NONLINEAR MODELS FOR GROUND-LEVEL OZONE FORECASTING

S. Bordignon, C. Gaetan, F. Lisi

2001.11

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Maggio 2001

Nonlinear models for ground-level ozone forecasting

Silvano Bordignon, Carlo Gaetan and Francesco Lisi
Dipartimento di Scienze Statistiche
Università di Padova

Summary

One of the main concerns in air pollution is excessive tropospheric ozone concentration. The aim of this work is to develop statistical models giving short-term prediction of future ground-level ozone concentrations. Since there are few physical insights about the dynamic relationship between ozone, precursor emissions and/or meteorological factors, a nonparametric and nonlinear approach seems promising in order to specify the prediction models. First, we apply four nonparametric procedures to forecast daily maximum 1-hour and maximum 8-hours averages of ozone concentrations in an urban area. Then, in order to improve the prediction performances, we combine the time series of the forecasts. This idea seems to give promising results.

Keywords: Ground-level ozone forecasting, Nonlinear time-series models, Combination of forecasts.

1 Introduction

Ground-level ozone is the primary constituent of photooxidative smog. It is recognized that ozone concentrations are increasing steadily on larger part of the northern hemisphere and that high concentrations of ozone have negative effects on vegetation, human health and various materials. The importance of ozone as an air quality parameter has induced most countries to adopt legislative measures establishing national air quality standards. Some of these include mandatory public warnings and traffic restrictions.

The institution of public information systems and of possible sanctions in case where ozone limits are exceeded has increased the demand for effective prediction models for maximum ozone concentrations.

More explicitly the goals of forecasting are to provide information in order:

1. to satisfy needs of public information;

2. to further reduce and prevent exposure;
3. to alert authorities, industries and the public to take short-term measures for emission reduction during smog-episodes;
4. to increase public support for structural measures for emission reductions.

These goals require reliable information and forecasts on a timely basis. Typically a forecast should be available at least for one-day in advance, since the time required to prepare emission reduction measures is at least one day and preferentially a few days, depending on the logistics.

Forecasting models for ozone concentrations could be based on deterministic equations derived from theories related to physical and chemical processes in the atmosphere (Seinfeld, 1986). However, such models are unsuitable in many operational settings because they require significant computer and staffing commitments as well as many complex chemical inputs. When these inputs are not known, the transfer and application of a model from one region to another is problematic. Further, reliable emission inventories are indispensable for these kind of models, but they are only partly existent and are limited in their regional validity. In addition, due to the influence of meteorological conditions on ozone concentrations and large uncertainty associated with input weather data, it is very difficult to obtain a good agreement between the prediction model and the observed data.

Because of these inherent difficulties, stochastic models mainly based on regression methods that include past values of ozone and ozone precursor, such as NO_2 and NO, and meteorological conditions as inputs have been widely employed as an alternative to deterministic models to forecast the ozone concentrations.

In early regression models a linear specification was often adopted (Milionis and Davies, 1994; Ryan 1995), and for this reason such models have not been proved always satisfactory, specially in providing accurate prediction in situations of forthcoming pollution episodes. Nowadays it is widely recognized that the relationship among ground-level ozone concentrations, ozone precursor and meteorological conditions, may be complex and highly nonlinear.

Some recent examples report results from nonlinear multiple regression (Coburn and Hubbard, 1999), artificial neural networks (Comrie, 1997), (Prybutok *et al.*, 2000), additive models (Niu, 1996), (Davis and Speckman, 1999), CART model (Burrows *al.*, 1995) and even fuzzy-logic-based models (Jorquera *et al.*, 1998).

Thus, according to this recent trend in the literature, in this paper we adopt a nonlinear point of view for modelling and predicting ozone concentration. However, rather than searching for specific nonlinear parametric models, for which the number and the importance of parameters can vary significantly with the specific site considered, or concentrating on a single nonlinear procedure, we explore various approaches based on nonlinear black-box modelling (Sjöberg *et al.*, 1995).

In particular, we construct nonparametric predictive models for ozone directly from time series data using some recent methods that combine broad approximation abilities and few specific assumptions according to a theory-poor and data-rich perspective.

The main features of our approach that distinguish it from previous studies on modelling and forecasting ozone concentrations are: (i) the development and application of different statistical nonparametric procedures to the same data sets (still quite infrequent in the literature); (ii) improvement of the forecasting results through combination of forecasts obtained from different models; (iii) the particular attention given to specification strategies in nonparametric modelling; (iv) the use, besides the classical criteria for evaluating predictive performances, of criteria specifically developed for evaluation of ozone forecasts.

The paper is organized as follows: the next section introduces the different nonparametric strategies used for obtaining ozone forecasts; in section 3 the main characteristics of the data uses for the forecasting exercise are briefly described; section 4 provides the results, while section 5 contains the conclusions and some indications for further developments.

2 Nonlinear and nonparametric models

Let Y represent a single response variable that depends on a vector of p predictor variables $X = (X_1, \dots, X_p)'$.

In our setup Y is either the daily maximum 1-hour average or the daily maximum 8-hour average of ozone concentrations, X contain current and lagged values of meteorological variables and precursors as well as lagged values of pollutants.

We assume that T sample units of Y and X, namely $\{Y(t), X(t)\}_{t=1}^{T}$, are given and that $Y(t)$ can be described by the nonlinear regression model

$$Y(t) = g(X(t)) + \varepsilon(t). \tag{1}$$

The function g reflects the 'true' but unknown relationship between Y and X. The random additive error variable $\varepsilon(t)$ is assumed to have mean zero and variance σ_ε^2. We suppose also that $\varepsilon(t)$ is independent of $X(t)$ so the optimal Mean Square Error (MSE) forecast $\hat{Y}(t)$, given $X(t)$, is $g(X(t))$. In the literature model (1) is known as NonLinear Autoregressive model with eXogenous variables (NLARX).

Unless the dimension p is very small, the general nonparametric approach suffers from the 'curse of dimensionality'. Briefly, because the nonlinear function in (1) is multidimensional, the analysis of such a model often requires multivariate smoothing. The virtue of nonparametric smoothing is that of making use of 'local properties' of the data; for a multivariate problem, a large sample is needed to obtain reliable local estimates. Consequently, for the sample size and the number of predictor involved in our problem, nonparametric estimates of model (1) can be associated with large variations (for more details, see Hastie and Tibshirani, 1990).

To avoid this problem, various parsimonious modelling strategies have been proposed (Sjöberg *et al.*, 1995; Härdle *et al.*, 1997).

From a general point of view we can assume that the function g can be written as

$$g(X) = \sum_{k=1}^{\infty} \alpha_k g_k(X), \tag{2}$$

where $\{g_k\}$ is a basis for g.

In this framework, the (true) relationship (2) can be approximated by a finite number of basis functions

$$h(X) = \sum_{k=1}^{n} \alpha_k g_k(X). \tag{3}$$

This setup covers various modelling procedures and some of these will be considered in the sequel.

2.1 Additive Models

The simplest case of a basis function is given by the Additive Model (AM) (Hastie and Tibshirani, 1990)

$$h(X) = \sum_{k=1}^{p} g_k(X_k). \tag{4}$$

The g_ks are assumed to be unknown and are estimated nonparametrically using only univariate smoothing splines. Note that the additive model encompasses linear models (for example the AutoRegressive with eXogenous variables model) and many interesting nonlinear models as special cases.

In order to estimate the model (4) we can use the back-fitting algorithm. The main idea of back-fitting is that if the additive model (4) is correct, then, for all k, $E(Y - \sum_{j\neq k} g_j(X_j)|X_k) = g_k(X_k)$. Consequently, we can treat $Y - \sum_{j\neq k} g_j(X_j)$ as the conditional response variable and use univariate smoothers to estimate g_k.

Since the g_ks are unknown, a starting value for all g_k is given and the estimates are iterated until the convergence is reached. The estimation procedure can be coupled with the selection of the smoothing parameters for each g_k using a generalized cross-validation criterion (for more details see Hastie and Tibshiran, 1990). Chen and Tsay (1993) have showed that this adaptive back-fitting, called BRUTO, is particularly useful in lag selection for AM.

2.2 Regression Trees and Multivariate Additive Regression Splines

Although in the additive modeling framework Hastie and Tibshirani (1990) suggest a number of ways of modeling interactions among predictors variables, other models, Regression Trees (RT) (Breiman *et al.*, 1984) and Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991), build these interactions directly.

A simple way to approximate g over D is splitting D into a (large) number M of disjoint hyper-rectangles $\{R_m\}_{m=1}^{M}$ and for each R_m using a constant α_m to estimate the value of g in R_m. A natural estimate for α_m is the average of those y values whose X values fall into R_m. In the recursive partitioning procedure (Breiman *et al.*, 1984) the hyper-rectangles R_m are determined starting with a single region $R_1 = D$, recursively splitting existing sub-regions into two halves and discharging parent sub-regions, until a large tree is developed with each terminal sub-region containing only a few observations. The over-sized tree is then pruned according to a cost complexity measure (Breiman *et al.*, 1984). In this work we use the `prune.tree` procedure, as described in Vanables and Ripley (1997, pag. 425).

If we consider n of these sub-regions, the approximating function looks like

$$h(X) = \sum_{k=1}^{n} \alpha_k B_k(X), \tag{5}$$

where $B_k(\cdot)$ is the indicator function of R_k. It is easy to see that this indicator function can be represented by a product of step functions

$$T(z) = 1, \quad \text{if } z \geq 0, \qquad T(z) = 0, \quad \text{if } z < 0.$$

While recursive partitioning is computationally fast and suitable to explore high dimensional approximation problems, there are some drawbacks. First the approximating function is necessarily discontinuous on the boundaries of the adjacent sub-regions. This is disconcerting if we believe g continuous. Further, recursive partitioning has an innate inability to adequately estimate functions that are linear or additive. To overcome these difficulties, Friedman (1991) proposed two important variants. The first is to replace step functions by truncated power splines of the first order $s_u(z) = (z-u)_+$, where u denotes a real number called knot. In the second variant the parent region R_k is not automatically eliminated for creating sub-regions but in subsequent iterations both the parent region and its corresponding sub-regions are eligible for further splitting. The final form of the MARS model is

$$h(X) = \sum_{k=1}^{n} \alpha_k S_k(X), \tag{6}$$

where S_k is the product basis function associated with the sub-region R_k.

Since for a given set of $\{B_k\}$ or $\{S_k\}$, the values of partition points or knots are fixed, MARS model is substantially a linear model where the parameters α_k may be determined by straightforward application of least squares algorithms. Similar to the recursive partitioning procedure, the process to build a MARS model consists of a first step in which a quite large number of product basis functions is used, followed by a selection step in which a generalized cross-validation criterion is used (Friedman, 1991).

2.3 Neural Networks

If we choose $g_k(X) = \sigma(\beta_k X + \gamma_k)$, where σ is an activation function and β_k is parameter vector of size p, we obtain

$$h(X) = \sum_{k=1}^{n} \alpha_k \sigma(\beta_k X + \gamma_k). \tag{7}$$

This model is referred to as a feed-forward Neural Networks (NN) with p input units, one hidden layer and one output unit. The most common choice in the NN literature (Hertz *et al.*, 1993) is $\sigma(z) = 1/(1 + e^{-z})$. NN are universal approximators (Hornik *et al.*, 1989) in the sense that (7) can approximate any continuous function on compact sets, by increasing the number n of the units in the hidden layer. The approximation results are non-constructive, and the parameters $(\alpha_k, \beta_k, \gamma_k)$ have to be chosen using the observed data (training phase). A common choice is to minimize the error function

$$E = \sum_{t=1}^{N} (Y(t) - h(X(t)))^2 .$$

In general the NN model is over parameterized and some reguralization technique should be used in order to restrict model complexity. In our work we have chosen

$$E + \lambda C(h),$$

where λ is a real positive number and

$$C(h) = \int \sum_i \frac{\partial^2 h(x)}{\partial x_i^2} dx.$$

2.4 Model identification

Let us give now some remarks about the identification procedure for the previous models. We have noted that for each model the estimation phase can be coupled with a model selection step following to an automatic criterion. But model validation forms the final stage of any model identification. This point has not been frequently considered in the environmental literature.

On the other hand, Billings and Voon (1986) have remarked that classical tests (Box and Jenkins, 1976) based on the autocorrelation function (ACF) of the residuals $\hat{\varepsilon}$ and on the cross-correlation function (CCF) between $\hat{\varepsilon}$ and the input variable X_i can provide incorrect information whenever nonlinear effects are present in the data. For this reason, Billings and Voon have proposed to inspect the estimated ACF and CCF

$$\rho(\hat{\varepsilon}), \quad \rho(\hat{\varepsilon}, X_i) \quad \rho(\hat{\varepsilon}, X_i^2), \quad \rho(\hat{\varepsilon}^2), \quad \rho(\hat{\varepsilon}, \hat{\varepsilon}X_i), \tag{8}$$

in order to find possible model inadequacies. For instance, significative cross-correlation at lag τ indicate that the variable $X_i(t-\tau)$ should be included in the model.

The model identification procedure we have adopted can be summarized in the following steps.

1. Perform a preliminary analysis of scatterplots between ozone concentrations and meteorological as well as precursors variables. Such analysis suggests which variables should be included in the model.
2. Estimate the suggested model.
3. We look for possible inclusions of omitted lags by estimating (8). If the model seems adequate, go to the next step, else return to step 2.
4. Try to simplify the model using the opportune selection criterion until the estimated (8) do not suggest any inadequacy.

2.5 Combined forecasts

In the previous subsections we have mentioned that the optimal MSE predictor for the model (1) is g. For nonparametric estimation of g *via* approximating models, various results (Yang and Barron, 1998) have shown that with appropriate model selection criteria, the resulting predictor converges in optimal fashions without knowledge of which approximating model is the best at the given sample size. However, when several forecasting procedures are available, a difficulty in applications is the choice of the right one for the data at hand. Combining forecasts (Clemen, 1989) is a well-established procedure to improve forecasting accuracy, which takes advantage of the availability of multiple resources for data intensive forecasting. Among various combining formulations of m forecasts $\hat{Y}_i(t)$, $i = 1, \ldots, m$, we have chosen to adopt the following

$$\hat{Y}_c(t) = w_0 + \sum_{i=1}^{m} w_i \hat{Y}_i(t).$$

The weights w_i are determined by least squares regression with the inclusion of a constant. Granger and Ramanthan (1984) have shown that if the individual forecasts, $\hat{Y}_i$ are biased, then the method will be superior to the optimal one which minimizes the error variance of the combination. This conclusion is supported also by empirical findings over large sample datasets.

3 Data and preliminary analysis

The data used in this study come from the air quality monitoring network of the Padova district, located in the Veneto region in the Northeast Italy. In particular, given our primary interest toward urban air pollution, we considered ground-level ozone measures taken from only three monitoring sites situated in the town of Padova.

Of the three monitoring stations, one, denoted S1, is placed in an area of the town characterized by high population density and intense vehicular traffic; the second station, S2, is located near the hospital in an area mainly affected by vehicular traffic; finally, the third monitoring site, S3, is situated in the industrial area of the town.

The considered monitoring stations provide also measures relative to other pollutants, usually considered as ozone precursors, like the various oxides of nitrogen (NO_x). Furthermore, from the same stations, with the exception of S2, we get the data relative to the meteorological variables.

Among these variables, the most relevant ones, and thus employed in the ozone modelling procedure, turned out to be, after some preliminary analysis, temperature (T), solar radiation (R) and wind speed (V).

The data have been placed at our disposal from ARPAV (Agenzia Regionale per la Protezione dell'Ambiente del Veneto) as hourly averages over the period 1 April 1992 - 30 September 1999. However, since high levels of ozone concentrations occur mainly in worm periods, only the so-called 'ozone season' for each year, running from 1 April to 30 September, was considered.

The EEC Directive on air pollution by ozone, which is currently in force in Italy, defines the threshold value for the ozone concentration. The threshold for human health protection has been set at 110 $\mu g/m^3$ for 8-h average measures. The relevant threshold values in the context of the Directive are the population information threshold value of 180 $\mu g/m^3$ and the population warning threshold value of 360 $\mu g/m^3$ as an hourly average.

Thus, according to these guidelines, we considered for each station two daily summary ozone series, the daily maximum 1-h average and the daily maximum 8-h average respectively.

Some descriptive statistics of the ozone time series are given in Table 1, while Figures 1 show the plots of the series relatively to the period considered.

From the examination of the Table 1 and Figure 1, one can observe

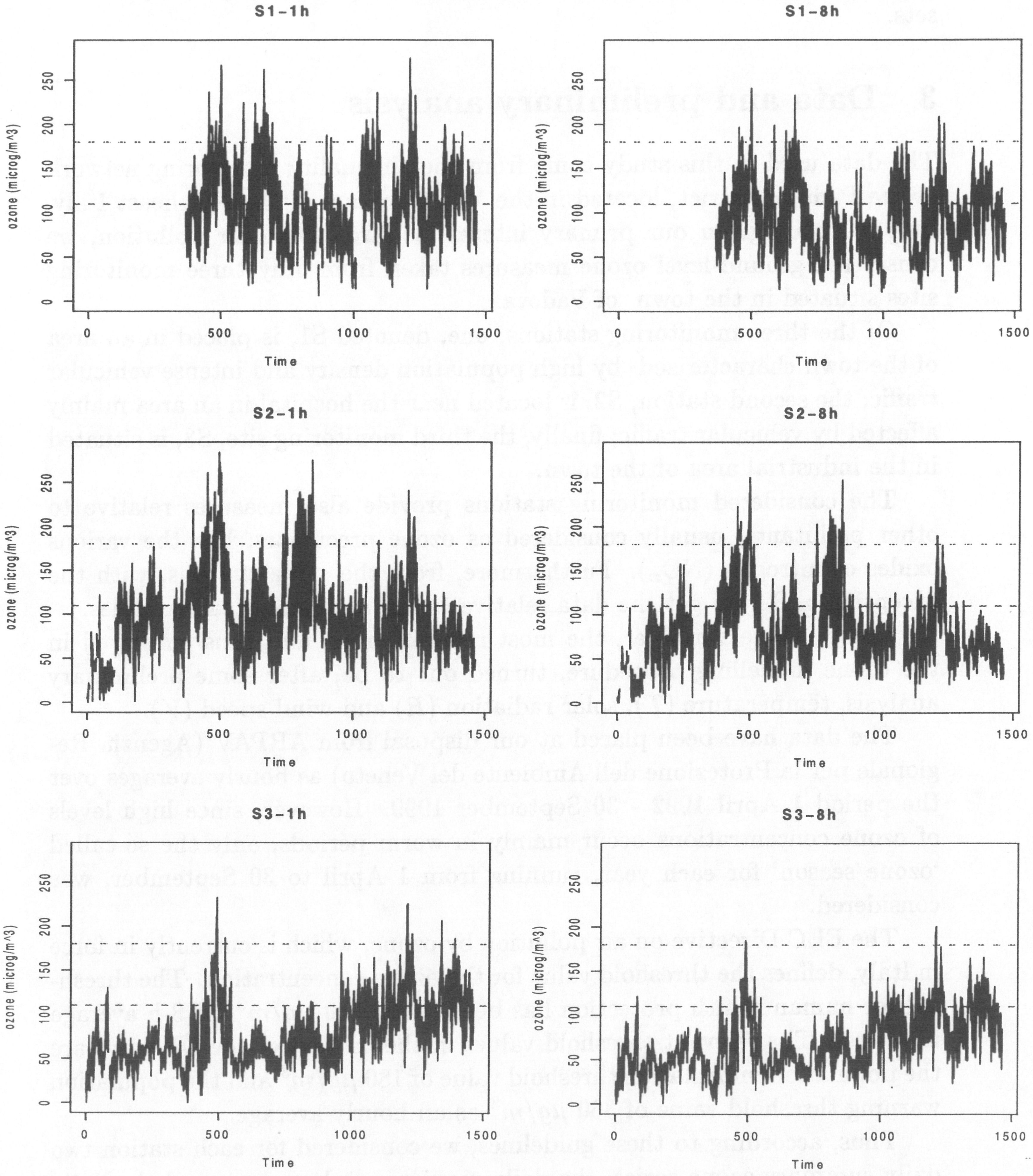

Figure 1: Time series plots of the ozone concentration in the three monitoring stations.

	n	Max	Mean	Med	S.D.
S1-1h	1099	275	110.0	107.0	45.8
S1-8h	1099	241	93.6	91.1	40.7
S2-1h	1464	284	97.0	90.0	47.7
S2-8h	1464	256	81.2	74.7	41.6
S3-1h	1464	233	78.6	71.0	36.3
S3-8h	1464	201	68.4	62.1	32.9

Table 1: Summary statistics for daily ozone concentration.

that: (i) the ozone mean level turns out to be higher in the monitoring sites located in areas with dense population and/or with intense vehicular traffic (S1 and S2). In the industrial area, S3, the ozone mean level is relatively low, although there is evidence of an increasing trend for the last year of the considered period, where more frequent exceedances over the threshold $110\mu g/m^3$ for the 8-h series are observed; (ii) as S1 and S2 are concerned, many exceedances over both the thresholds 110 $\mu g/m^3$ and 180 $\mu g/m^3$ are signaled by the 8-h and 1-h series respectively; the exceedances are concentrated mainly in the warmest months (July and August) of the ozone season.

Finally, in order to obtain from the data further useful information for the subsequent modelling procedure, we have examined the scatterplots between ozone and its potential inputs. The main results lead toward an ineffective relationship between ozone and its precursors (specially NO_2) and significant and probably nonlinear relationships between ozone and meteorological variables, in particular T, R and V (see, for example, Figure 2).

4 Results

To compare the forecasting ability of the mentioned models, the period 1/4/1992-30/9/1998 has been used as training set for model identification and estimation, while the period 1/4/1999-30/9/1999 has been left as testing set for the comparison between observed and predicted values.

The forecasting ability in the testing set has been evaluated estimating the models every month and predicting the values for the next month. As the objectives of a forecast may widely differ, and there is not a single evaluation procedure, several performance indicators have been considered. A first set is intended to evaluate the numerical information. This includes classical measures as (i) the mean error (ME), that indicates how much observed concentrations are over or under-predicted; (ii) the root mean squared error (RMSE); (iii) the mean absolute error (MAE), which has the benefit of being not sensitive to outliers; (iv) the correlation coefficient between observed and

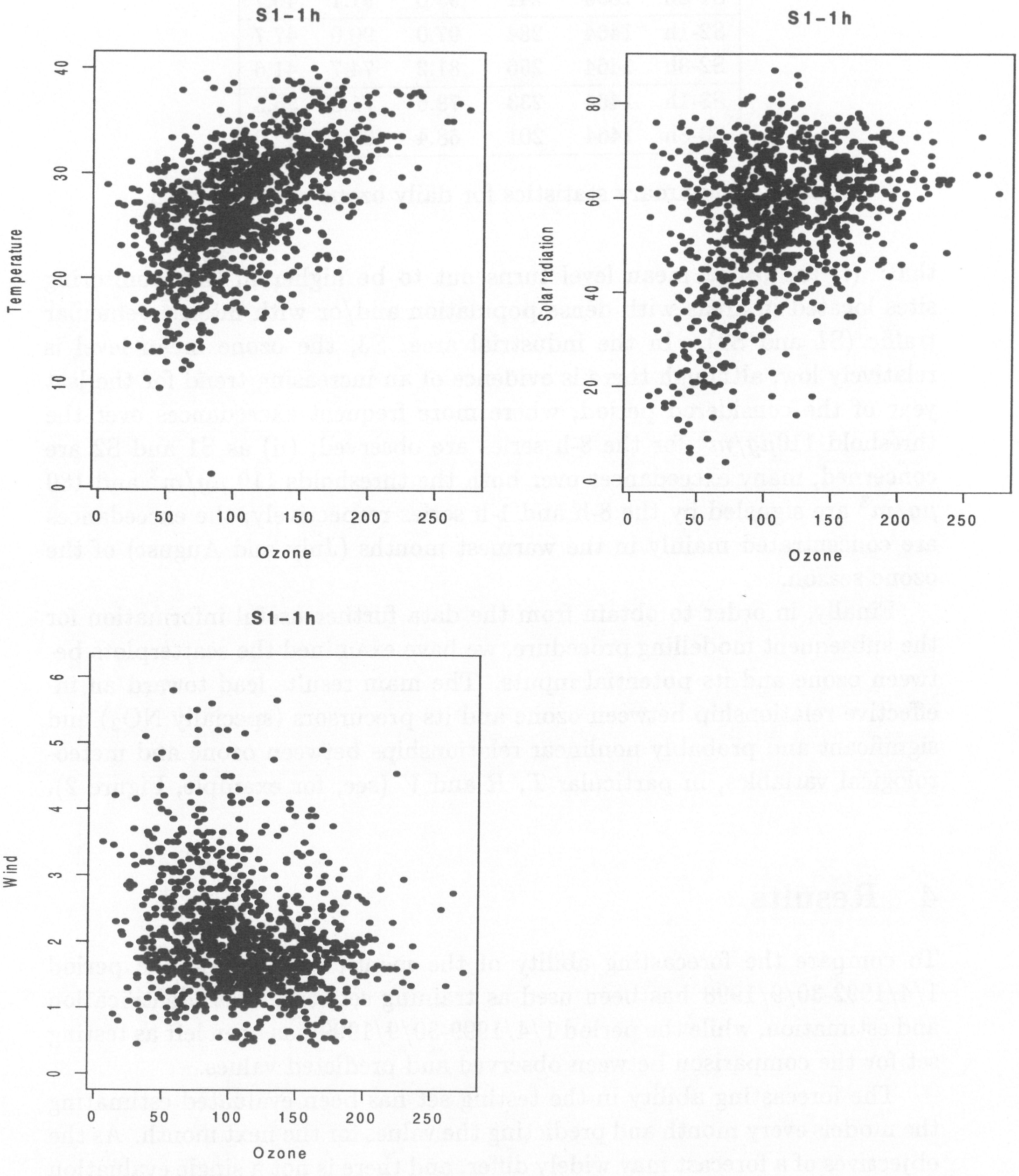

Figure 2: Scatterplots between ozone 1-h average concentration and meteorological variables (S1 monitoring station).

predicted values (CORR).

A second set of indicators has been considered, designed for evaluating qualitative information such as if forecasts/measurements are above/below a threshold value. In fact, in evaluating an environmental warning system, there are two key points to consider: (i) the proportion of exceedances or events correctly predicted by the model; (ii) the number of false alarms. Large percentage of successful forecasts jointly with a percentage of false alarms as small as possible are desirable.

Denoting f the total number of forecasted events, m the total number of observed exceedances and a the correctly forecasted events, the following criteria are suggested (De Leeuwe, 2000):

1. the percentage of correct forecast events

$$SP = \frac{a}{m} \cdot 100\%,$$

2. the percentage of realized forecast events

$$SR = \frac{a}{f} \cdot 100\%,$$

3. the skill of correctly forecasting non exceedances

$$CN = \frac{N + a - m - f}{N - m} \cdot 100\%,$$

4. the ratio of correct forecast events and total potential risk events

$$ST = \frac{a}{(m + f - a)} \cdot 100\%,$$

5. the ability of a correct forecast of the exceedances

$$SI = \frac{a + (N + a - m - f)}{N} \cdot 100\%.$$

Prediction horizons of 1, 2 and 3 days ahead are considered, both for daily maximum 1-hour and 8-hours mean ozone concentrations. However, since conclusions for different horizons are basically the same, here only one-day ahead results are reported.

Models identification has been carried out in a step sequence using the procedure described in subsection 2.4. This means to analyze pictures like Figure 3.

The whole procedure led to identify the models in Table 2, specified for the different classes of models, both for the daily maxima 1-hour and 8-hours mean ozone concentrations. In Table 2, for example, the expression

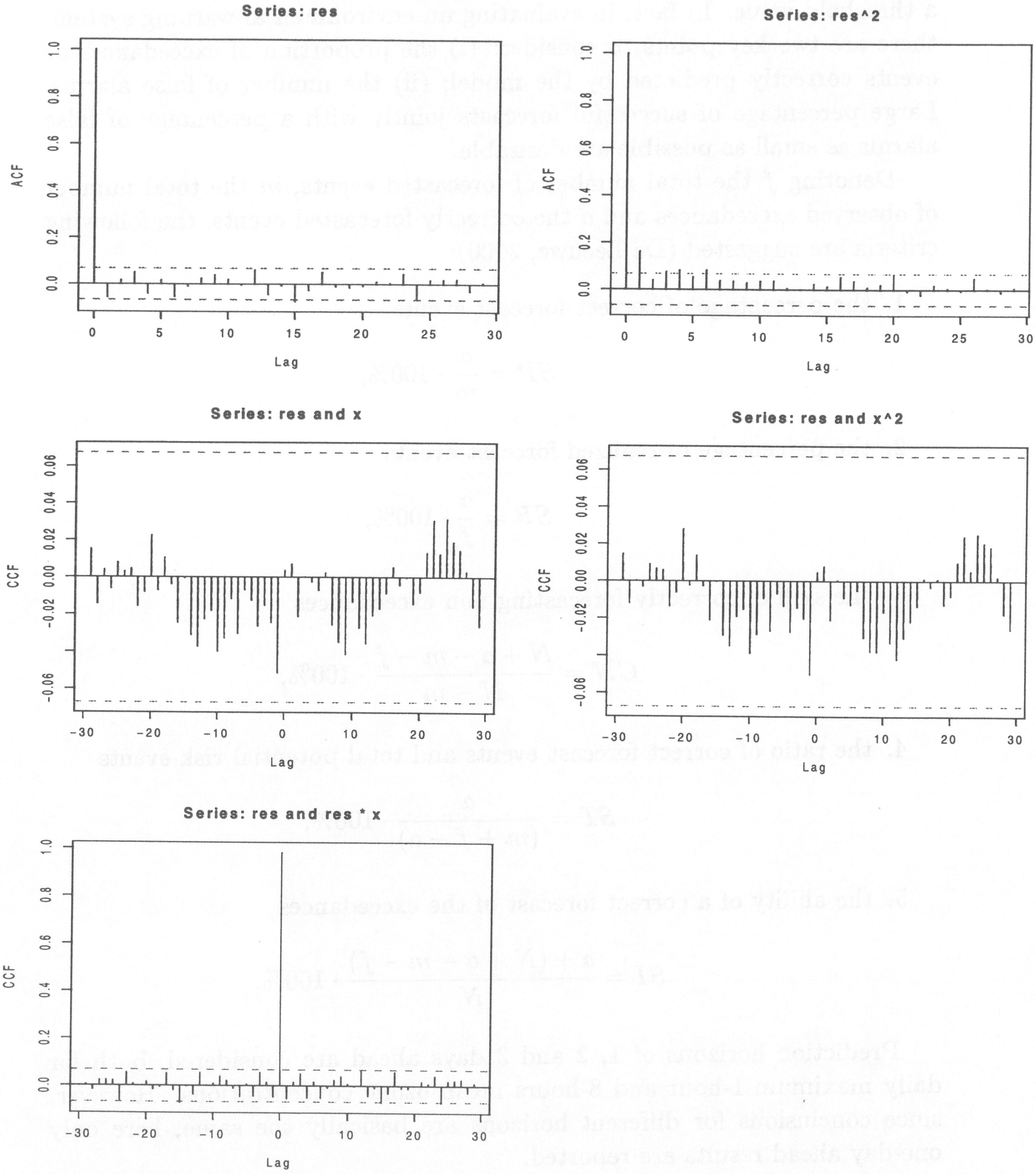

Figure 3: Examples of estimated ACF and CCF.

S1-1h	
ARX	O3(1,6,12,23), t(0,1), r(0,1), v(0,1)
AM	O3(1,6,11,23), t(0,1), r(0), v(0)
MARS	O3(1,6,11,23), t(0,1), r(0,1), v(0)
NN	O3(1,6,23), t(0,1), r(0,1), v(0)
RT	O3(1), t(0,1), r(0), v(0)
COMB	ARX+AM
S1-8h	
ARX	O3(1,6), t(0,1), r(0,1,7), v(0)
AM	O3(1,6), t(0,1), r(0,7), v(0)
MARS	O3(1,6), t(0,1), r(0,1,7), v(0)
NN	O3(1,6), t(0,1), r(0,1,7), v(0)
RT	O3(1,6), t(0,1), r(0,1,7), v(0)
COMB	ARX+AM+NN+RT
S2-1h	
ARX	O3(1,2,6,7,14), t(0,1,7), r(0,1,6,7), v(0,1)
AM	O3(1,2,6,7,14), t(0,1,8), r(0,1,8), v(0,1)
MARS	O3(1,2,3,6,7,13,14), t(0,1,6), r(0,1,6), v(0)
NN	O3(1,2,3,6,7,14), t(0,1,8), r(0,1,6,8), v(0,1)
RT	O3(1,2,6,7), t(0,1,3), r(0,7), v(0)
COMB	ARX+AM+MARS
S2-8h	
ARX	O3(1,3,6,7,14), t(0,1,8), r(0,8), v(0,5)
AM	O3(1,3,4,5,6,7,13), t(0,1,6), r(0,8), v(0,1)
MARS	O3(1,3,7), t(0,1,7), r(0), v(0)
NN	O3(1,3,4,6,7), t(0,1), r(0), v(0,1)
RT	O3(1,3,4,7), t(0,1,3), r(0), v(0)
COMB	ARX+AM+RT
S3-1h	
ARX	O3(1,2,6,7,11,14), t(0,1,7), r(0,1,2,5,7), v(0)
AM	O3(1,2,6,7,14), t(0,1,7), r(0,1,6), v(0)
MARS	O3(1,3,6,14), t(0,1,12), r(0,1,6), v(0)
NN	O3(1,2,3,4,6,14), t(0,1,8), r(0,1,6), v(0,1)
RT	O3(1,7,14), t(0,1,4), r(0,1), v(0)
COMB	ARX+AM+RT
S3-8h	
ARX	O3(1,2,6,7,14), t(0,1,7), r(0,1,6,7), v(0,1)
AM	O3(1,2,6,7,14), t(0,1,8), r(0,1,8), v(0,1)
MARS	O3(1,2,3,6,7,13,14), t(0,1,6), r(0,1,8), v(0)
NN	O3(1,2,3,6,7,14), t(0,1,8), r(0,1,6,8), v(0,1)
RT	O3(1,2,6,7), t(0,1,3), r(0,7), v(0)
COMB	ARX+AM

Table 2: Identified models with predictors variables.

O3(1,6,12,23) means that the delays 1, 6, 12 and 23 of the ozone variable enter the model as predictors.

In the training set these models resulted, in the whole, sufficiently suitable to describe the data. The only unsatisfactory point is the permanence of some slight significative autocorrelations for squared residuals, pointing out a probable heteroskedasticity in the data.

Concerning the predictive abilities of these models, a summary of the results provided by the indicators, referred to the entire 1999, is contained in Table 3. It shows that, in general and except for RT, nonlinear predictors behave better than the linear one and that, among nonlinear predictors, AM gives relatively better results. These findings are consistent with other recent work referred to the Italian context (for example Finzi *et al.*, 1999) and confirm the effectiveness of nonlinear modelling for ozone forecasting. Furthermore, if we compare these results with those obtained from the benchmark persistent model, as suggested by De Leeuwe, (2000), it is manifest that all models behave clearly better.

However, it must also be noted that there is not a unique model that definitely outperforms the others with respect to all indicators. For this reason, and to further improve the forecasts, a linear combination of the different predictors (denoted by COMB) has been considered. The results are given in Table 3 together with the models that entered the best combination. In the whole, the combined forecasts produced good results.

Let us consider, for example, the S3 station, which is interesting because in 1999 it shows high levels of ozone concentrations and several threshold excedancees. For this station the best model resulted to be a linear combination of ARX, AM e RT for the daily maximum 1-hour average ozone concentrations, and of ARX e AM for daily maximum 8-hours average. It is worth noting that the models entering the combination giving the best prediction are not those who perform better separately. Further, the combined forecasts give for most of indicators the best performance or a performance very near to the best.

If we consider again the S3 station, the human health protection level has been exceeded several times. Vice versa, the attention level has never been reached. For this reason, in our experiments, the threshold has been put at 140 $\mu g/m^3$, which is a value close to the mean of the whole series plus twice its standard deviation.

The results of qualitative information, related to S1 and S3, are given in Table 4. The lack of results for S2 is due to the absence of exceedances for this station during 1999. Table 4 shows that, also with respect to threshold exceedances, the best performances are those related to the combinations previously described. Individually, instead, the models giving the best results are MARS e AM.

S1-1h	ME	MSE	MAE	CORR
RW	-0.142	28.611	22.472	0.615
ARX	1.984	19.994	15.624	0.794
AM	0.816	18.788	14.994	0.822
MARS	1.027	19.271	15.534	0.812
NN	1.763	20.645	15.978	0.789
RT	1.112	21.276	17.027	0.774
COMB=ARX+AM	0.869	18.964	15.076	0.814
S1-8h	ME	MSE	MAE	CORR
RW	-0.644	25.942	20.673	0.623
ARX	1.717	17.931	13.807	0.819
AM	1.537	17.394	13.414	0.831
MARS	1.668	17.317	13.134	0.834
NN	1.105	17.664	13.862	0.837
RT	1.144	17.826	13.798	0.837
COMB=ARX+AM+NN+RT	0.724	16.700	13.008	0.855
S2-1h	ME	MSE	MAE	CORR
RW	-0.493	21.043	16.363	0.613
ARX	-1.662	13.851	10.769	0.767
AM	-3.276	13.930	11.225	0.799
MARS	-2.278	13.934	11.315	0.790
NN	-2.712	13.967	11.300	0.772
RT	-4.924	16.826	12.628	0.687
COMB=ARX+AM+MARS	-3.221	14.180	11.407	0.803
S2-8h	ME	MSE	MAE	CORR
RW	-0.558	16.780	12.963	0.630
ARX	-0.151	10.430	8.099	0.788
AM	-0.512	9.411	7.464	0.841
MARS	-1.162	11.306	8.627	0.767
NN	-1.868	11.540	8.589	0.740
RT	-3.285	11.602	8.822	0.798
COMB=ARX+AM+RT	-1.158	9.487	7.235	0.844
S3-1h	ME	MSE	MAE	CORR
RW	0.115	27.224	21.961	0.544
ARX	2.473	18.441	14.674	0.774
AM	2.697	16.659	13.334	0.822
MARS	3.381	17.108	13.826	0.813
NN	2.452	17.958	14.367	0.790
RT	4.048	18.043	14.050	0.803
COMB=ARX+AM+RT	2.211	15.155	11.986	0.855
S3-8h	ME	MSE	MAE	CORR
RW	-0.143	25.057	20.184	0.550
ARX	2.919	14.778	12.071	0.845
AM	2.290	13.917	11.409	0.857
MARS	2.302	14.535	11.961	0.847
NN	3.141	15.425	12.931	0.829
RT	5.073	15.504	12.063	0.847
COMB=ARX+AM	2.187	13.146	10.353	0.874

Table 3: Prediction results for daily maximum 1-hour and 8-hours mean ozone concentrations.

S1-1h	SP	SR	CN	ST	SI
ARX	58.6	89.5	97.4	54.8	86.6
AM	62.1	90	97.4	58.1	87.7
MARS	65.6	95	98.7	63.3	89.6
NN	60	90	97.6	56.2	87.6
RT	77.1	65.8	86.9	55.1	84.5
COMB	62.1	90	97.3	58.1	87.5
S1-8h	SP	SR	CN	ST	SI
ARX	58.6	89.5	97.4	54.8	86.6
AM	62.1	90	97.4	58.1	87.7
MARS	65.6	95	98.7	63.3	89.6
NN	60	90	97.6	56.2	87.6
RT	77.1	65.8	86.9	55.1	84.5
COMB	62.1	90	97.3	58.1	87.5
S3-1h	SP	SR	CN	ST	SI
ARX	52	86.7	98.1	48.1	89.0
AM	56	87.5	98.1	51.8	89.8
MARS	64	94.1	99.0	61.5	92.1
NN	62.5	93.7	99.0	60	92.0
RT	51.8	100	100	51.8	90.2
COMB	68	94.4	99.0	65.4	92.9
S3-8h	SP	SR	CN	ST	SI
ARX	64.4	82.9	92.5	56.9	82.4
AM	65.2	76.9	88.9	54.5	80.3
MARS	63.0	78.4	90.9	53.7	81.3
NN	58.7	77.1	90.9	50	79.8
RT	68.9	86.1	94.2	62	85.4
COMB	71.1	82.1	91.0	61.5	83.7

Table 4: Prediction results of qualitative indicators.

5 Concluding remarks

In this work we have been concerned with the problem of short-term prediction of ground-level ozone concentration in an urban environment. Given the complexity of the dynamic relationship between ozone, meteorological variables and/or precursor emissions, nonparametric and nonlinear approaches, that combine broad approximation abilities and few specific assumptions, have been preferred. In particular, rather than concentrating on a single procedure, the predictive performances of several nonparametric statistical procedures are compared . To improve the forecasting results also the combination of the individual forecasts obtained from the different models has been considered.

The forecasting abilities of the mentioned procedures have been evaluated through several performance indicators, which include classical summary measures of the numerical information provided by the forecasts (such as ME, RMSE, MAE and the correlation coefficient between observed and predicted values) and other measures specifically designed for evaluating the qualitative information of an air pollution warning system.

The main results obtained applying the previous procedures to the same data sets, consisting of the daily maximum 1-h and maximum 8-h averages drawn from the monitoring network of the Padova district, are briefly summarized as follows.

Generally nonlinear procedures provide better forecasting performances when compared with the benchmark persistence model and a properly specified linear regression model. These findings align with the recent trend in the literature and confirm the effectiveness of nonlinear modelling for ozone forecasting. Turning now to the comparison among the nonlinear procedures, there is no evidence of a unique model definitely outperforming the others with respect to all indicators. As a whole, it seems that the nonlinear additive model provides slightly better forecasting performances. Significant improvements with respect to the whole set of indicators are however obtained when considering the forecasts combination procedure.

This encouraging result shows that a sensible way to construct an effective ozone forecasting system could be based on the combination of forecasting procedures.

In this connection, further extensions should be explored and we are currently undertaking this task turning our efforts toward the experimentation of more recent ways to combine forecasts, based on boosting procedures (see Freund and Schapire, 1997) and artificial neural networks.

Acknowledgments

This work was supported by MURST under grant n. 9813200379: *Statistical methods for monitoring the environment and the health–environment*

interactions.

References

Billings, S.A. and Voon, W.S.F. (1986), Correlation based model validity tests for non-linear models. *International Journal of Control*, **44**, 235–244.

Box, G.E.P. and Jenkins, G.M. (1976), *Time Series analysis: Forecasting and Control* (second edition) San Francisco: Holden Day.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Burrows, W.R., Benjamin, M., Beauchamp, S., Lord, E.R., McCollor, D. and Thomson, B. (1995), CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of Applied Meteorology*, **34**, 1848–1862.

Chen, R. and Tsay, R.S. (1993), Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**, 955–967.

Clemen, R.T. (1989), Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583.

Cobourn, W.G. and Hubbard, M.C. (1999), An enhanced ozone forecasting model using air mass trajectory analysis. *Atmospheric Environment*, **33**, 4663–4674.

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

Davis, J.M. and Speckman, P. (1999), A model for predicting maximum and 8h average ozone in Houston. *Atmospheric Environment*, **33**, 2487–2500.

De Leeuwe, F.A.A.A. (2000), Criteria for evaluation of smog forecast systems. *Environmental Monitoring and Assessment*, **60**, 1–14.

Finzi, G., Bergoli, D. and Volta, M. (1999), Modelli per la previsione di episodi critici di ozono troposferico in accordo alle linee guida europee. *Atti del Convegno SCO99*, 399–404.

Freund, Y. and Schapire, R.E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman, J.H. (1991), Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–50.

Granger, C.W.J. and Ramanathan, R. (1984), Improved methods of forecasting. *Journal of Forecasting*, **3**, 197–204.

Härdle, W., Lütkepohl, H. and Chen, R. (1997), A review of nonparametric time series analysis. *International Statistical Review*, **65**, 49–72.

Hastie, T. J., Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Hertz, J. Krogh, A. and Palmer, R. (1991), *Introduction to the Theory of Neural Computation*, Reading MA: Addison-Wesley.

Hornik, K., Stinchcombe, M. and White, H. (1989), Multilayer feedforward networks. *Neural Networks*, **4**, 251–257.

Jorquera, H., Perez, R., Cipriano, A., Espejo, A., Letelier, M.V. and Acuna, G. (1998), Forecasting ozone daily maximum level at Santiago, Chile. *Atmospheric Environment*, **32**, 3415–3424.

Milionis A.E. and Davies, T.D. (1994), Regression and stochastic models for air pollution - I. Review, comments and suggestions. *Atmospheric Environment*, **28**, 2801–2810.

Niu, X. (1996), Nonlinear additive models for environmental time series with applications to ground-level ozone data analysis. *Journal of the American Statistical Association*, **91**, 1310–1321.

Prybutok, V.R., Yi, J. and Mitchell, A. (2000), Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. *European Journal of Operational Research*, **122**, 31–40.

Ryan, W.F. (1995), Forecasting severe ozone episodes in the Baltimore metropolitan area, *Atmospheric Environment*, **29**, 2387–2398.

Seinfeld, J.H. (1986)., *Atmospheric Chemistry and Physics of Air Pollution*, New York: Wiley.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P., Hjalmarsson, H. and Juditsky, A. (1995), Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, **31**, 1691–1724.

Venables, W.N. and Ripley, B.D. (1997), *Modern Applied Statistics with S-Plus* (second edition), New York: Springer.

Yang, Y. and Barron, A.R. (1998), An asymptotic property of model selection criteria. *IEEE Trans. on Information Theory*, **44**, 95–116.