

The inclusion of exogenous variables in functional autoregressive ozone forecasting

Julien Damon^{1‡} and Serge Guillas^{2*†§}

¹*Université Paris 6 (Pierre et Marie Curie) & Médiamétrie*

²*Université Paris 6 (Pierre et Marie Curie) & Ecole des mines de Douai*

SUMMARY

In this article, we propose a new technique for ozone forecasting. The approach is functional, that is we consider stochastic processes with values in function spaces. We make use of the essential characteristic of this type of phenomenon by taking into account theoretically and practically the continuous time evolution of pollution. One main methodological enhancement of this article is the incorporation of exogenous variables (wind speed and temperature) in those models. The application is carried out on a six-year data set of hourly ozone concentrations and meteorological measurements from Béthune (France). The study examines the summer periods because of the higher values observed. We explain the non-parametric estimation procedure for autoregressive Hilbertian models with or without exogenous variables (considering two alternative versions in this case) as well as for the functional kernel model. The comparison of all the latter models is based on up-to-24 hour-ahead predictions of hourly ozone concentrations. We analyzed daily forecast curves upon several criteria of two kinds: functional ones, and aggregated ones where attention is put on the daily maximum. It appears that autoregressive Hilbertian models with exogenous variables show the best predictive power. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: autoregressive; functional data; exogenous variables; ozone; prediction; ARHX

1. INTRODUCTION

The prediction of atmospheric pollutants is a problem studied by a large community of researchers working in various scientific areas. Due to the inner complexity of the spatiotemporal phenomena and the parsimony of data, one may have a look at statistical techniques. Since basic models did not satisfy experts enough, statisticians created numerous models which can roughly be classified into three main approaches.

Times series and regression models have been naturally and widely used, involving ARMA models with time-varying coefficients (Barrat *et al.*, 1990), threshold autoregressive models (Mélard and Roy, 1988), the two latter models with exogenous variables (Bauer *et al.*, 2001), GARCH models

*Correspondence to: Serge Guillas, Université Paris 6 (Pierre et Marie Curie), Laboratoire de Statistique Théorique et Appliquée, 175 rue du Chevaleret, 75013 Paris, France.

[†]E-mail: guillas@ccr.jussieu.fr

[‡]Also with Médiamétrie.

[§]Also with Ecole des mines de Douai.

(Graf-Jacottet and Jaunin, 1998), linear regression models (Comrie and Diem, 1999), non-parametric regression models with eventually a linear combination of exogenous variables (Gonzalez-Manteiga *et al.*, 1993) and Prada-Sanchez *et al.*, 2000), non-parametric discriminant analysis and multivariate adaptive regression splines—MARS—(Silvia *et al.*, 2001), generalized additive models—GAM—(Davis and Speckman, 1999), and chaotic time series (Kocak *et al.*, 2000).

Another way of making predictions is to use neural networks. Boznar *et al.* (1993), Yi and Prybutok (1996), Gardner and Dorling (1999) and Pérez *et al.* (2000) considered the multilayer perceptron, while Ruiz-Suarez *et al.* (1995) chose bidirectional associative memory and holographic associative memory models.

Classification and regression trees models (Breiman *et al.*, 1984) showed good capacities in this field. Ghattas (1999) and Gardner and Dorling (2000) illustrated this approach.

The idea of considering functional models comes from the observation of those phenomena. From a physical point of view, it is clear that the processes involved in the production of ozone are of continuous time type, therefore continuous time stochastic processes should be accurate to model the evolution of pollutants. The enhancement in the modelling procedure is the incorporation of exogenous variables in the AutoRegressive Hilbertian Model of order one—denoted by the acronym ARH(1) or simply ARH—giving as a result an AutoRegressive Hilbertian Model of order one with exogenous variables denoted by the acronym ARHX(1) or simply ARHX.

Such Hilbertian processes have been studied because they can handle, theoretically and practically (for example by use of smoothing splines, see Besse and Cardot (1996)), a large number of continuous time processes. Our approach is a functional one, as curves are our object of study; see Ramsay and Silverman (1997) for a review of ‘functional data analysis’, and Rice and Wu (2001) for work in this field relatively close to ours but with a fixed basis of spline functions.

In the next section, we present the data and how we can measure correctness of forecasts. Then, we will introduce the various models and explain how the non-parametric estimation procedures are working, in particular for cross-validation. Finally, a comparison will permit evaluation of the qualities of the diverse approaches.

1.1. Pollution and weather data; associated criteria of accuracy for predictions

The data came from the so-called AREMARTOIS air quality authority for the Artois area in the north of France. Six years of data were used: the period range is from 1 January 1995 to 31 December 2000. The monitoring station collected information about ozone every 15 min, but the available data were the hourly averages. One weather monitoring station collected hourly measurements of temperature, wind speed and wind direction.

Notice that missing values were essentially missing during entire months for technical reasons; only a few were missing for one or two hours for maintenance purposes. Therefore, our choice was not to replace the missing values with interpolated ones.

We fitted the various models to data ranging from 1 January 1995 to 31 December 1999. We analyzed the predictions on the remaining year. Due to the missing values, which affect the various variables differently, the set of prediction days are slightly different, depending on the model used, but the comparisons are made on the days where each method provided results. The horizon of prediction is of 24 h, but not in the usual sense. Every 11 p.m., we forecast the 24 values of the following day. This choice was made for simplicity purposes, but the model is relatively flexible regarding the hour when the prediction starts and the number of predicted values, as functional models can.

The criteria used in the sequel are of functional type: we want to see if the entire predicted curve for one day is close to the real one. Considering our pollutant (ozone), let $X_{i,j}$ be its concentration in $\mu\text{g}/\text{m}^3$ of day j at hour i and $\widehat{X}_{i,j}$ the prediction of $X_{i,j}$.

To compare the curves, we compute for integers $p = 1, 2$, respectively, the following empirical L^p -errors on a sample of n days based on the discretization scheme:

$$\begin{aligned}\|\widehat{X} - X\|_{L^1} &= \frac{1}{n} \sum_{j=1}^n \frac{1}{24} \sum_{i=0}^{23} |\widehat{X}_{i,j} - X_{i,j}|, \\ \|\widehat{X} - X\|_{L^2} &= \frac{1}{n} \sum_{j=1}^n \sqrt{\frac{1}{24} \sum_{i=0}^{23} (\widehat{X}_{i,j} - X_{i,j})^2}.\end{aligned}$$

The L^2 -error is much more sensitive to large errors over one day as might be possible during peak days of ozone. For $p = \infty$, we recall that the L^∞ -error is calculated as

$$\|\widehat{X} - X\|_{L^\infty} = \frac{1}{n} \sum_{j=1}^n \sup_{i=0, \dots, 23} |\widehat{X}_{i,j} - X_{i,j}|.$$

It is clearly possible to compute with this type of data and predictions the daily maximum or the 8 h average level as required by the National Ambient Air Quality Standards in the U.S.A. In France, a so-called ATMO index for ozone is calculated. Instead of announcing the maximum value of the daily maximum concentration of ozone, the authorities decided to give to the public a simplified index between 1 and 10. This index is associated with the interval containing the daily maximum concentration (on an hourly basis, in $\mu\text{g}/\text{m}^3$) from the following list: $[0, 30[$, $[30, 55[$, $[55, 80[$, $[80, 105[$, $[105, 130[$, $[130, 150[$, $[150, 180[$, $[180, 250[$, $[250, 360[$, $[360, +\infty[$.

The more classical criteria we will use for these daily maxima are presented below. Data are real numbers and not of functional type as above, so if we denote $(X_i)_{i=1, \dots, n}$ the observations and $(\widehat{X}_i)_{i=1, \dots, n}$ the forecasts, we can compute:

- the mean squared error (MSE),

$$MSE = \frac{1}{n} \sum_{i=1}^n (\widehat{X}_i - X_i)^2,$$

and the root mean squared error (RMSE),

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\widehat{X}_i - X_i)^2};$$

- the mean absolute error (MAE),

$$MAE = \frac{1}{n} \sum_{i=1}^n |\widehat{X}_i - X_i|;$$

- the mean relative error (MRE),

$$MRE = \frac{1}{n} \sum_{i=1}^n (\hat{X}_i - X_i)/X_i;$$

- the mean relative absolute error (MRAE),

$$MRAE = \frac{1}{n} \sum_{i=1}^n |\hat{X}_i - X_i|/X_i.$$

Large errors of prediction clearly exert a bigger influence on the two squared errors MSE and RMSE than on the MAE. If the forecasts from one model are almost always correct but rarely very bad, this could be better for MAE but not for MSE and RMSE than another model with daily prediction errors of homogeneous size, which makes worse predictions in general. Moreover, the relative errors must be regarded carefully when data are relatively close to 0.

Since the ‘persistence method’—that is the naïve prediction which makes today’s curve the predicted one for tomorrow—was clearly not accurate on the curves, we only focused on the following methods.

2. THE MODELS

Let H be a real and separable Hilbert space, e.g. H is $L^2[0,24]$ in our application. Denoting by $(x_t)_{t \in \mathbb{R}}$ the continuous time ozone process, we will consider the associated H -valued process $(X_n)_{n \in \mathbb{Z}}$ defined by

$$X_n(t) = x_{24n+t}, \quad t \in [0, 24].$$

In this way, we place the problem in a simpler discrete time context. (X_n) will be the variable of interest in the following models, and we want to predict X_{n+1} .

2.1. Autoregressive Hilbertian process

Let ρ be a bounded linear operator on H . Let $(\varepsilon_n)_{n \in \mathbb{Z}}$ be a strong Hilbertian white noise (SWN) that is a sequence of i.i.d. H -valued random variables satisfying

$$E\varepsilon_n = 0, \quad 0 < E \|\varepsilon_n\|_H^2 = \sigma^2 < \infty \quad n \in \mathbb{Z}.$$

(X_n) is an ARH process defined as the unique stationary solution of

$$X_n = \rho(X_{n-1}) + \varepsilon_n. \quad (1)$$

In order to ensure the stationarity of (X_n) , we assume the standard hypothesis on ρ , that is $\sum_{n=0}^{\infty} \|\rho^n\| < \infty$. We recall that, under such conditions, limit theorems and consistent estimation are obtained (Bosq, 2000).

To produce one-step-ahead forecasts we need to estimate ρ . Notice that this statistical model is non-parametric since ρ is an infinite dimensional parameter. The technique—as exposed in Bosq (2000)—proceeds as follows. Since the empirical estimator C_n of the covariance operator is not invertible in general, ρ ’s empirical estimator ρ_n is computed in a subspace spanned by the k_n eigenvectors of C_n associated with its k_n greatest eigenvalues.

2.2. Autoregressive Hilbertian model with exogenous variables

Keeping the same notations, let us introduce a_1, \dots, a_q bounded linear operators on H . We consider the following autoregressive Hilbertian with exogenous variables of order one model, denoted by ARHX(1):

$$X_n = \rho(X_{n-1}) + a_1(Z_{n,1}) + \dots + a_q(Z_{n,q}) + \varepsilon_n, \quad n \in \mathbb{Z}, \tag{2}$$

where $Z_{n,1}, \dots, Z_{n,q}$ are q zero-mean autoregressive of order one–ARH(1)–exogenous variables associated respectively with operators u_1, \dots, u_q and strong white noises $(\eta_{n,1}), \dots, (\eta_{n,q})$, i.e.

$$Z_{n,i} = u_i(Z_{n-1,i}) + \eta_{n,i}. \tag{3}$$

We assume that the noises $(\varepsilon_n), (\eta_{n,1}), \dots, (\eta_{n,q})$ are independent and similar hypotheses on the various operators as in the previous section to ensure existence, limit theorems and consistent estimation.

We choose to study ARHX(1) models in order to assess the influence of exogenous variables, considering this way an extension of the Granger causality in function spaces; see, for example, Guillas (2000) for theoretical results and Pitard and Viel (1999) for an illustration in an epidemiologic field with a selection of exogenous variables using Granger-causality tests.

One may write ARHX(1) models with exogenous processes taking their values in various spaces, i.e. H_i -valued $Z_{n,i}$. In this paper, we prefer to use H -valued $Z_{n,i}$ for simplicity, knowing that proofs are similar to the general case.

We will use the autoregressive representation of (2) in a product space in order to compute our estimates.

2.2.1. Autoregressive representation. The following construction enables us to manage ARHX processes as ARH processes, and thereby adapt the technique of estimation. As Mourid (1995) did for ARH(p) processes, we consider the Cartesian product H^{q+1} of $q + 1$ copies of H equipped with the scalar product

$$\langle (x_1, \dots, x_{q+1}), (y_1, \dots, y_{q+1}) \rangle_{q+1} := \sum_{j=1}^{q+1} \langle x_j, y_j \rangle.$$

H^{q+1} is then a separable Hilbert space.

Let us denote

$$T_n = \begin{pmatrix} X_n \\ Z_{n+1,1} \\ \vdots \\ Z_{n+1,q} \end{pmatrix}, \varepsilon'_n = \begin{pmatrix} \varepsilon_n \\ \eta_{n,1} \\ \vdots \\ \eta_{n,p} \end{pmatrix} \quad \text{and} \quad \rho' = \begin{pmatrix} \rho & a_1 & \dots & \dots & a_q \\ 0 & u_1 & 0 & \dots & 0 \\ 0 & 0 & u_2 & 0 & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & 0 & 0 & u_q \end{pmatrix}.$$

Let (X_n) be an ARHX(1) defined by (2); then it can easily be proved that (T_n) is an H^{q+1} -valued ARH(1) process (it is the unique stationary solution to the following equation):

$$T_n = \rho'(T_{n-1}) + \varepsilon'_n \tag{4}$$

In practice, we will compute estimators of eigenvalues and eigenvectors of the covariance operator C_n^T concerning the H^{q+1} -valued ARH(1) process (T_n) .

2.2.2. Two variants of the ARHX model. As stated in the introduction, we want to improve the ARH model by incorporating exogenous variables. Thus, we consider (with $q = 2$) the ARHX model (2), where $X_n, Z_{n,1}$ and $Z_{n,2}$ represent centered ozone concentration, temperature and wind speed, respectively. Two approaches are at least possible to estimate this ARHX.

The first one is simply to apply the theoretical techniques exposed previously, that is represent the process and the exogenous variables in a vector of H^{q+1} as shown in (4). This model will be denoted by the acronym ARHX(a).

The second is an empirical improvement of the previous one. It aims to get rid of the important reproductive behavior of the ARH model. In this way, we want to take into account better the exogenous variables' influence on our variable of interest.

To estimate an ARHX, our approach is to consider a finite subspace in which we project the observations, in order to inverse the covariance operator C_n^T . Indeed, there is no obligation to use the k_n eigenvectors of C_n^T associated with the k_n greatest eigenvalues, so we propose here an alternative choice for this subspace. Rather than considering T_n as a simple ARH model, we adapt the estimation to its known inner structure.

To do so, we decompose H^{q+1} in $q + 1$ spaces H . In each space H , linked to a variable (X , or a Z_i , $i = 1, \dots, q$), we choose (as in the ARH model) the subspace generated by the eigenvectors of the appropriate covariance operator (of X , or a Z_i , $i = 1, \dots, q$) associated with the greatest eigenvalues. The resulting basis for our subspace in H^{q+1} is therefore spanned by vectors of the following form:

$$\begin{pmatrix} v_1^1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} v_1^{k_n^{(1)}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_1^i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_1^i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_1^{q+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_1^{q+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In this way, we get $k_n^{(1)}$ vectors with non-zero coefficients in line 1, $k_n^{(2)}$ vectors with non-zero coefficients in line 2, \dots , $k_n^{(q+1)}$ vectors with non-zero coefficients in line $q + 1$. Hence, we control the evolution space of our different variables: we can then isolate the effect of Z_i in the estimated matrix ρ'_n of ρ' and adapt the subspace independently for each variable using the $k_n^{(i)}$. If we had not done so, the subspace obtained (by the eigenvectors of C_n^T) would have been drastically designed by the autoregressive part of the interest variable. The practical disadvantage of this approach is that we now have to choose $q + 1$ parameters $k_n^{(1)}, \dots, k_n^{(q+1)}$, and the cross-validation procedure is much more complicated.

This model will be denoted by the acronym ARHX(b).

2.3. Functional kernel model

Time series of ozone exhibit changes of scale, in particular when a peak appears (usually levels then remain high for a few days). Hence, it could be interesting to consider more general models. Although the ARH model is non-parametric, it assumes the linearity of ρ . To get rid of such an assumption, we applied a functional kernel model as follows. An alternative approach is to consider a local ARH predictor where local estimates of covariance and cross-covariance operators are computed using kernel techniques as in Besse *et al.* (2000).

One non-parametric way to deal with the conditional expectation $\rho(x) = E[X_i | X_{i-1} = x]$, where (X_i) is a H -valued process, is to consider a predictor inspired by the classical kernel regression, as in Nadaraja (1964) and Watson (1964). Notice that we choose to include no exogenous variable in this model, and therefore the results should be compared in the sequel essentially with the ARH model.

Let K denote the Gaussian kernel defined by

$$K(x) = (\sqrt{2\pi})^{-1} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

We used the following functional kernel estimator of ρ :

$$\hat{\rho}_{h_n}(x) = \frac{\sum_{i=1}^{n-1} X_{i+1} \cdot K\left(\frac{\|X_i - x\|_H}{h_n}\right)}{\sum_{i=1}^{n-1} K\left(\frac{\|X_i - x\|_H}{h_n}\right)}, \quad (5)$$

where h_n is the bandwidth, $\|\cdot\|_H$ is the norm of H and x belongs to H . Using the norm allows the use of a tool a priori devoted to finite dimensional valued processes.

Hence we get the predicted value of X_{n+1} given by

$$\hat{X}_{n+1} = \hat{\rho}_{h_n}(X_n),$$

where h_n is obtained by cross-validation, that is

$$h_n = \arg \min_h \sum_{i=n-r}^{n-1} \|\hat{\rho}_{h, n-r}(X_i) - X_{i+1}\|_{L^2}^2,$$

where $\hat{\rho}_{h, n-r}$ is the functional kernel estimator of ρ based on $(X_i)_{i=1, \dots, n-r}$, written as in (5) replacing $n-1$ by $n-r$, and h_n by $h(r = \lfloor n/5 \rfloor$ in our applications). To be coherent with the norm selected to compare the curves, H is chosen to be $L^2([0; 24])$.

The ARH model is a subset of the functional kernel model, but the associated estimators are drastically different. As a matter of fact, the first step of the ARH method for estimating ρ consists in making a projection of the subspace H_n spanned by the k_n eigenvectors of the empirical estimator C_n associated with its k_n greatest eigenvalues. This stage is crucial for a good estimation in the ARH model, while in the functional kernel model it is missing since the estimator is built using the Hilbertian norm. The alternative of considering a multivariate (finite dimensional) kernel model on H_n

seems inappropriate. Indeed, we will face the so-called ‘curse of dimensionality’ problem because the optimal k_n may be quite large (see Section 3.1). The inclusion of exogenous variables would make matters worse since the dimensions are higher.

3. RESULTS

In this section we use the various models for ozone forecasting, and compare the predictions on our real data set. We developed one specific library in the statistical software R (see Ihaka and Gentleman (1996)), which might soon be submitted to the CRAN (<http://cran.r-project.org>).

Since we noticed that our series were not stationary, we decided as usual to remove trend and seasonality. Of course, the daily seasonality is kept, because functional models can cope with it. We only dropped the annual trend and seasonality. We observed a relatively large heteroscedasticity over a year, but since we focused our study on summers, this problem disappeared.

The ozone data we studied, over the period from 1995 to 1999, are distributed as shown in Figure 1. The annual means of the series considered are presented in Table 1.

Table 1. Trend of the series computed as annual means

	Ozone ($\mu\text{g}/\text{m}^3$)	Temperature ($^{\circ}\text{C}$)	Wind speed (m/s)
1995	29	12	4.7
1996	26	9.1	4.1
1997	31	12	4
1998	36	12	4.3
1999	36	13	4.2

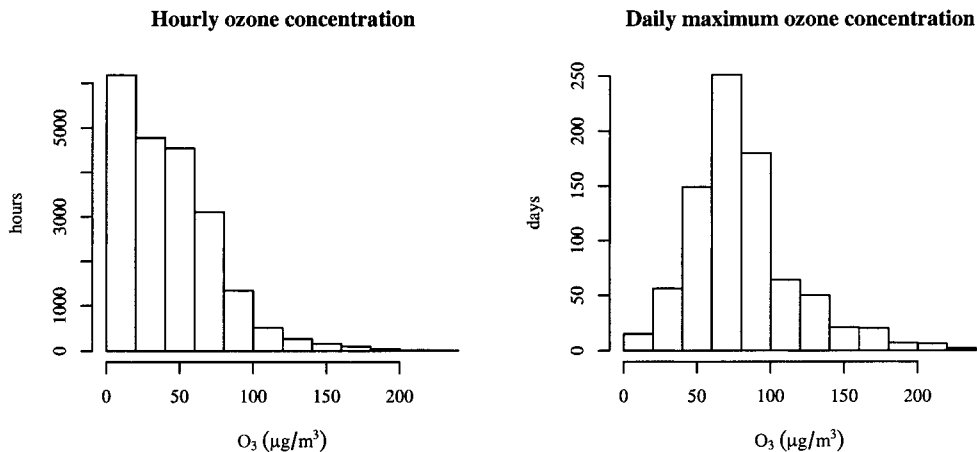


Figure 1. Ozone concentrations

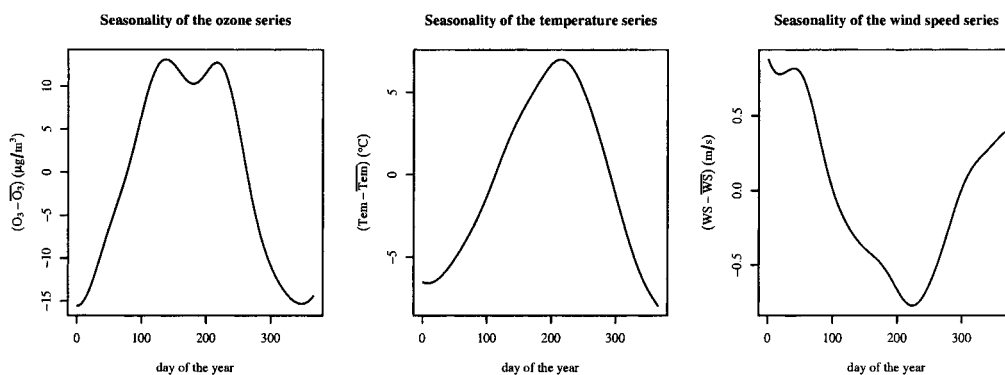


Figure 2. Centered seasonalities

Because modeling the trend with such a small sample with unequally distributed missing values may be hazardous, we simply report the 1999 estimated trends for year 2000. Our choice is confirmed by the lack of a clear evolution in Table 1.

To evaluate the seasonality of ozone, temperature and wind speed series, we performed a simple Nadaraya–Watson regression (see Nadaraja, 1964 and Watson, 1964) of the series with time, with a cross-validation procedure leading to a bandwidth of 25 days. Figure 2 presents the smoothed seasonalities we obtained. Notice that the wind speed seasonality is less important compared with the observed levels (with a range reaching only 39% of the annual mean against 89% for ozone and 128% for temperature).

We focused on the summer period for data, both in estimation and forecast procedures. Indeed, because of higher values of ozone, the authorities are more interested in summer study. This choice is also relevant for modeling, and corresponds to our previous remarks on heteroscedasticity. Note that with this selection in the data, the forecasts were slightly better in the summer merely because of better fitted data. Nevertheless, overall criteria were higher since the ozone concentrations—and consequently the possible errors—are much lower in the winter, except for relative errors because over many days in the winter the ozone concentrations are very close to 0.

Our period of interest was from day 120 to day 270 in the year, which makes 151 days and not exactly a summer. The predictions were only available on 77 days, of which 6 days could not be used for comparison purposes because of missing values. The mean value of ozone during summer 2000 was approximately $81 \mu\text{g}/\text{m}^3$.

3.1. Estimation

In order to tune the various models well, we chose to perform cross-validation procedures on summer 1999, which is a fifth of the training data set:

1. For the functional kernel, the optimal bandwidth was 105 for L^2 errors. This high number explains the smoothness of the prediction for this model. Maybe an L^∞ cross-validation would have provided different behavior.
2. For the ARH model, the procedure led to $k_n = 12$ for L^1 , L^2 and L^∞ errors as shown in Table 2.
3. Concerning the ARHX(a) model, the cross-validation procedure led to $k_n = 35$ for L^2 and L^∞ errors (and $k_n = 36$ for L^1 errors). It may seem large, but is relatively fitted to the size of the model (72 components because of 3 variables).

Table 2. Cross-validation for the ARH model

k_n	L^1 norm	L^2 norm	L^∞ norm	k_n	L^1 norm	L^2 norm	L^∞ norm
1	18.603	21.645	39.851	13	15.774	19.119	37.313
2	17.873	20.975	39.491	14	15.748	19.114	37.354
3	17.075	20.262	38.621	15	15.758	19.131	37.429
4	16.407	19.648	38.021	16	15.725	19.105	37.456
5	16.223	19.446	37.611	17	15.754	19.148	37.610
6	16.101	19.324	37.435	18	15.741	19.128	37.541
7	15.921	19.206	37.412	19	15.778	19.163	37.67
8	15.835	19.186	37.319	20	15.847	19.249	37.704
9	15.791	19.152	37.347	21	15.858	19.258	37.758
10	15.743	19.113	37.282	22	15.774	19.132	37.529
11	15.732	19.078	37.18	23	15.92	19.296	37.821
12	15.711	19.047	37.147	24	15.925	19.326	37.898

4. The ARHX(b) model was the hardest model to fit because of the high combinatorial choice of the parameters ($24^3 = 13\,824$ choices). Nevertheless, the computation yielded

$$\left(k_n^{(1)}, k_n^{(2)}, k_n^{(3)}\right) = (7, 8, 7),$$

where the variables are respectively ozone, temperature and wind speed. Figure 3 shows how the L^2 norm is disrupted when $k_n^{(1)} = 7$ and the other parameters vary.

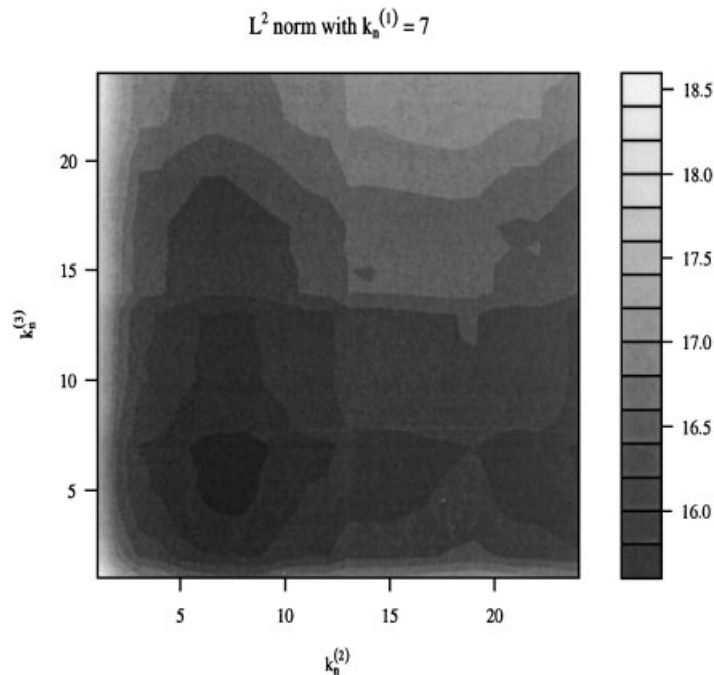


Figure 3. Levels of the L^2 error with varying parameters for temperature and wind speed

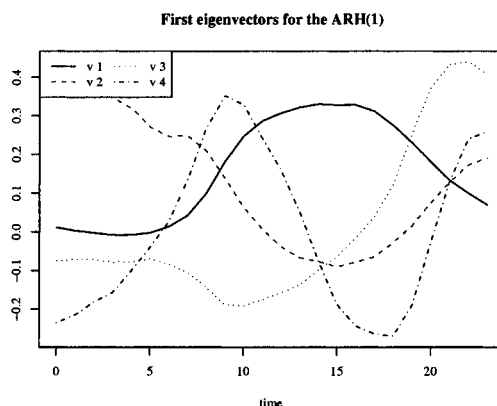


Figure 4. First four functional eigenvectors of the covariance operator in the ARH(1) model

Although the ARH model is not parametric, one may give interpretations of the eigenvectors associated with the first eigenvalues and the coefficients of $\hat{\rho}_n$. For example, we can guess in Figure 4 that the mean daily seasonality is in the space spanned by the eigenvector associated with the first eigenvalue. Indeed, if we consider the centered Hilbertian process, $Y_i = X_i - \mu$, then we get the following decomposition of the autocovariance operator of the stationary process (X_i) :

$$\begin{aligned} C(x) &= E[\langle X_0, x \rangle X_0] = E[\langle Y_0 + \mu, x \rangle (Y_0 + \mu)] \\ &= E[\langle Y_0, x \rangle Y_0] + E[\langle Y_0, x \rangle \mu] + E[\langle \mu, x \rangle Y_0] + E[\langle \mu, x \rangle \mu] \\ &= E[\langle Y_0, x \rangle Y_0] + \langle \mu, x \rangle \mu. \end{aligned}$$

μ is the only eigenvector associated with the operator T defined by $T(x) = \langle \mu, x \rangle \mu, x \in H$, with eigenvalue $\|\mu\|^2$. If the eigenvalues of the autocovariance operator of the process (Y_i) are smaller than $\|\mu\|^2$, then μ may appear as the eigenvector of C associated with the first eigenvalue. It should be interesting to link this information to the chemical processes involved in ozone evolution.

3.2. Comparison

The various models aim to assess the dynamics of ozone creation or dispersion. In Figure 5 it is possible to see that the two ARHX models are relatively good for that purpose: the increase of 24 July and the decrease the next day are relatively well predicted, as well as the stability during 16 September. The prediction of the high level of 24 August is only slightly underestimated, and the exit of the high ozone episode the next day is well predicted. We are a little disappointed with the behavior of ARHX models on 23 August which might be explained by the fact that our model does not take into account some meteorological variables such as solar radiation or cloud cover and possible explanatory variables related to human activity such as the day of the week.

In terms of functional criteria for measuring the accuracy of the forecasts, Table 3 shows that the modifications suggested to improve the ARHX model in its ARHX(b) version is relevant. Figure 6 illustrates the daily behavior of L^2 errors.

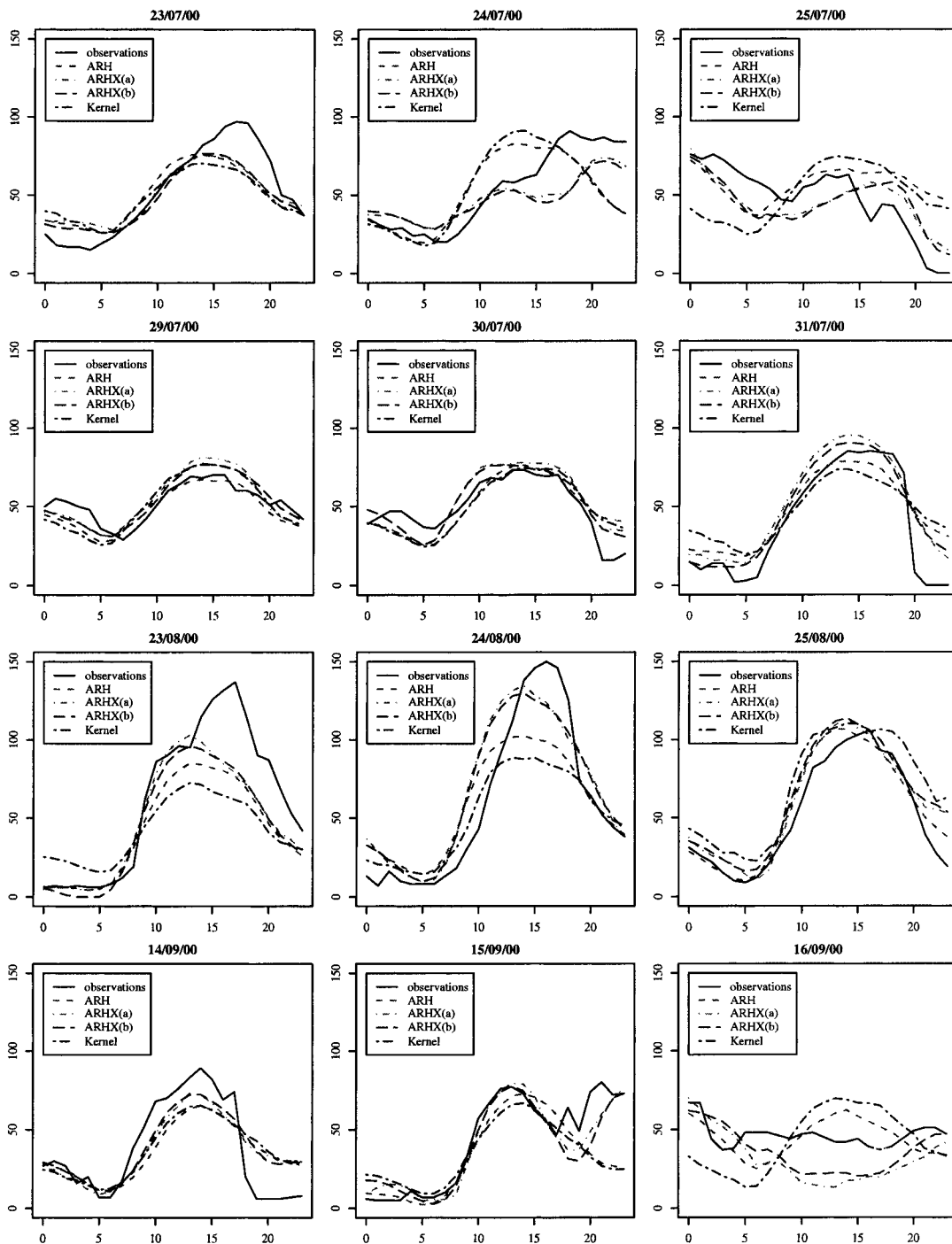
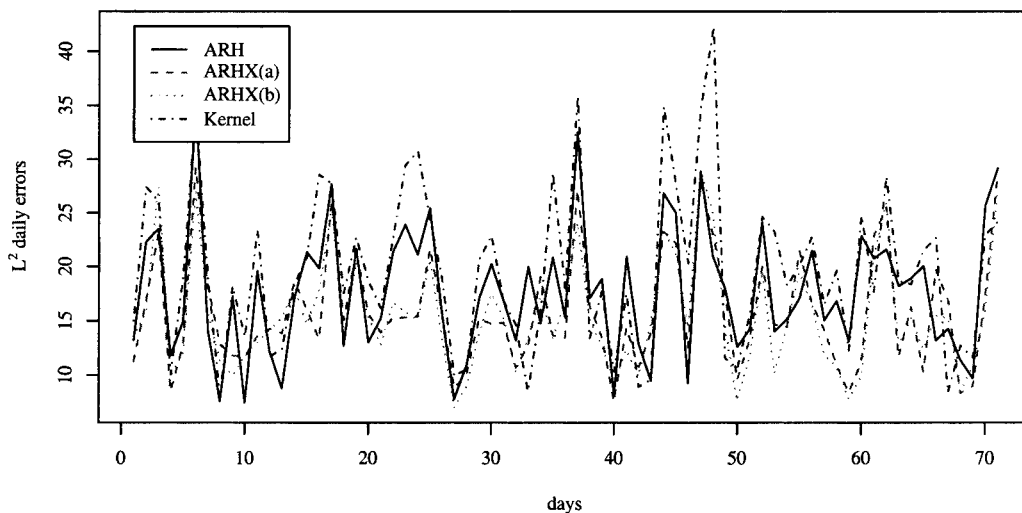


Figure 5. Comparison of the various model for predicting ozone (in $\mu\text{g}/\text{m}^3$)

Table 3. Mean functional errors of the models during summer 2000

	ARH	ARHX(a)	ARHX(b)	Kernel
L^1 norm	14.29	13.23	12.79	16.37
L^2 norm	17.73	15.86	15.45	19.75
L^∞ norm	35.5	32.15	30.88	38.94

Comparison of the daily L^2 errors of the modelsFigure 6. Comparison of L^2 errors of the various models during summer 2000 (missing values are excluded)

Some ozone standards are calculated upon the daily maximum. Therefore, we used this complementary approach to compare the behavior of the predictions. The generalized additive models (GAM)—see Hastie and Tibshirani (1990)—appeared to be very competitive in forecasting the peaks in ozone as shown in Davis and Speckman (1999). That is why we decided to add two new alternative models, already discussed in this context, denoted respectively by GAM1 and GAM2 and defined roughly by

$$O_3 \text{ max} = f_1(O_3 \text{ max lag}_1) + f_2(T \text{ max}) + f_3(W \text{ mean}), \quad (6)$$

$$\log(O_3 \text{ max}) = g_1(O_3 \text{ max lag}_1) + g_2(T \text{ max}) + g_3(W \text{ mean}), \quad (7)$$

where:

- $O_3 \text{ max}$ stands for the daily maximum of ozone
- $O_3 \text{ max lag}_1$ stands for the daily maximum of ozone the day before
- $T \text{ max}$ stands for the daily maximum of temperature
- $W \text{ mean}$ stands for the daily mean of wind speed
- and f_1, f_2, f_3, g_1, g_2 and g_3 are unspecified functions to be estimated by the GAM procedure.

Comparison of the daily maximum obtained by the different models

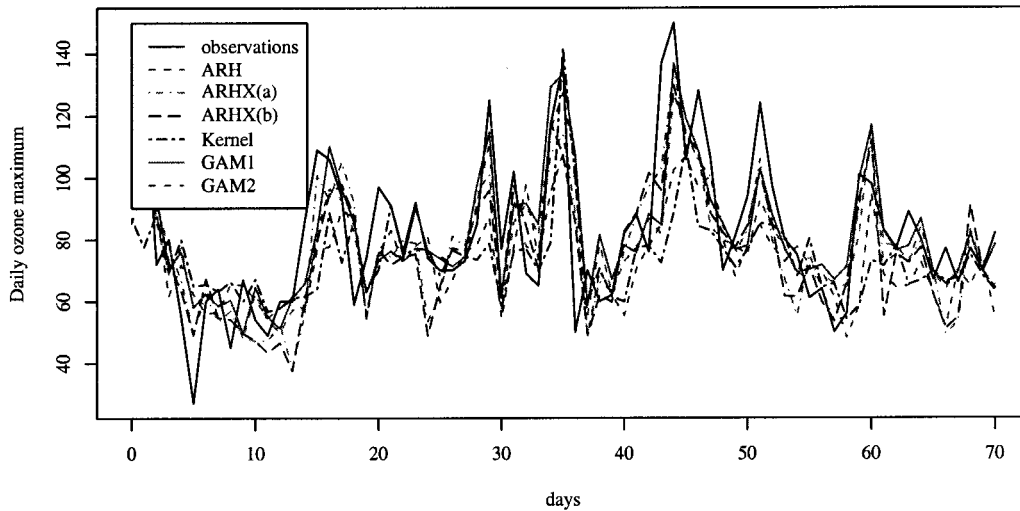


Figure 7. Comparison of predicted daily maxima of ozone versus real ones of the various model during summer 2000 (missing values are excluded)

Although ARHX models do not aim to forecast curve maxima, they show good predictive skills in this field, outperforming the other functional models, as seen in Table 4. The two versions of the ARHX model reveal somehow different qualities: the (a) version performs better for predicting the daily maximum, and the (b) version performs better for predicting the entire daily curve. Besides, the RMSE is lower for the ARHX(a) than for the GAM models. The MAE, MRE and MRAE are lower for the GAM models. Figure 7 illustrates the daily behavior of each model for predicting the daily maximum.

Another point of view is to consider the distribution of errors in terms of classes. First, if we look at $10 \mu\text{g}/\text{m}^3$ long intervals, we obtain Table 5, in which the percentages of the errors falling into each interval are presented for each method. We see that only 2.817% of the errors between the daily maxima and the forecasted ones made by the ARHX(b) method during summer 2000 were larger than $30 \mu\text{g}/\text{m}^3$. This method makes the smallest number of large errors.

Secondly, if we look at errors in terms of the difference between the real ATMO indexes and the forecasted ones, we obtain Table 6, in which the number of days where the difference was of zero, one, two or three indexes are presented. Notice that not one of these methods made errors of four or more indexes. The success rate is computed as follows: we say that there is a success when the error is of zero

Table 4. Errors on the daily maximum of the various models during summer 2000

	MSE	RMSE	MAE	MRE	MRAE
ARH	445.6	21.11	16.62	-0.04828	0.2092
ARHX (a)	239.4	15.47	12.89	-0.02443	0.1702
ARHX (b)	275.4	16.59	13.55	-0.04191	0.1751
Kernel	527.6	22.97	16.56	-0.0353	0.2034
GAM1	249.2	15.78	11.80	0.04587	0.1655
GAM2	274	16.55	12.59	0.0104	0.1694

Table 5. Distribution of errors for the maximum of the day during summer 2000

	ARH	ARHX(a)	ARHX(b)	Kernel	GAM1	GAM2
[0;10]	42.25%	39.44%	45.07%	40.85%	46.48%	49.3%
[10;20]	21.13%	39.44%	28.17%	28.17%	35.21%	32.39%
[20;30]	22.54%	15.49%	23.94%	15.49%	14.08%	12.68%
>30	14.08%	5.634%	2.817%	15.49%	4.225%	5.634%

Table 6. Errors on the ozone ATMO index during summer 2000

	ARH	ARHX(a)	ARHX(b)	Kernel	GAM1	GAM2
0 classes	28	31	34	33	35	34
1 class	35	38	35	30	33	33
2 classes	7	2	2	5	2	3
3 classes	1	0	0	3	1	1
Success rate	88.73%	97.18%	97.18%	88.73%	95.77%	94.37%

or one index. We see that the ARHX(a) and ARHX(b) methods showed the best success rates, but we may notice also that the GAM methods made the largest number of 0-class errors among all the methods. The success rate of ATMO index prediction is computed as the number of zero and one class errors divided by the number of predictions.

4. CONCLUSION

The ozone forecasting problem is both a health issue and a difficult problem. The statistical methods presented above show good skills for this purpose. The functional approach seems relevant. The autoregressive Hilbertian model with exogenous variables in its two versions was the best model with respect to functional or daily maximum criteria.

Several improvements should be developed. The first are methodological ones. Indeed, in the pollution field, practitioners are interested in specific hours and horizons of prediction, for instance in the afternoon of the day before in order to inform the population better, or in the morning of the day considered in order to increase the accuracy of the forecasts. To do so, we may use overlapping intervals to construct our variable of interest, considering, for instance, $X_n(t) = x_{24n+t}$, with t in $[0, 36]$ and n in \mathbb{N} . Furthermore, since peaks are not easy to predict, the methodology could change slightly for that purpose by putting weights on highly polluted days during the estimation procedure.

The second are practical. It seems clearly interesting to take into account more information about the meteorological situation or human activities. Hence, exogenous variables such as solar radiation, humidity, precipitation, pressure, cloud cover and day of the week should be inputs of the model. Those variables are not always easy to measure or predict, and a careful use has to be made.

ACKNOWLEDGEMENT

We are grateful to the AREMARTOIS association for supplying the data. We would like to thank the referee for many valuable comments which improved the presentation of the paper.

REFERENCES

- Barrat M, Lecluse Y, Slamani Y. 1990. Etude comparative de différents modèles mathématiques pour la prédiction des niveaux de pollution atmosphérique, analyse univariante. *R.A.I.R.O. APII* **3**: 283–298.
- Bauer G, Deistler M, Scherrer W. 2001. Time series models for short term forecasting of ozone in the eastern part of Austria. *Environmetrics* **12**: 117–130.
- Besse P, Cardot H. 1996. Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Revue Canadienne de la Statistique/Canadian Journal of Statistics* **24**: 467–487.
- Besse P, Cardot H, Stephenson D. 2000. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27**(4): 673–687.
- Bosq D. 2000. *Linear Processes in Function Spaces: Theory and Applications, (Lecture Notes in Statistics, Vol. 149)*. Springer-Verlag: New York.
- Boznar M, Lesjak M, Mlakar P. 1993. A neural network-based method for short-term predictions of ambient SO₂ concentrations in highly polluted industrial areas of complex terrain. *Atmospheric Environment* **27B**(2): 221–230.
- Breiman L, Friedman J, Olshen R, Stone R. 1984. *Classification and Regression Trees*. Wadsworth: Belmont, CA.
- Comrie AC, Diem JE. 1999. Climatology and forecast modeling of ambient carbon monoxide in phoenix, arizona. *Atmospheric Environment* **33**: 5023–5036.
- Davis JM, Speeckman P. 1999. A model for predicting maximum and 8 h average ozone in houston. *Atmospheric Environment* **33**: 2487–2500.
- Gardner MW, Dorling SR. 1999. Neural network modelling and prediction of hourly NO_x and NO₂ concentrations in urban air in London. *Atmospheric Environment* **33**: 709–719.
- Gardner MW, Dorling SR. 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment* **34**: 21–34.
- Ghattas B. 1999. Prévisions des pics d'ozone par arbres de régression simples et agrégés par bootstrap. *Revue de Statistique Appliquée XLVII*(2): 85–98.
- Gonzalez-Manteiga W, Prada-Sanchez JM, Cao R, Garcia-Jurado I, Febrero-Bande M, Lucas-Dominguez T. 1993. Time series analysis for ambient concentrations. *Atmospheric Environment* **27A**(2): 153–158.
- Graf-Jacottet M, Jaunin M-H. 1998. Predictive models for ground ozone and nitrogen dioxide time series. *Environmetrics* **9**: 393–406.
- Guillas S. 2000. Non-causalité et discrétisation fonctionnelle, théorèmes limites pour un processus ARHX(1). *C. R. Acad. Sci. Paris Sér. I Math* **331**: 91–94.
- Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models*. Chapman & Hall: New York.
- Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *Journal of Graphical and Computational Statistics* **5**: 299–314.
- Kocak K, Saylan L, Sen O. 2000. Nonlinear time series prediction of O₃ concentration in Istanbul. *Atmospheric Environment* **34**: 1267–1271.
- Mélaud G, Roy R. 1988. Modèles de séries chronologiques avec seuils. *Revue de Statistique appliquée XXXVI*(4): 5–24.
- Mourid T. 1995. *Contribution à la statistique des processus autorégressifs à temps continu*. D. Sc. Thesis, Université Paris 6.
- Nadaraja EA. 1964. On estimating regression. *Theory Probab. An.* **9**: 141–142.
- Pérez P, Trier A, Reyes J. 2000. Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment* **34**: 1189–1196.
- Pitard A, Viel JF. 1999. A model selection tool in multi-pollutant time series: the Granger-causality diagnosis. *Environmetrics* **10**: 53–65.
- Prada-Sanchez JM, Febrero-Bande M, Cotos-Yanez T, Gonzalez-Manteiga W, Bermudez-Cela L, Lucas-Dominguez T. 2000. Prediction of SO₂ pollution incident near a power station using partially linear models and an historical matrix of predictor-response vectors. *Environmetrics* **11**: 209–225.
- Ramsay J, Silverman B. 1997. *Functional Data Analysis*. Springer-Verlag.
- Rice J, Wu C. 2001. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**: 253–259.
- Ruiz-Suarez JC, Mayora-Ibarra OA, Torres-Jimenez J, Ruiz-Suarez LG. 1995. Short-term ozone forecasting by artificial neural networks. *Advances in Engineering Software* **23**: 143–149.
- Silvia C, Pérez P, Trier A. 2001. Statistical modelling and prediction of atmospheric pollution by particulate material: two nonparametric approaches. *Environmetrics* **12**: 147–149.
- Watson GS. 1964. Smooth regression analysis. *Sankhya Ser. A* **26**: 359–372.
- Yi J, Prybutok VR. 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution* **3**: 349–357.