Metodología y aplicaciones de la Estadística al control Medioambiental

Manuel Febrero Bande , Wenceslao González Manteiga *y María Piñeiro Lamas*

06/02/2009
Oviedo

Congreso de la Real Sociedad Matemática 2009
Sesión especial de transferencia Matemática

Esquema

1. *Un problema medioambiental.*
2. *Modelos de predicción.*
3. *Nuevas aportaciones a la predicción medioambiental: el fenómeno de la cointegración.*

Un problema medioambiental

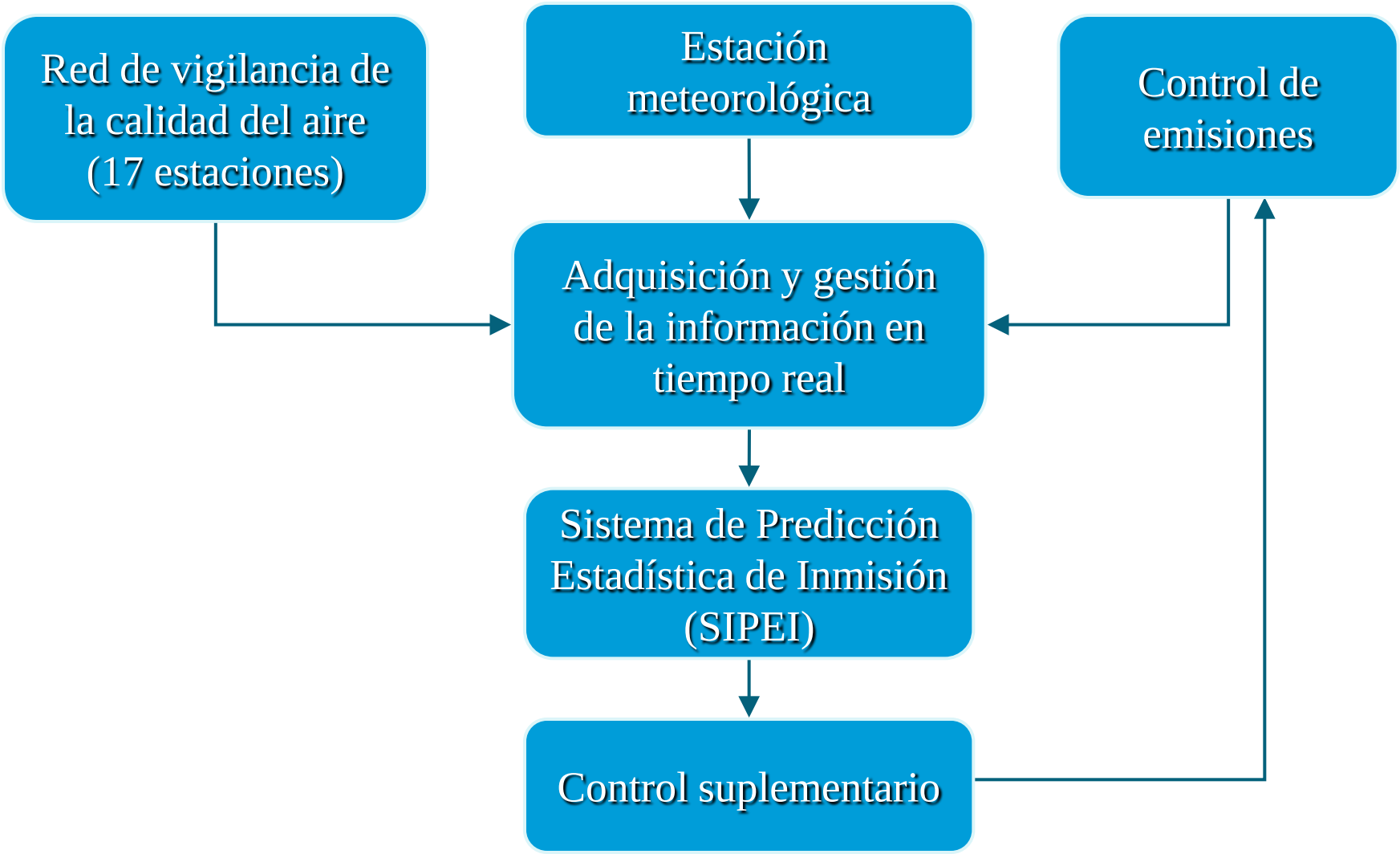
- La Unidad de Producción Térmica (UPT) de As Pontes constituye uno de los centros productivos propiedad de Endesa Generación S.A., situado en el municipio de As Pontes, al noroeste de la provincia de A Coruña.


As Pontes de García Rodríguez
Image © 2009 DigitalGlobe
© 2009 AND
© 2009 Tele Atlas
© 2009 Cnes/Spot Image
Google
43°27'29.86" N 7°53'14.95" O
elev. 264 m
12 Oct. 2003
Alt. ojo 15.67 km

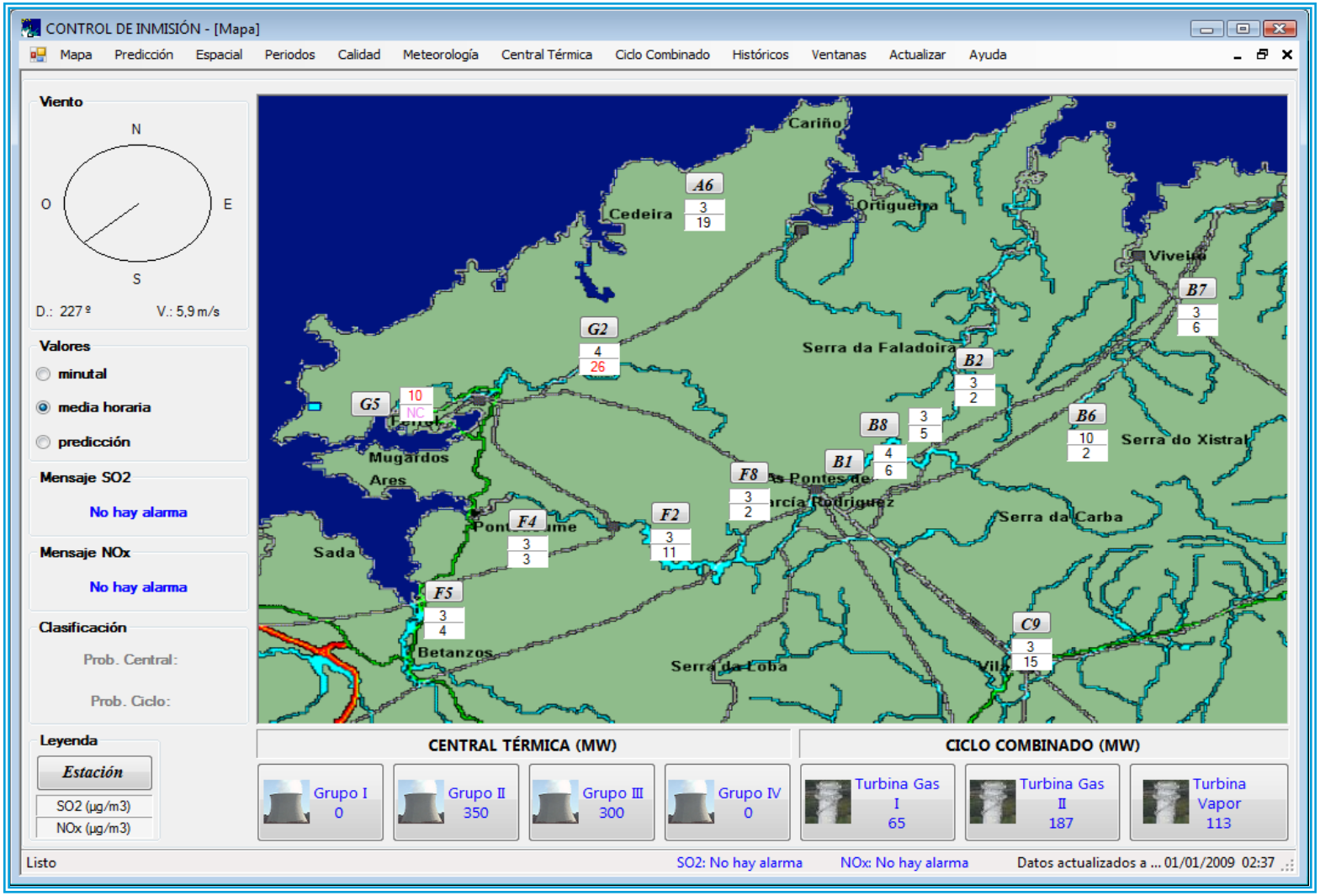
Descripción general

- Inició su actividad en 1976 con la puesta en marcha de un grupo de generación de energía, disponiendo en la actualidad de cuatro.
- Fue diseñada para utilizar los lignitos extraídos de la mina a cielo abierto situada en sus proximidades con alto contenido de azufre.
- En el período 1993-1996 fue transformada con el objetivo de utilizar mezclas de lignito local con carbones subbituminosos de importación caracterizados por sus bajos contenidos en azufre.
- Actualmente ya se ha finalizado una nueva adaptación para consumir carbón subbituminoso de importación como combustible principal (período de transformación: 2005-2008).

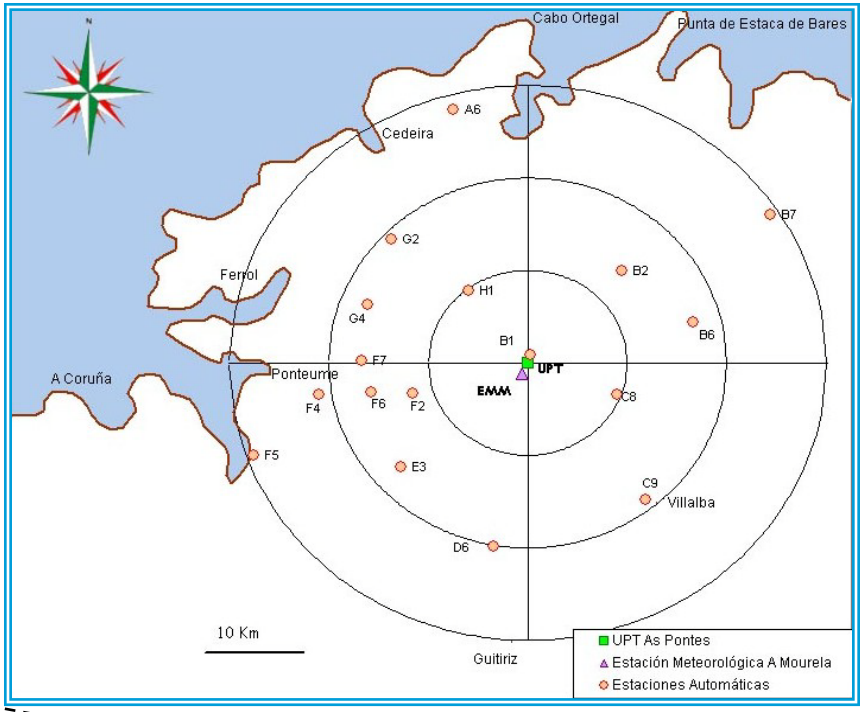
Sistema de Seguimiento y Control de la Calidad Atmosférica



SIPEI 2008: Pantalla principal



Sistema de Seguimiento y Control de la Calidad Atmosférica



Problema

- Existe un foco emisor donde se monitoriza en continuo distintas medidas de calidad de aire en su entorno, e interesa predecir valores futuros de la calidad de aire. El principal interés inicial es la predicción de valores futuros de SO_2.

- Los modelos estadísticos de predicción son una herramienta eficaz para obtener estas predicciones y sugerir una línea de actuación para intentar evitar los episodios de calidad de aire.

Nueva problemática

- Cercano al emplazamiento de la Central Térmica de As Pontes se ha construido una nueva Central de Ciclo Combinado de gas natural.

Descripción general

- Consiste en un grupo generador de electricidad formado por dos turbinas de gas y una turbina de vapor (dos focos emisores).
- Está diseñado para utilizar gasóleo como combustible de emergencia.
- Como cualquier instalación de generación de energía por combustión debe disponer de un Sistema de Control de las Emisiones que ha sido integrado en el Sistema de Seguimiento y Control de la Calidad Atmosférica .

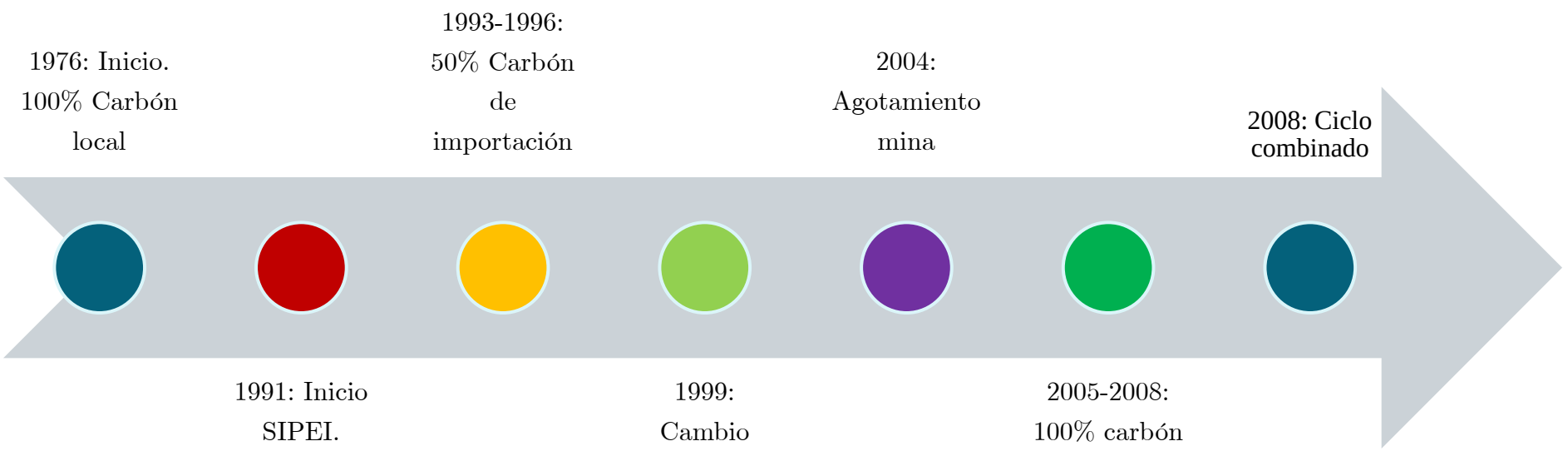
Nuevos retos

En el período 2007-2008 las emisiones se han visto afectadas por dos actuaciones clave:

- La Central Térmica actual consume únicamente carbón de importación. → ***Disminución de las emisiones de SO_2 (alrededor del 95%)***

- La nueva Central de Ciclo Combinado entró en funcionamiento. → ***Necesidad de predecir las emisiones de NO_x***

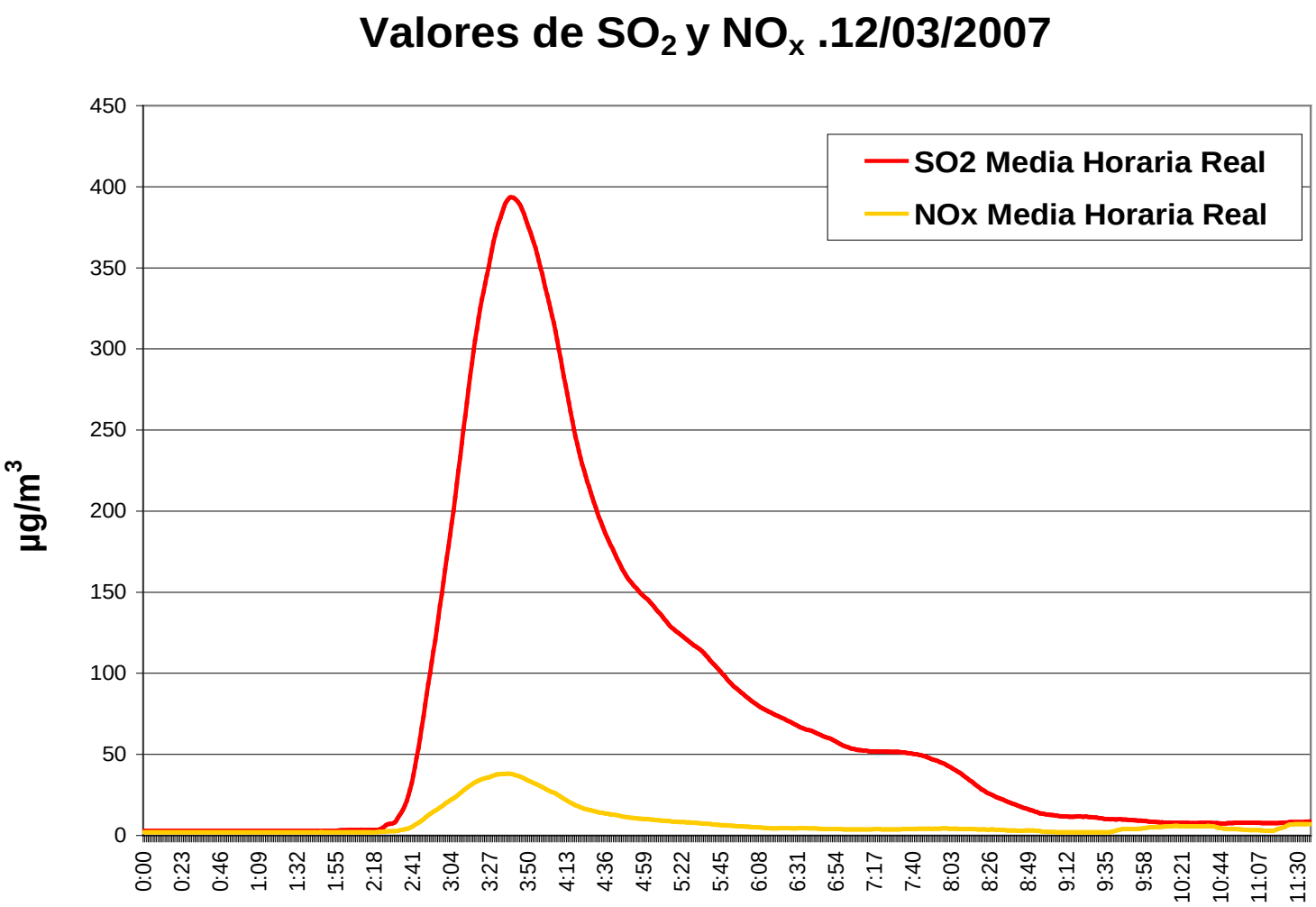
Resumen histórico



Los datos

- Actualmente, se reciben datos con frecuencia minutal y pentaminutal en tiempo real.
- Se va a considerar la media horaria arrastrada, tanto de los valores de SO_2 como de NO_x, para obtener las predicciones de valores futuros.
- La serie de valores medios horarios de SO_2 tiene un comportamiento muy característico:
  - Toma valores próximos a cero durante largos períodos de tiempo.
  - En ocasiones crece de manera brusca y repentina (episodios).
  - Estos episodios están muy espaciados en el tiempo.
- La serie del NO_x tiene un comportamiento similar pero a menor escala.

Los datos


Valores de SO_2 y NO_x .12/03/2007
SO2 Media Horaria Real
NOx Media Horaria Real
µg/m3
450
400
350
300
250
200
150
100
50
0
0:00
0:23
0:46
1:09
1:32
1:55
2:18
2:41
3:04
3:27
3:50
4:13
4:36
4:59
5:22
5:45
6:08
6:31
6:54
7:17
7:40
8:03
8:26
8:49
9:12
9:35
9:58
10:21
10:44
11:07
11:30

Modelos de predicción

Modelo Semiparamétrico

- En los primeros años de desarrollo la frecuencia de envío de datos era pentaminutal y la legislación vigente establecía límites para las medias bihorarias.
- El modelo usa la serie temporal de medias bihorarias arrastradas

$$x_t = \frac{1}{24}\sum_{i=0}^{23} SO_2\left(t-i\right)$$

donde $\mathrm{SO}_2(t)$ representa la concentración de SO_2 en el instante t (pentaminutal), medida en $\mu g/m^3$.

- El modelo semiparamétrico generaliza los modelos Box-Jenkins de la siguiente manera:

$$X_{t+k} = \varphi\left(X_t, X_{t-l}\right) + Z_{t+k}$$

Modelo Semiparamétrico

- En cada instante t se estima la función de regresión

  $E(X_{t+6} \,/\, X_t, X_{t-1})$ con el estimador tipo núcleo Nadaraya-Watson y la matriz histórica.

- Se calcula la serie de residuos $\hat{Z}_{t-64},\ldots,\hat{Z}_t$ (6 últimas horas), donde $\hat{Z}_i := X_i - \hat{E}(X_i \,/\, X_{i-6}, X_{i-7})$ y se ajusta un modelo ARIMA adecuado.
- Se obtiene la predicción Box-Jenkins de $\hat{Z}_{t+6}$.

- La predicción final propuesta es:

$$\hat{X}_{t+6} = \hat{E}\left(X_{t+6} \,/\, X_t, X_{t-1}\right) + \hat{Z}_{t+6}$$

Modelo Semiparamétrico

García-Jurado I., González-Manteiga W., Prada-Sánchez J.M., Febrero-Bande M. and Cao R. ***Predicting using Box-Jenkins, Nonparametric and Bootstrap Techniques.*** Technometrics 1995; 37: 303-310.

Matrices Históricas

- La clave del buen funcionamiento de todos los modelos de predicción diseñados está en un mecanismo de memoria diseñado en los primeros años de desarrollo.
- La serie temporal de interés está formada en su mayor parte por valores próximos a cero y pocos valores de episodios de inmisión.
- Se estratifican los datos disponibles de SO_2 según un rango de valores que representan razonablemente a los episodios pasados con objeto de mantener la información interesante de los episodios.
- A lo largo de los años este concepto se ha adaptado a las distintas técnicas estadísticas utilizadas.

Matrices Históricas

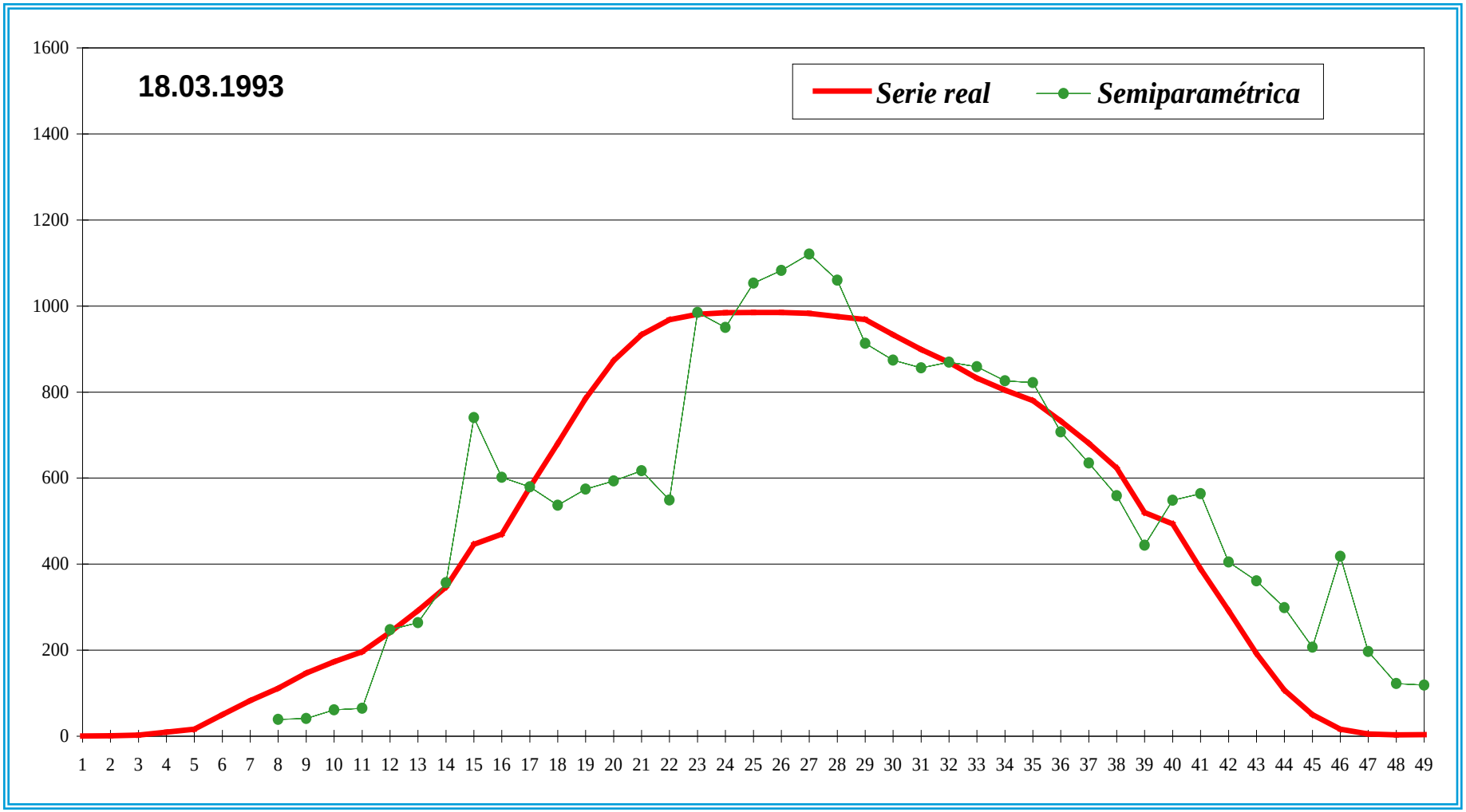
- ~~Serie temporal~~
- Matriz histórica:
  - Un número grande de registros.
  - Divididos en estratos.
  - Se asigna un rango de valores de X_{t+k} a cada clase.
  - Cada vector se asigna a la clase a la que pertenece X_{t+k} .

$$\begin{matrix} (X^1_{t-l}, X^1_t, X^1_{t+k}) \\ \\ \\ (X^i_{t-l}, X^i_t, X^i_{t+k}) \\ \\ \\ \vdots \\ \\ (X^M_{t-l}, X^M_t, X^M_{t+k}) \end{matrix}$$

Matrices Históricas

Prada-Sánchez, J.M. and Febrero-Bande M. ***Parametric, Non-Parametric and Mixed approaches to prediction of sparsely distributed pollution incidents: a case study.*** Journal of Chemometrics 1997; 11: 13-32.

Modelo Semiparamétrico (Matriz histórica)



Modelos Parcialmente Lineales

- La información utilizada por los modelos semiparamétricos es el pasado de la propia serie.
- Se puede introducir información adicional: variables meteorológicas y variables de emisión.
- Los modelos parcialmente lineales utilizan estas variables y amplían el horizonte de predicción a una hora.
- Se consideran datos de la forma $(V_l,\ Z_l,\ Y_l)$ donde V_l es un vector de variables exógenas, $Z_l = (X_l,\ X_{l-3})$ e $Y_l = X_{l-12}$, siendo $\{X_l\}$ la serie de medias bihorarias de SO_2.
- Se asume que se ajustan al modelo parcialmente lineal:

$$Y_l = V_l'\beta + \varphi\left(Z_l\right) + \varepsilon_l$$

Modelos Parcialmente Lineales

Prada-Sánchez J.M., Febrero-Bande M., Cotos-Yáñez T., González-Manteiga W., Bermúdez-Cela J.L. and Lucas-Domínguez T. ***Prediction of SO_2 pollution incidents near a power station using partially linear models and a historical matrix of predictor-response vectors.*** Environmetrics 2000; 11: 209-225.

Modelos de Predicción binaria

- El objetivo de estos modelos es estimar la probabilidad de que la serie de medias bihorarias de SO_2 supere un cierto nivel con una hora de antelación.

- Se pretende predecir:

$$p\left(Z_t\right) = p\left(X_{t+12} > r \mid Z_t\right)$$

donde

$$Z_t = (X_t, X_t - X_{t-3})$$

- Para ello se van a utilizar *modelos aditivos generalizados con función link desconocida* (*G-GAM*)

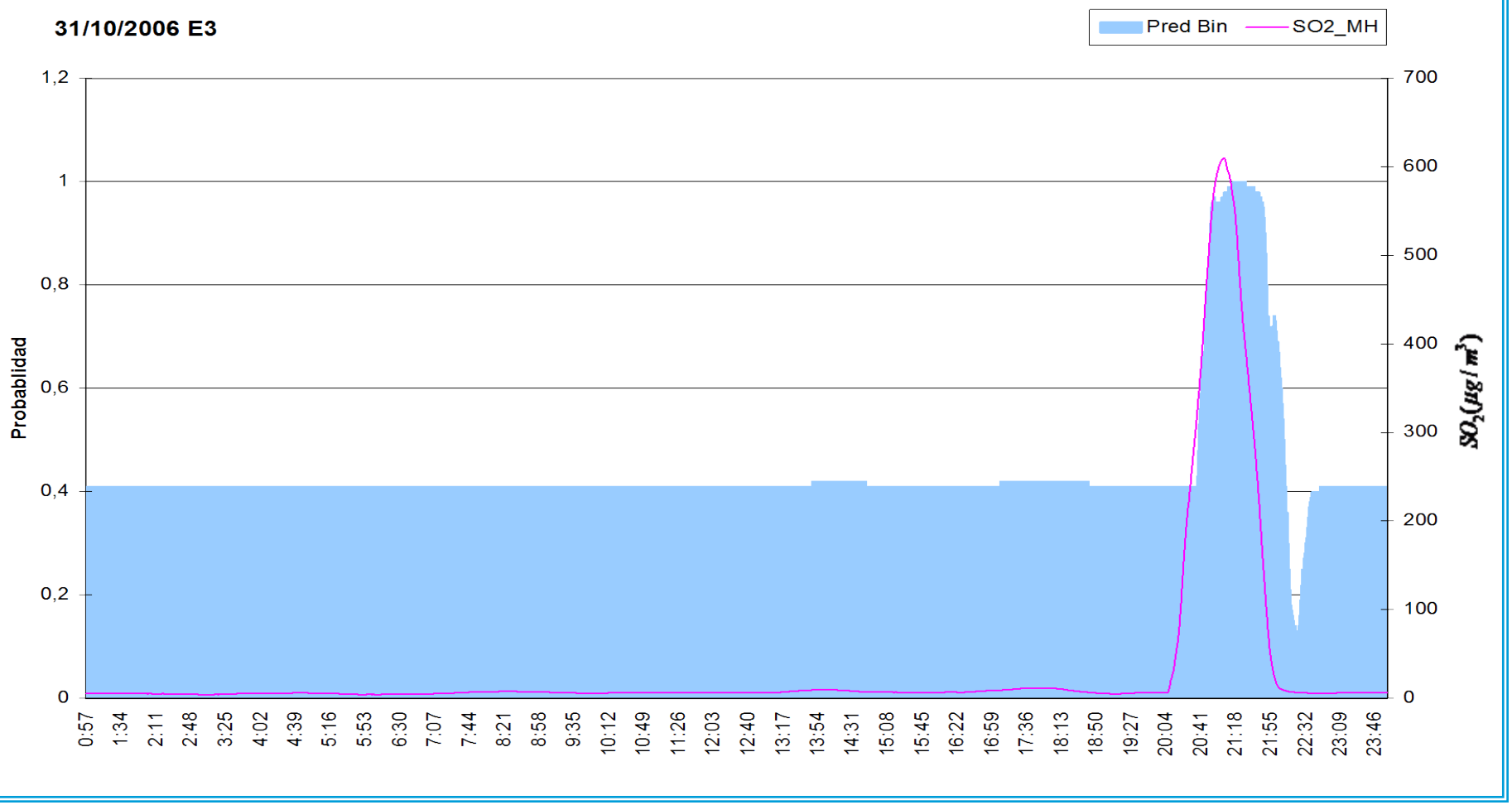
Modelos de Predicción binaria

- Sea Y una variable con respuesta binaria, y $Z=(1,Z_1,\ldots,Z_p)$ el vector de covariables asociado.
- Se denota $p\left(Z\right)=p\left(Y=1\,|\,Z\right)$.
- El *modelo G-GAM* tiene la siguiente expresión:

$$p\left(Z\right)=H\left(\eta_z\right)=H\left(\beta_0+\sum_{j=1}^{p}f_j\left(Z_j\right)\right)$$

donde f_j son funciones suaves desconocidas y la función link H es monótona creciente.

Modelos de Predicción binaria



Modelos de Predicción binaria

Roca-Pardiñas, J., González-Manteiga W., Febrero-Bande M., Prada-Sánchez J.M. and Cadarso-Suárez C. ***Predicting binary time series of SO_2 using generalized additive models with unknown link function.*** Environmetrics 2004; 15: 729-742.

Modelos de Predicción binaria

- También se han considerado modelos GAM incluyendo términos de interacción de segundo orden

$$p(Z) = H(\eta_z) = H\left(\beta_0 + \sum_{j=1}^{p} f_j(Z_j) + \sum_{1 \le j \le k \le p} f_{jk}(Z_j, Z_k)\right)$$

donde f_j son funciones unidimensionales desconocidas, las $\{f_{jk}\}_{1 \le j \le k \le p}$ son un conjunto de funciones bidimensionales desconocidas y H es la función link monótona y conocida.

- Se ha visto que el GAM con interacciones detecta el inicio de los episodios con mayor antelación de lo que lo hace el GAM puro.

Modelos de Predicción binaria

Roca-Pardiñas, J., Cadarso-Suárez C. and González-Manteiga W. ***Testing for interactions in generalized additive models: Application to SO2 pollution data.*** Statistics and Computing 2005; 15: 289-299.

Modelos de Redes Neuronales

- La entrada en vigor del Directiva Europea 1999/CE/30, provoca un cambio en la serie de interés, de medias bihorarias a medias horarias.

$$x_t = \frac{1}{12}\sum_{i=0}^{11} SO_2\left(t-i\right)$$

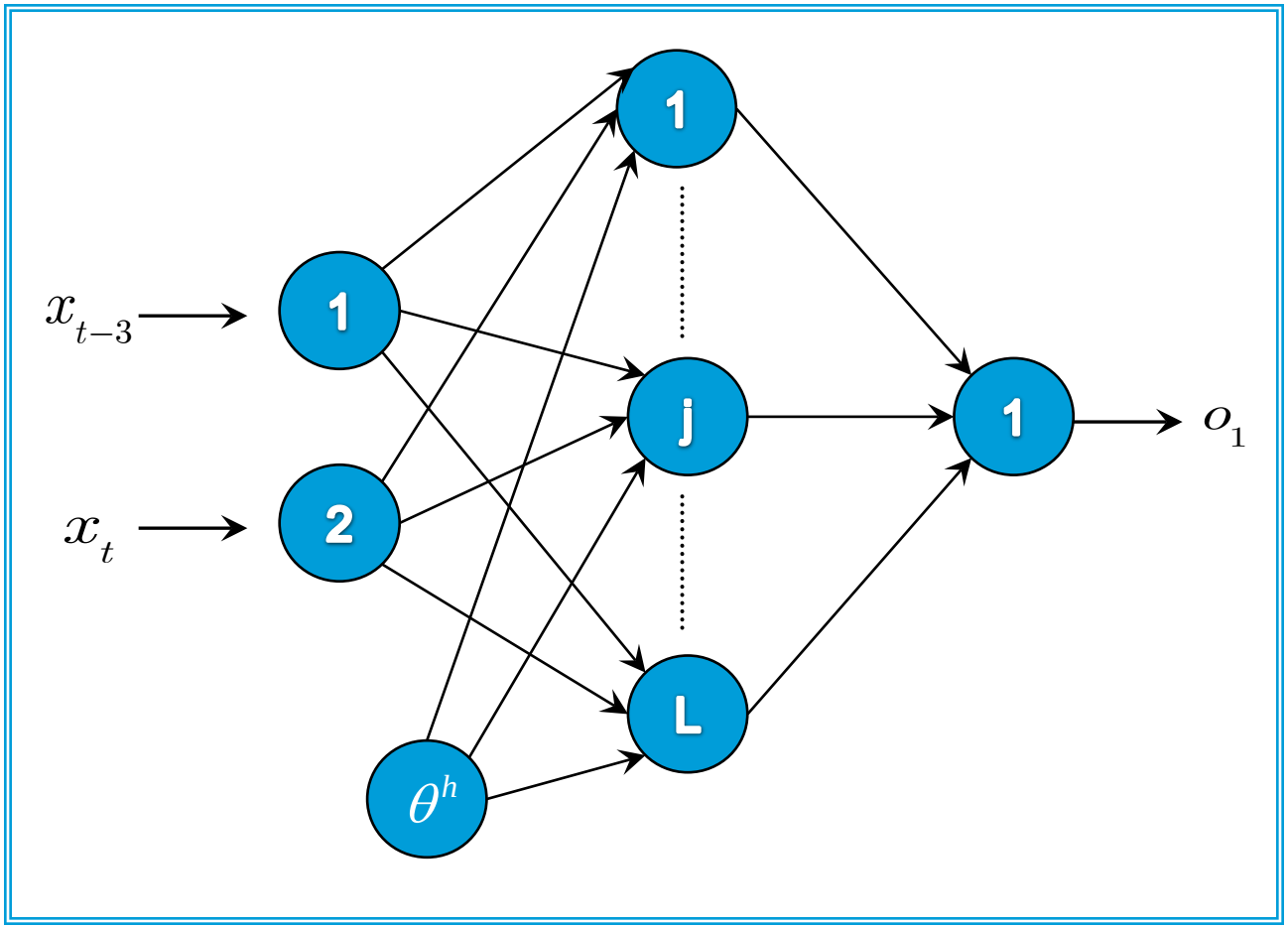
donde $\mathrm{SO}_2(t)$ representa la concentración de SO_2 en el instante t (pentaminutal), medida en $\mu g/m^3$.

- Se adaptó el modelo semiparamétrico para trabajar con la nueva serie pero se observó un considerable aumento de la variabilidad.

- Para mejorar las predicciones se desarrollaron modelos de redes neuronales.

Modelos de Redes Neuronales

- La topología de la red diseñada es la siguiente:



Modelos de Redes Neuronales

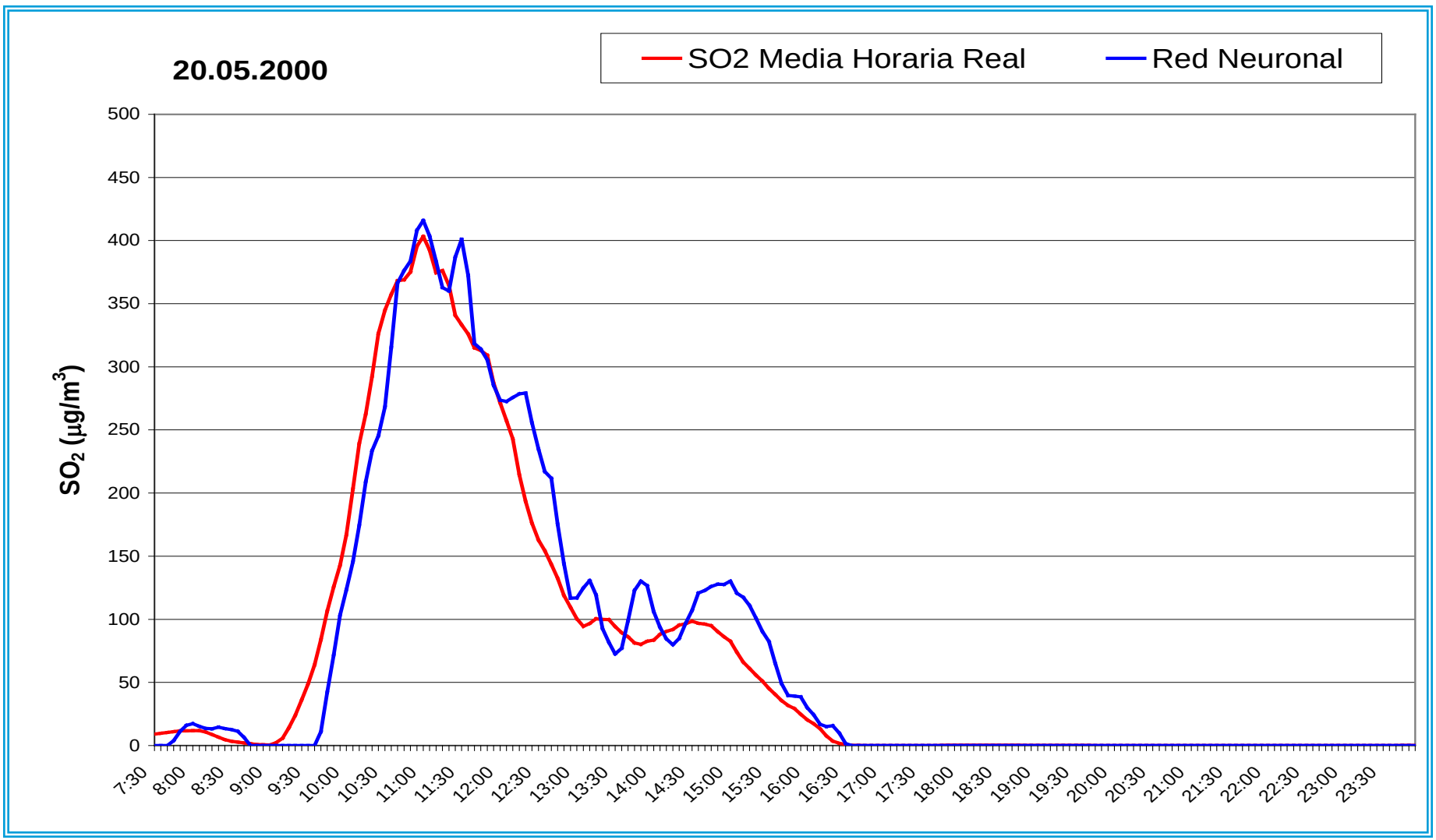
- El predictor dado por la red neuronal tiene la siguiente expresión:

$$\hat{x}_{t+6} = o_1 = \sum_{j=1}^{L} \omega_1^o f_j^h \left(\theta_j^h + \omega_{j1}^h x_{t-3} + \omega_{j2}^h x_t \right)$$

donde

- $f_j^h(z) = \dfrac{1}{1+e^{-z}}$ para todos los nodos j de la capa oculta.
- Los pesos $\left\{ \omega_{j1}^h, \omega_{j2}^h ; j = 1, \ldots, L \right\}$ y las tendencias $\left\{ \theta_j^h ; j = 1, \ldots, L \right\}$ serán determinados durante el proceso de entrenamiento.
- El número L de nodos de la capa oculta se determinará en el entrenamiento como el valor cuya red neuronal proporcione mejores resultados, tras haber entrenado redes idénticas con distintos valores de L.

Modelos de Redes Neuronales

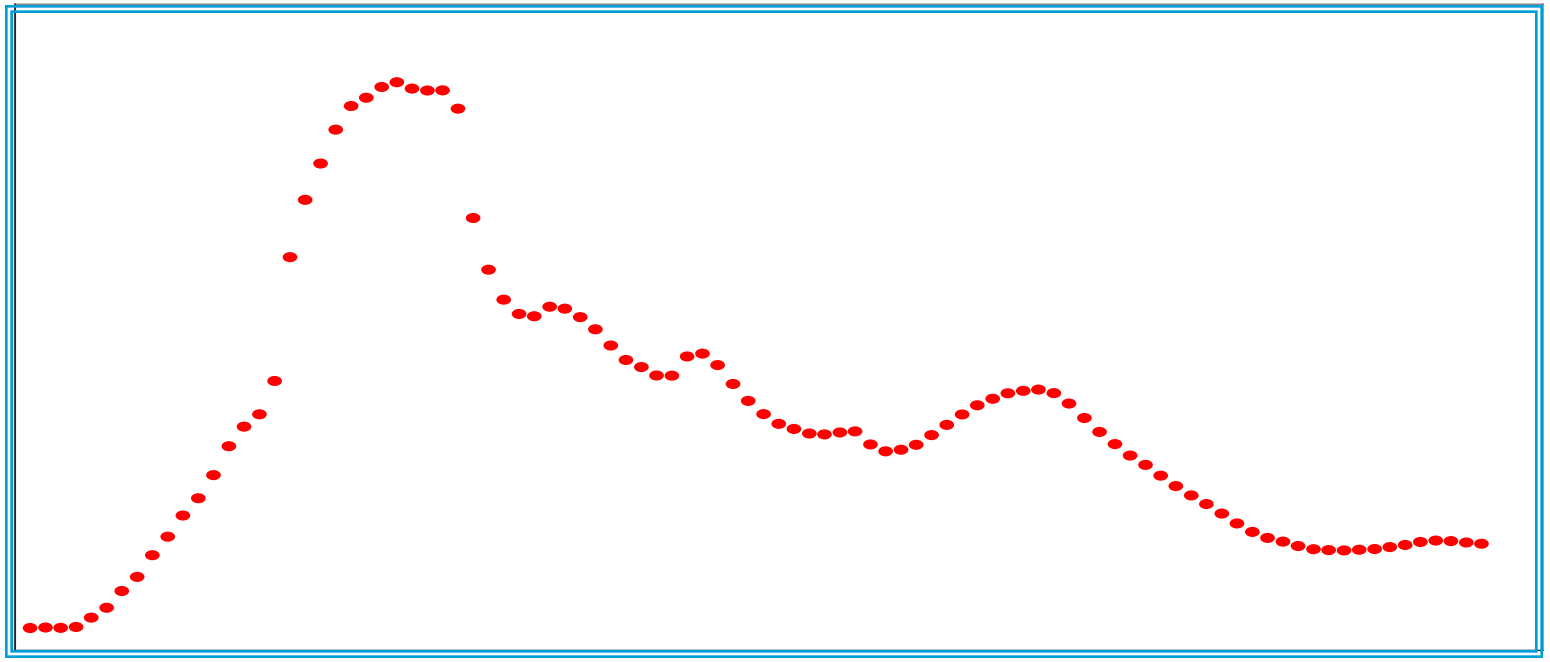


Modelos de Redes Neuronales

Fernández de Castro B.M., Prada-Sánchez J.M., González-Manteiga W., Febrero-Bande M., Bermúdez-Cela J.L. and Hernández Fernández J.J. ***Prediction of SO_2 levels using neural networks.*** Journal of the Air and Waste Management Association 2003; 53: 532-538.

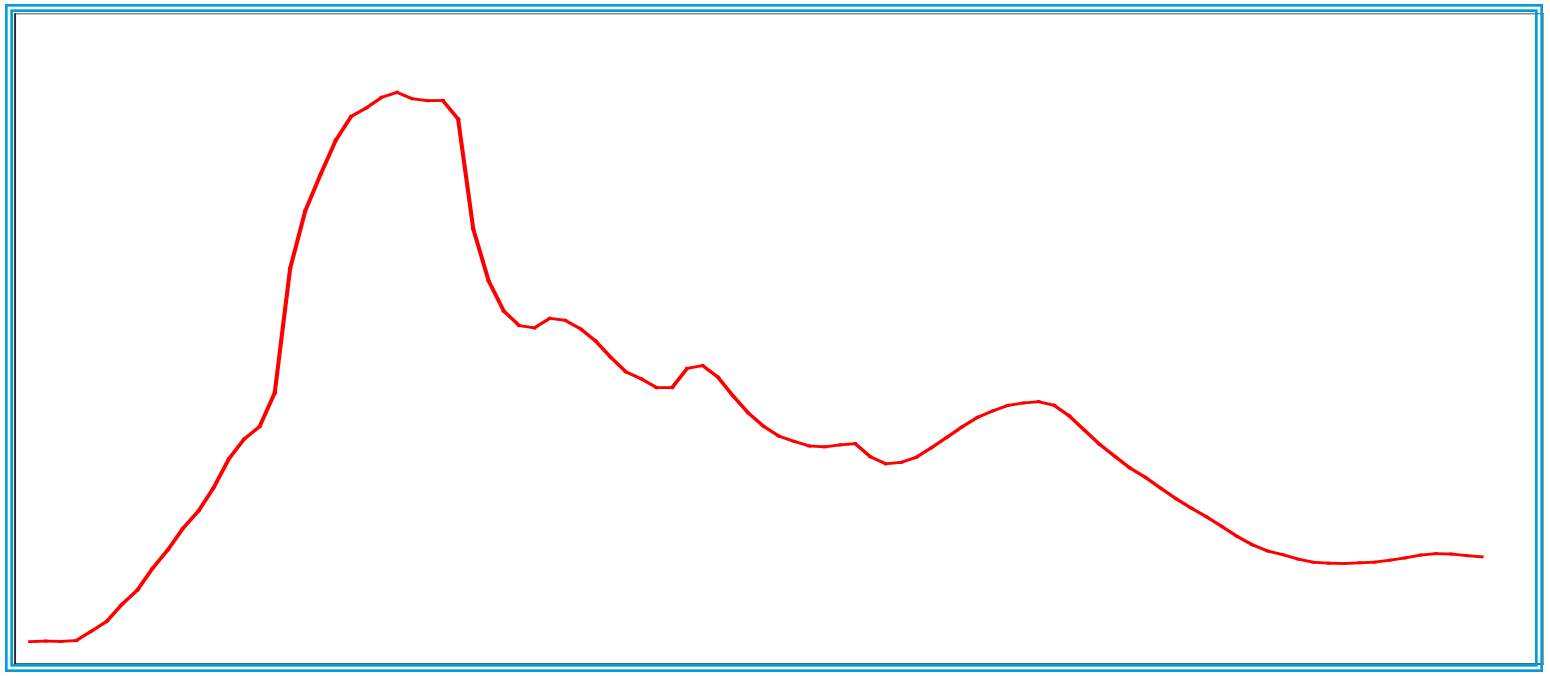
Modelos Funcionales

- Idea central:



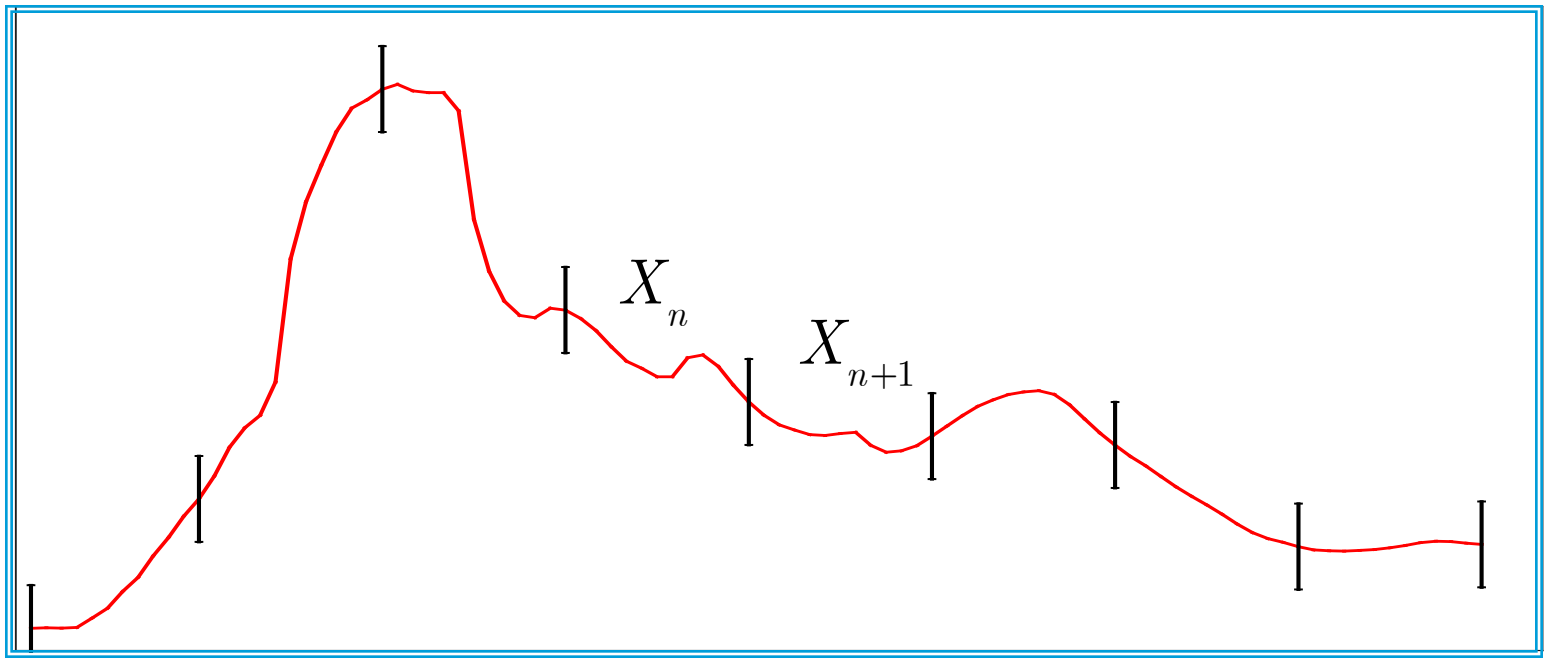
Modelos Funcionales

- Idea central:



Modelos Funcionales

- Idea central:



Modelos Funcionales

- Se considera la serie de datos como observaciones de un proceso estocástico en tiempo continuo que modeliza los niveles medios horarios de SO_2.

- Se consideran porciones de dicho proceso estocástico que representan media hora.

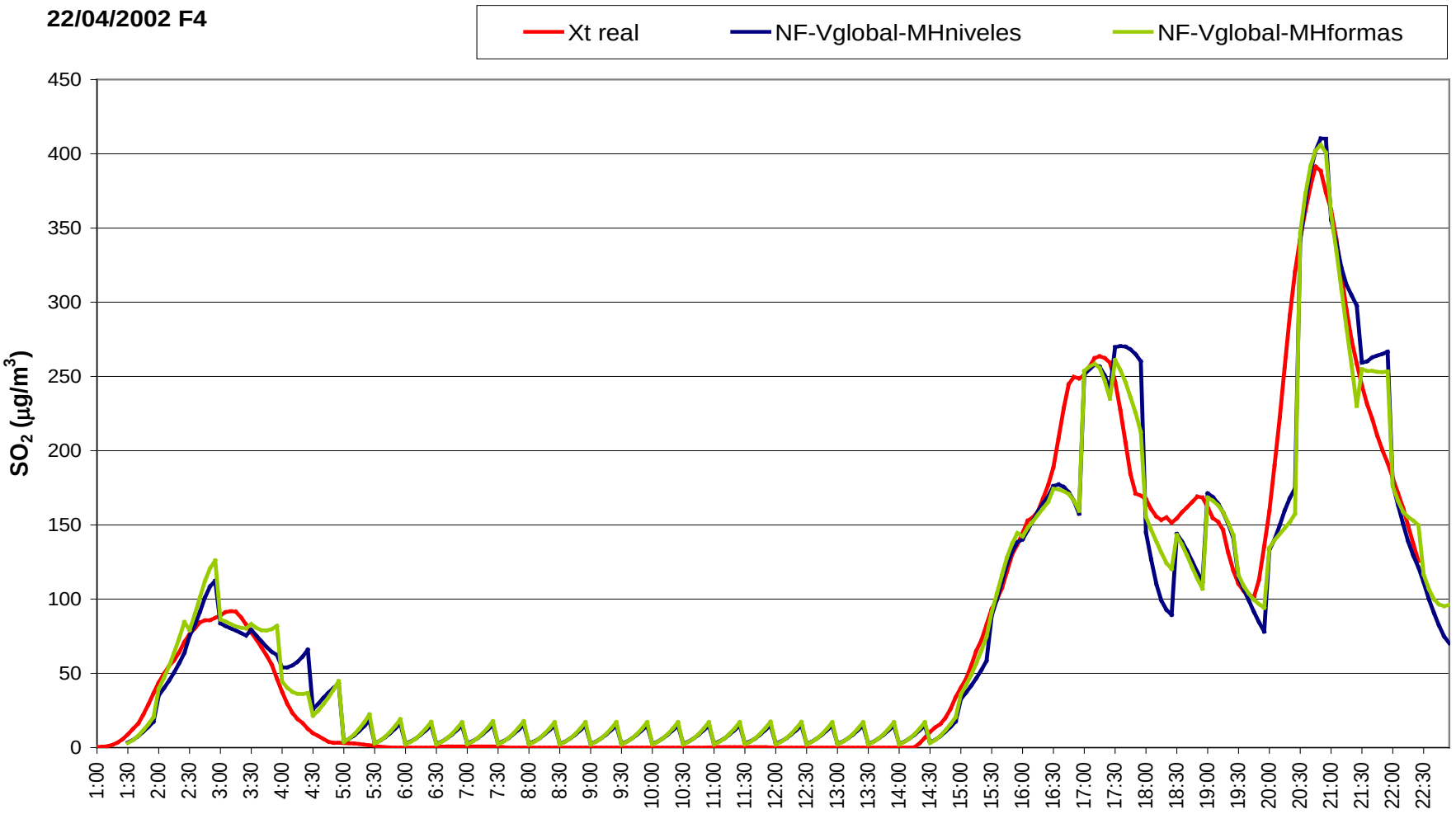
- En consecuencia se consideran variables aleatorias que toman valores en

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

de la forma:

$$X_n\left(u\right) = x\left(6n + u\right)$$

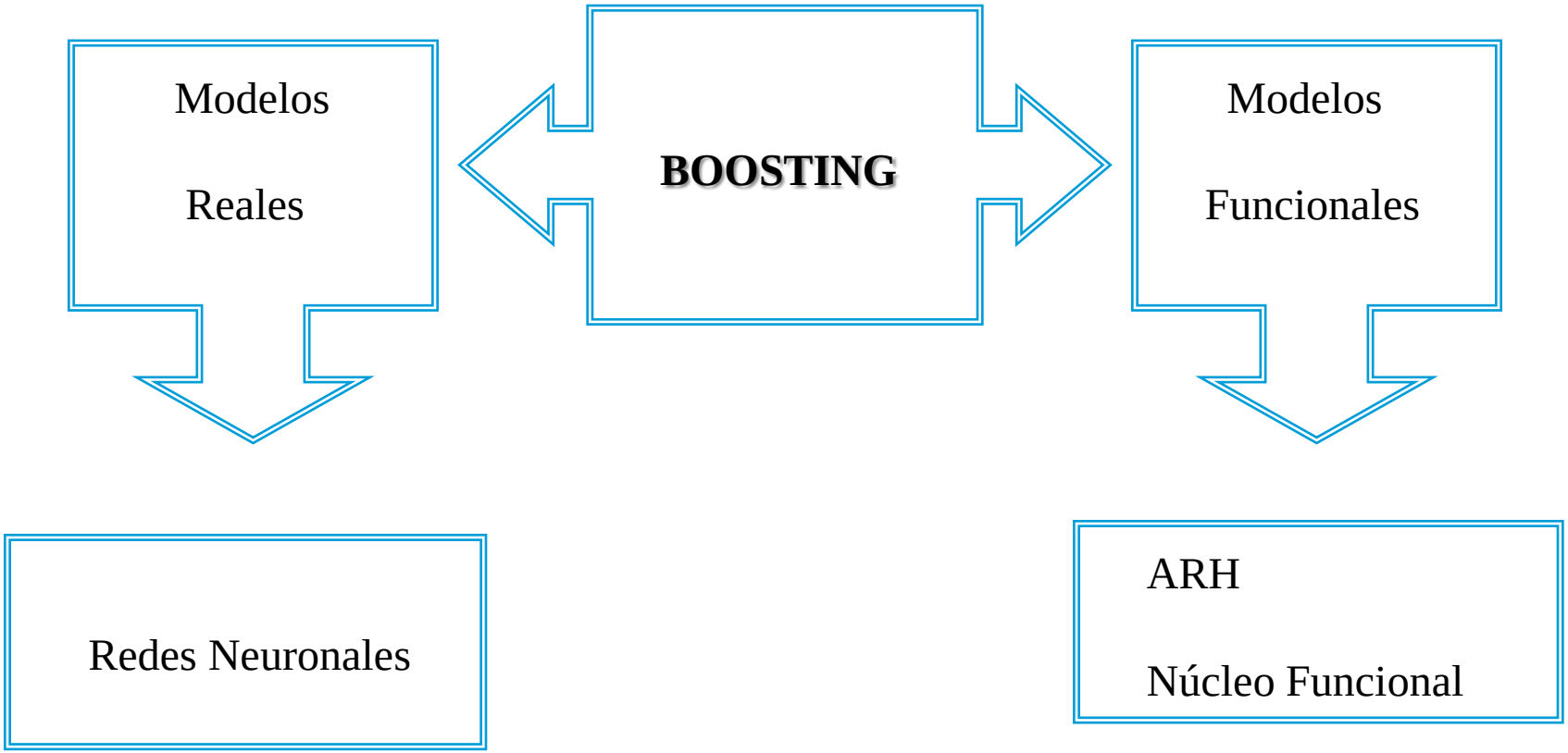
Modelos Funcionales



Modelos Funcionales

Fernández de Castro B.M., Guillas S. and González-Manteiga W. ***Functional Samples and Boostrap for Predicting Sulfur Dioxide Levels.*** Technometrics 2005. 47: 212-222.

Boosting


Modelos
Reales
BOOSTING
Modelos
Funcionales
Redes Neuronales
ARH
Núcleo Funcional

Boosting

Fernández de Castro B.M. and González-Manteiga W. ***Boosting for real and functional samples: an application to an environmental problem.*** Stochastic Environmental Research and Risk Assesment 2008. 22: 27-37.

Nuevas aportaciones a la predicción: el fenómeno de la cointegración

- El nuevo objetivo de los modelos de predicción será modelizar el NO_x , para poder predecir sus valores con media hora de antelación, así como seguir obteniendo las predicciones del SO_2.

- La idea es generalizar el enfoque semiparamétrico unidimensional propuesto por García Jurado, *et al.* (1995), teniendo en cuenta la estructura de correlación entre las series que se pretenden predecir: las relaciones de cointegración.

- Se van a considerar dos posibles enfoques:
  - Cointegración lineal con respuesta multidimensional.
  - Cointegración no paramétrica unidimensional.

Estructura de correlación:Cointegración

- Granger (1983) y Engle & Granger (1987) desarrollaron formalmente el concepto de procesos cointegrados.

- Sea $Z_t = (z_{1t}, \ldots, z_{nt})'$ un vector de n series de tiempo *I(1)*.

  Z_t está *cointegrada* si existe un vector $\beta = \left(\beta_1, \ldots, \beta_n\right)'$ tal que

$$\beta' Z_t = \beta_1 z_{1t} + \ldots + \beta_n z_{nt} \sim \mathrm{I}\left(0\right)$$

- β recibe el nombre de *vector de cointegración*. Dicho vector no es único; se suele considerar la normalización $\beta = \left(1, -\beta_2, \ldots, -\beta_n\right)'$ y así la relación de cointegración se puede expresar:

$$z_{1t} = \beta_2 z_{2t} + \ldots + \beta_n z_{nt} + u_t$$

  donde $u_t \sim \mathrm{I}\left(0\right)$, se denomina *error de desequilibrio* o *residuo de cointegración*.

Cointegración: Procedimiento de Johansen

- Se considera el modelo autorregresivo vectorial de orden p (VAR*(p)*) para el vector de series temporales Z_t

$$Z_t = \Phi D_t + \Pi_1 Z_{t-1} + \ldots + \Pi_p Z_{t-p} + \varepsilon_t, \; t = 1, \ldots, T$$

donde D_t contiene los términos determinísticos.

- Supóngase que Z_t es integrada de orden 1 (I*(1)*) con posibilidad de estar cointegrada.

- La representación VAR no es la más adecuada para el análisis: las relaciones de cointegración no aparecen explícitamente.

Cointegración: Procedimiento de Johansen

- Se transforma el VAR en un *modelo de corrección de errores vectorial* (*VECM*)

$$\Delta Z_t = \Phi D_t + \Pi Z_{t-1} + \Gamma_1 \Delta Z_{t-1} + \ldots + \Gamma_{p-1} \Delta Z_{t-p+1} + \varepsilon_t$$

donde $\Pi = \Pi_1 + \ldots + \Pi_p - I_n$ $\quad \Gamma_k = -\sum_{j=k+1}^{p} \Pi_j, \quad k = 1, \ldots, p-1$

- El rango de la matriz singular Π proporciona la información del número de relaciones de cointegración existentes, es decir, el rango de cointegración.

Modelo Semiparamétrico Multidimensional con Cointegración en los errores.

- Se considera el modelo

$$Z_l = \varphi(Y_l) + e_l$$

donde e_l tiene una estructura $VECM(p)$ independiente de Y_l .

- La predicción se define por:

$$\hat{Z}_t = \hat{\varphi}(Y_t) + \dot{e}_t$$

donde $\hat{\varphi}(Y_t)$ es un estimador no paramétrico y $\dot{e}_t$ es la predicción dada por el $VECM$ de la serie residual

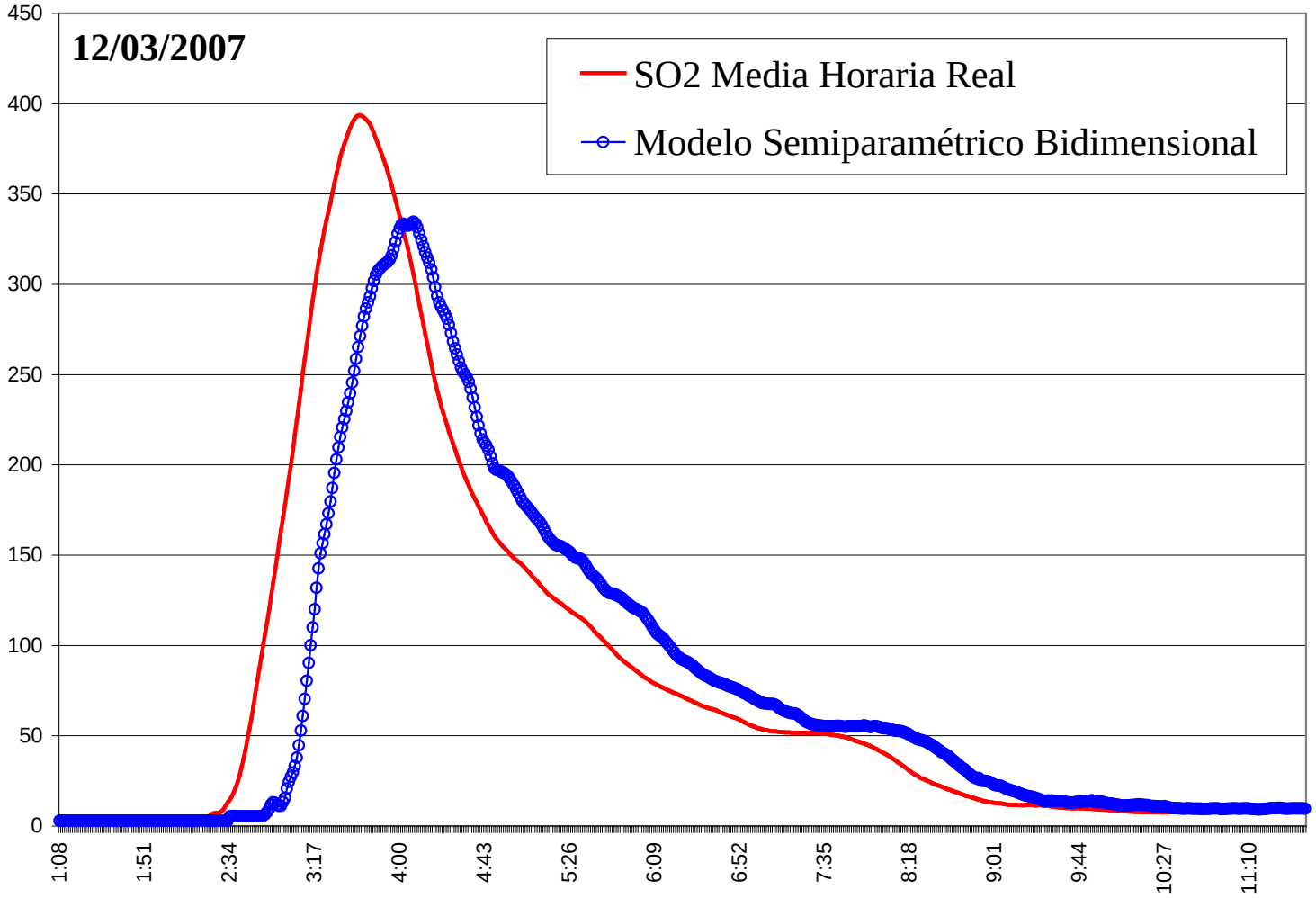
$$\hat{e}_t = Z_t - \hat{\varphi}(Y_t)$$

Aplicación al problema medioambiental

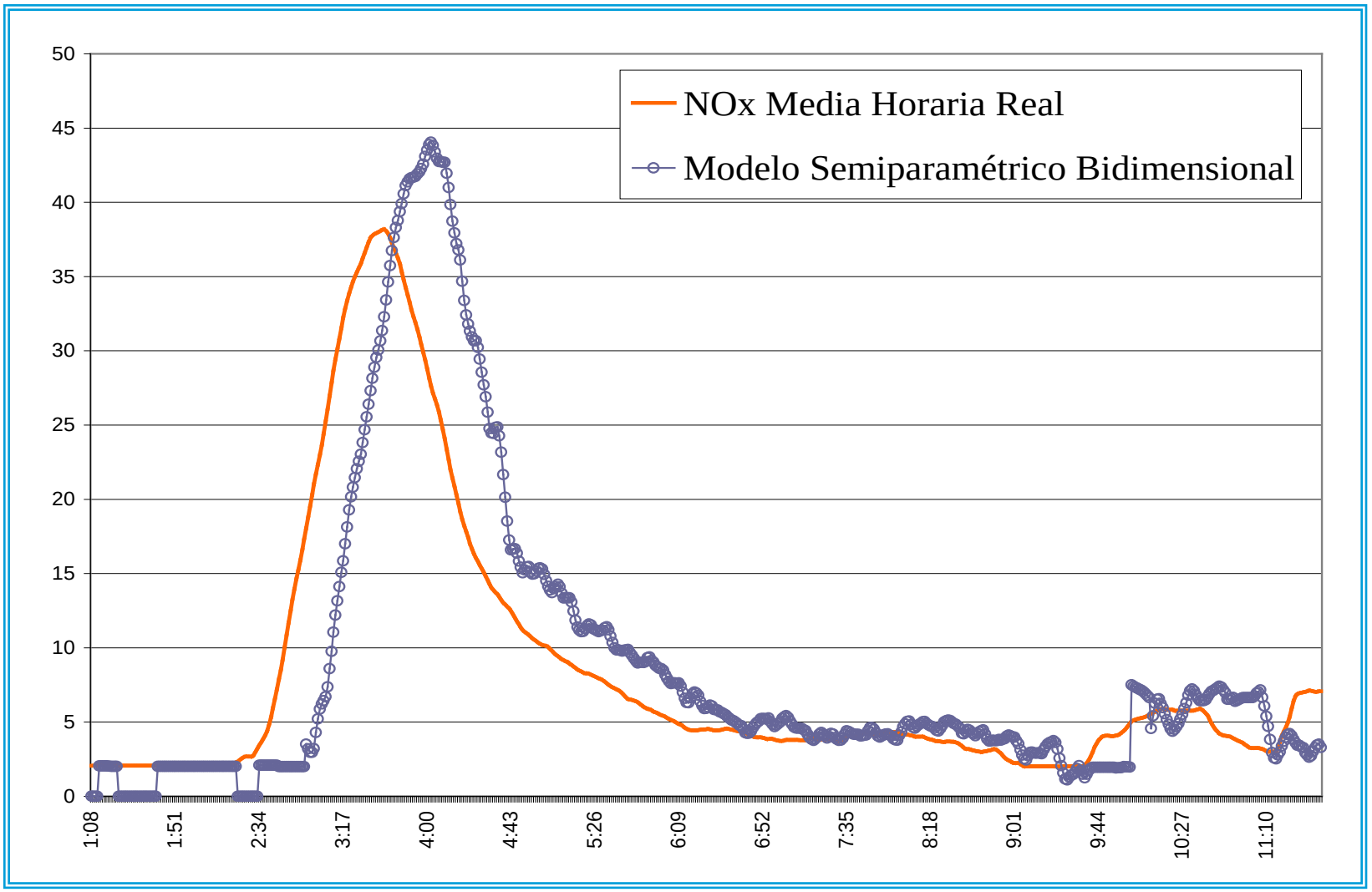
- Sea X_t la serie bidimensional formada por las series de valores medios horarios de SO_2 y NO_x relativa a las últimas cuatro horas.
- Se pretende predecir X_{t+30} utilizando el modelo anterior:
  - Se supone que $X_t = \varphi(X_{t-1}, X_{t-1} - X_{t-6}) + e_t$.
  - En cada instante se estima $\varphi(X_{t-1}, X_{t-1} - X_{t-6})$ con modelos aditivos, de forma independiente para cada componente.
  - Se construye la serie de residuos $\hat{e}_{t-6}, \ldots, \hat{e}_t$ donde $\hat{e}_i := X_i - \hat{\varphi}(X_{i-1}, X_{i-1} - X_{i-6})$ y se ajusta un VECM adecuado.
  - Se obtiene la predicción $\dot{e}_{t+30}$.
  - La predicción final propuesta es:

$$\hat{X}_{t+30} = \hat{\varphi}(\hat{X}_{t+29}, \hat{X}_{t+29} - \hat{X}_{t+24}) + \dot{e}_{t+30}$$

Aplicación al problema medioambiental



Aplicación al problema medioambiental



Bibliografía adicional

- Engle R.F. and Granger C.W.J. *Co-Integration and Error Correction: Representation, Estimation and Testing*. Econometrica 1987; 55: 251-276.

- Granger C.W.J. *Co-Integrated Variables and Error-Correcting Models*. Unpublished University of California, San Diego 1983; Disscusion Paper: 83-13.

- Hastie T.J., Tibshirani R.J. and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York 2001.

- Johansen S. *Statistical Analysis of Cointegration Vectors*. Journal of Economic Dynamics and Control 1988; 12: 231-254.


USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA
endesa generación
U. P. T. As Pontes