

BOOTSTRAP CON DATOS FUNCIONALES. APLICACIÓN A LA PREDICCIÓN DE NIVELES DE SO₂.

B. M. Fernández de Castro¹, S. Guillas², W. González Manteiga¹.

¹Departamento de Estadística e Investigación Operativa
Universidade de Santiago de Compostela

²Center for Integrating Statistical and Environmental Science
University of Chicago

RESUMEN

La aparición de técnicas de medida en continuo, por ejemplo de niveles de contaminación, hacen posible el planteamiento de modelos de datos funcionales. En este trabajo se presentan dos técnicas bootstrap para predicción con datos funcionales. Por un lado, una extensión del bootstrap de residuos, utilizando la estructura de dependencia dada por el modelo ARH. Y un bootstrap general, enfocado a las predicciones dadas por un modelo tipo núcleo funcional. La aplicación de dichas técnicas a la predicción de niveles de contaminación de SO₂ nos permite generar regiones de predicción, utilizando el concepto de profundidad de Fraiman y Muniz.

Palabras clave: datos funcionales, ARH, núcleo funcional, predicción de contaminación, bootstrap, profundidad.

1. INTRODUCCIÓN

En este trabajo se presenta la predicción de niveles de SO₂ en el entorno de una central térmica, con un horizonte de media hora. La naturaleza de este contaminante es local, si lo comparamos, por ejemplo, con el ozono. Los repentinos cambios en la serie de datos de medidas de SO₂ ilustran tal afirmación. García Jurado, *et al.* (1995), Lapenna, *et al.* (1996), Schlink, *et al.* (1997), Andretta *et al.* (2000), y Fernández de Castro *et al.* (2003) ya han tratado la predicción de niveles de SO₂ utilizando metodologías de series temporales o redes neuronales.

El objetivo del estudio es la aplicación a estos datos de modelos de núcleo funcional y modelos autorregresivos de Hilbert de orden 1, ARH(1), para proporcionar nuevas y mejores herramientas para la predicción. Para tener en cuenta información relevante sobre episodios de alteración de la calidad de aire, se ha adaptado al contexto funcional el concepto de matriz histórica. Se han construido regiones de predicción funcionales. Para ello se presentan distintos mecanismos bootstrap y, utilizando el concepto de profundidad de Fraiman y Muniz (2001), se establece un orden para las curvas respuesta obteniendo la región de predicción generada por las $(1-\alpha)\%$ curvas más profundas.

2. DATOS

Se utilizan modelos funcionales para predecir niveles de SO₂ en el entorno de la Central Térmica de As Pontes, propiedad de Endesa Generación S.A. Dicha central posee un Sistema de Control Suplementario de la Contaminación Atmosférica (Bermúdez Cela *et al.* 2002) que incluye una Red de Control de la Calidad de Aire compuesta por 17 estaciones de medida automáticas situadas en un radio de 30 kilómetros. En estas estaciones se miden, en continuo, concentraciones de dióxido de

Azufre, óxidos de Nitrógeno, partículas en suspensión y, en algunas de ellas, Ozono. En este trabajo se estudian los valores de dióxido de Azufre (SO_2).

El principal objetivo es prevenir episodios de alteración de calidad de aire. Los valores de SO_2 que se registran en el entorno de la Central son en general próximos a cero. En condiciones meteorológicas desfavorables para la difusión del penacho, estos valores pueden crecer bruscamente, produciendo un episodio de alteración de la calidad de aire. El objeto de las predicciones es proporcionar a los operadores información fiable sobre los niveles de SO_2 en el entorno de la instalación, con la antelación suficiente para intervenir en el proceso de combustión y evitar el episodio de alteración de la calidad de aire.

La legislación vigente obliga a instalaciones como la Central Térmica de As Pontes a controlar valores horarios de concentración de SO_2 . El sistema de adquisición de datos de la instalación proporciona un nuevo dato cada 5 minutos, por lo que es posible calcular la media horaria cada 5 minutos. Así la serie de interés es la media horaria arrastrada de concentraciones de SO_2 . El interés es predecir lo que ocurrirá en la próxima media hora, por lo que cada dato funcional representará, en este análisis, un periodo de media hora.

3. METODOLOGÍA

3.1. MODELOS FUNCIONALES

Se consideran los datos como observaciones de un proceso estocástico en tiempo continuo que modela los niveles de SO_2 en media horaria. Para predecir valores futuros $(x(t), t \geq T)$ de un proceso estocástico en tiempo continuo, se utilizará la información contenida en un número infinito de variables del pasado $(x(t), t \leq T)$ considerando porciones del proceso estocástico como curvas. Distintos autores han trabajado en la predicción de curvas. Bosq (2000) presenta un estudio teórico de procesos lineales que toman valores en espacios funcionales.

Cada dato funcional representa media hora y está caracterizado por 6 puntos. Por tanto, se consideran variables aleatorias en el espacio $H = L^2([0,6])$ en la siguiente forma:

$$X_n(t) = x(6n+t).$$

Sea e_n un ruido blanco fuerte de Hilbert: variables aleatorias i.i.d. con valores en H verificando $Ee_n = 0$, $0 < E\|e_n\|_H^2 = s^2 < \infty$, $n \in \mathbb{Z}$.

Se consideran modelos de orden 1. Así, el modelo estadístico será:

$$X_n = r(X_{n-1}) + e_n$$

donde $r : H \rightarrow H$ es el funcional a estimar.

Se ha adaptado el concepto de matriz histórica introducido en García Jurado *et al.* (1995) al caso funcional. Se han construido matrices históricas utilizando los datos reales del año 2001. Cada matriz está formada por 1500 registros de la forma (X_t, X_{t+1}) , donde cada curva $X_t \in H$. Se ha considerado un criterio de clasificación para introducir los datos en las matrices históricas. Se han considerado 5 clases basadas en la forma de la curva X_{t+1} : “crecimiento”, “decrecimiento”, “planos”, “cambios” y “el resto”.

Los resultados de estudios empíricos preliminares muestran claras ventajas para los modelos contruidos utilizando la información de la matriz histórica frente a otro tipo de muestras.

3.1.1. MODELO AUTORREGRESIVO DE HILBERT

Sea r un operador lineal acotado en H . El modelo estimado es $\hat{X}_n = \hat{r}(X_{n+1})$, donde el estimador $\hat{r}$ se calcula a partir de la relación existente entre las matrices de varianzas cruzadas D y varianzas-covarianzas C : $D = rC$ (Bosq 2000).

3.1.2. NÚCLEO FUNCIONAL

Besse *et al.* (2000) proponen la extensión del estimador núcleo de Nadaraya-Watson para la regresión al contexto funcional. En este caso el operador estimado sería

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

con K núcleo gaussiano. Se ha utilizado validación cruzada para la elección de la ventana h de forma local.

3.2. TÉCNICAS BOOTSTRAP

Desde un punto de vista estadístico, y con el fin de tomar buenas decisiones, es importante conocer, en la medida que sea posible, el rango de variación de las predicciones. Cao (1999) presenta una revisión de diversas técnicas bootstrap para series temporales de valores reales. Hemos adaptado alguno de esos métodos al caso funcional, con el fin de obtener regiones de confianza para las predicciones.

3.2.1. BOOTSTRAP PARA NÚCLEO FUNCIONAL

El método propuesto proporciona una idea sobre el rango de las predicciones, suponiendo que la serie temporal es un proceso de Markov. Los pasos a seguir para generar p predicciones bootstrap $Y_{m+1,1}^*, \dots, Y_{m+1,p}^*$ en el punto Y_m se describen a continuación:

1. Con los datos recogidos en la matriz histórica se construyen bloques de dimensión 2 de la forma:

$$B_j = \{X_j, X_{j+1}\}, j = 1, \dots, n.$$

2. Se calcula la probabilidad asociada a cada bloque:

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

con h ventana local óptima, elegida por validación cruzada.

3. De manera aleatoria se seleccionan p bloques con pesos iguales a sus probabilidades asociadas $\hat{p}_j$, y se extrae de ellos su segundo elemento, obteniendo así la secuencia de réplicas $Y_{m+1,1}^*, \dots, Y_{m+1,p}^*$.

3.2.2. BOOTSTRAP PARA ARH

Cuando se dispone de una estructura de dependencia específica, como la proporcionada por el operador autorregresivo de orden 1, es posible hacer uso de otros métodos más apropiados para obtener las réplicas bootstrap. Los pasos del mecanismo propuesto son los siguientes:

1. Se calculan los residuos de las predicciones para $i = 2, \dots, n+1$

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

y su versión corregida:

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

donde

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

2. Se hace un Análisis de Componentes Principales (ACP) de $\hat{a}'_i$ de la siguiente forma:

$$\hat{a}'_i = c_1^i V_1 + \dots + c_{k_n}^i V_{k_n}$$

3. Para cada coordenada c_l se calcula su función de distribución empírica $F_n^{c_l}, l = 1, \dots, k_n$
4. Se generan c_l^* , con función de distribución $F_n^{c_l}, l = 1, \dots, k_n$, y se construyen los residuos bootstrap:

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

5. Finalmente se generan las réplicas bootstrap:

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

Se utiliza el ACP para generar un ruido funcional, ya que es más sencillo generar valores de variables aleatorias univariantes c_l que multivariantes. Además, así se evita la correlación existente entre los puntos de cada curva.

3.2.3. ORDEN FUNCIONAL

Una vez obtenidas las réplicas bootstrap, trabajando con series de tiempo reales, es entonces posible calcular su función de distribución empírica y calcular intervalos de confianza. Al tratar con datos funcionales, no son posibles estos cálculos. Se podría calcular la función de distribución empírica en cada punto de muestreo t en el que se tiene un valor real del dato funcional. Pero las curvas obtenidas de esta forma no miden realmente la dispersión funcional.

Fraiman y Muniz (2001) definen un concepto de profundidad para el caso funcional, que permite obtener la mediana de un conjunto de curvas y ordenar el conjunto en función de la profundidad de cada curva. Se utiliza este concepto de profundidad para obtener la mediana de las réplicas bootstrap: la curva más profunda, y obtener un orden en las réplicas:

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

Así se utilizará un subconjunto de réplicas bootstrap, el $(1-a)\%$ de las curvas más profundas, para calcular su envoltura convexa y así obtener una región límite para las predicciones.

Por tanto, utilizando esta noción de profundidad se obtiene una predicción bootstrap, la mediana de las réplicas bootstrap, y una región que describe las curvas más próximas a la mediana.

4. RESULTADOS

Para mostrar los resultados de las predicciones se utilizan los datos reales del día 24 de abril de 2002, en el que ha tenido lugar un episodio de alteración de la calidad de aire en la estación F4 de la Red de Vigilancia de Calidad de Aire de la Central Térmica de As Pontes. Este episodio no ha sido incluido en la matriz histórica, construida con datos de 2001.

En las gráficas se presenta la curva real, la predicción dada por el modelo para esa curva, la mediana de las réplicas bootstrap y la envoltura convexa generada por el 90% de curvas más profundas.

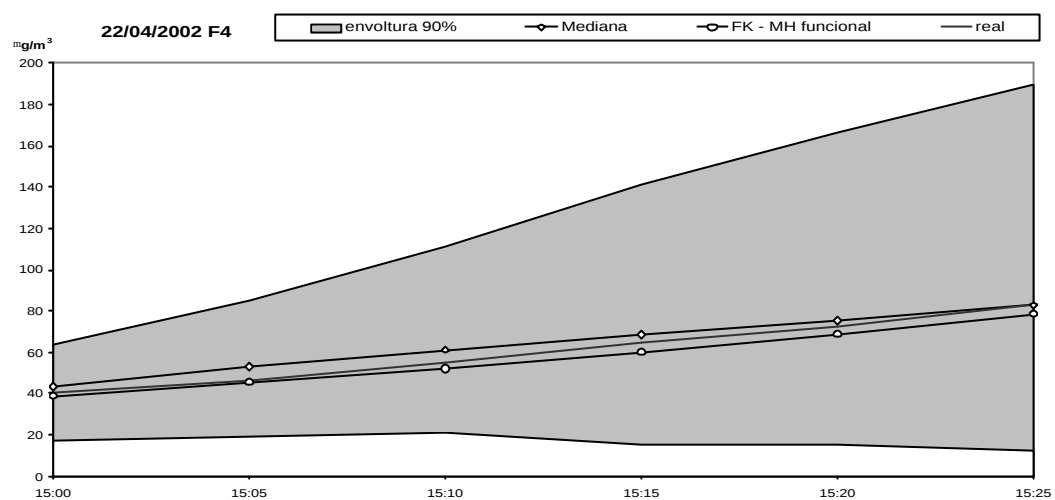


Figura 1: Bootstrap y predicción del modelo de Núcleo Funcional, modelo estimado con matriz histórica funcional. Datos del día 22 de Abril de 2002, estación F4.

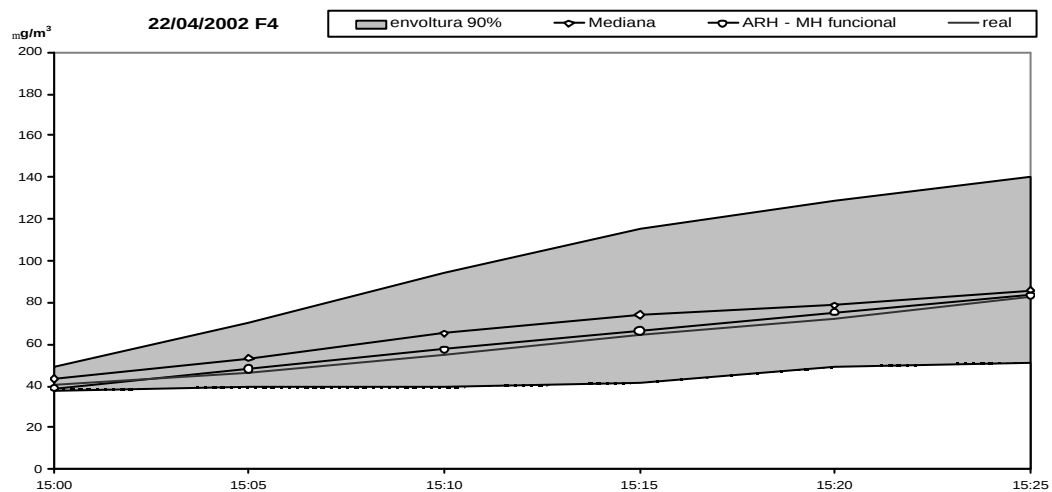


Figura 2: Bootstrap y predicción del modelo ARH(1), modelo estimado con matriz histórica funcional. Datos del día 22 de Abril de 2002, estación F4.

5. CONCLUSIONES

En este trabajo se proponen ideas sobre técnicas bootstrap con datos funcionales. Por un lado, utilizando la información contenida en la matriz histórica se propone un método de remuestreo para las predicciones dadas por un modelo de núcleo funcional. Por otro lado, utilizando la estructura de dependencia dada por el modelo ARH, se propone un mecanismo bootstrap basado en los residuos de las predicciones. Además, siguiendo la idea de intervalos de confianza para datos reales, se obtienen regiones que describen las predicciones utilizando el concepto de profundidad de Fraiman y Muniz para ordenar las réplicas bootstrap y obtener su mediana.

6. AGRADECIMIENTOS

Investigación parcialmente financiada por la subvención BFM2002-03213 (financiación Europea FEDER incluida) del MCyT y por la subvención PGIDT01MAM04E de la Xunta de Galicia para el primer y tercer autores.

Investigación parcialmente financiada por la subvención R-82940201-0 de U.S. E.P.A para el segundo autor.

7. REFERENCIAS

- 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.
- Bermúdez Cela, J.L., Fernández de Castro, B.M., Souto González, J.A. (2002) Sistemas de Control Suplementario de la Contaminación Atmosférica. *Revista Galega de Economía*. Vol. 11, n 2, 77-88.
- 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.
- 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*, vol. 53, n 2, 532-539.
- 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.
- 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.
- 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.