Functional Samples and Bootstrap for Predicting Sulfur Dioxide Levels

B. FERNÁNDEZ DE CASTRO

Department of Statistics and Operations Research
University of Santiago de Compostela
15782 Santiago de Compostela, Spain
(*fdcastro@usc.es*)

S. GUILLAS

CISES, University of Chicago, Chicago, IL 60637

Now at: School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
(*guillas@math.gatech.edu*)

W. GONZÁLEZ MANTEIGA

Department of Statistics and Operations Research
University of Santiago de Compostela
15782 Santiago de Compostela, Spain
(*wenceslao@usc.es*)

In this article we give enhancements of several functional techniques to forecast sulfur dioxide levels near a power plant. The data are considered as a time series of curves. Assuming a lag-one dependence, the predictions are computed using the functional kernel (with local bandwith) and the linear autoregressive Hilbertian model. We carry out the estimation with a so-called "historical matrix," which is a subsample that emphasizes uncommon shapes. We introduce a bootstrap method to evaluate the range of the forecasts, which uses Fraiman and Muniz's order for functional data. Finally, we compare our functional techniques with neural networks and semiparametric methods, and find that the former models are often more effective.

KEY WORDS: Air pollutant; Bootstrap methods; Data depth; Functional data; Time series.

1. INTRODUCTION

In various fields, including environmental science, finance, the study of traffic, and biology, large datasets are available essentially by real-time monitoring. Nowadays, computers can manage such databases. The object of statistical study can then be curves and not numbers (or vectors). The central idea is to aggregate consecutive discrete recordings and to view them as sampled values of a random curve. Functional data analysis (Ramsay and Silverman 1997), which has as a fundamental tool principal components analysis (PCA) (Rice and Silverman 1991; Silverman 1996), appears to be an efficient framework for dealing with such statistical elements. This approach becomes more and more popular in the statistical community because of its ability to aid understanding of the whole evolution of a stochastic process. For instance, Besse and Cardot (1996) and Besse, Cardot, and Stephenson (2000) compared forecasts of traffic and the so-called *el Niño* climatic variation based on seasonal autoregressive integrated moving average (SARIMA) and functional autoregressive models (Bosq 2000). The latter models showed better predictive skills on these datasets than the classical SARIMA model. Indeed, using the great bulk of information entails better predictions, especially when the stochastic process has some functional property, such as some particular smoothness (a Sobolev space is then a natural framework) or some typical variation. Here we actually did not prescribe any smoothness—because preliminary tests showed that our predictions are likely to be flawed in the case of sharp variations, which are of great interest—but we voluntarily selected the shapes of our curves by considering a subsample of increases, decreases, plateaus, changes, and everything else.

The article is oriented toward prediction, and the adequacy of the model may be viewed as a byproduct of its excellent results compared with some techniques that already outdo ARIMA models (García Jurado, González Manteiga, Febrero Bande, Prada Sánchez, and Cao 1995) on this type of data. So far, there is no test in the literature addressing the diagnostic issue. Theoretical studies should be carried out in the model-checking direction (e.g., independence tests for functional data), to confirm the adequacy of such a functional model. But this work would be beyond the scope of this article.

In this article we deal with ground-level sulfur dioxide (SO_2) around a power plant. We want to generate a 30-minute forecast of the concentration at a particular monitoring station, using previously recorded data. The nature of such pollution is very local, in comparison with ozone data, for instance, and sudden changes in the measured SO_2 concentrations illustrate that point. García Jurado et al. (1995), Lapenna, Macchiato, Cosmi, Ragosta, and Serio (1996), Schlink, Herbarth, and Tetzlaff (1997), Andretta et al. (2000), and Fernández de Castro et al. (2003) have addressed the forecasting issue of ground-level SO_2 by means of time series methods, as well as neural networks for the two latter references.

The goals of our study were to examine the predictive power of the kernel-based method and of the autoregressive of order-one Hilbertian model [ARH(1)] on our dataset, and also to furnish some new practical tools for a better and more reliable prediction. Following García Jurado et al. (1995),

TECHNOMETRICS, MAY 2005, VOL. 47, NO. 2
DOI 10.1198/004017005000000067

we considered a subsample of the original sample—a so-called "historical matrix"—to avoid the overrepresentation of repeated data and better take into account some very informative episodes. We show that this kind of estimation improves the forecasts. Moreover, when we classified the data in the historical matrix by shapes and not by levels, we obtained even better forecasts. To deal with the confidence of our predictions, we chose to give some bootstrap prediction functional intervals. More precisely, we present a range for the possible curves instead of confidence curves. We constructed this as the convex envelope generated by the deepest curves according to a functional distance of Fraiman and Muniz (2001). The deepest curves were those closest to the median curve.

The article is organized as follows. Section 2 describes our dataset of SO_2 concentrations. Section 3 is devoted to the presentation of the methodology used in the sequel, that is, the functional models and the bootstrap procedures. Section 4 evaluates the quality of the functional forecasts and of the bootstrap replications, and carries out a comparison with some real data techniques used in recent years, including semiparametric and neural networks methods.

2. DATA

We use these functional models to forecast SO_2 levels around a power plant located in As Pontes in Northwest Spain. An air-quality control system was installed to measure the amount of pollution that the power plant emits in the area. This system includes 17 monitoring stations located within a 30-km radius. Among other pollutants, the stations measure SO_2 levels and continuously send the values to the power plant.

A major concern is to prevent air-quality episodes caused by high levels of SO_2 on the ground. Figure 1 displays a plot and a histogram of the measurements obtained from April 2002 to July 2002. The SO_2 values registered around the power plant are generally near zero for long time periods. During unfavorable meteorologic conditions, these levels can quickly rise and cause an air-quality episode. Our main interest is to forecast

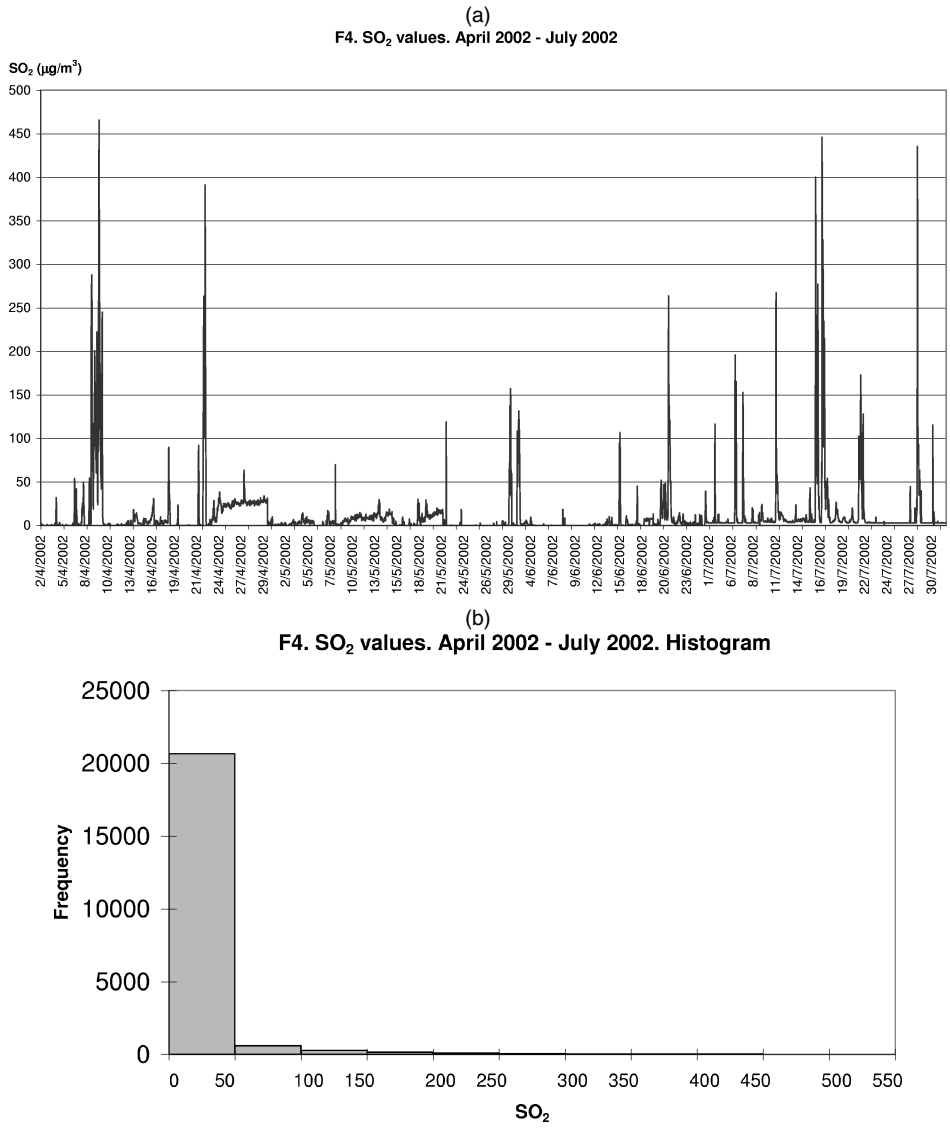

Figure 1. Plot (a) and Histogram (b) of SO_2 Values at Station F4. Measurements every 5 minutes, April–July 2002.

these episodes. The aim of such forecasts is to provide useful information to the power plant staff, so that they can try to reduce the SO_2 level by exchanging high-sulfur coals for low-sulfur coals in the power plant furnace.

Current legislation forces the power plant to control hourly averaged SO_2 values. The communication system in use at the power plant records a new SO_2 datum, for each station, every 5 minutes. Thus we can recalculate the SO_2 average over the last hour at 5-minute intervals. We are interested in forecasting what will happen to those SO_2 concentrations during the next half-hour, so each functional datum represents a half-hour period for our analysis. We restrict our study to one station, denoted by F4 in the sequel.

The meteorologic conditions are quite important for determining when and where an air-quality episode will occur. The power plant's smokestack is 356 meters high, but measurements are available for temperature, wind speed, and wind direction at only 10 and 80 meters. These meteorologic variables do not contain enough information to help substantially in forecasting such episodes. Therefore, we rely only on past SO_2 levels, which contain most of the information. Furthermore, when examining the time series, we did not know whether any changes in the composition of the coal were made. Accordingly, some episodes showing a sudden falloff were difficult to forecast.

3. METHODOLOGICAL ASPECTS

3.1 Functional Models

We treat our half-hour samples, consisting of six consecutive data, as observations of the continuous-time stochastic process that models the SO_2 levels. When attempting to forecast the future values $[x(u),\ u \geq T]$ of a continuous-time stochastic process, the idea is to try to use the information contained in the infinite number of variables of the past $[x(u),\ u \leq T]$, by considering portions of this stochastic process as curves. Several authors have already worked in the specific area of curve prediction. Bosq (2000) provided a theoretical study of linear processes that take their values in function spaces. Some progress was made in this area by Besse and Cardot (1996) and Besse et al. (2000) using smoothing splines. We chose not to include any smoothness restriction for the curves, because our intent was to precisely forecast sharp increases or decreases. We restricted our analysis to a lag-one dependence because our curves will be sampled with six points, that is, every 5 minutes per half hour, which seems sufficient to cover the evolution of the process. We thus considered random variables with values in $H = L^2([0; 6])$ in the following way: $\mathbf{X}_n(u) = x(6n + u)$, for $u \in [0; 6]$, $n \in 1, 2, \ldots$. With more than six data points per half-hour, the curves could be described with more detail. In the future, the power plant will be able to provide a measurement every minute, and thus our functional methods extend beyond what is usually seen as the multivariate framework.

Note that Mourid (2002) investigated the case of order p, $p > 1$, autoregressive processes, proposing to look at them as a particular case of ARH(1) processes. Indeed, by replacing the base Hilbert space H by the product H^p, it is possible to carry out estimation and prediction procedures with the same techniques presented later for the ARH(1) model. Damon and Guillas (2002) added explanatory variables to the ARH(1) model to make predictions of ozone levels. Recently, Guillas (2002) studied an ARH with nonadditive inclusion of such variables.

Now, let $(\boldsymbol{\varepsilon}_n)$ be a Hilbertian strong white noise (SWN), that is, a sequence of iid H-valued random variables satisfying

$$E\boldsymbol{\varepsilon}_n = \mathbf{0}, \qquad 0 < E\|\boldsymbol{\varepsilon}_n\|_H^2 = \sigma^2 < \infty, \qquad n \in \mathbb{Z}.$$

We consider the statistical model

$$\mathbf{X}_n = \rho(\mathbf{X}_{n-1}) + \boldsymbol{\varepsilon}_n, \tag{1}$$

where $\rho : H \to H$ is a function to be estimated. The operator ρ and its estimators are functions from H to H in theory, but the estimation resorts to a finite sample, as is common in many nonparametric procedures. The estimation technique takes into account these model assumptions. The resulting estimators are then applied to arguments that are in $\mathbb{R}^6$.

In the context of real-valued time series, García Jurado et al. (1995) introduced the notion of historical matrix (HM) to carry out the estimation more efficiently. There are only a few high-level episodes in history, often quite sparse in time. Taking into account that our main purpose is to give accurate predictions for air-quality episodes, we need a mechanism for storing those limited data intelligently. The idea of historical matrix allows us to have data of the whole range of variation of the variable of interest at every instant. García Jurado et al. (1995) embodied in a matrix vectors of the form (x_{t-1}, x_t, x_{t+6}), where they predicted x_{t+6} using x_{t-1} and x_t. They classified those vectors into 10 bins according to the value of the response x_{t+6}. The design of the historical matrix is completely heuristic and is based on the experience with this type of data. The notion of historical matrix has already been used for semiparametric models (García Jurado et al. 1995) and for neural networks (Fernández de Castro et al. 2003), obtaining good results in prediction.

We followed the previous approach for real data, adjusting it to our curve data. For our study, the original sample contained all of the data from 2001. We looked through the entire year and stored curves with relevant information. Our matrices are filled with 1,500 vectors of the form $(\mathbf{X}_n, \mathbf{X}_{n+1})$, where now each data $\mathbf{X}_n$ is composed of six consecutive SO_2 measures. Each vector $(\mathbf{X}_n, \mathbf{X}_{n+1})$ from 2001 is classified into a bin according to two different classification procedures:

1. An "ordinary" classification. We build this matrix following the approach made in the real case (García Jurado et al. 1995). This functional historical matrix is divided into 10 bins. Each bin is associated with a range of real SO_2 levels for the last value of $\mathbf{X}_{n+1}$: $[0, 30)$, $[30, 60)$, $[60, 100)$, $[100, 200), \ldots, [600, 700)$, and $[700, \infty)$. Most values in history belong to the first two intervals. If we define the first bin for values between 0 and 100, then most functional vectors in that bin will have values near 0, and we will not have enough information on functional vectors whose values are around 70 or 80 $\mu g/m^3$. Thus we divide the range into small intervals for low values. As we increase the SO_2 level, the number of real data in history decreases, so we define intervals of length 100 for values larger than 100 $\mu g/m^3$. Almost no air-quality episodes exceed 700 $\mu g/m^3$, so the last bin consists of vectors associated with values higher than 700 $\mu g/m^3$. As a result of

the way we built the intervals, we obtain 10 bins. With this design, we are sure to capture information on the whole range of variation of the variable. Given a functional vector $(\mathbf{X}_n, \mathbf{X}_{n+1})$ in our sample, we look at the last real value of the response $\mathbf{X}_{n+1}$ and we place it in the bin associated with this value. This design appears to cover the information needed for estimation. We fix a maximum length for the bins, allowing 200 functional vectors in each bin. We go through data of year 2001, building the functional vectors $(\mathbf{X}_n, \mathbf{X}_{n+1})$, and classifying them into the described bins. Once a bin is full, the new vector replaces the oldest one. In this case, with data from year 2001, this results in a set of approximately 1,500 functional vectors, because there are not enough data to fill the last bins. We refer to this matrix as the "historical matrix of levels."

2. A "functional" classification. The functional historical matrix is divided into five bins, each bin of which has an associated curve shape. We distinguish five curve shapes: increases, decreases, plateaus, changes, and everything else. To determine which class $\mathbf{X}_n = (X_n^1, \ldots, X_n^6)$ belongs to, we compute the differences,

$$(X_n^2 - X_n^1, \ldots, X_n^6 - X_n^5).$$

When the absolute value of a difference is strictly <5, it is regarded symbolically as a 0. When a difference is >5, it is regarded as a "+". When a difference is <-5, then it is regarded as a "−". That is, each curve $\mathbf{X}_n$ has a vector of five associated differences; then those differences are replaced by signs and/or 0's. The five bins of the historical matrix are defined as follows: five + for an increase, five − for a decrease, five 0 for a plateau, at least one + and one −, and no 0 for a change. The changes' category is difficult to fill out because there were only a few in the sample, but it contains a great amount of information. Every functional vector $(\mathbf{X}_n, \mathbf{X}_{n+1})$ is placed into the bin according to the shape of the response $\mathbf{X}_{n+1}$, replacing the oldest functional vector in that bin. Thus five classes of 300 couples of the form $(\mathbf{X}_n, \mathbf{X}_{n+1})$ compose our historical matrix, using the entire curve $\mathbf{X}_{n+1}$ to classify them. It is quite difficult to classify all of the different shapes that can be found in real data. We summarize all of those cases in five types of shapes, including the most informative ones for the present problem. We refer to this matrix as the "historical matrix of shapes."

At the end of this procedure, the historical matrices, of either levels or shapes, are sets of the form $\{(\mathbf{X}_{i_1}, \mathbf{X}_{i_1+1}), (\mathbf{X}_{i_2}, \mathbf{X}_{i_2+1}), \ldots, (\mathbf{X}_{i_N}, \mathbf{X}_{i_N+1})\}$. For instance, with the aforementioned choices in the case of the "functional" classification, and when the historical matrix is complete, $N = 1{,}500$, the first 300 couples are in the first category, the next 300 couples are in the second category, and so on. We use these matrices as samples for the various procedures of estimation. We estimate the operator ρ using the historical matrix and forecast on a validation dataset distinct from the historical matrix.

Results of preliminary empirical studies without historical matrix suggested that such a choice in our sample was relevant (see Sec. 4 and the corresponding figure therein). Let $\hat{\mathbf{X}}$ denote the forecast for the random variable $\mathbf{X}$. Our procedures use the following empirical L^p-errors, for integers $p = 1, 2$, on a sample of n half-hours (or six 5-minute datum):

$$\|\hat{\mathbf{X}} - \mathbf{X}\|_{L^p} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{6}\sum_{j=1}^{6}|\hat{X}_i^j - X_i^j|^p\right]^{1/p}.$$

We also examine the L^∞-error, which is calculated as

$$\|\hat{\mathbf{X}} - \mathbf{X}\|_{L^\infty} = \frac{1}{n}\sum_{i=1}^{n}\sup_{j=1,\ldots,6}|\hat{X}_i^j - X_i^j|.$$

3.1.1 Autoregressive Hilbertian Model. Here ρ is a bounded linear operator on H. We may call this model the "linear" one, in comparison with the more general model (1). Bosq (2000) proved that under mild assumptions, this autoregressive model has a unique stationary solution. The estimation of the operator ρ relies on the relation $D = \rho C$, where D is the cross-covariance operator given by $D(\mathbf{x}) = E[\langle \mathbf{X}_0, \mathbf{x}\rangle \mathbf{X}_1]$ and $C(\mathbf{x}) = E[\langle \mathbf{X}_0, \mathbf{x}\rangle \mathbf{X}_0]$. Because C generally is not invertible, a projections on a subspace of finite dimension k_n yields an estimate of ρ. Here k_n is linked to the decreasing rate of C's eigenvalues (Bosq 2000; Guillas 2001). Estimation of ρ is carried out in several steps after withdrawing the mean from the process:

1. Compute, by PCA, empirical estimators of the eigenelements of the covariance operator C associated with $(\mathbf{X}_n)$.
2. Project the relation between the cross-covariance operator, D and C,

$$D = \rho C,$$

   in the subspace spanned by the first k_n eigenvectors of associated with the k_n greatest empirical eigenvalues of C.
3. Get a consistent estimator ρ_n of ρ using this projected relation whenever it is possible, because of invertibility conditions. Several assumptions can ensure this, especially by taking k_n of a small size in comparison with the sample size n. Depending on the decay rate of the eigenvalues of C, the usual (and optimal) k_n's are of $\log(n)$ or $n^{1/\alpha}$, $\alpha > 1$ types (see Bosq 2000 for the almost-sure mode of convergence and Guillas 2001 for the L^2 mode). Our number of sample points was small (equal to six), our sample size was very large (i.e., 1,500), and the k_n was chosen by cross-validation on 200 curves in the historical matrix was 6. Keep in mind that k_n cannot be greater than this sample size, because in the computing process we have only information for vectors of size 6.

Note that this statistical model is nonparametric, because ρ is an infinite-dimensional parameter.

We carried out estimation of C and D using data from the historical matrix $\{(\mathbf{X}_{i_1}, \mathbf{X}_{i_1+1}), (\mathbf{X}_{i_2}, \mathbf{X}_{i_2+1}), \ldots, (\mathbf{X}_{i_N}, \mathbf{X}_{i_N+1})\}$. Each functional vector of the historical matrix is interpreted as a piece of time series: two consecutive curves $(\mathbf{X}_i, \mathbf{X}_{i+1})$. Once the operator ρ is estimated, we perform predictions over an air-quality episode not included in the historical matrix.

3.1.2 Functional Kernel. It may sound too restrictive to consider only linear operators for such dependent samples of curves. Besse et al. (2000) proposed extending the classical Nadaraya–Watson kernel regression estimator to a functional context, and this was our approach. Note that in the case of a regression on functional data with a scalar response, Ferraty and Vieu (2000, 2002) and Ferraty, Goia, and Vieu (2002) established asymptotic results linked to the fractal dimension of the functional process.

Then ρ can be estimated by the following functional kernel estimator (Besse et al. 2000), using the n functional vectors $(\mathbf{X}_i, \mathbf{X}_{i+1})$ of the historical matrix given by $\{(\mathbf{X}_{i_1}, \mathbf{X}_{i_1+1}), (\mathbf{X}_{i_2}, \mathbf{X}_{i_2+1}), \ldots, (\mathbf{X}_{i_N}, \mathbf{X}_{i_N+1})\}$:

$$\hat{\rho}_{h_N}(\mathbf{x}) = \frac{\sum_{j=1}^{N} \mathbf{X}_{i_j+1} \cdot K(\|\mathbf{X}_{i_j} - \mathbf{x}\|/h_N)}{\sum_{j=1}^{N} K(\|\mathbf{X}_{i_j} - \mathbf{x}\|/h_N)}, \tag{2}$$

where K denotes a kernel (our choice was the Gaussian kernel), N is the sample size, h_N is the bandwidth, and $\mathbf{x}$ belongs to H. The global bandwith was chosen by cross-validation on a subsample of 200 points. We also decided to look at local bandwidths, because they should yield more precise weights in (2). Indeed, a local bandwidth selection can lead to using a large number of curves in the computations of the regression weights when a curve is somehow odd, whereas it allows one to use a small number of very near curves to avoid very different ones when a curve has many neighbors in the $L^2([0; 6])$ metric. Our method was then pointwise.

We chose to perform the so-called "leave-one-out" cross-validation procedure for local bandwith determination. Let $\mathbf{X}_{i+1}$ be the data that we want to forecast from $\mathbf{X}_i$. We picked up in the historical matrix a relatively small number of data (*nvc* ones), which are the closest to the current one $\mathbf{X}_i$ in the L^2 norm. We took *nvc* equal to 100. For each point in this cross-validation sample, we computed the optimal bandwith using the other $nvc - 1$ points of this sample, by calculating the prediction errors made on those $nvc - 1$ points. We selected the optimal bandwith for $\mathbf{X}_i$ as the optimal bandwith associated with the lowest prediction error.

3.2 Bootstrap Techniques

From a statistical standpoint and to take a good decision, it is important to provide an idea of the range of the forecasts whenever possible. Bootstrap techniques in our context of dependent Hilbert space-valued random variables have been investigated by Politis and Romano (1994) as a means of calculating standard errors of estimators and constructing confidence regions for parameters. But our purpose is to give confidence regions for our predictions. Cao (1999) presented several bootstrap methods for real-valued time series prediction with or without an explicit dependence structure. We adapted those methods to our functional data framework as follows. We denote by $\mathbf{X}_i$ the curves stored in the historical matrix and by $\mathbf{Y}_i$ the curves for which we want to forecast $\mathbf{Y}_{i+1}$. Clearly, the two samples are distinct, because the $\mathbf{X}_i$'s are from year 2001 and the $\mathbf{Y}_i$'s are not. We denote by N the number of pairs $(\mathbf{X}_i, \mathbf{X}_{i+1})$ in our historical matrix. We make use of $p = 100$ bootstrap forecasts in the two following cases.

3.2.1 Bootstrap for Kernel-Based Predictions. As described by Cao (1999), the proposed resampling method will give us an idea of the range of our predictions, assuming that our time series is a Markov process. The algorithm proceeds as follows to draw p bootstrap one-step-ahead forecasts $\mathbf{Y}^*_{m+1,1}, \ldots, \mathbf{Y}^*_{m+1,p}$ at point $\mathbf{Y}_m$:

1. From the historical matrix, construct the sample blocks of length two,
$$\mathbf{B}_j = \{\mathbf{X}_j, \mathbf{X}_{j+1}\}, \qquad j = 1, \ldots, N.$$
2. Compute the probabilities
$$\hat{p}_j = \frac{K(\|\mathbf{X}_j - \mathbf{Y}_m\|/h)}{\sum_{i=1}^{N} K(\|\mathbf{X}_i - \mathbf{Y}_m\|/h)},$$
with h the optimal bandwith (global or local)
3. Randomly toss p blocks $\{\mathbf{Y}^*_{m,i}, \mathbf{Y}^*_{m+1,i}\}$ with probability $\hat{p}_j$ of choosing $\{\mathbf{Y}^*_{m,i}, \mathbf{Y}^*_{m+1,i}\} = \{\mathbf{X}_j, \mathbf{X}_{j+1}\}$, and extract from them the second element $\mathbf{Y}^*_{m+1,i}$, obtaining in this way a sequence (with possible repetitions) of replications $\mathbf{Y}^*_{m+1,1}, \ldots, \mathbf{Y}^*_{m+1,p}$.

Note that the weights $\hat{p}_j$ computed in the second step measure the proximity of the current data with the data considered in the historical matrix. The forecasts $\mathbf{Y}^*_{m+1,1}, \ldots, \mathbf{Y}^*_{m+1,p}$ are given merely by past observations $\mathbf{X}_{j+1}$'s and are drawn at random from the sample, according to the probabilities $\hat{p}_j$. With real time series, it is then possible to compute the empirical distribution function on the replications, and therefore the confidence intervals. Unfortunately, because we deal with functional data, these calculation do not give insightful results. We may evaluate the empirical distribution function in each sampling point t where we have a real value of our functional data. But the curves built with these points (e.g., those composed of values produced as limits of the 90% confidence intervals) will not really measure the scatter. Indeed, Fraiman and Muniz (2001) (see also Ramsay and Silverman 1997) pointed out that for functional data, using finite-dimensional techniques is not accurate. They propose the following notion of depth, which we use to calculate the proximity of a curve to the median one. Hence, we use curves that are relatively close to this median (e.g., the 90% less distant), and treat their convex envelope as possible limiting curves for our predictions.

Let us recall how the functional depth of Fraiman and Muniz (2001) is constructed. The usual univariate depth D is a measure of centrality. Let F be the cumulative distribution function of a certain real random variable, and let x be a data point. Then D is defined by

$$D(x) = 1 - \left|\tfrac{1}{2} - F(x)\right|.$$

The depth measures the nearness to the median *med*, for which we have $D(med) = 1$, the maximum value of D. Multivariate depth can be defined (e.g., Tuckey's depth or simplicial depth), but are not easy to compute empirically for high dimensions and are not adapted to functional data, which show some particular features. Consider a functional dataset $\mathbf{X}(t), t \in [a, b]$, and a sample $\mathbf{X}_1(t), \ldots, \mathbf{X}_p(t)$ following the same distribution as $\mathbf{X}(t)$. For each sample point t, we calculate the empirical distribution

$$F_{p,t}(\mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \mathbb{1}_{\mathbf{X}_j(t) \le \mathbf{x}}. \tag{3}$$

Actually, we need only use $F_{p,t}(\mathbf{X}_i(t))$ in our calculations. Let D_p be the empirical univariate depth,

$$D_{p,t}(\mathbf{x}) = 1 - \left|\tfrac{1}{2} - F_{p,t}(\mathbf{x})\right|.$$

Fraiman and Muniz proposed looking at the following integrated index:

$$I_i = \int_a^b D_{p,t}(\mathbf{X}_i(t))\, dt.$$

Here I_i globally measures the nearness to the empirical median, which can be defined here as the curve for which the index is maximum. We put our replications in a descending order with respect to this index, obtaining the order statistics

$$Y^*_{m+1,1:1}, \ldots, Y^*_{m+1,p:p}.$$

3.2.2 Bootstrap for ARH Predictions. When a specific dependence structure is available, such as the lag-one autoregression operator, more accurate methods enable us to obtain bootstrap replications faster and in a more appropriate way. We followed Cao, Febrero Bande, González Manteiga, Prada Sánchez, and García Jurado (1997) (see also Cao 1999). Here are the steps:

1. Compute the forward residuals for $i = 2, \ldots, n+1$,

$$\hat{\mathbf{a}}_i = \mathbf{X}_i - \hat{\rho}\mathbf{X}_{i-1},$$

and their corrected version,

$$\hat{\mathbf{a}}'_i = \hat{\mathbf{a}}_i - \bar{\mathbf{a}},$$

where

$$\bar{\mathbf{a}} = \frac{1}{n}\sum_{i=2}^{n+1} \hat{\mathbf{a}}_i.$$

2. Conduct a PCA of the $\hat{\mathbf{a}}'_i$ in the following manner:

$$\hat{\mathbf{a}}'_i = \mathbf{c}^i_1 V_1 + \cdots + \mathbf{c}^i_{k_n} V_{k_n}.$$

3. Derive for each coordinate $\mathbf{c}_l$ its empirical distribution function, $F_n^{\mathbf{c}_l}$, $l = 1, \ldots, k_n$.
4. Generate $\mathbf{c}^*_l$, with distribution function $F_n^{\mathbf{c}_l}$, $l = 1, \ldots, k_n$, and construct the bootstrap residuals,

$$\hat{\mathbf{a}}^*_i = \mathbf{c}^*_1 V_1 + \cdots + \mathbf{c}^*_{k_n} V_{k_n}.$$

5. Generate bootstrap replications,

$$\mathbf{Y}^*_{m+1,i} = \hat{\rho}\mathbf{Y}_m + \hat{\mathbf{a}}^*_i.$$

We carried out a PCA to simulate a functional noise. We exploited the fact that it is easier to simulate several uncorrelated random variables than a multivariate one. Indeed, the random coordinates $\mathbf{c}_l$ are not correlated, whereas the sample points values of a curve are strongly correlated. We then classified our replications in the same way as in Section 3.2.1.

4. RESULTS

4.1 Predictions

We worked on two different days: April 22, 2002 and June 21, 2002. Both days show a peak, but not at the same level, and with different ways of going to the peak. The forecasts are displayed in Figures 2–5. One has to look carefully at those figures, because every 30 minutes we are joining the last predicted real value, $\hat{X}^6_{n-1}$, at time $n-1$ and the first predicted real, value, $\hat{X}^1_n$, at time n. Hence there is a piece of line that is not really predicted. Note that when the levels are around zero, the forecast is typically an increase, which is usually a wrong prediction but is not a serious error. Indeed, the interesting part of the curve is when the concentration exceeds a threshold of 150 $\mu g/m^3$, because it roughly corresponds to a level of intervention for the plant staff. This part is quite well forecasted, in comparison to reincreasing parts that triggered large errors. Furthermore, the peaks are better predicted when using a shape-fitted historical matrix.

For April 22, 2002, Table 1 indicates that on the episode (i.e., when the levels are high), the best model is the ARH

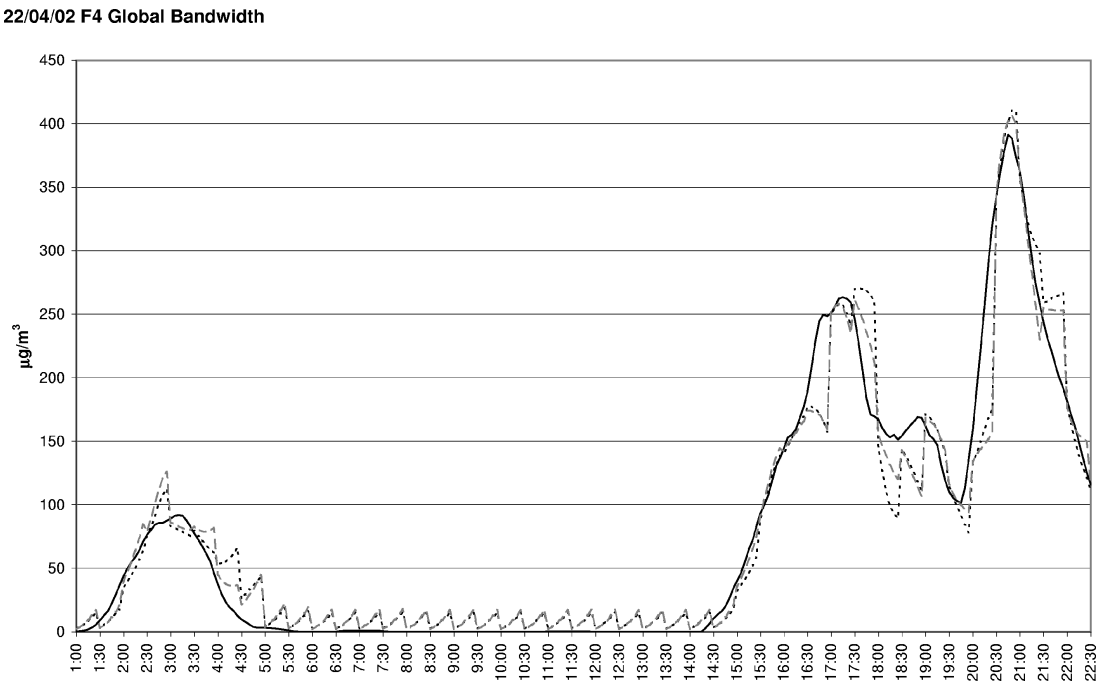

Figure 2. Functional Kernel Forecasts, Global Bandwidths, April 22, 2002. (——, real X_t; ·····, Pred2 with classical historical matrix; – – –, Pred4 with the historical matrix built using shapes.)

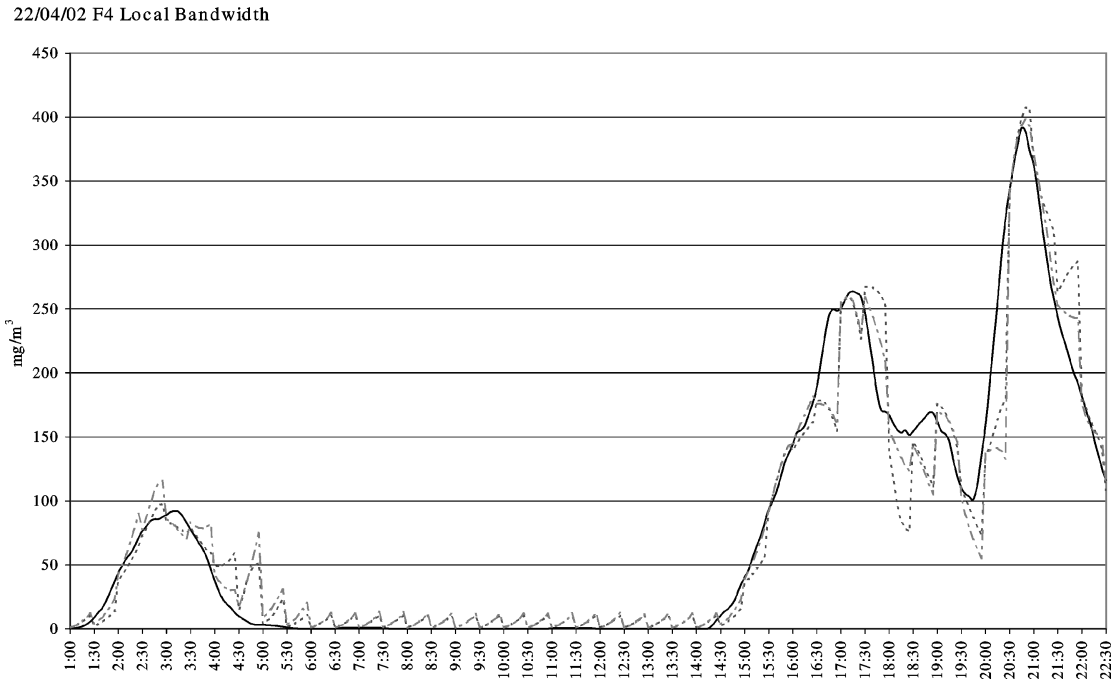

Figure 3. Functional Kernel Forecasts, Local Bandwidths, April 22, 2002. (——, real X_t; ·····, Pred1 with classical historical matrix; – · – ·, Pred3 with the historical matrix built using shapes.)

one with a shape-fitted historical matrix. Without the historical matrix, the ARH model performs the best, because its answers are zero when the levels are low, but as explained before, this is less important than being able to predict well high levels (which is not the case for this particular model). Moreover, the functional kernel gives systematically more accurate forecasts using the shape-fitted historical matrix.

Figure 6 shows how much the local bandwith is linked to the level. Indeed, the higher the level, the fewer the number of close curves, and therefore the larger the bandwith.

In Figure 7 presents an example of our half-hour forecasts together with the real future curve, the empirical median $Y^*_{m+1,1:1}$ of our bootstrap replications, and the convex envelope generated by the 90% close replications, which are the first 90% ones in the ordered list.

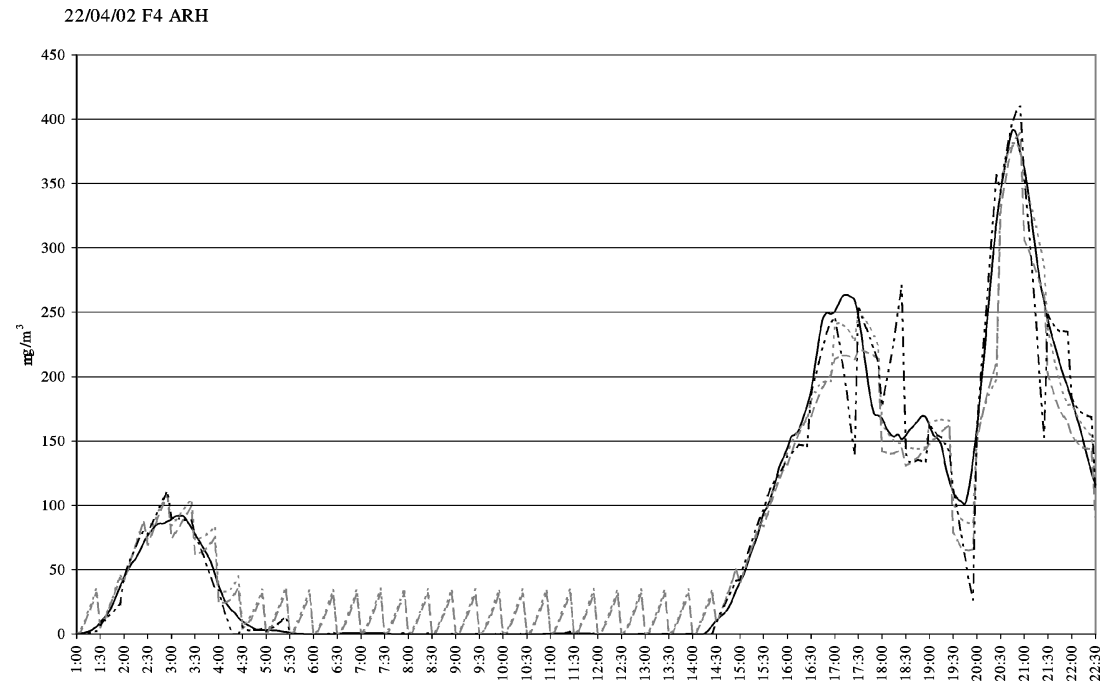

Figure 4. Autoregressive Hilbertian Forecasts, April 22, 2002. (——, real X_t; ·····, ARH-2 with the historical matrix built using shapes; – – –, ARH-1 with the historical matrix built using levels; – · · –, ARH-0 without historical matrix.)

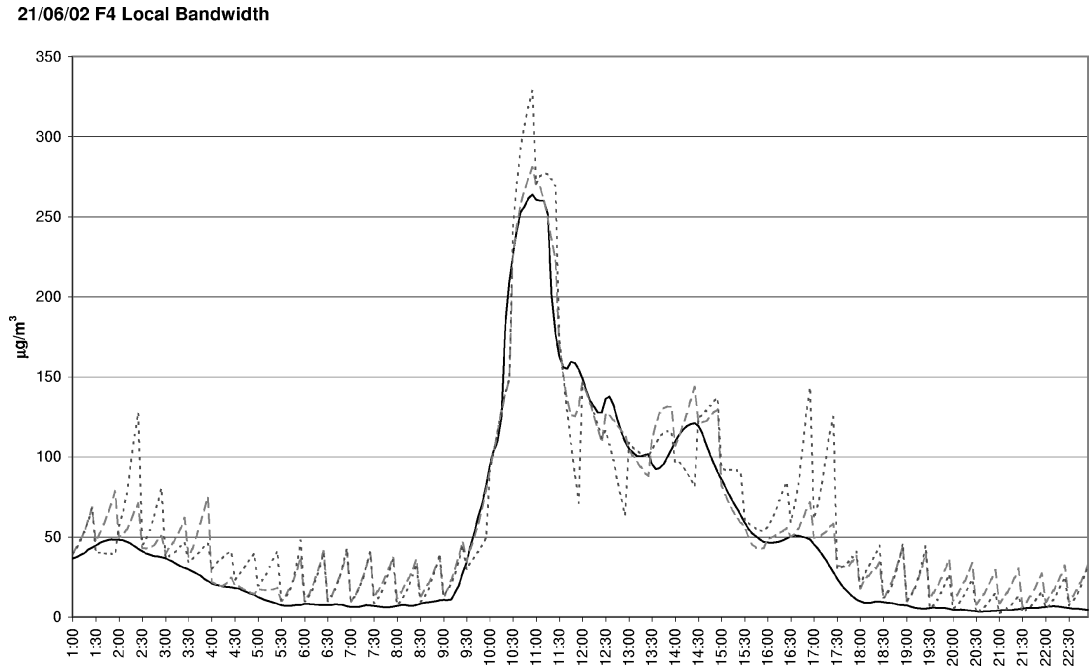

Figure 5. Functional Kernel Forecasts, Local Bandwidths, June 21, 2002. (——, real X_t; ·····, Pred1 with classical historical matrix; – – –, Pred3 with the historical matrix built using shapes.)

Table 1. Prediction Errors at Station F4 on April 22, 2002

	Error			
Model	L^1	L^2	L^∞	L^2 *on episode*
FK local bandwidth, HM-level	16.14	18.27	28.12	32.88
FK global bandwidth, HM-level	16.66	18.65	28.52	30.03
FK local bandwidth, HM-shape	14.61	16.78	26.96	26.70
FK global bandwidth, HM-shape	15.26	17.36	27.60	26.04
ARH, no HM	10.84	12.88	21.31	28.61
ARH, HM-level	16.57	19.65	31.67	25.85
ARH, HM-shape	15.24	18.75	31.74	20.42

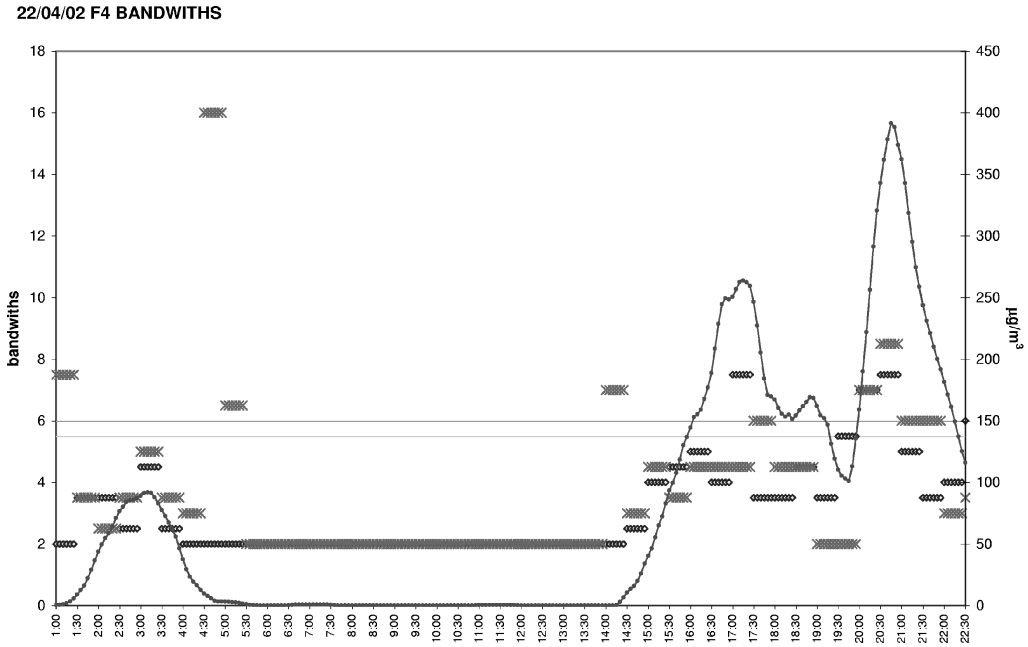

Figure 6. Local and Global Bandwidths for Functional Kernel Forecasts on April 22, 2002. (◇, Pred1 bandw; - - - -, Pred2 bandw; ×, Pred3 bandw; — – —, Pred4 bandw; —•—, real X_t.)

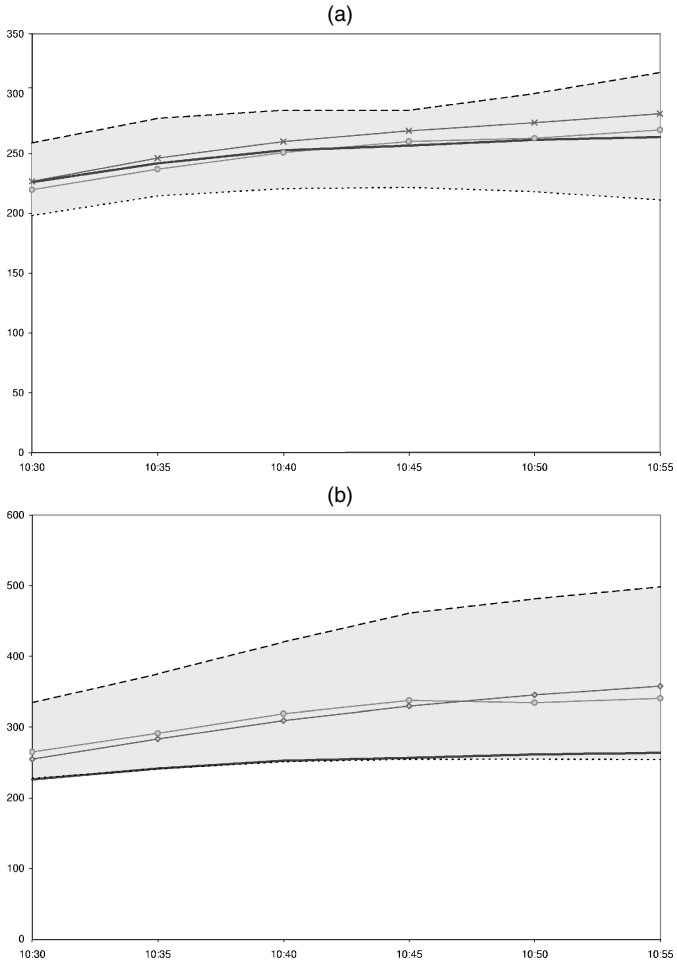

Figure 7. Bootstrap for (a) the Functional Prediction Method, Local Bandwidth, Shaped-Based Historical Matrix (—×—, Pred4), and (b) the ARH Model (—◦—), June 21, 10:30, in $\mu g/m^3$. For both panels: —◦—, median; ——, real; - - -, max; ·····, min.

ARH replications are model-based, whereas the functional kernel replications are model free. Sherman (1998) addressed this issue and proved that in the case of the variance estimation by a bootstrap procedure, then model-based estimators are asymptotically superior to model-free estimators in the class of real-valued ARMA(p, q) processes. Indeed, by looking at several examples, it appears that the bootstrap procedure works better for the ARH model than for the functional kernel model.

4.2 Comparison

We compared our forecasts with those obtained by two methods used in the past: neural networks (Fernández de Castro et al. 2003), and semiparametric models (García Jurado et al. 1995). Note that semiparametric models surpass ARIMA models (García Jurado et al. 1995) on this type of data. To do so, we contrasted the 30-minutes-ahead forecasts every 5 minutes. The mean absolute errors (MAEs) and mean squared errors (MSEs) of each method are summarized in Tables 2 and 3. As one may notice, the functional methods presented here are very competitive. Actually, except for the MSE during April 22, 2002, the functional kernel showed more accurate forecasts. This can be seen from Figure 8, where it is possible to discern the great variability of the neural network and semiparametric predictions, especially for the peak.

Table 2. Thirty-Minutes-Ahead Prediction Errors at Station F4 on April 22, 2002

	Error	
Model	*MAE*	*MSE*
FK local bandwidth,HM-shape	24.12	1,305.76
FK global bandwidth, HM-shape	25.41	1,303.01
ARH, HM-shape	31.14	1,372.57
Neural network	27.57	1,156.64
Semiparametric	25.30	1,650.85

5. CONCLUSION

In this article we have proposed a new way of building a historical matrix focusing on the particularity of functional data, that is, classifying our data according to the shape instead of the level. This idea helped us select data with enough interesting information to estimate a model. We examined the predictions of the ARH and the functional kernel model. To find an optimal bandwidth, we computed global and local bandwidths. We compared the forecasts obtained with such functional models with those given by the semiparametric methods and neural networks models considered before for our particular problem of forecasting SO_2. These functional models appeared to be very competitive options. We exposed some ideas to the use of bootstrap techniques with such functional data. On the one hand, using the information contained in the historical matrix, we supplied a resampling method for functional kernel predictions. On the other hand, assuming the ARH dependence structure, we proposed a bootstrap technique using the residuals of the predictions. The latter technique, based on generating new residuals, seemed to yield better results than the former based only on resampling within the historical matrix. Moreover, we built a sample of extremal predicted curves, following the idea of confidence intervals for real data. To do so, we used the concept of functional depth to establish an order between our bootstrap replications. With that order, we were able to get a sample of curves that are far (in terms of depth) from the median of the replications.

In the near future, it should be interesting to include more informative variables (e.g., temperature at an altitude of > 80 meters, wind speed, wind direction) if they could be measured with a sufficient precision and accuracy. Moreover, integrating numerical methods for the assessment and prediction of such meteorologic variables combined with our statistical approach may be fruitful. Several articles have been written in that direction. For instance, Vautard, Beekmann, Roux, and Gombert (2001) built a hybrid statistical-deterministic chemistry transport model to forecast ground-level ozone, and Gelpke and

Table 3. Thirty-Minutes-Ahead Prediction Errors at Station F4 on June 21, 2002

	Error	
Model	*MAE*	*MSE*
FK local bandwidth, HM-shape	27.33	1,068.47
FK global bandwidth, HM-shape	25.11	834.64
ARH, HM-shape	31.05	1,190.64
Neural network	28.83	1,131.34
Semiparametric	28.28	1,965.26

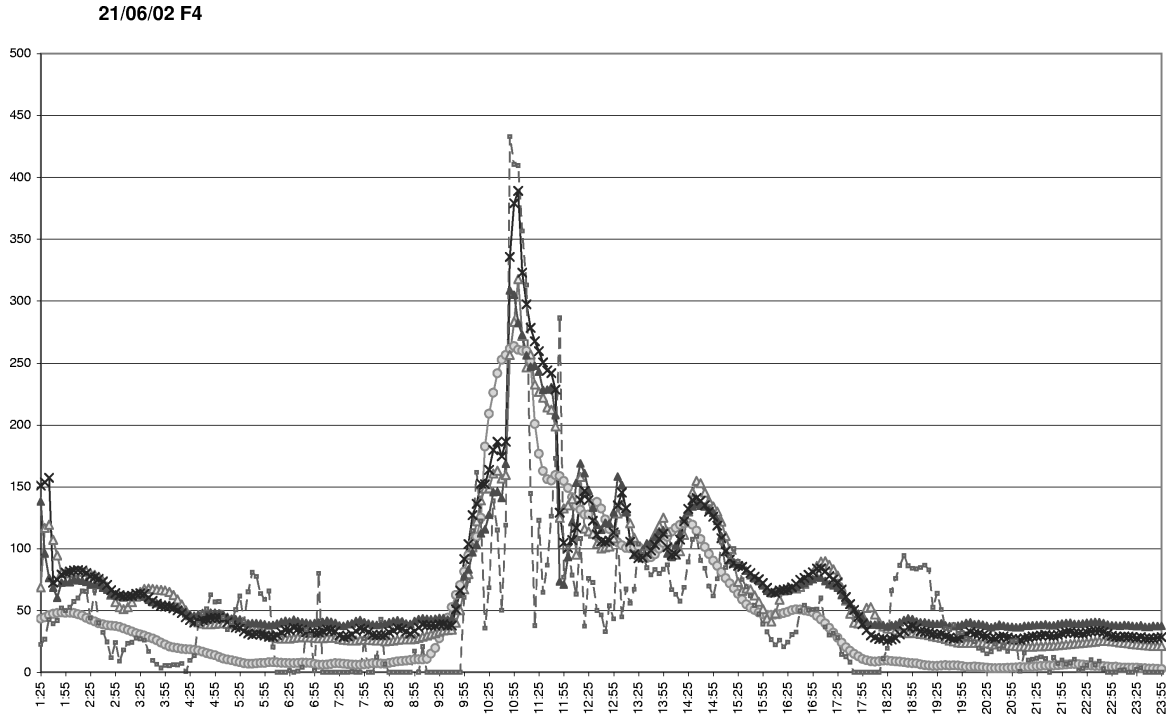

Figure 8. Comparisons of 30-Minutes-Ahead Forecasts for Functional Kernel With Global Bandwith, Shape-Based Historical Matrix (Pred4), ARH Model, Shape-Based Historical Matrix (ARH-2), Neural Network, and Semiparametric Models, in $\mu g/m^3$. (—○—, real; —△—, Pred4; —▲—, ARH-2; —×—, neural network; – – ▪ – –, semiparam.)

Künsch (2001) combined statistical procedures and partial differential equations to model the dynamic of the stratospheric ozone and measure its day-to-day variability. Furthermore, in the future the SO_2 measurements will be made every minute instead of every 5 minutes, so that the curves will be sampled at 30 points. Such an improvement is likely to entail much more precise functional data estimation, and forecasts at a larger horizon could then possibly be considered.

ACKNOWLEDGMENTS

The authors thank the editor, associate editor, and referees for their comments. They also thank Endesa Generación, U.P.T. As Pontes, for the cooperation, during last years, with the Department of Statistics and O. R. of the University of Santiago de Compostela. This research was partially supported by MCyT Grant BFM2002-03213, Xunta de Galicia Grant PGIDIT03PXIC20702PN, U.S. EPA Grant R-82940201-0.

[Received February 2003. Revised July 2004.]

REFERENCES

Andretta, M., Eleuteri, A., Fortezza, F., Manco, D., Mingozzi, L., Serra, R., and Tagliaferri, R. (2000), "Neural Networks for Sulphur Dioxide Ground-Level Concentrations Forecasting," *Neural Computing and Applications*, 9, 93–100.

Besse, P., and Cardot, H. (1996), "Spline Approximation of the Prediction of a First-Order Autoregressive Functional Process," *Canadian Journal of Statistics*, 24, 467–487.

Besse, P., Cardot, H., and Stephenson, D. (2000), "Autoregressive Forecasting of Some Functional Climatic Variations," *Scandinavian Journal of Statistics*, 27, 673–687.

Bosq, D. (2000), *Linear Processes in Function Spaces*, New York: Springer-Verlag.

Cao, R. (1999), "An Overview of Bootstrap Methods for Estimating and Predicting in Time Series," *Test*, 8, 95–116.

Cao, R., Febrero Bande, M., González Manteiga, W., Prada Sánchez, J. M., and García Jurado, I. (1997), "Saving Computer Time in Constructing Consistent Bootstrap Prediction Intervals for Autoregressive Processes," *Communication in Statistics, Part B—Simulation and Computation*, 26, 961–978.

Damon, J., and Guillas, S. (2002), "The Inclusion of Exogenous Variables in Functional Autoregressive Ozone Forecasting," *Environmetrics*, 13, 759–774.

Fernández de Castro, B. M., Prada Sánchez, J. M., González Manteiga, W., Febrero Bande, M., Bermúdez-Cela, J. L., and Hernández-Fernández, J. J. (2003), "Prediction of SO_2 Level Using Neural Networks, *Journal of the Air and Waste Management Association*, 53, 532–539.

Ferraty, F., Goia, A., and Vieu, P. (2002), "Functional Nonparametric Model for Times Series: A Fractal Approach for Dimension Reduction," *Test*, 11, 317–344.

Ferraty, F., and Vieu, P. (2000), "Fractal Dimensionality and Regression Estimation in Seminormed Vector Spaces," *Comptes Rendus de l'Académie des Sciences, Série I, Mathématique*, 330, 139–142.

——— (2002), "The Functional Nonparametric Model and Application to Spectrometric Data," *Computational Statistics*, 17, 545–564.

Fraiman, R., and Muniz, G. (2001), "Trimmed Means for Functional Data," *Test*, 10, 419–440.

García Jurado, I., González Manteiga, W., Febrero Bande, M., Prada Sánchez, J., and Cao, R. (1995), "Predicting Using Box–Jenkins, Nonparametric and Bootstrap Techniques," *Technometrics*, 37, 303–310.

Gelpke, V., and Künsch, H. R. (2001), "Estimation of Motion From Sequences of Images: Daily Variability of Total Ozone Mapping Spectrometer Ozone Data," *Journal of Geophysical Research-Atmosphere*, 106, 11825–11834.

Guillas, S. (2001), "Rates of Convergence of Autocorrelation Estimates for Autoregressive Hilbertian Processes," *Statistics & Probability Letters*, 55, 281–291.

——— (2002), "Doubly Stochastic Hilbertian Processes," *Journal of Applied Probability*, 39, 566–580.

Lapenna, V., Macchiato, M., Cosmi, C., Ragosta, M., and Serio, C. (1996), "Predictability Analysis of SO_2 Time Series by Linear and Non-Linear Forecasting Approaches," *Environmetrics*, 7, 525–535.

Mourid, T. (2002), "Estimation and Prediction of Functional Autoregressive Processes," *Statistics*, 39, 125–138.

Politis, D. N., and Romano, J. P. (1994), "Limit Theorems for Weakly Dependent Hilbert Space–Valued Random Variables With Application to the Stationary Bootstrap," *Statistica Sinica*, 4, 461–476.

Ramsay, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society*, Ser. B, 53, 233–243.

Schlink U., Herbarth, O., and Tetzlaff, G. (1997), "A Component Time-Series Model for SO_2 Data: Forecasting, Interpretation and Modification," *Atmospheric Environment*, 31, 1285–1295.

Sherman, M. (1998), "Efficiency and Robustness in Subsampling for Dependent Data," *Journal of Statistical Planning and Inference*, 75, 133–146.

Silverman, B. W. (1996), "Smoothed Functional Principal Components Analysis by Choice of Norm," *The Annals of Statistics*, 24, 1–24.

Vautard, R., Beekmann, M., Roux, J., and Gombert, D. (2001), "Validation of a Hybrid Forecasting System for the Ozone Concentrations Over the Paris Area," *Atmospheric Environment*, 35, 2449–2461.