The inclusion of exogenous variables in functional autoregressive ozone forecasting

Julien DAMON et Serge GUILLAS
The inclusion of exogenous variables in functional autoregressive ozone forecasting

Julien DAMON et Serge GUILLAS

*Université Paris VI, ISUP-LSTA*
*Tour 45-55, Boîte 158*
*4, place Jussieu, 75252 Paris Cedex 05*

Juin 2001
The inclusion of exogenous variables in functional autoregressive ozone forecasting

J. Damon  S. Guillas
Université Paris 6  Université Paris 6
Ecole des Mines de Douai

Juin 2001

Abstract

In this paper, 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 phenomena by taking into account theoretically and practically the continuous time evolution of pollution. One main methodological enhancement of this paper 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 showed the best predictive power.

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 (Graf-Jacquet and Jaunin (1998)), linear regression models (Comrie and Dier (1999)), nonparametric regression models with eventually
a linear combination of exogenous variables (Gonzalez-Manteiga et al. (1993) and Prada-Sánchez et al. (2000)), nonparametric discriminant analysis and Multivariate Adaptive Regression Splines - MARS - (Silvia et al. (2001)), Generalized Additive Models - GAM- (Davis and Speeckman (1999)), and chaotic time series (Kocak et al. (2000)).


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 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 a work in this field relatively close to ours but with a fixed basis of spline functions.

In the next section, we will 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 to evaluate 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 January 1st 1995 to December 31st 2000. The monitoring station collected information about ozone every 15 minutes, 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 during one hour or two 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 January 1st 1995 to December 31st 1999. We analyzed the predictions on the remaining year. Due to the missing
values which affect differently the various variables, 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 hours, 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 $\mu g/m^3$ of day $j$ at hour $i$ and $\hat{X}_{i,j}$ the prediction of $X_{i,j}$.

In order 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

$$\| \hat{X} - X \|_{L^1} = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{1}{24} \sum_{i=0}^{23} | \hat{X}_{i,j} - X_{i,j} | \right),$$

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

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

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

It is clearly possible to compute with this type of data and predictions the daily maximum or the 8h average level as required by the National Ambient Air Quality Standards in the USA. In France, a so-called ATMO index is calculated. It is the maximum of the partial ATMO indexes associated to $O_3$, $NO_2$, $SO_2$ and particulate material, computed as a number between 1 and 10 representing intervals of maximum daily concentrations.

The 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 $(X_i)_{i=1,\ldots,n}$ the observations and $(\hat{X}_i)_{i=1,\ldots,n}$ the forecasts, we can compute

- The Mean Squared Error (MSE)

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

and the Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{ \frac{1}{n} \sum_{i=1}^{n} (\hat{X}_i - X_i)^2 }.$$
• The Mean Absolute Error (MAE)

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{X}_i - X_i| \]

• 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)

\[ MRAE = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{X}_i - X_i|}{X_i} \]

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 - appeared clearly not accurate on 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 \( \rho \) 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

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

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

\[ X_n = \rho(X_{n-1}) + \varepsilon_n. \]
In order to ensure the stationarity of \((X_n)\) and, we assume the standard hypothesis on \(\rho\), that is \(\sum_{n=0}^{\infty} \|\rho^n\| < \infty\). We recall that under such conditions, limit theorems and consistent estimation are obtained (see Bosq (2000)).

To produce one step ahead forecasts, we need to estimate \(\rho\). 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, \(\rho\)'s empirical estimator \(\rho_n\) is computed in a subspace spanned by the \(k_n\) eigenvectors of \(C_n\) associated to its \(k_n\) greatest eigenvalues.

### 2.2 Autoregressive Hilbertian Model with exogenous variables

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

\[
X_n = \rho(X_{n-1}) + a_1(Z_{n,1}) + \ldots + a_q(Z_{n,q}) + \varepsilon_n, n \in \mathbb{Z}
\]  

(2)

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

\[
Z_{n,i} = u_i(Z_{n-1,i}) + \eta_{n,i}.
\]

(3)

We assume that the noises \((\varepsilon_n), (\eta_{n,1}), \ldots, (\eta_{n,q})\) 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 e.g. Guillas (2000) for theoretical results and Pitard and Viel (1999) for an illustration in an epidemiologic field with 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 the equation (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

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

\(H^{q+1}\) is then a separable Hilbert space.
Let us denote

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

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 \(C_T^n\) 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 (4) where \(X_n, Z_{n,1}\) and \(Z_{n,2}\) represent respectively centered ozone concentration, temperature and wind speed. Two approaches are at least possible, in order 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 equation (4). This model will be denoted by the acronym \(ARHX(a)\).

The second is a 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 an finite subspace in which we project the observations, in order to inverse the covariance operator \(C_T^n\). Indeed, there is no obligation to use the \(k_n\) eigenvectors of \(C_T^n\) associated to 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 \(Z_i, i = 1, \ldots, q)\), we choose (as in the \(ARH\) model) the subspace generated by the eigenvectors of the appropriate covariance operator (of \(X,\) or a \(Z_i, i = 1, \ldots, q)\) associated to the greatest eigenvalues. The resulting basis for our subspace in \(H^{q+1}\) is therefore spanned
by vectors of the following form

\[
\begin{pmatrix}
    v_1^1 \\
    0 \\
    \vdots \\
    \vdots \\
    \vdots \\
    \vdots \\
    0 \\
\end{pmatrix}, \ldots,
\begin{pmatrix}
    v_1^{k_n(1)} \\
    0 \\
    0 \\
    0 \\
    0 \\
    \vdots \\
    \vdots \\
    0 \\
\end{pmatrix}, \ldots,
\begin{pmatrix}
    0 \\
    \vdots \\
    \vdots \\
    0 \\
    0 \\
    v_1^{k_n(q+1)} \\
    \vdots \\
    \vdots \\
\end{pmatrix}, \ldots,
\begin{pmatrix}
    v_1^{q+1} \\
    0 \\
    \vdots \\
    \vdots \\
    \vdots \\
    \vdots \\
    \vdots \\
\end{pmatrix}, \ldots,
\begin{pmatrix}
    0 \\
    \vdots \\
    \vdots \\
    0 \\
    v_1^{k_n(q+1)} \\
    \vdots \\
    \vdots \\
    \vdots \\
\end{pmatrix}, \ldots,
\begin{pmatrix}
    0 \\
    \vdots \\
    \vdots \\
    \vdots \\
    \vdots \\
    \vdots \\
    0 \\
\end{pmatrix}
\]

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, ..., \( k_n(q+1) \) 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 \( \rho_n' \) of \( \rho' \) 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 \( C_n^2 \)) 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 \( k_n(1), \ldots, k_n(q+1) \), and cross validation procedure is much more complicated.

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

### 2.3 Functional kernel model

Although the \( ARH \) model is non parametric, it assumes the linearity of \( \rho \). To get rid of such an assumption, we applied a functional kernel modelling 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 nonparametric 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].

Let \( K \) denote the Gaussian kernel defined by

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

We used the following functional kernel estimator of \( \rho \)

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

where \( h_n \) is the bandwidth, \( \|.\|_H \) is the norm of \( H \) and \( x \) belongs to \( H \). Using the norm allows to use a tool \( a \ priori \) devoted to finite dimensional valued processes.
Hence we get the predicted value of $X_{n+1}$ given by

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

where $h_n$ is obtained by cross validation, that is

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

where $\hat{\rho}_{h,n-r}$ is the functional kernel estimator of $\rho$ based on $(X_i)_{i=1,...,n-r}$, written as in (4) 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 $L^2([0; 24])$.

### 3 Results

In this section, we will 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 be soon submitted to the CRAN (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, during 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>
<thead>
<tr>
<th>Year</th>
<th>Ozone ($\mu g/m^3$)</th>
<th>Temperature (°C)</th>
<th>Wind speed (m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>29</td>
<td>12</td>
<td>4.7</td>
</tr>
<tr>
<td>1996</td>
<td>26</td>
<td>9.1</td>
<td>4.1</td>
</tr>
<tr>
<td>1997</td>
<td>31</td>
<td>12</td>
<td>4.0</td>
</tr>
<tr>
<td>1998</td>
<td>36</td>
<td>12</td>
<td>4.3</td>
</tr>
<tr>
<td>1999</td>
<td>36</td>
<td>13</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 1: Trend of the series computed as annual means

Thinking that modelling 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. We are comforted in our choice 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 of the series on the time, with a cross-validation procedure leading to a bandwidth of 25 days. Figure 2 present the smoothed seasonalties we obtained. It may be observed that the wind speed seasonality is not really significative compared to the observed levels.
Hourly ozone concentration

Daily maximum ozone concentration

Figure 1: Ozone concentrations

Seasonality of the ozone series

Seasonality of the temperature series

Seasonality of the wind-speed series

Figure 2: Centered seasonalities
We focused on the summer period for data, both in estimation and forecast procedures. Indeed, because of higher values of ozone, authorities are more interested in summer study. This choice is also relevant for modelling, and corresponds to our previous remarks on heteroscedasticity. We may mention that with this selection in the data, the forecasts were slightly better in the summer merely because of better fitted data, but overall criteria were higher since the ozone concentrations - and consequently the possible errors - are much lower in the winter.

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 µg/m³.

3.1 Estimation

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

- 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 a $L^\infty$ cross-validation would have provided a different behavior.

- For the ARH model, the procedure led to $k_n = 12$ for $L^1$, $L^2$ and $L^\infty$ errors as shown in table 2.

- Concerning the ARHX(a) model, the cross-validation procedure led to $k_n = 35$ for $L^2$ and $L^\infty$ 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).

- The ARHX(b) model was the hardest model to fit, because of the high combinatorial choice of the parameters ($24^3 = 13824$ 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.

Although the ARH model is not parametric, one may give interpretations of the eigenvectors associated to 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 to the first eigenvalue. It should be interesting to link those informations to the chemical processes involved in ozone evolution.
3.2 Comparison

The various models aim at assessing 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 July 24th and the decrease the next day are relatively well predicted, as well as the stability during September 16. The prediction of the high level of August 24th is only slightly underestimated, and the exit of the high ozone episode the next day is well predicted. We are a little bit disappointed with the behavior of ARHX models during August 23rd, which might be explained by a lack of meteorological variables.

In terms of functional criteria for measuring the accuracy of the forecasts, the 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.

<table>
<thead>
<tr>
<th>$k_n$</th>
<th>$L^1$ norm</th>
<th>$L^2$ norm</th>
<th>$L^\infty$ norm</th>
<th>$k_n$</th>
<th>$L^1$ norm</th>
<th>$L^2$ norm</th>
<th>$L^\infty$ norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>17.075</td>
<td>20.262</td>
<td>38.621</td>
<td>15</td>
<td>15.758</td>
<td>19.131</td>
<td>37.429</td>
</tr>
<tr>
<td>5</td>
<td>16.223</td>
<td>19.446</td>
<td>37.611</td>
<td>17</td>
<td>15.754</td>
<td>19.148</td>
<td>37.610</td>
</tr>
<tr>
<td>7</td>
<td>15.921</td>
<td>19.206</td>
<td>37.412</td>
<td>19</td>
<td>15.778</td>
<td>19.163</td>
<td>37.67</td>
</tr>
<tr>
<td>8</td>
<td>15.835</td>
<td>19.186</td>
<td>37.319</td>
<td>20</td>
<td>15.847</td>
<td>19.249</td>
<td>37.704</td>
</tr>
<tr>
<td>9</td>
<td>15.791</td>
<td>19.152</td>
<td>37.347</td>
<td>21</td>
<td>15.858</td>
<td>19.258</td>
<td>37.758</td>
</tr>
<tr>
<td>10</td>
<td>15.743</td>
<td>19.113</td>
<td>37.282</td>
<td>22</td>
<td>15.774</td>
<td>19.132</td>
<td>37.529</td>
</tr>
<tr>
<td>11</td>
<td>15.732</td>
<td>19.078</td>
<td>37.18</td>
<td>23</td>
<td>15.92</td>
<td>19.296</td>
<td>37.821</td>
</tr>
<tr>
<td>12</td>
<td>15.711</td>
<td>19.047</td>
<td>37.147</td>
<td>24</td>
<td>15.925</td>
<td>19.326</td>
<td>37.898</td>
</tr>
</tbody>
</table>

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

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,
Figure 3: Levels of the $L^2$ error with varying parameters for temperature and wind speed denoted respectively by $GAM1$ and $GAM2$ and defined roughly by

$$O_{3\text{ max}} = f_1 (O_{3 \text{ max lag}_1}) + f_2 (T_{\text{ max}}) + f_3 (W_{\text{ mean}})$$  \hspace{1cm} (6)

$$\log (O_{3 \text{ max}}) = g_1 (O_{3 \text{ max lag}_1}) + g_2 (T_{\text{ max}}) + g_3 (W_{\text{ mean}})$$  \hspace{1cm} (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.

Although $ARHX$ models do not aim at forecast curves maxima, they showed 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
Figure 4: First four functional eigenvectors of the covariance operator in the $ARH(1)$ model

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 4 illustrates the daily behavior of each model for predicting the daily maximum.

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

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

Another point of view is to consider the distribution of errors in terms of classes. Firstly, if we look at 10 $\mu g/m^3$ long intervals, we obtain table 5. Secondly, if we look at error in term of the number of classes in the so-called ATMO index for ozone, we obtain table 6. Let us recall before what are the ATMO classes for daily maximum of ozone in $\mu g/m^3$:

$[0, 30], [30, 55], [55, 80], [80, 105], [105, 130], [130, 150], [150, 180], [180, 250], [250, 360], [360, \infty]$. 

The success rate of ATMO index prediction is computed as the number of 0 an 1 class error divided by the number of predictions.
Figure 5: Comparison of the various model for predicting ozone (in $\mu g/m^3$)
Figure 6: Comparison of $L^2$ errors of the various model during summer 2000 (missing values are excluded)

Figure 7: Comparison of predicted daily maxima of ozone versus real ones of the various model during summer 2000 (missing values are excluded)
### Table 5: Classes of errors for the maximum of the day during summer 2000

<table>
<thead>
<tr>
<th></th>
<th>ARH</th>
<th>ARHX(a)</th>
<th>ARHX(b)</th>
<th>Kernel</th>
<th>GAM1</th>
<th>GAM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0;10]</td>
<td>42.25 %</td>
<td>39.44 %</td>
<td>45.07 %</td>
<td>40.85 %</td>
<td>46.48 %</td>
<td><strong>49.3 %</strong></td>
</tr>
<tr>
<td>[10;20]</td>
<td>21.13 %</td>
<td>39.44 %</td>
<td>28.17 %</td>
<td>28.17 %</td>
<td>35.21 %</td>
<td>32.39 %</td>
</tr>
<tr>
<td>[20;30]</td>
<td>22.54 %</td>
<td>15.49 %</td>
<td>23.94 %</td>
<td>15.49 %</td>
<td>14.08 %</td>
<td>12.68 %</td>
</tr>
<tr>
<td>&gt;30</td>
<td>14.08 %</td>
<td>5.634 %</td>
<td><strong>2.817 %</strong></td>
<td>15.49 %</td>
<td>4.225 %</td>
<td>5.634 %</td>
</tr>
</tbody>
</table>

### Table 6: Errors on the ozone ATMO index during summer 2000

<table>
<thead>
<tr>
<th></th>
<th>ARH</th>
<th>ARHX(a)</th>
<th>ARHX(b)</th>
<th>Kernel</th>
<th>GAM1</th>
<th>GAM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 class</td>
<td>28</td>
<td>31</td>
<td>34</td>
<td>33</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>1 class</td>
<td>35</td>
<td>38</td>
<td>35</td>
<td>30</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>2 classes</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3 classes</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Success Rate</td>
<td>88.73 %</td>
<td><strong>97.18 %</strong></td>
<td><strong>97.18 %</strong></td>
<td>88.73 %</td>
<td>95.77 %</td>
<td>94.37 %</td>
</tr>
</tbody>
</table>

### 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 version 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, practicians 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. Accordingly, modifications of our model are necessary to cope with such a problematic by considering overlapping intervals for example. 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.

### References

Barrat M, Lecluse Y, Slamani Y. 1990. Etude comparative de différents modèles mathématiques pour la prédiction des niveaux de pollution atmosphérique, analyse uni-


LISTE DES RAPPORTS TECHNIQUES

N°1 Février 1999  Autoregressive Hilbertian Process
  Denis BOSQ

N°2 Avril 1999  Using the Entringer numbers to count the alternating permutations according a new parameter
  Christiane POUPAD

N°1 Mars 2000  Noncausality and functional discretization, limit theorems for an ARHX(1) process
  Serge GUILLAS

N°2 Juin 2000  Sieve estimates via neural network for strong mixing processes
  Jian tong ZHANG

N°3 Juillet 2000  Technical tools to evaluate Hausdorff-Besicovitch measure of random fractal sets
  Alain LUCAS

N°1 Avril 2001  Rates of convergence of autocorrelation estimates for autoregressive Hilbertian processes
  Serge GUILLAS

N°2 Avril 2001  Sur l’estimation de la densité marginale d’un processus à temps continu par projection orthogonale
  Arnaud FRENAY

N°3 Juin 2001  Rate-optimal data-driven specification testing in regression models
  Emmanuel GUERRE et Pascal LAVERGNE
The inclusion of exogenous variables in functional autoregressive ozone forecasting

Julien DAMON et Serge GUILLAS