathrm{o}) + \mathrm{Mean}\left(C_\mathrm{p}\right)}$
vi	Geometric variance. Quantifies the geometric variance.	$\mathrm{VG} = \mathrm{e}^{\mathrm{Mean}(\ln(C_\mathrm{o}) - \ln(C_\mathrm{p})^2)}$
vii.	Factor of two (FAC2). Quantifies the percentage of forecasted cases in which the values of the ratio $C_\mathrm{o}/C_\mathrm{p}$ were in the range [0.5, 2].	$0.5 \le \frac{C_\mathrm{o}}{C_\mathrm{p}} \le 2$
viii.	Fractional Variance. Normalized measurement for comparing the difference between predicted variance and observed variance. A model with FV = 0 is a model whose variance is equal to the variance of the observed values.	$\mathrm{FV} = 2\frac{\sigma_{C_\mathrm{o}} - \sigma_{C_\mathrm{p}}}{\sigma_{C_\mathrm{o}} + \sigma_{C_\mathrm{p}}}$
ix.	Index of agreement (d_2). Indicates the congruence between forecasted and observed data, taking into account the degree of freedom.	$d_2 = 1 - \frac{\sum_{i=1}^{n}\left(C_\mathrm{p} - C_\mathrm{o}\right)^2}{\sum_{i=1}^{n}\left(\left\lvert C_\mathrm{p} - \mathrm{Mean}(C_\mathrm{p})\right\rvert + \left\lvert C_\mathrm{o} - \mathrm{Mean}(C_\mathrm{o})\right\rvert\right)^2}$
x.	Mean Absolute Error. Quantifies residual errors.	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert C_\mathrm{o} - C_\mathrm{p}\right\rvert$
xi.	Mean Bias Error. Provides information about underestimation or overestimation of a model.	$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(C_\mathrm{o} - C_\mathrm{p}\right)$

C_p corresponds to forecasted values; C_o represents observed values and Mean is the arithmetic mean value.

1991. It is now a practical tool frequently used to evaluate statistical model performance (European Commission, 1994; Chang and Hanna, 2004, 2005). Table 1 describes the measurements used.

4. Materials and methods

4.1. Data set sources

In the area studied, previous publications have largely described the relevant variables for habitual contaminants, which are worth taking into consideration (Watson and Chow, 2001; Mukerjee, 2001; Mukerjee et al., 2001; US-EPA, 2000a,b,c), in addition to the physical–chemical mechanisms highlighted in Section 1 of this study.

Most of the data used to develop the prediction models were provided by the California Air Resources Board (CARB) and the US Environmental Protection Agency (US-EPA). In particular, the databases on atmospheric and meteorological contaminants were obtained in an hourly data format from monitoring station 060250005. This monitor is part of the Salton Sea Air Basin and administered by the California Air Resources Board (CARB), being geographically situated at latitude 32°40′34″, longitude 115°28′59″, and 6 m asl in Calexico City, California, US.

4.2. The data matrix

To develop these models, two data matrices with different characteristics were prepared. One matrix (Matrix A) was formed by the mean of atmospheric and meteorological predictor values from the preceding day: ozone (O_{3t}), temperature (T_t), nitrogen dioxide (NO_{2t}), nitrogen monoxide (NO_t), nitrogen oxides (NO_{xt}), resultant wind speed (RWS_t), resultant wind direction (RWD_t), carbon monoxide (CO_t), barometric pressure (BP_t), solar radiation (SR_t) and maximum ozone levels for the 24 h of the previous day (O_{3maxt}). Hourly data were used and then the daily means were calculated for the years 1999–2004 (except 2001, as the information provided was insufficient). Both meteorological and atmospheric variables are daily mean values observed in time t, with the exception of O_{3maxt}, which corresponds to the maximum concentration of ozone during the previous day (t).

A second matrix (Matrix B) of data was formed by the mean values for the first 6 h (0–5 h) of each day between 1997 and 2005 (except 2001, as noted). The matrix consists of ozone (O_3), temperature (T), nitrogen dioxide (NO_2), nitrogen monoxide (NO), carbon monoxide (CO), resultant wind speed (RWS), relative humidity (RH) and maximum ozone (O_{3max}). In this case, the matrix has no barometric pressure or wind direction variables as there are insufficient hourly data, and solar radiation data present zero values during the first hours of the day, as is to be expected.

It is important to mention that the decision to use this second matrix (Matrix B) was taken because data analysis revealed a behavior pattern in the first hourly concentrations of the different atmospheric and meteorological parameters. These first six values revealed a certain influence for the maximum value of ozone concentration during the day. For this reason, it was considered appropriate to develop prediction models using Matrix B.

Different methods exist for the treatment of absent information in a data matrix. Dixon (1979) proposes interesting skills for a correct treatment of these cases. For this investigation, every line (row) with data holes into the data matrix was eliminated. The final data matrix denominated "Matrix A" has 1343 lines and the "Matrix B" has 2367 lines.

4.3. Data sets and pruning

With a view to construct models with MLP, part of the work involved selecting the best structure for the network. To do so, five data groups were first formed, with extraction from each group of the training, test and validation data subsets. Each data group was selected randomly without replacement from every data matrix, so none of the five data groups has exactly the same data order and, consequently, the different training, test and validation subsets do not have exactly the same data values or the same data order, although the five data groups are extracted using the same data matrix. Fig. 4 can be helpful to show the data grouping process.

Matrix A used 850 data for training, 426 for testing and 67 for validation. Matrix B used 1499 training data, 750 test data and 118 data for validation. Once formed, the different data subsets for training, validation and testing from the five data groups were trained into a neural network, and its behavior

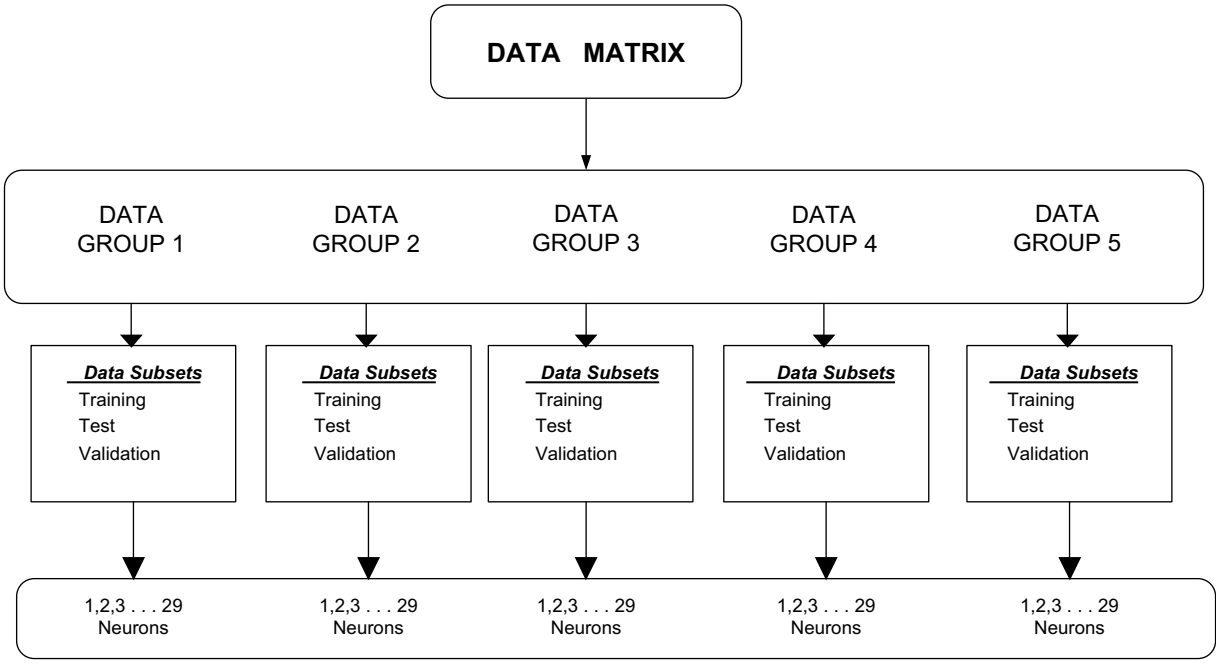

Fig. 4. Data grouping process to design MLP prediction models.

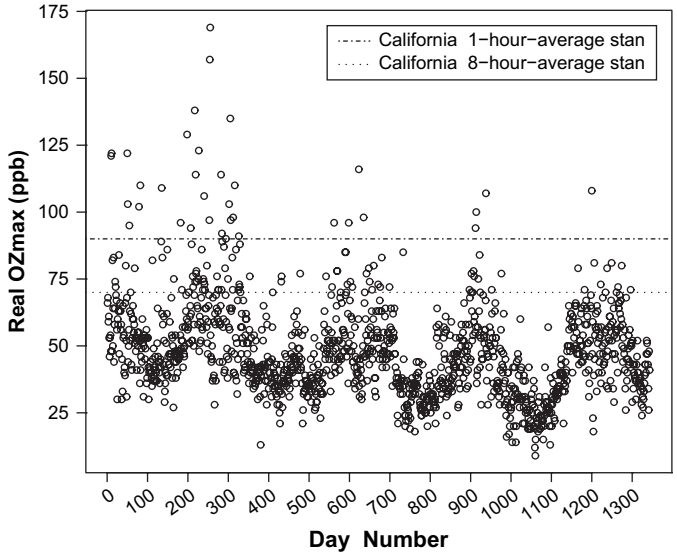

Fig. 5. Maximum daily ozone registered from 1999 through 2004 (except 2001), showing 130 violations of the 8 h California standard (above 70 ppb) and 35 violations of the 1 h California limit (above 90 ppb).

was evaluated with the presence of 1–29 neurons, in order to finally select the structure with the best performance.

4.4. The tools

R project software was basically used for data and evaluation processing and for developing the Linear and Ridge Regression models. Stuttgart Neural Network Simulator (SNNS) software was used throughout MLP and ELMAN network trainings. Each MLP neural network training structure took approximately 0.35 h, with a total of 290 different neural network structures being formed. The ELMAN network required 41% less time than its MLP counterparts.

5. Results and discussion

5.1. Ozone behavior

The plot in Fig. 5 shows maximum daily ozone from 1999 to 2004 according to the database on ozone concentrations. This period registered at least 130 violations of the 8 h California standard (above 70 ppb) and 35 violations of the 1 h California limit (above 90 ppb). It is important to bear in mind the absence of data for 2001. In view of this information gap, the total number of concentrations exceeding California limits is higher than that registered.

5.2. Prediction models

Following data segmentation into learning, validation and model quality measurement subsets, and using the validation data whenever appropriate, there were no surprises in the results for the Persistence model, despite their positive statistical association with their own past or future ozone values. Linear models and Ridge Regression fit coefficients and their qualities were determined by their coefficient confidence intervals (Ihaka and Gentleman, 1996). Ridge Regression models use a lambda = 18 for data Matrix A, and lambda = 4

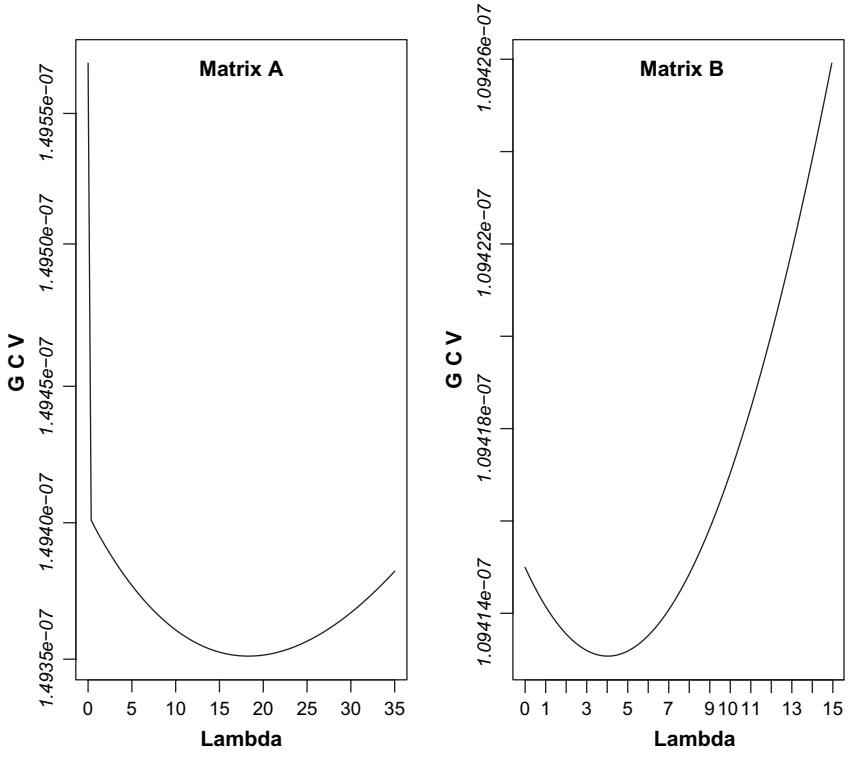

Fig. 6. Lambda evaluation for Ridge Regression models. The best lambda values selected were 4 and 18 for the trainings according to the corresponding data matrix.

Table 2
Best neural network structures according to MLP training error

Neurons	Matrix A	Matrix B
Inputs	10	8
Hidden layer	19	20
Output	1	1

for data Matrix B, selected according to the lower GCV value, see Fig. 6.

A number of tests were performed in the case of the non-parametric models to induce the random start-up of the weights of these models and to study the evolution of their predictions. Early stopping was used to avoid overtraining and five different data groups were taken to select the group with the best performance according to training error results.

The RMSE error was calculated during each training to select the adequate number of neurons in the hidden layer and find the best network structure. The selected number of neurons in the hidden layer corresponds to the smaller RMSE. As a result, the best structure for data Matrix A corresponds to 19 neurons in the hidden layer, and the use of 20 neurons in the hidden layer for data Matrix B gave the best structure. Consistent with these results, the better network structures appear in Table 2.

Finally, six prediction models were developed using data Matrix A and six models using data Matrix B, giving a total of 12 prediction models formulated. For practical purposes, the work performed to determine the best models for predicting maximum daily ozone concentrations is summarized in Tables 3 and 4. As can be seen, the calculated results corresponded to the performance measurements implemented in the Model Validation Kit (Chang and Hanna, 2005).

5.3. Analysis of results according to the data matrix used

As seen from the results shown in Tables 3 and 4, almost all the models developed with data Matrix A recorded a better performance, with the exception of the SVM with data Matrix B, which recorded a relatively insignificant improvement in R and d_2 compared to the model using Matrix A. These results showed that the influence of the behavior of atmospheric and meteorological contaminants during the first 6 h of the day was unable to prompt an improvement in the maximum ozone concentration on that day in comparison with the models using data corresponding to contaminant concentrations during the previous 24 h.

5.4. Analysis of results according to the prediction models

The prediction models developed (Persistence, Linear, Ridge Regression, MLP, SVM and ELMAN) for each data matrix clearly showed that the best performance levels were obtained by the Artificial Neural Network models, and specifically by the model that applied the Multilayer Perceptron (MLP) technique, followed by modeling with the ELMAN network, which was only slightly better than the Support Vector Machine (SVM); the performances achieved by the Linear model, Ridge Regression and the Persistence model were notably worse, as expected due to the highly non-linear and disperse behavior of the data.

According to the behavior of developed models using Ridge Regression and Linear Regression techniques, it is observed that the model of Ridge Regression was no better than its Linear Regression counterpart. Accordingly, it can be concluded that no forceful collinearity is reflected between the predictor variables in the case under study. This situation is reflected in the dispersion and correlation graphs in Figs. 7 and 8.

Low collinearity between the ozone maximum and the different predictor variables during data processing is also reflected in the Ridge Regression and Linear models, which are similar. Table 5 shows the result coefficients for both models (Ridge Regression and Linear) using Matrix B. The coefficient values are similar, but there are some details to observe, such as the slightly different trends of smaller values in the Regression Ridge model vs. Linear model for O_3m, NO_2, NOm, RHm and COm variables.

The following graphs (Figs. 9–11) show the behavior of the best three models corresponding, as mentioned previously,

Table 3
Results of the performance measurements of the models developed using data Matrix A

Performance measures	Persistence	Linear	R. ridge	MLP	SVM	ELMAN
RMSE	17.1547	14.0516	14.0568	*9.4303*	11.4345	10.8929
R	0.5351	0.6124	0.6124	*0.7417*	0.6050	0.6743
FB	−0.0004	0.0000	0.0000	*0.0153*	0.0028	0.0186
MG	0.9997	0.9624	0.9624	*1.0000*	0.9892	1.0000
NMSE	0.1295	0.0870	0.0869	*0.0430*	0.0624	0.0554
VG	1.1026	1.0760	1.0700	*1.0410*	1.0601	1.0520
FAC2	0.9590	0.9799	0.9881	*1.0000*	1.0000	1.0000
MAE	11.4430	9.8307	9.8379	*7.5261*	8.3845	8.2839
MBE	0.0208	0.0000	0.0000	*−0.6959*	−0.1308	−0.8599
d_2	0.7221	0.7250	0.7249	*0.8511*	0.7561	0.8218
FV	−0.0009	0.4800	0.4800	*0.1209*	0.2837	−0.1031

The values in italics mean "best performance" and correspond to the best model developed.

Table 4
Results of the performance measurements of the models developed using data Matrix B

Performance measures	Persistence	Linear	R. ridge	MLP	SVM	ELMAN
RMSE	19.5819	16.0386	16.0386	*13.7856*	14.6613	14.3106
R	0.5309	0.5976	0.5233	*0.6922*	0.6339	0.6012
FB	−0.0002	0.0000	0.0000	*−0.0190*	0.0169	−0.0429
MG	1.0000	0.9610	0.9610	*0.9431*	0.9884	0.9215
NMSE	0.1407	0.0891	0.0890	*0.0679*	0.0784	0.0766
VG	1.1280	1.0850	1.0855	*1.0800*	1.0815	1.0750
FAC2	0.9427	0.9768	0.9805	*0.9831*	0.9797	0.9718
MAE	13.0718	11.2663	11.2663	*10.2666*	10.1266	11.0234
MBE	0.0000	0.0000	0.0000	*0.0010*	−0.8858	0.0022
d_2	0.7213	0.7115	0.7115	*0.7875*	0.7615	0.7282
FV	0.0004	0.5033	0.5488	*0.4312*	0.3606	0.2242

The values in italics mean "best performance" and correspond to the best model developed.

to the MLP, ELMAN and SVM models of data Matrix A. These graphs offer a clear view of the structure of the real and predicted data for each best prediction model developed.

The graph in Fig. 9 shows that the MLP model in this case study was the method closest to predicting peak ozone concentrations, when compared with the SVM and ELMAN models. The predictions obtained by the SVM model (Fig. 10) were

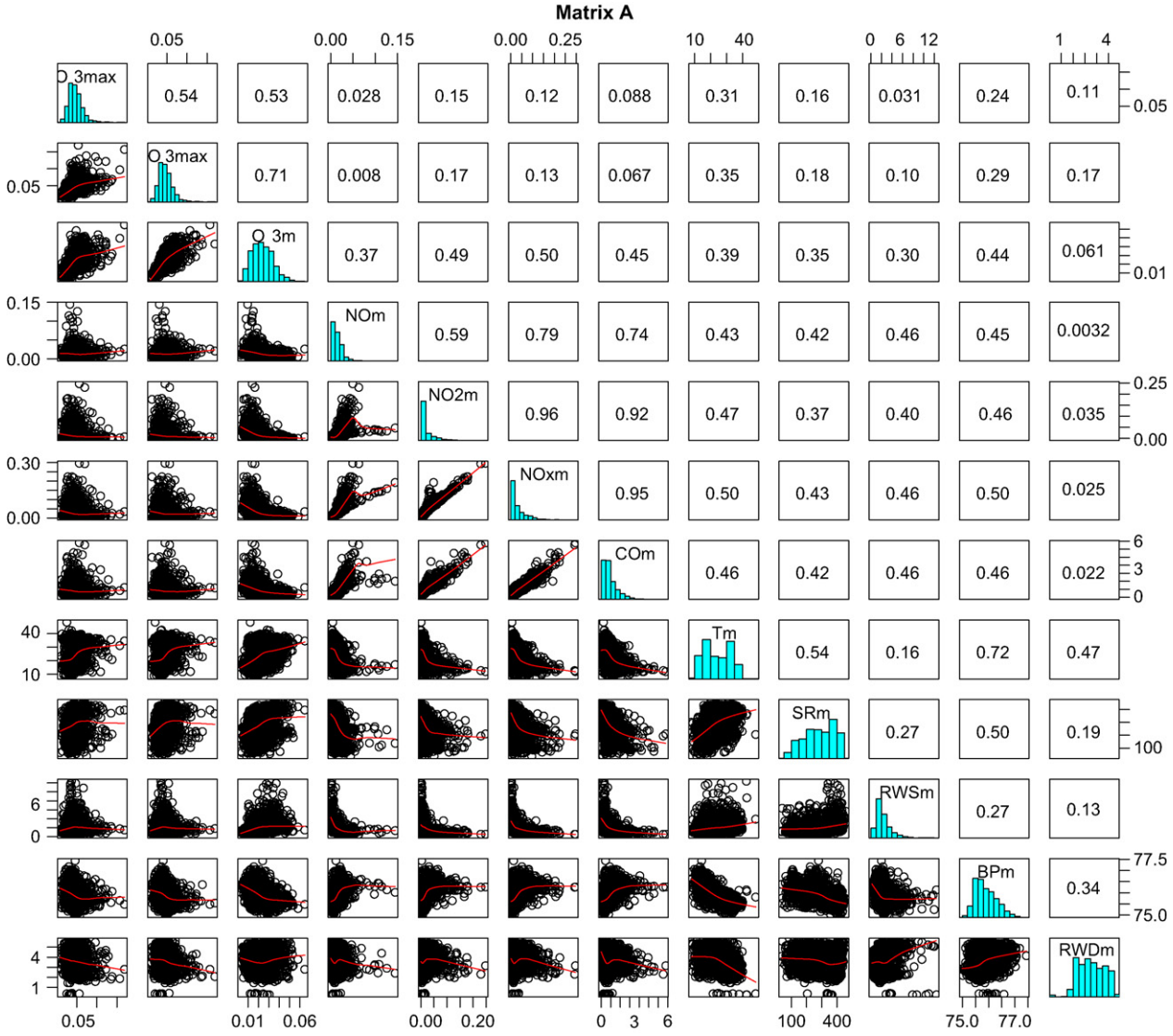

Fig. 7. The plot shows a low and moderate correlation between the predictor variables of data Matrix A.

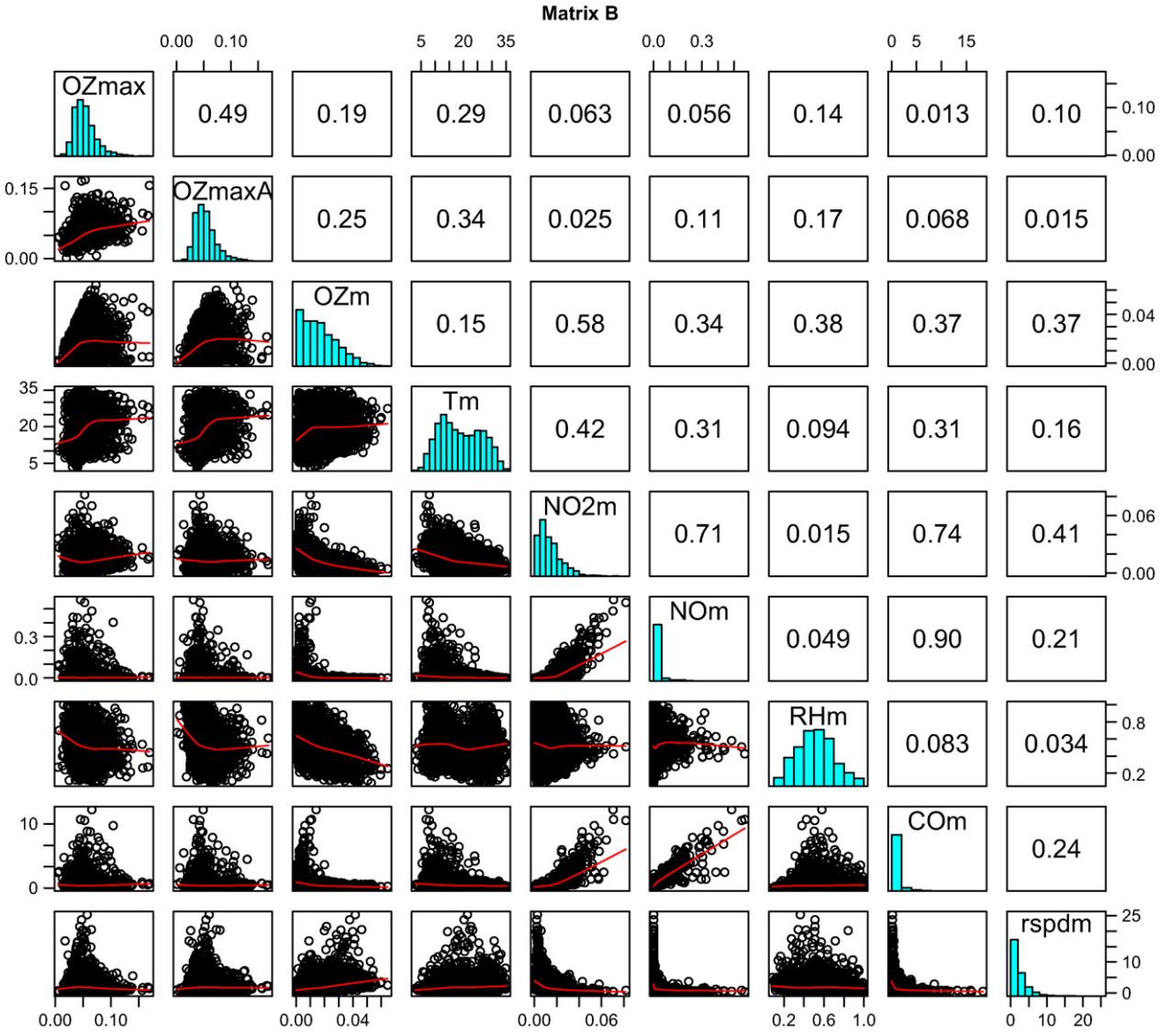

Fig. 8. The plot shows an overly low and moderate correlation between the predictor variables of data Matrix B.

slightly out of phase, whereas the ELMAN model (Fig. 11) started very well before tailing off with respect to real values shortly before reaching the halfway stage. However, it is important to highlight one characteristic in favor of ELMAN networks, namely, that in this trial they required 41% less time for trainings than MLP networks.

Figs. 12–14 show scatter plots of forecasted and observed values corresponding to the best three prediction models. The graph reveals the difficulty of the problem and the presence of types of behaviors that were difficult for all the models to predict, such as the behavior indicated by the value above 100 ppb, which was poorly managed by all the models, although the MLP model predicted this better than the ELMAN network, which was, in turn, better than the SVM.

6. Conclusions

A wide range of parametric and non-parametric models have been developed showing that, in the prediction of atmospheric contaminants such as ozone, the results of non-parametric

Table 5
Regression Ridge and Linear model coefficients

	O_{3max}	O_3m	Tm	NO_2	Nm	RHm	COm	RWSm
Ridge	0.28528	0.62731	0.000881	0.92472	−0.12469	0.00910	0.00216	−0.000924
Linear	0.28411	0.63358	0.000887	0.93566	−0.12672	0.00929	0.00219	−0.000924

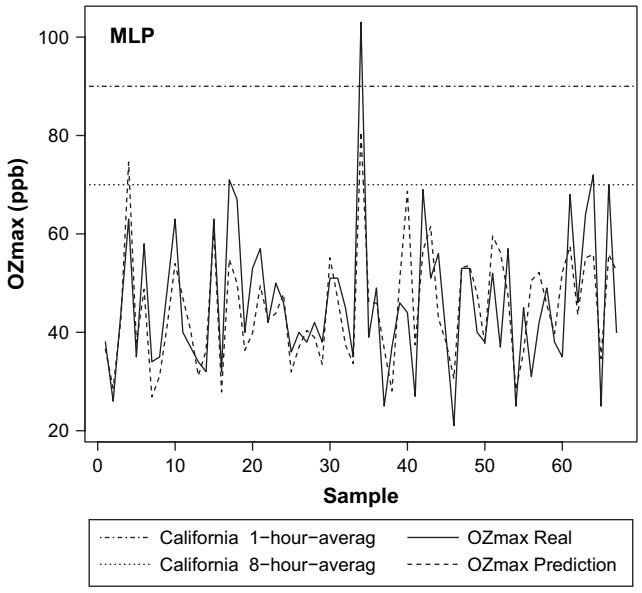

Fig. 9. Maximum ozone prediction sample of the best model: MLP.

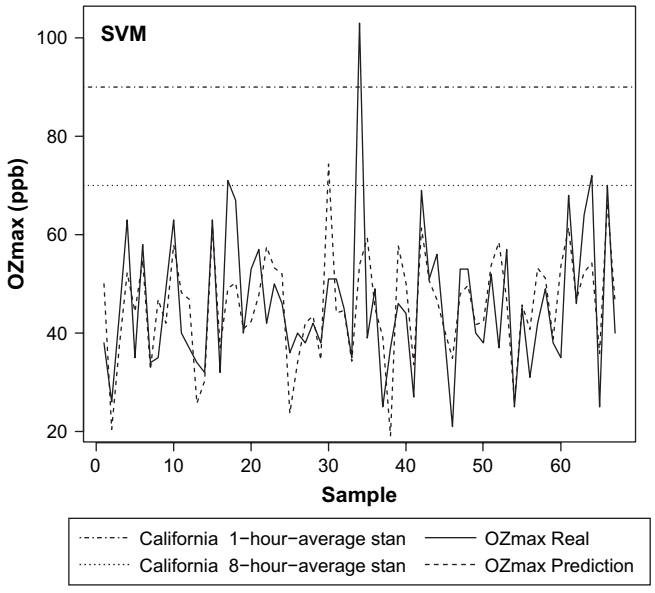

Fig. 11. Maximum ozone prediction sample of the third best model: SVM.

models are better than those obtained with parametric and semiparametric techniques. Specifically, the MLP model of data Matrix A displayed a prediction capacity with better performance measurements than those reported in the literature for similar cases (Sokhi et al., 2006; Aguirre-Basurko et al., 2006; Sousa et al., 2007; Dudot et al., 2007; Han, 2007). Accordingly, the MLP model developed for this case study could be used as part of a strategy to manage ozone precursor sources and thus helps to improve the environment for a significant population on both sides of the frontier.

The precision of the predictions in this study was better with information 24 h in advance than with predictors using only 6 h of advance information. Furthermore, it is notable that the SVM model was less tolerant to noise than the MLP model and produced average quality models. Something similar occurred with the ELMAN network model, although the training time was around 41% less than with MLP; MLP offered better prediction quality, ELMAN did not match the precision achieved by the best MLP, situated in an intermediate area between these and the SVM and the generalized Linear models.

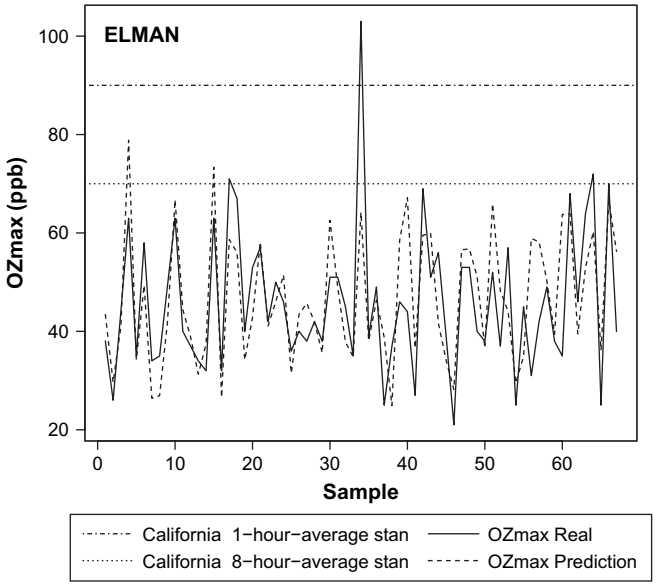

Fig. 10. Maximum ozone prediction sample of the second best model: ELMAN.

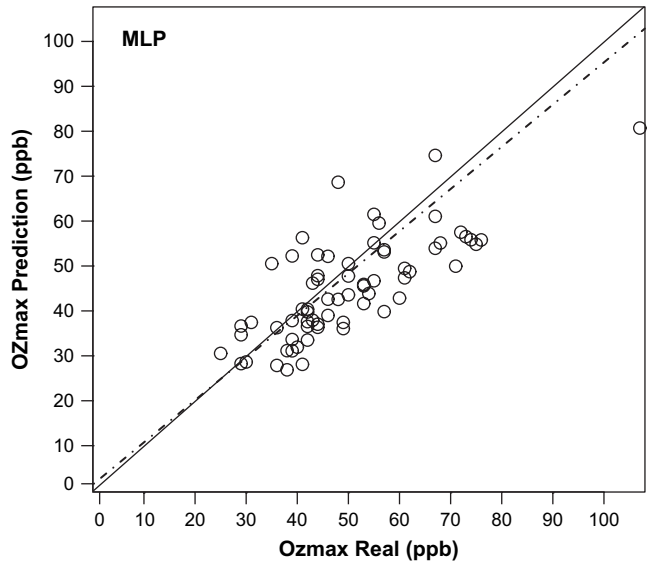

Fig. 12. Scatter plots of a sample of forecasted and observed values corresponding to the MLP model.

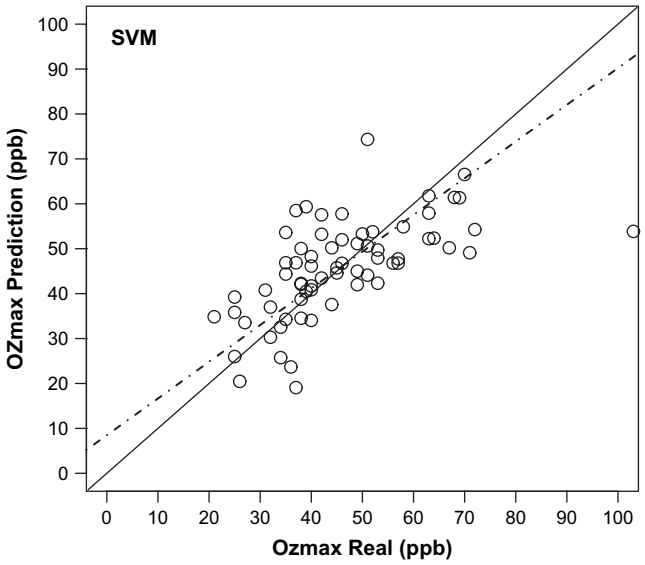

Fig. 13. Scatter plots of a sample of forecasted and observed values corresponding to the SVM model.

Models using the Ridge Regression method did not significantly improve on results. Data preprocessing eliminates the ozone correlations with the most important variables, like NO_x, and for this reason the presence of collinearity was very low.

The development of effective ozone concentration prediction models poses a major challenge. The management of ozone control and public protection activities requires accurate forecasts. Although many ozone prediction models have been developed and some of them are in use, there is a pressing need for accurate models capable of determining the relative importance of environmental variables.

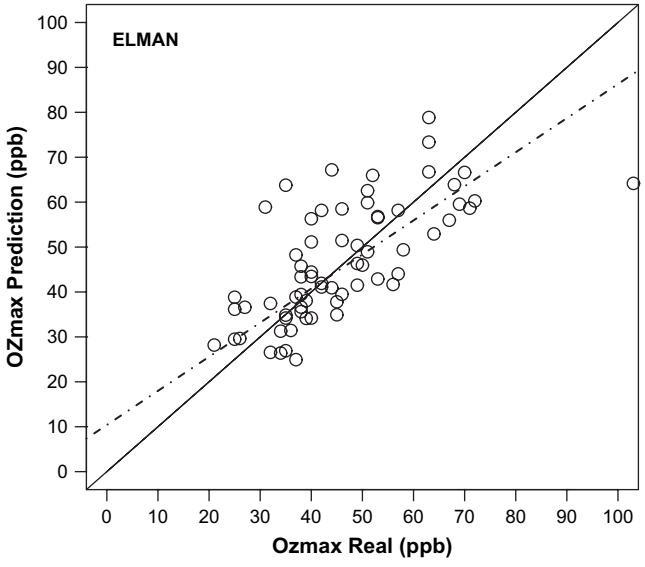

Fig. 14. Scatter plots of a sample of forecasted and observed values corresponding to the ELMAN model.

Acknowledgements

We gratefully acknowledge the financial support of the Spanish Government (DPI2007-61090), and Mexico's *Sistema Nacional de Educación Superior Tecnológica*, *Dirección General de Educación Tecnólogica* and the SES-ANUIES 2007 agreement for their financial support, as well as the US-EPA, Cal/EPA, ARB, Bob Weller, Virginia Ambrose and Gabriel Ruiz for the data provided, which have made this work possible.

References

Aguirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environmental Modelling and Software 21, 430–446.

Bishop, C.M., 1997. Neural Networks for Pattern Recognition. Oxford University Press Inc., New York.

CARB, 2005. California Air Resources Board 2005. Available from: <http://arb.ca.gov/research/aaqs/caaqs/ozone/ozone.htm>.

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 Industriales. Universidad de La Rioja, España.

Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. Meteorology and Atmospheric Physics 87, 167–196.

Chang, J.C., Hanna, S.R., 2005. Model Validation Kit (MvK). BOOT Statistical Model Evaluation Software Package. National Environmental Research Institute, Ministry of the Environment, Denmark.

Chelani, A.B., Chalapati Rao, C.V., Phadke, K.M., Hasan, M.Z., 2002. Prediction of sulphur dioxide concentration using artificial neural networks. Environmental Modelling & Software 17, 161–168.

Clapp, L.J., Jenkin, M.E., 2001. Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NO_x in the UK. Atmospheric Environment 35 (36), 6391–6405.

Davis, D.D., Orendovici, T., 2006. Incidence of ozone symptoms on vegetation within National Wildlife Refuge in New Jersey, USA. Environmental Pollution 143 (3), 555–564.

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

Dudot, A.L., Rynkiewicz, J., Steiner, F.E., Rude, J., 2007. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. Environmental Modelling and Software 22, 1261–1269.

European Commission, 1994. The Evaluation of Models of Heavy Gas Dispersion. Model Evaluation Group Seminar. Office for Official Publications of the European Communities L-2985, Luxemburg.

Filleul, L., Cassadou, S., Médina, S., Fabres, P., Lefranc, A., Eilstein, D., Le Tertre, A., Ledrans, M., 2006. The relation between temperature, ozone, and mortality in nine French cities during the heat wave of 2003. Environmental Health Perspectives 114 (9), 1344–1347.

Fischer, P., Hoek, G., Brunekreef, B., Verhoeff, A., Van Wijnen, J., 2003. Air pollution and mortality in the Netherlands: are the elderly more at risk? European Respiratory Journal Supplement 21 (40), 34S–38S.

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.

Frost, G.J., Trainer, M., Allwine, G., Buhr, M.P., Calvert, J.G., Cantrell, C.A., Fehsenfeld, F.C., Williams, E.J., 1998. Photochemical ozone production in the rural southeastern United States during the 1990 Rural Oxidants in the Southern Environment (ROSE) program. Journal of Geophysical Research D: Atmospheres 103 (D17), 22491–22508.

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.

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

Gardner, M.W., Dorling, S.R., 2000a. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmospheric Environment 34 (1), 21–34.

Gardner, M.W., Dorling, S.R., 2000b. Meteorologically adjusted trends in UK daily maximum surface ozone concentrations. Atmospheric Environment 34 (2), 171–176.

Guicherit, R., Roemer, M., 2000. Tropospheric ozone trends. Chemosphere – Global Change Science 2 (2), 167–183.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London, ISBN 9780412343902.

Han, Z.W., 2007. A regional air quality model: evaluation and simulation of O_3 and relevant gaseous species in East Asia during spring 2001.

Haykin, S., 1994. Neural Networks. Macmillan College Publishing Company, Inc, New York.

Haykin, S., 1999. 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.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software 21 (5), 602–614.

Kleinman, L.I., 2000. Ozone process insights from field experiments – part II: observation-based analysis for ozone production. Atmospheric Environment 34 (12–14), 2023–2033.

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

Kongtip, P., Thongsuk, W., Yoosook, W., Chantanakul, S., December 2006. Health effects of metropolitan traffic-related air pollutants on street vendors. Atmospheric Environment 40 (37), 7138–7145.

La Paz Agreement, 1983. Agreement between the US and Mexico on Cooperation for the Protection and Improvement of the Environment in the Border Area, 14 August 1983. La Paz Baja California Sur, Mexico, TIAS No. 10827.

Lee, D.S., Holland, M.R., Falla, N., 1996. The potential impact of ozone on materials in the UK. Atmospheric Environment 30 (7), 1053–1065.

Martínez-De-Pisón, F.J., Alba-Elías, F., Castejón-Limas, M., González-Rodríguez, J.A., 2006. Improvement and optimisation of hot dip galvanising line using neural networks and genetic algorithms. Ironmaking and Steelmaking 33 (4), 344–352.

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

Monks, P.S., 2000. A review of the observations and origins of the spring ozone maximum. Atmospheric Environment 34 (21), 3545–3561.

Mukerjee, S., 2001. Selected air quality trends and recent air pollution investigations in the US–Mexico border region. 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. Science of the Total Environment 276, 205–224.

Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E., 2005. Neural network prediction model for fine particulate matter (PM 2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environmental Modelling and Software 20 (5), 547–559.

Pernía-Espinoza, A., Castejón-Limas, M., González-Marcos, A., Lobato-Rubio, V., 2005. Steel annealing furnace robust neural network model. Ironmaking and Steelmaking 32 (5), 418–426.

Pires, J., Martins, F., Sousa, S., Alvim-Ferraz, M., Pereira, M., 2008. Selection and validation of parameters in multiple linear and principal component regressions. Environmental Modelling and Software 23 (1), 50–55.

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.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, New York.

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.

Scebba, F., Canaccini, F., Castagna, A., Bender, J., Weigel, H.-J., Ranieri, A., 2006. Physiological and biochemical stress responses in grassland species are influenced by both early-season ozone exposure and interspecific competition. Environmental Pollution 142 (3), 540–548.

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E., 2006. Statistical models to assess the health effects and to forecast ground-level ozone. Environmental Modelling and Software 21, 547–558.

Sillman, S., 1999. The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. Atmospheric Environment 33 (12), 1821–1845.

Sokhi, R.S., San Jose, R., Kitwiroon, N., Fragkou, E., Perez, J.L., Middleton, D.R., 2006. Prediction of ozone levels in London using the MM5-CMAQ modelling system. Environmental Modelling and Software 22 (1), 97–103.

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

The U.S.–Mexican border environment: air quality issues along the U.S.–Mexican border. In: SCERP Monograph Series, No. 6, 2003. San Diego State University Press.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment 35, 617–630.

Trainer, M., Parrish, D.D., Goldan, P.D., Roberts, J., Fehsenfeld, F.C., 2000. Review of observation-based analysis of the regional factors influencing ozone concentrations. Atmospheric Environment 34 (12–14), 2045–2061.

US-EPA and Secretaría de Desarrollo Urbano y Ecología SEDUE, 1991. Integrated Environmental Plan for the Mexican–US Border Area, First Stage, 1992–1994. Washington, DC.

US-EPA, October 1996. U.S.–Mexico Border XXI Program Framework Document. EPA 160-R-96-003. Washington, DC.

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

US-EPA, 2000b. Summary of Selected Environmental Indicators From the US–Mexico Border XXI Program: Progress Report 1996–2000. EPA 909-R-00-002. 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.

US-EPA and Secretaría de Medio Ambiente y Recursos Naturales SEMARNAT, 2003. Border 2012/Frontera 2012: U.S.–Mexico Environmental Program. Washington, DC.

US-EPA, 2006. National Ambient Air Quality Standards (Clean Air Act). Available from: http://www.epa.gov/air/criteria.html.

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

Weschler, C.J., 2006. Ozone's impact on public health: contributions from indoor exposures to ozone and products of ozone-initiated chemistry. Environmental Health Perspectives 114 (10), 1489–1496.

Wilks, D.S., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press, San Diego, CA, 467 pp.