FUNCTIONAL SAMPLES AND BOOTSTRAP FOR PREDICTING OF SO_2 LEVELS

B.M. Fernández de Castro

S. Guillas

W. González Manteiga

INTRODUCTION

Bosq (2000): theoretical study of linear processes with values in function spaces.

Besse and Cardot (1996): traffic study

Besse et al. (2000): climatic variation *El Niño*

Damon and Guillas (2002): ozone levels

Our work: ground level Sulfur Dioxide (SO_2) around a power plant.

INTRODUCTION

➢ The legislation in use in Spain forces to control air quality.

➢The power plant has an Atmospheric Pollution Supplementary Control System.

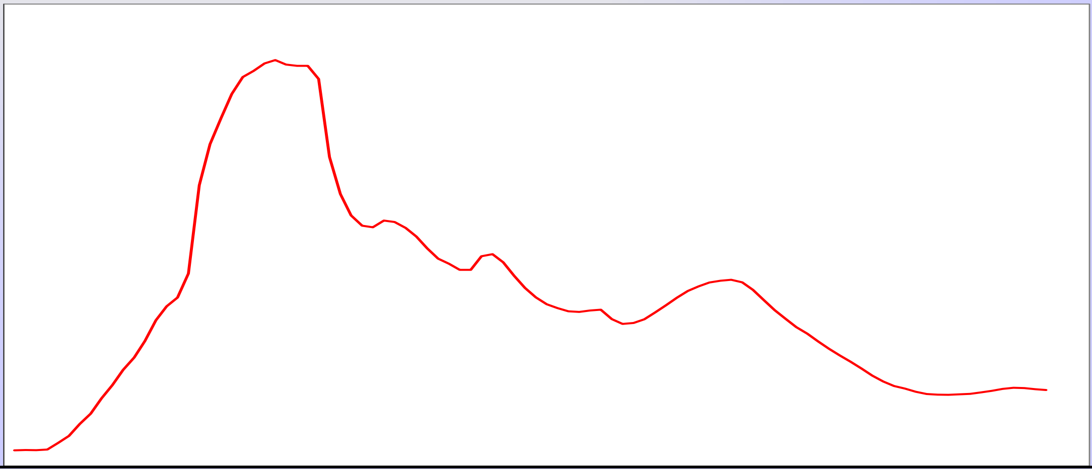
➢Prediction tools are necessary to make these systems effective.

➢An indicator of Air Quality is the SO2 level.

endesa generación
U. P. T. As Pontes

INTRODUCTION

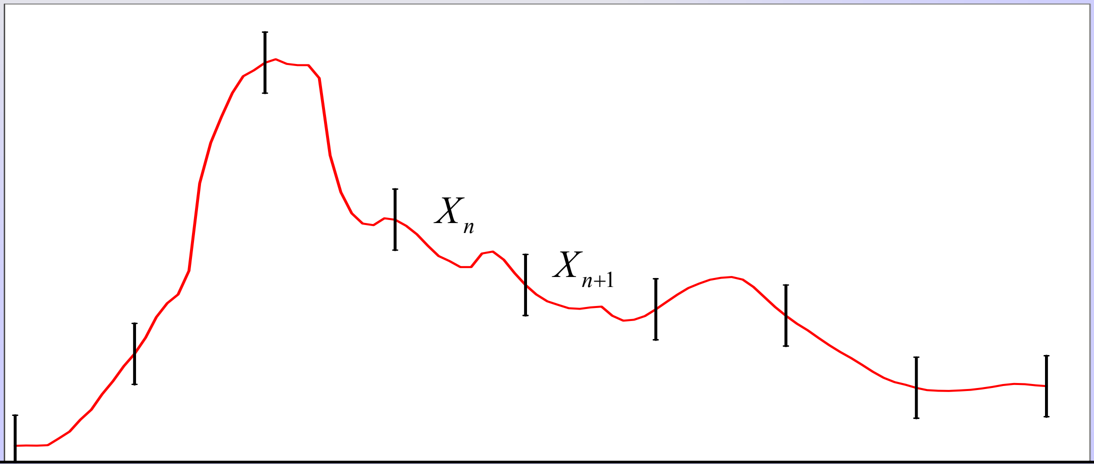
- Major Concern: **prevent air quality level episodes**.
- Legislation in use in Spain (*Real Decreto 1073/2002*) forces to control hourly average SO_2 values.
- The power plant needs at least half an hour ahead predictions.
- We will look at the time series of SO2 values as observations of the continuous-time stochastic process which models the SO2 levels.



INTRODUCTION

- Our interest are half an hour predictions.
- The communication system at the power plant gives us a new datum every 5 minutes.

We will consider random variables with values in $H = L^2([0,6])$ in the following way: $$X_n(t) = x(6n+t)$$



INTRODUCTION

1. Methodology
    1. ARH
    2. Functional Kernel
    3. Bootstrap
        - Data Depth
        - Bootstrap for kernel based predictions
        - Bootstrap for ARH predictions
2. Application
    1. Historical Matrix
    2. Results
3. Comparison
4. Conclusion

We will consider random variables with values in:

$$H = L^2\left(\left[0,6\right]\right)$$

We will forecast future values $x(t), t \geq T$ of the continuous-time stochastic process, using the information contained in a infinite number of variables of the past:

$$x(t), t \leq T$$

Let ε_n be a strong Hilbertian white noise ⇒

⇒ i.i.d. H-valued random variables with

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

Consider the statistical model:

$$X_n = \rho\left(X_{n-1}\right) + \varepsilon_n$$

Where $\rho : H \to H$ is the function to be estimated.

Our procedures use the following empirical L^p-errors, for $p=1, 2$:

$$\left\| \hat{X} - X \right\|_{L^p} = \frac{1}{n} \sum_{t=1}^{n} \left[\frac{1}{6} \sum_{j=1}^{6} \left| \hat{X}_t^j - X_t^j \right|^p \right]^{1/p}$$

And:

$$\left\| \hat{X} - X \right\|_{L^\infty} = \frac{1}{n} \sum_{t=1}^{n} \sup_{j=1,\dots,6} \left| \hat{X}_t^j - X_t^j \right|$$

n: sample size

Autoregressive Hilbertian Model (ARH)

ρ is a bounded linear operator on H.

Steps:

0. Withdraw mean from the process.

1. Using PCA, compute empirical estimators of eigenelements of C.

2. Project the relation $D = \rho C$ in the subspace spanned by k_n eigenvectors.

3. Get a consistent estimator ρ_n using the projected relation.

k_n Selection: We use cross-validation.

Functional Kernel Model (FK)

It may be too restrictive to consider only linear operators.

Besse, et al. (2000) proposed to extend the Nadaraya-Watson kernel regression estimator to the functional context.

Then ρ can be estimated by:

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

K Gaussian kernel, h_n bandwidth, n sample size, x in H.

***h_n* Selection:** Global and Local bandwidths using cross-validation.

Bootstrap

It is interesting to provide an idea of the range of the forecasts.

In the context of dependent Hilbert space valued random variables: *Politis and Romano* (1994): confidence regions for parameters.

We are looking for confidence regions for predictions.

We extend two different bootstrap methods for real valued time series (*Cao*, 1999) to functional data.

Let X_i be the curves in the sample and Y_i the curves for which we want to forecast Y_{i+1} .

At point Y_m we will draw p bootstrap one step ahead forecasts:

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

Data Depth

We use Fraiman and Muniz (2001) concept of data depth for functional data.

They measure the nearness of a sample of curves to their median.

A sample of functional data: $X_1(t),...,X_p(t)$ from the same distribution.

For each sample point t: $F_{p,t}(x) = \frac{1}{p}\sum_{j=1}^{p} 1_{X_j(t)\le x}$

The empirical univariate depth:

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

They propose to look at the integrated index:

$$I_i = \int_a^b D_{p,t}\left(X_i(x)\right)dt$$

The **median** is the curve with the maximum index.

We can order our curves using this index.

Bootstrap for Kernel based predictions

We propose a resampling method based on the bootstrap for prediction of a general stationary process (no parametric dependence structure is known).

Algorithm

For each point Y_m

1. Construct the sample blocks of length 2: $B_j = \left\{ X_j, X_{j+1} \right\}, j = 1, \ldots, n$
2. Compute probabilities

$$\hat{p}_j = \frac{K\left(\frac{\left\| X_j - Y_m \right\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\left\| X_i - Y_m \right\|}{h}\right)}$$

Where h is the bandwidth (global or local).

3. Randomly toss *p* blocks with those probabilities. Extract from them the second element.
4. The sequence of replications:

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

5. Order the replications using F-M depth:

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

6. The median is the curve with maximum value:

$$Y^*_{m+1,1:1}$$

7. Chose the % of replications less distant to the median
8. Draw the envelope generated by the selected curves.

Bootstrap for ARH predictions

We use the dependence structure given by the ARH model.

Algorithm

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

$$\hat{a}_i = X_i - \hat{\rho} X_{i-1}$$

And their corrected version

$$\hat{a}_i' = \hat{a}_i - \overline{a}$$

Using: $\overline{a} = \frac{1}{n}\sum_{i=2}^{n+1} \hat{a}_i$

2. Make PCA in the following manner: $\hat{a}_i' = c_1^i V_1 + \ldots + c_{k_n}^i V_{k_n}$

3. For each coordinate c_l derive its empirical distribution:

$$F_n^{c_l}, l = 1,\ldots,k_n$$

METHODOLOGY

ARH FK **BOOTSTRAT**

APPLICATION

4. Using that distribution generate c_l^*

5. Construct bootstrap residuals:

$$\hat{a}_i^* = c_1^* V_1 + \ldots + c_{k_n}^* V_{k_n}$$

6. Generate bootstrap replications, $i=1,\ldots,p$:

$$Y_{m+1,i}^* = \hat{\rho} Y_m + \hat{a}_i^*$$

7. Order the replications using F-M depth:

$$Y_{m+1,1:1}^*, \ldots, Y_{m+1,p:p}^*$$

8. The median is the curve with maximum value:

$$Y_{m+1,1:1}^*$$

9. Chose the % of replications less distant to the median

10. Draw the envelope generated by the selected curves.

García Jurado et al. (1995) introduced the notion of **HISTORICAL MATRIX** in the context of real valued time series.

Fernández de Castro et al. (2003) has used that idea to build training sets for neural networks.

We adapt the historical matrix to functional data.

We must fill the historical matrix with vectors of the form:

$$(X_n, X_{n+1})$$

where: $X_n = (X_n^1, \ldots, X_n^6)$

Matrix of levels

An “ordinary” classification can be done, based on the last real value of X_{n+1}.

We use 10 classes in this matrix.

Matrix of shapes

A “functional” classification based on shapes of data.

We establish 5 classes.

We compute: $(X_{n+1}^{2} - X_{n+1}^{1}, \ldots, X_{n+1}^{6} - X_{n+1}^{5})$

And we look at the sign:

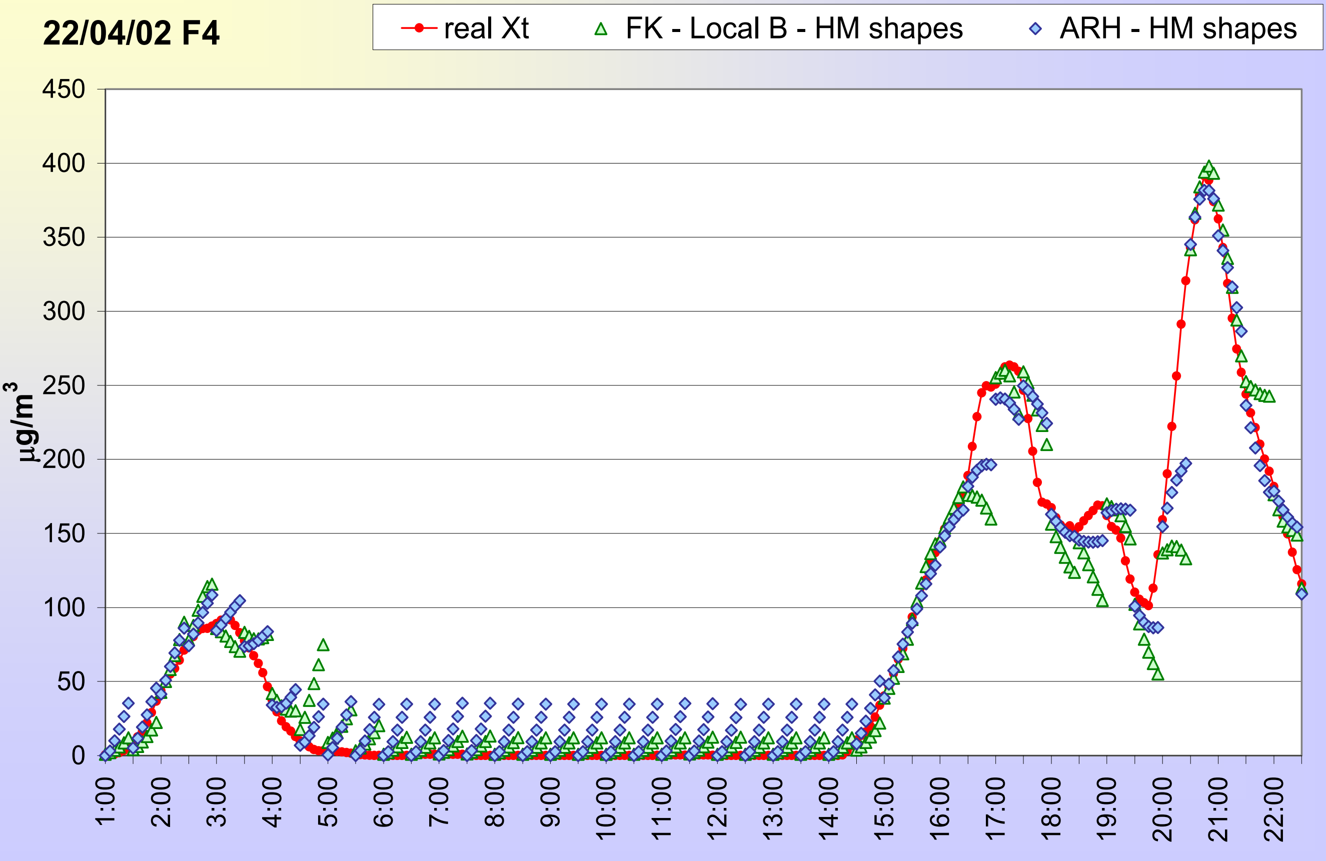
Increase	Decrease	Plateaus	Change	Anything else
(+,+,+,+,+)	(-,-,-,-,-)	(0,0,0,0,0)	At least one + and one -	else

METHODOLOGY

APPLICATION

HISTORICAL MATRIX RESULTS

22/04/02 F4



Prediction errors at F4 station on April, 22th 2002

Model	Error		
	L^1	L^2	L^∞
FK local bandwidth, HM-levels	16.14	18.27	28.12
FK global bandwidth, HM-levels	16.66	18.65	28.52
FK local bandwidth, HM-shape	**14.61**	**16.78**	**26.96**
FK global bandwidth, HM-shape	15.26	17.36	27.60
ARH, HM-levels	16.57	19.65	31.67
ARH, HM-shape	15.24	18.75	31.74

Prediction errors at F4 station on April, 22th 2002.

Episode period: 14:00 – 22:30

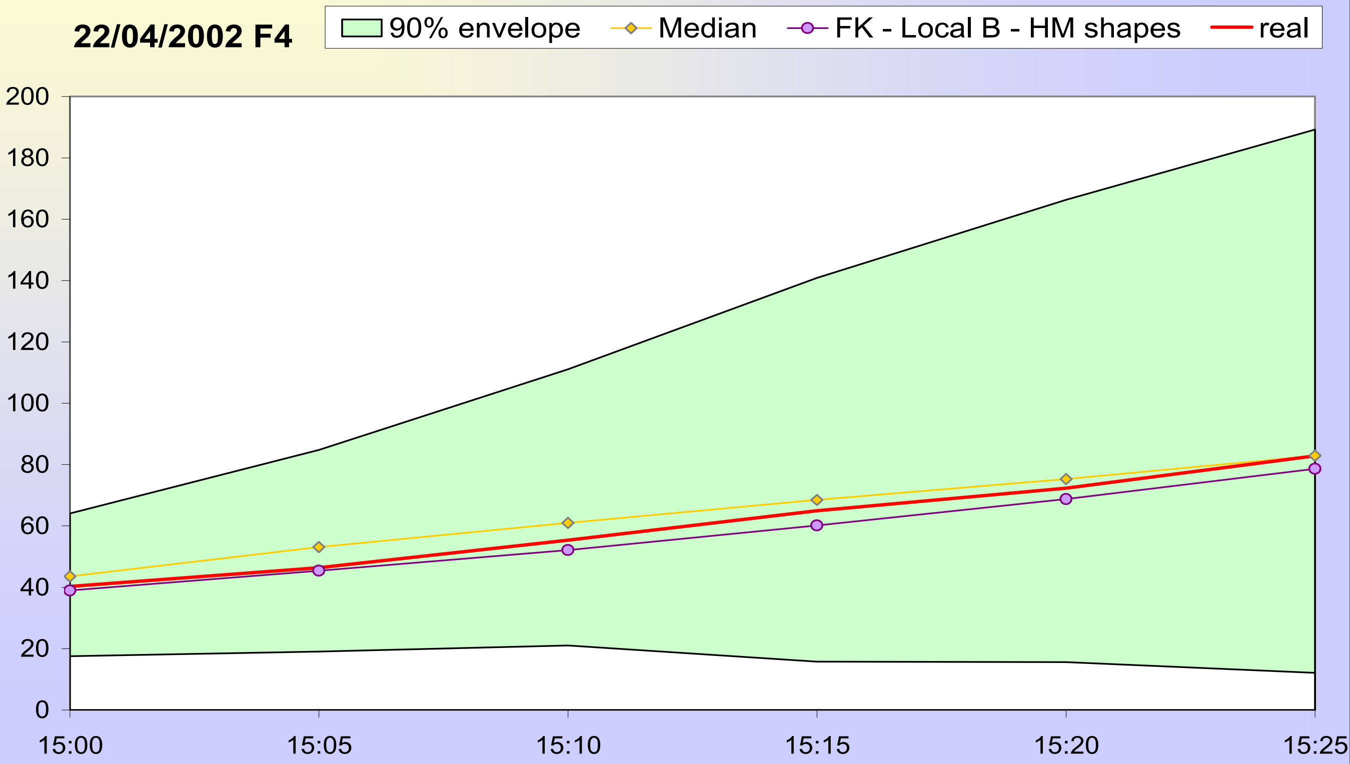
Model	Error		
	L^1	L^2	L^∞
FK local bandwidth, HM-levels	29.15	32.88	49.91
FK global bandwidth, HM-levels	26.76	30.03	45.38
FK local bandwidth, HM-shape	23.63	26.70	41.76
FK global bandwidth, HM-shape	23.13	26.04	40.15
ARH, HM-levels	23.49	25.85	37.20
ARH, HM-shape	**17.55**	**20.42**	**32.00**

METHODOLOGY

APPLICATION

HISTORICAL MATRIX RESULTS

Bootstrap Results

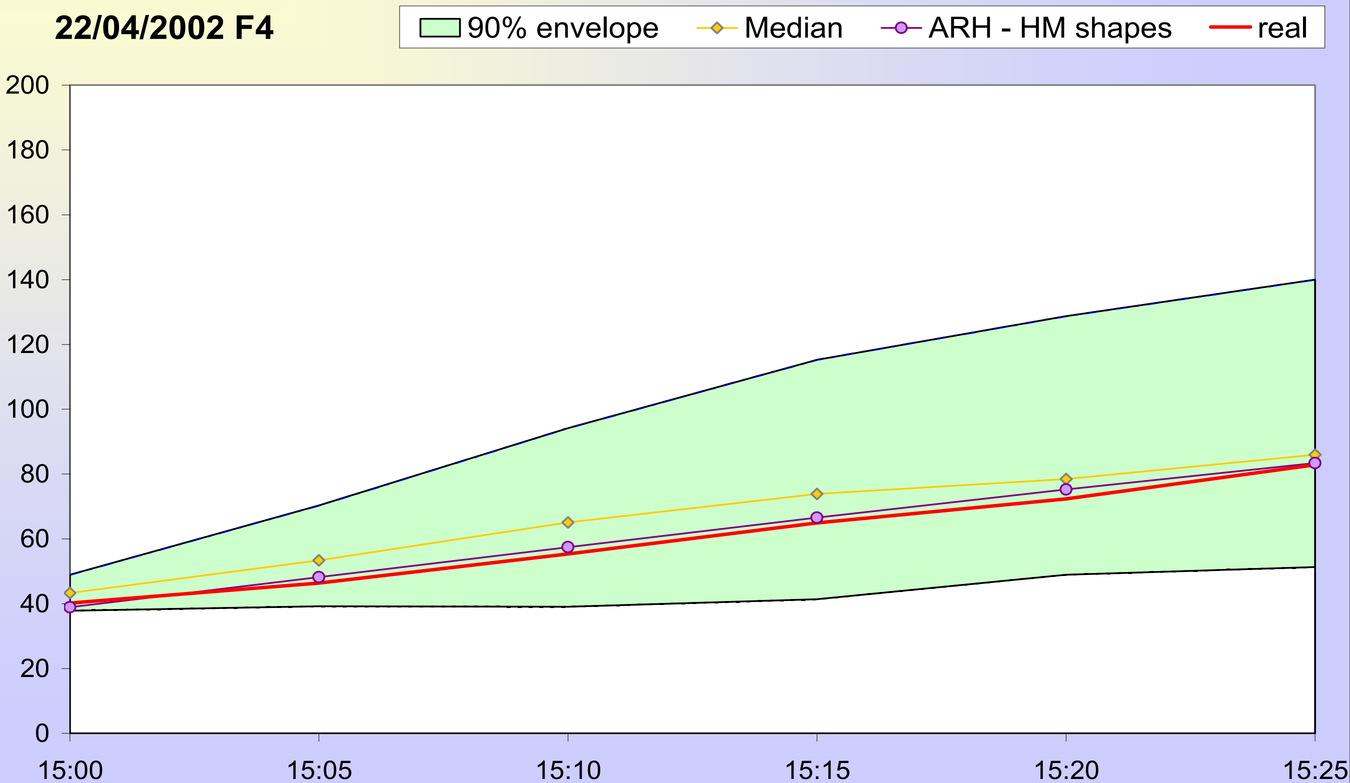


METHODOLOGY

APPLICATION

HISTORICAL MATRIX RESULTS

Bootstrap Results



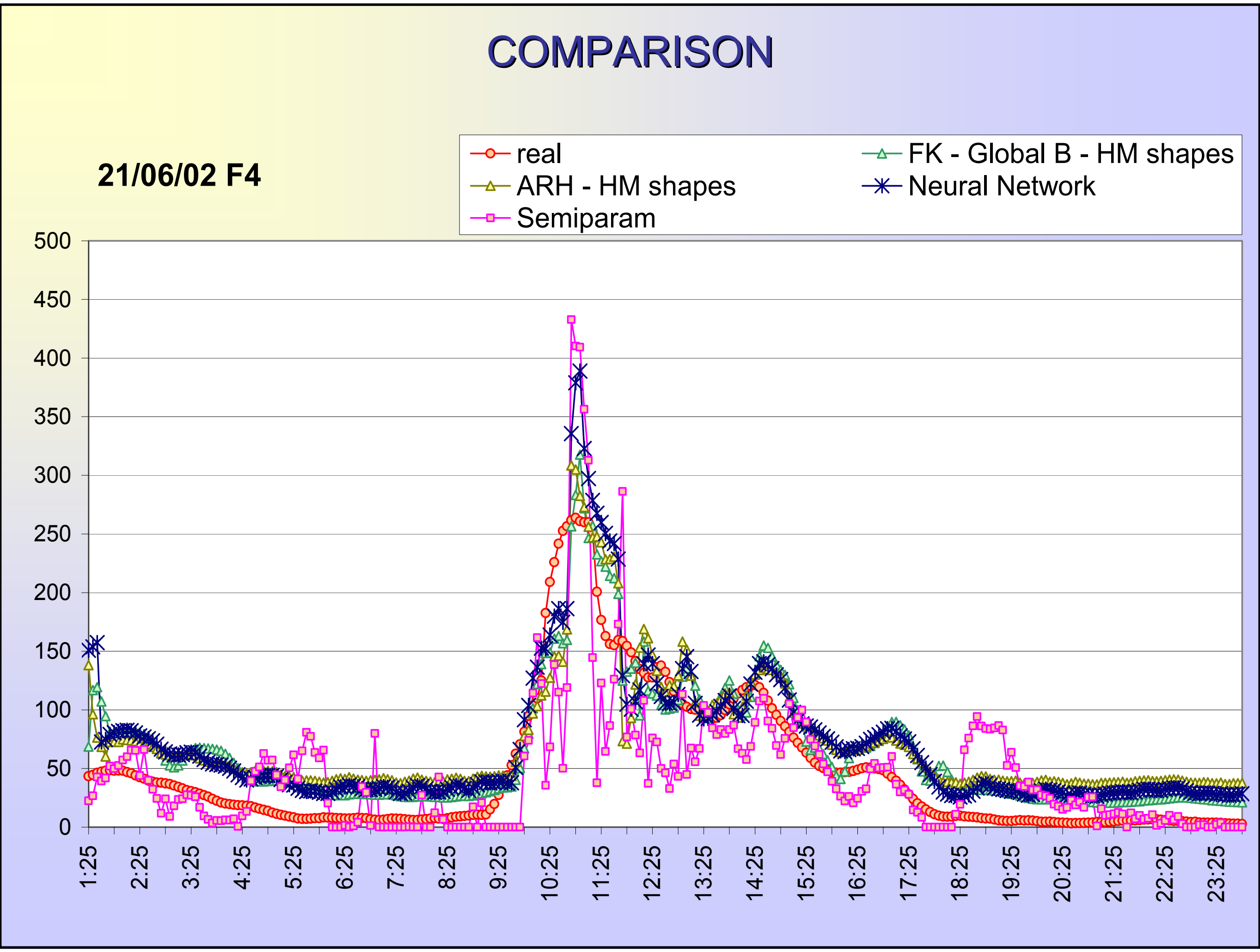
COMPARISON

We compared our forecasts to those obtained by two methods used in the past:

neural networks (Fernández de Castro *et al*., 2003)

semi-parametric models (García Jurado *et al*., 1995)

We contrasted the 30 minutes ahead forecasts every five minutes.



COMPARISON
21/06/02 F4
real
ARH - HM shapes
Semiparam
FK - Global B - HM shapes
Neural Network
500
450
400
350
300
250
200
150
100
50
0
1:25
2:25
3:25
4:25
5:25
6:25
7:25
8:25
9:25
10:25
11:25
12:25
13:25
14:25
15:25
16:25
17:25
18:25
19:25
20:25
21:25
22:25
23:25

COMPARISON

30 minutes ahead prediction errors at F4 station on April 22, 2002

Model	Error	
	MAE	MSE
FK local bandwidth,HM-shape	24.12	1305.76
FK global bandwidth, HM-shape	25.41	1303.01
ARH, HM-shape	31.14	1372.57
Neural Network	27.57	1156.64
Semi-paremetric	25.30	1650.85

CONCLUSION

- Proposed a new way of building an historical matrix focusing on functional data: classifying our data according to the shape.
- We examined the predictions of the ARH and the functional kernel model, with global and local bandwidths.
- These functional models appeared to be a very competitive option to solve our problem.
- We exposed some ideas to use bootstrap techniques with such functional data.
- Using the concept of functional depth to establish an order between our bootstrap replications, we build a region of predicted curves, following the idea of confidence intervals for real data.

REFERENCES

Andretta M, Eleuteri A., Fortezza F., Manco D., Mingozzi L., Serra R., 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, *Canad. J. Statist.*, 24, 467-487.

Besse, P., Cardot, H., and 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*, Springer.

Cao, R. (1999). An overview of bootstrap methods for estimating and predicting in time series. *Test*, 8, no. 1, 95-116.

Cao, R., Febrero-Bande, M., Gonzalez-Manteiga, W., Prada-Sanchez, J. M., and Garcia-Jurado, I. (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. *Comm. Statist. Simulation Comput.* 26, no. 3, 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., Hernández Fernández, J.J. (2003), Prediction of SO2 level using neural networks, *Journal of the air and waste management association*, to appear.

REFERENCES

Ferraty, F., Goia, A., and Vieu, P. (2002), Functional nonparametric model for times series: a fractal approach for dimension reduction, *Test*, 11, no. 2, 317-344.

Ferraty, F., and Vieu, P. (2000), Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. (French) [Fractal dimensionality and regression estimation in seminormed vector spaces] *C. R. Acad. Sci. Paris Sér. I Math*, 330, 2, 139-142.

Ferraty, F., and Vieu, P. (2002), The functional nonparametric model and application to spectrometric data, *Computational Statistics*, 17, 4, 545-564.

Fraiman, R. and Muniz, G. (2001), Trimmed means for functional data. *Test*, 10, 2, 419-440.

García Jurado, I., Gonzalez 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 Kunsch H. R. (2001), Estimation of motion from sequences of images: Daily variability of Total Ozone Mapping Spectrometer ozone data, *Journal of Geophysical Researche-Atmosphere*, 106, D11, 11825-11834.

Guillas, S. (2001), Rates of convergence of autocorrelation estimates for autoregressive Hilbertian processes, *Statist. Probab. Lett.*, 55, 281-291.

Guillas, S. (2002), Doubly stochastic Hilbertian processes, *Journal of Applied Probability*, 39, 566-580.

REFERENCES

Lapenna V., Macchiato M., Cosmi C., Ragosta M. and Serio C. (1996), Predictability analysis of SO2 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. *Statist. Sinica*, 4, 461-476.

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

Rice, J. A. and Silverman, B. W. (1991), Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc*. Ser. B, 53, no. 1, 233-243.

Schlink U., Herbarth O. and Tetzlaff G. (1997), A component time-series model for SO2 data: forecasting, interpretation and modification, *Atmospheric Environment*, 31, 9, 1285-1295.

Sherman, M. (1998), Efficiency and robustness in subsampling for dependent data, *J. Statist. Plann. Inference* 75, 1, 133--146.

Silverman, B. W. (1996), Smoothed functional principal components analysis by choice of norm. *Ann. Statist*. 24 , 1, 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, 14, 2449-2461.

CONTACT US

B. M. Fernández de Castro[1]
Department of Statistics and Operations Research
University of Santiago de Compostela (Spain)
fdcastro@usc.es

S. Guillas[2]
Center for Integrating Statistical and Environmental Science
University of Chicago (USA)
guillas@uchicago.edu

W. González Manteiga[1]
Department of Statistics and Operations Research
University of Santiago de Compostela (Spain)
wenceslao@usc.es

1. Partially supported by MCyT Grant BFM2002-03213 (European FEDER support included) and Xunta de Galicia Grant PGIDT01MAM04E.
2. Partially supported by U.S. E.P.A. Grant R-82940201-0.