UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Departamento de Estatística e Investigación Operativa

Trabajo de Investigación Tutelado

MODELOS DE PREDICCIÓN MEDIOAMBIENTAL

María Piñeiro Lamas

Santiago de Compostela, Julio 2008

Se presenta el trabajo tutelado “Modelos de predicción medioambiental”, realizado bajo la dirección de D. Wenceslao González Manteiga y D. Manuel Febrero Bande, ambos Catedráticos del Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela, como Trabajo de Investigación Tutelado para la obtención de los 12 créditos del periodo de investigación dentro de la etapa de formación del Programa de Doctorado Interuniversitario “Estatística e Investigación Operativa” (Bienio 2006-2008).

Santiago de Compostela, 30 de Julio de 2008.

Asdo. Wenceslao González Manteiga

Asdo. Manuel Febrero Bande

Asdo. María Piñeiro Lamas

Índice

Introducción

Hace unas décadas era impensable que el desarrollo económico pudiera afectar tan negativamente a la naturaleza como para llegar a representar un serio problema. Sin embargo, el acelerado crecimiento de la población humana y el consumo incontrolado de los recursos naturales han hecho mella en el aire, el agua y el suelo. Hoy en día la contaminación provoca importantes daños en la salud humana, los seres vivos y el entorno.

Este modelo económico impuesto por el sistema capitalista basado en un desarrollo sin límites deberá reorientar sus objetivos hacia un desarrollo sostenible coherente con el respeto a la naturaleza y a la biodiversidad. Es necesario tomar conciencia de que estos problemas medioambientales existen, intentando compaginar el desarrollo industrial y tecnológico que nos permita seguir avanzando, con una actitud responsable hacia el planeta.

En los últimos años, tanto los gobiernos y las organizaciones ecologistas como los científicos buscan soluciones para evitar los problemas ambientales. Las acciones políticas para la protección del Medio Ambiente obligan a las empresas a desarrollar planes medioambientales que permitan prevenir daños ecológicos, aunque esto suponga reducciones en sus beneficios.

Por su parte, las investigaciones científicas intentan aportar soluciones desde sus diferentes ámbitos. Entre la infinidad de campos de aplicación de la Estadística, se encuentra el Medio Ambiente.

Este trabajo recoge alguno de los modelos de predicción desarrollados a lo largo de los años, gracias a la amplia colaboración entre el Departamento de Estadística e Investigación Operativa de la USC y la Sección de Medio Ambiente de la Central Térmica de As Pontes, así como los nuevos modelos propuestos ante las modificaciones que se están produciendo actualmente en la Central.

En el capítulo 1 se presenta la problemática ambiental que motiva el desarrollo de modelos estadísticos. En el siguiente capítulo se hace una revisión de los modelos utilizados en los últimos años y, por último, en el capítulo 3 se presentan los modelos elaborados ante la nueva situación de la Central, además de estudiar la estructura de

correlación entre las dos series objeto del estudio. También se van a hacer comparaciones entre los distintos modelos utilizados.

El objetivo final es obtener predicciones fiables que permitan incorporar las técnicas desarrolladas en un Sistema de Control de la Calidad de Aire, como una herramienta efectiva para prevenir la contaminación en el entorno de la Central, y aportar así, una pequeña ayuda al Medio Ambiente.

Capítulo 1. Un problema medioambiental

1.1 Descripción general

La Unidad de Producción Térmica (UPT) de As Pontes constituye uno de los centros productivos propiedad de Endesa Generación S.A. en la península Ibérica. Está situada en el municipio de As Pontes de García Rodríguez, al noreste de la provincia de A Coruña.

La Unidad fue diseñada y construida para hacer uso racional de los lignitos pardos extraídos de la Mina a cielo abierto situada en sus proximidades. Este combustible sólido se caracteriza por sus elevados contenidos en humedad y azufre, así como por su bajo poder calorífico. El conjunto de ambas instalaciones, Mina y Central Térmica, constituían el Complejo Mineroeléctrico de As Pontes.

La instalación inició su actividad en 1976 con la puesta en marcha de un grupo; disponiendo en la actualidad de 4 grupos de generación de energía eléctrica con una potencia nominal de aproximadamente 350 MW cada uno, que funcionan de forma independiente.

En el año 1993 se inició un proceso de transformación de la Central, con objeto de utilizar, con la máxima eficiencia, mezclas de lignito local con carbones subbituminosos de importación caracterizados por sus bajos contenidos en azufre y cenizas. Esta transformación, finalizada en el año 1996, ha permitido alcanzar una reducción global en las emisiones de dióxido de azufre (SO_2) superior al 40 %, cumpliendo compromisos previamente establecidos con las Administraciones Central y Autonómica. Con esta actuación se ha asegurado también la explotación del lignito de As Pontes hasta el final del yacimiento, alargando por tanto su vida útil.

En la actualidad Endesa Generación S.A. ha finalizado una nueva adaptación de la UPT de As Pontes para consumir, como combustible principal, carbón subbituminoso de importación. Esta actuación se ha llevado a cabo sucesivamente en los cuatro grupos generadores en el período comprendido entre 2005 y 2008.

Los objetivos fundamentales de esta adaptación son los siguientes:

- Dar cumplimiento a los requisitos establecidos en la Directiva 2001/80/CE del Parlamento Europeo y del Consejo, del 23 de octubre de 2001, sobre limitación de las emisiones a la atmósfera de determinados agentes contaminantes procedentes de grandes instalaciones de combustión.

- Prolongar la vida útil de la Central Térmica más allá del agotamiento y cierre de la explotación del lignito local, realizado el 1 de enero de 2008.

La legislación vigente y la localización de la Central Térmica próxima a enclaves naturales de alto valor ecológico, como el Parque Natural de “As Fragas do Eume”, uno de los bosques atlánticos mejor conservados de Europa, declarado espacio natural protegido en 1997, o la sierra de “A Capela”, de especial interés geológico, hacen que desde sus inicios haya existido una gran preocupación por su impacto en el entorno.

En 1976, primer año de producción de la instalación, se creó también una Red de Vigilancia de la Calidad de Aire con 32 estaciones de medida manuales en el entorno de la central. Esta Red ha ido evolucionando con el paso del tiempo, llegando a disponer de 40 estaciones manuales y 17 estaciones automáticas.

En la actualidad, la Central Térmica posee un Sistema de Seguimiento y Control de la Calidad Atmosférica formado por: un Sistema de Control de Emisiones, una Red de Vigilancia y Control de la Calidad Atmosférica, una estación meteorológica y una Unidad Central de Adquisición de Datos y Gestión de la Información.

El Sistema de Control de Emisiones está constituido por analizadores automáticos de dióxido de azufre, óxidos de nitrógeno, partículas en suspensión, temperatura y oxígeno en cada uno de los grupos generadores. Este sistema se encuentra situado en la planta 215 de chimenea, muy alejado de cualquier perturbación del flujo de gases y, por tanto, en una zona óptima de medida. Los datos que, en tiempo real, transmiten estos analizadores, suministran la información necesaria para mantener en todo momento la calidad de los gases dentro de los límites legalmente establecidos.

La Red de Vigilancia y Control de la Calidad Atmosférica (Figura 1.1) está formada por 17 estaciones automáticas, distribuidas en un radio de 30 km y comunicadas en tiempo real, vía radio, con la Central Térmica. Las estaciones automáticas proporcionan una medición continua de las concentraciones de dióxido de azufre,

óxidos de nitrógeno, partículas en suspensión y, en algún caso, ozono, y parámetros meteorológicos.

La Estación Meteorológica, denominada E. M. de A Mourela, situada aproximadamente a un kilómetro de la instalación, está dotada de un mástil de 80 metros de altura con medidas de temperatura, velocidad y dirección de viento a distintos niveles, así como de sensores de humedad relativa, precipitación, radiación solar y presión atmosférica a nivel del suelo.

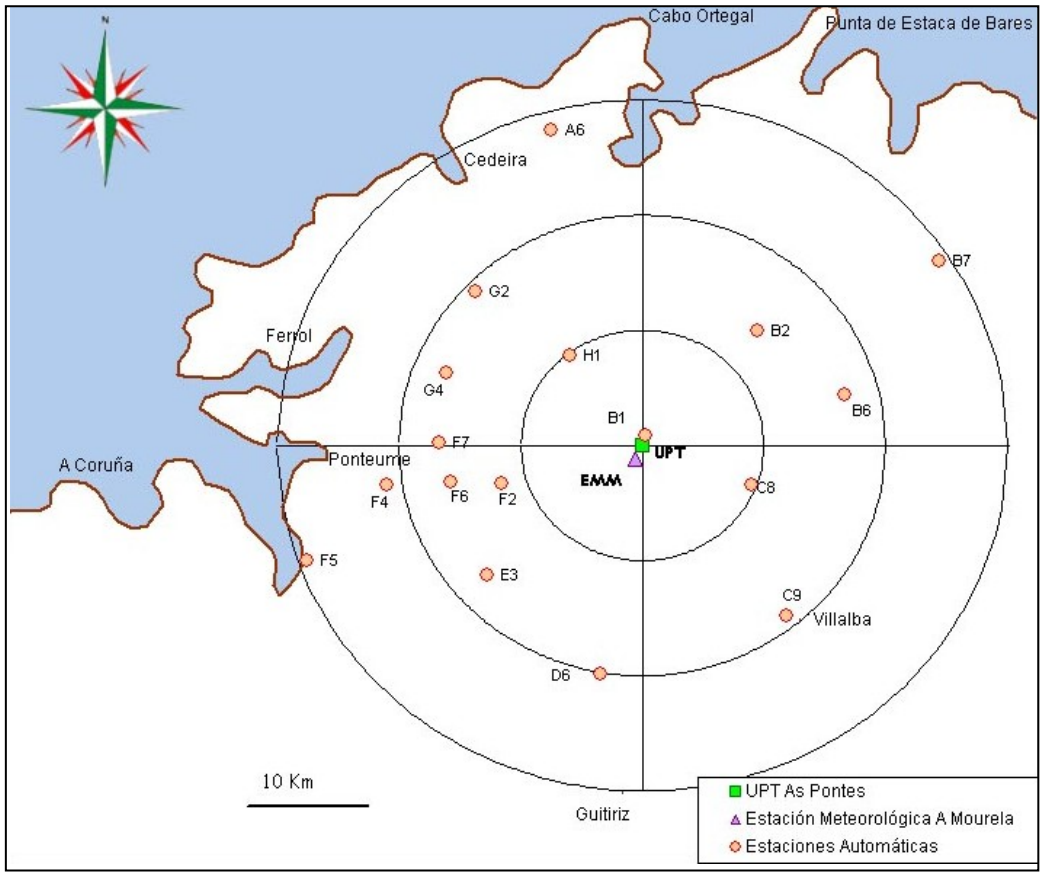

Figura 1.1: Red de Vigilancia y Control de la Calidad Atmosférica de la Central Térmica de As Pontes, propiedad de Endesa Generación S. A. (Noroeste de España)

La Unidad Central de Adquisición de Datos y Gestión de la Información, situada en la Central Térmica, recibe, gestiona y almacena los datos enviados por el Sistema de Control de Emisiones, todas las estaciones de la Red de Vigilancia y Control de la Calidad Atmosférica y la Estación Meteorológica. La comunicación entre la Unidad

Central y las distintas estaciones se realiza vía radio. Esta Unidad posee además periféricos en las Salas de Control de la instalación.

Los lignitos pardos extraídos de la mina tienen un alto contenido en azufre, como ya se ha comentado, por lo que en el proceso de combustión del carbón para generar energía eléctrica se obtienen como residuos diversos óxidos, entre los que se encuentra el azufre (en estado gaseoso), que son emitidos a la atmósfera por la chimenea, de 350 metros de altura, de la Central. El penacho emitido, concretamente el dióxido de azufre que contiene, puede provocar episodios de alteración de la calidad del aire en el entorno de la instalación, en condiciones desfavorables para su dispersión. La legislación española mediante normas y decretos fija las concentraciones máximas que se pueden alcanzar de estos gases en un determinado período de tiempo. En particular, para esta Central el único límite susceptible de ser rebasado alguna vez, es aquel que se establece sobre la media horaria arrastrada de la concentración de SO_2 en el suelo, en el valor de 350 $\mu g/m^3$.

A lo largo de los años se ha investigado para poder anticiparse a estos episodios y así poder prevenirlos. Además, gracias a las transformaciones efectuadas en la Central, estos episodios son cada vez menos frecuentes y, se espera que con la última adaptación de la Central para utilizar únicamente carbón subbituminoso de importación apenas se produzcan.

Es posible determinar qué condiciones físicas no favorecen la dispersión del penacho emitido, debido a que se desplaza con las masas de aire, además de los parámetros propios que caracterizan las emisiones atmosféricas (concentración del contaminante, velocidad de salida de los gases, temperatura, altura de la chimenea,…) y diversas variables meteorológicas, como el gradiente térmico, la velocidad de viento, su dirección o la altura de la capa de mezcla. Por tanto, es posible caracterizar teóricamente las situaciones en las que se produce un episodio de la calidad de aire.

El inconveniente surge al intentar clasificar situaciones meteorológicas reales, ya que el penacho se rige por las condiciones que se producen en altura (generalmente a alturas de 600 a 1300 metros sobre el nivel del mar) y los medidores existentes están a niveles que oscilan entre los 10 y 150 metros de altura. Así, las situaciones son conocidas pero no son medibles, es decir, no es posible, con los medios disponibles, establecer las condiciones que se están produciendo en un determinado momento.

También es importante señalar que los episodios de contaminación asociados a cada estación son distintos.

Esta dificultad hace que para poder determinar la ocurrencia de un episodio los datos que se van a utilizar son los datos pasados de cada estación de medida y la única suposición que se va a hacer es pensar que el comportamiento en el pasado se va a parecer al comportamiento futuro.

En resumen, el problema que se plantea es poder predecir la citada media horaria de los niveles de SO_2, a partir de la información que se recibe en continuo de las estaciones de muestreo y la información pasada de dichas medidas. El objetivo de estas predicciones es conseguir activar los sistemas de reducción de emisiones con tiempo suficiente para evitar episodios de alteración de calidad de aire. Esta reducción de emisiones se conseguía, antes de la última adaptación, mediante la modificación de la mezcla de carbón consumido, aumentando la proporción de carbón importado, con un menor contenido en azufre. Ahora, la opción será quemar una cantidad menor de combustible.

Los modelos estadísticos de predicción son la clave para obtener estas predicciones y sugerir una línea de actuación a los operadores de la Central, para intentar evitar los episodios de calidad de aire.

1.2 Nueva problemática

En el emplazamiento de la Central Térmica de As Pontes, se ha construido una nueva Central de Ciclo Combinado de Gas Natural, también propiedad de Endesa Generación S. A., actualmente, en período de prueba.

La nueva Central de Ciclo Combinado consiste en un grupo generador de electricidad de tecnología de Ciclo Combinado, formado por dos turbinas de gas y una turbina de vapor de potencia nominal en el entorno de 800 MW, diseñado para emplear gasóleo como combustible líquido de emergencia. El Ciclo Combinado sólo operará con gasóleo cuando se produzca un fallo en el suministro de gas natural y el Operador Nacional del Sistema considere que la Planta debe aportar energía a la red

en esa situación. El gasóleo a emplear como combustible de emergencia tendrá un contenido máximo de azufre del 0,005% en peso.

El Ciclo Combinado consiste en una configuración 2 x 1 de la siguiente manera:

- Dos turbinas de gas con potencia de 270 MW cada una.
- Dos calderas de recuperación de calor de disposición horizontal con 3 niveles de presión y recalentamiento.
- Una turbina de vapor con potencia bruta de 297 MW.

La potencia bruta de la instalación en el emplazamiento es de aproximadamente 837 MW.

La Central tiene dos focos de emisión (chimeneas) con una altura de 80 m y un diámetro de 6999 mm, asociadas a cada turbina.

El Sistema de Control de Emisiones proporciona medidas en continuo de los siguientes contaminantes: dióxido de azufre, óxidos de nitrógeno, partículas, monóxido de carbono, oxígeno, compuestos orgánicos volátiles, presión y temperatura.

Como consecuencia de la instalación de la nueva Central de Ciclo Combinado, así como de la transformación de los cuatro grupos de la Central Térmica, comentada en la sección anterior, se hace necesaria la redefinición de la Red de Inmisión con objeto de asegurar la adecuada vigilancia de la calidad de aire del entorno. Esto supone la ubicación de puntos de medida más próximos a los focos emisores y la medición de nuevos contaminantes (monóxido de carbono y partículas en suspensión de menor tamaño, 2,5 µm de diámetro).

Los combustibles que van a ser utilizados hacen que el principal interés recaiga en predecir los valores de los óxidos de nitrógeno (NO_x), para así, evitar superar los niveles límite fijados por la legislación .También será interesante seguir prediciendo los valores del SO_2.

Se plantea entonces un nuevo problema: predecir las concentraciones medias horarias de dióxido de azufre y de los óxidos de nitrógeno, medidas en el entorno de las dos instalaciones, con el fin de evitar episodios de alteración de la calidad de aire.

Además, también será necesario determinar cuál es el origen de los episodios de alteración de calidad de aire: la Central Térmica, el Ciclo Combinado u otros posibles focos, como por ejemplo el tráfico o las actividades agrícolas de la zona.

Ante este nuevo planteamiento, los modelos estadísticos de predicción vuelven a ser una herramienta eficaz para intentar evitar episodios de alteración de la calidad de aire.

1.3 Marco legislativo actual

El establecimiento de los valores límite que garanticen una calidad de aire satisfactoria, precisa de un desarrollo para las distintas sustancias contaminantes. Este desarrollo se concreta, dentro de la Unión Europea, en la *Directiva 1999/30/CE del Consejo, de 22 de abril de 1999, relativa a los valores límite de dióxido de azufre, dióxido de nitrógeno y óxidos de nitrógeno, partículas y plomo en el aire ambiente*, y en la *Directiva 2000/69/CE del Parlamento Europeo y del Consejo, de 16 de noviembre de 2000, sobre los valores límite para el benceno y el monóxido de carbono en el aire ambiente*.

Ante la obligación de los países miembro de incorporar al derecho interno las normas comunitarias, las Directivas anteriormente citadas han sido transpuestas a la legislación española. Así, el *Real Decreto 1073/2002, de 18 de octubre, sobre evaluación y gestión de la calidad del aire ambiente en relación con el dióxido de azufre, dióxido de nitrógeno, óxidos de nitrógeno, partículas, plomo, benceno y el monóxido de carbono*, vigente en la actualidad, establece objetivos de calidad de aire con condiciones particulares para cada sustancia contaminante.

Concretamente, los valores límite y umbrales de alerta, que el citado Real Decreto establece para el SO_2, son los siguientes:

- Valor límite horario para la protección de la salud humana:

- Período de promedio: 1 hora.

- Valor límite: 350 µg/m^3, valor que no podrá superarse en más de 24 ocasiones por año civil.

- Margen de tolerancia: 90 µg/m^3, a la entrada en vigor del Real Decreto, reduciendo el 1 de Enero de 2003 y posteriormente cada 12 meses 30 µg/m^3, hasta alcanzar el valor límite el 1 de enero de 2005

- Valor límite diario para la protección de la salud humana:

  - Período de promedio: 24 horas.

  - Valor límite: 125 µg/m^3, valor que no podrá superarse en más de 3 ocasiones por año civil.

- Valor límite para la protección de los ecosistemas:

  - Período de promedio: Año civil e invierno (del 1 de octubre al 31 de marzo).

  - Valor límite: 20 µg/m^3.

- Umbral de alerta: 500 µg/m^3 registrados durante tres horas consecutivas en lugares representativos de la calidad del aire.

Los valores límite y umbrales de alerta, que establece el Real Decreto 1073/2002 para el dióxido de nitrógeno (NO_2) y los óxidos de nitrógeno (NO_x), son los siguientes:

- Valor límite horario para la protección de la salud humana:

  - Período de promedio: 1 hora.

  - Valor límite: 200 µg/m^3 de NO_2 que no podrán superarse en más de 18 ocasiones por año civil.

  - Margen de tolerancia: 80 $\mu g/m^3$ a la entrada en vigor del Real Decreto, reduciendo el 1 de Enero de 2003 y posteriormente cada 12 meses 10 $\mu g/m^3$, hasta alcanzar el valor límite el 1 de enero de 2010.

- Valor límite anual para la protección de la salud humana:

  - Período de promedio: 1 año civil.

  - Valor límite: 40 $\mu g/m^3$ de NO_2.

  - Margen de tolerancia: 16 $\mu g/m^3$, a la entrada en vigor del Real Decreto, reduciendo el 1 de Enero de 2003 y posteriormente cada 12 meses 2 $\mu g/m^3$, hasta alcanzar el valor límite el 1 de enero de 2010.

- Valor límite anual para la protección de la vegetación:

  - Período de promedio: 1 año civil.

  - Valor límite: 30 $\mu g/m^3$ de NO_x.

- Umbral de alerta: 400 $\mu g/m^3$ registrados durante tres horas consecutivas en lugares representativos de la calidad del aire.

1.4 Los datos

La Unidad Central de Adquisición de Datos y Gestión de la información de la Central Térmica de As Pontes, crea bases de datos minutales y pentaminutales a partir de las medidas de emisión, meteorología y calidad de aire que recibe. Un sistema de comunicación mediante archivos permite el acceso a dicha base de datos, siendo posible, por tanto, disponer de datos reales cada minuto y cada cinco minutos.

La legislación vigente y la disponibilidad de datos con frecuencia minutal en tiempo real, nos hacen considerar la media horaria arrastrada tanto de los valores de SO_2 como del NO_x, para obtener las predicciones de los valores futuros de ambos contaminantes. Así vamos a construir dos series temporales x_t e y_t, para las que el subíndice t representa un instante minutal, y cada valor se va a obtener como promedio de los valores reales correspondientes a la hora anterior:

$$x_t = \frac{1}{60}\sum_{i=0}^{59} SO_2(t-i)$$

$$y_t = \frac{1}{60}\sum_{i=0}^{59} NO_x(t-i)$$

donde $SO_2(t)$ y $NO_x(t)$ representan la concentración de SO_2 y NO_x en el instante t, medida en $\mu g/m^3$, respectivamente.

Cada uno de los puntos de medida, es decir, cada una de las 17 estaciones de la Red de Vigilancia y Control de la Calidad Atmosférica, se va a tratar de forma individual. Por tanto, vamos a disponer de 34 series temporales, dos para cada una de las estaciones de medida. Toda la metodología se ha desarrollado de forma global, pero a la hora de obtener y mostrar los resultados, hablaremos de las estaciones de forma individual.

La serie de valores medios horarios de SO_2 tiene un comportamiento bastante peculiar, muy influenciado por las condiciones meteorológicas y la topografía local. Toma valores próximos a cero durante largos períodos de tiempo, y crece de manera repentina en condiciones meteorológicas desfavorables para la dispersión del penacho. Estos valores altos de concentración, se mantienen hasta que la reducción de emisiones llevada a cabo en la Central Térmica surte efecto en el lugar de impacto del episodio de alteración de la calidad del aire, o bien, las condiciones meteorológicas varían favoreciendo la dispersión del penacho. Los episodios suelen durar entre 1 y 4 ó 5 horas, dependiendo de la distancia entre la Central y el punto de impacto, y la dirección y velocidad de viento. Además ocurren en su mayoría en los meses de primavera y verano, ya que es en estos meses cuando se dan las condiciones meteorológicas desfavorables para la dispersión del penacho. En la Figura 1.2 se puede observar la frecuencia en la que ocurren los distintos valores de la serie de medias horarias de SO_2. En esta figura se representa una estimación de la densidad de los valores minutales de los meses de junio, julio y agosto de 2006,

correspondientes a la estación B6. Como se puede ver, la mayoría de los valores están entre 0 y 50 µg/m^3.

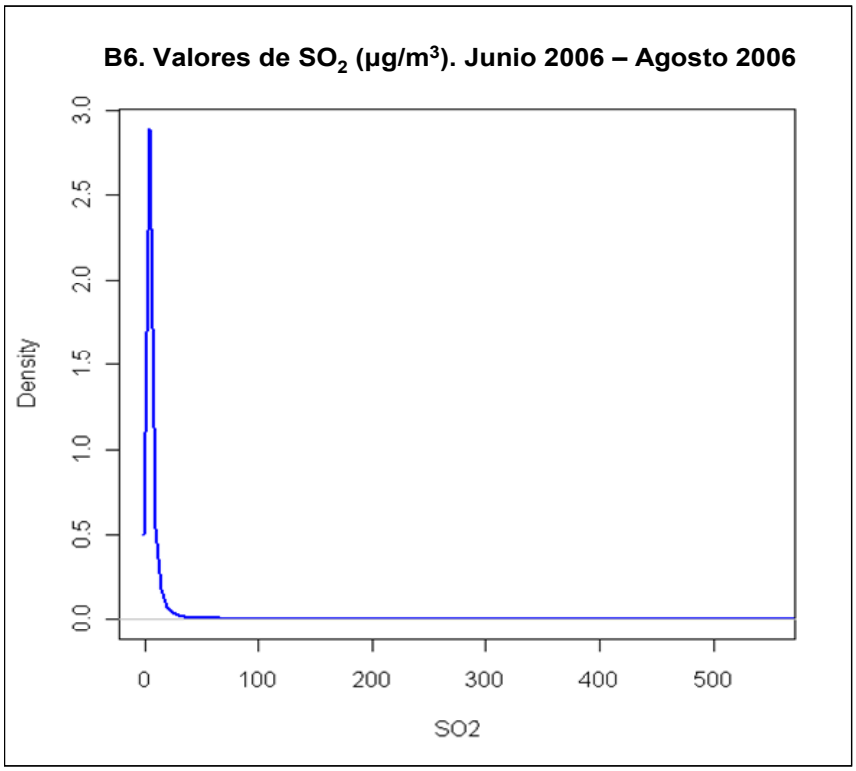

Figura 1.2: Estimación de la densidad de los valores medios horarios de SO_2. Estación B6. Período: Junio 2006–Agosto 2006

En la actualidad, la serie de valores medios horarios de NO_x tiene un comportamiento similar a la del SO_2 pero a menor escala, es decir, los episodios de NO_x no alcanzan niveles de concentración tan altos como los de SO_2 y son simultáneos a los de dicho contaminante, salvo en algunas excepciones (sólo en tres ocasiones en el 2006) que son debidos a otros focos de contaminación ajenos a la Central Térmica.

En la Figura 1.3 se puede observar la evolución de un episodio de alteración de la calidad del aire para el SO_2 ocurrido en una de las estaciones de medida el 12 de marzo del año 2007. También se muestran los valores de NO_x de ese mismo día.

La puesta en marcha del Ciclo Combinado y la total transformación de la Central Térmica van a provocar un cambio en esta situación: los episodios de SO_2 serán mucho más esporádicos y los episodios que realmente tendrán importancia serán los

del NO_x. Se espera que los episodios de NO_x tengan un comportamiento similar al que tienen ahora los del SO_2.

El principal objetivo de los modelos estadísticos desarrollados es predecir los episodios de alteración de la calidad del aire, por lo que nuestro interés se centra en los valores que menos ocurren a lo largo de la serie temporal. Además, la predicción ha de darse con la anticipación suficiente para que los operarios de la Central puedan activar los sistemas de reducción de emisiones y se consiga minimizar el número de superaciones de los valores de referencia establecidos por la legislación vigente.

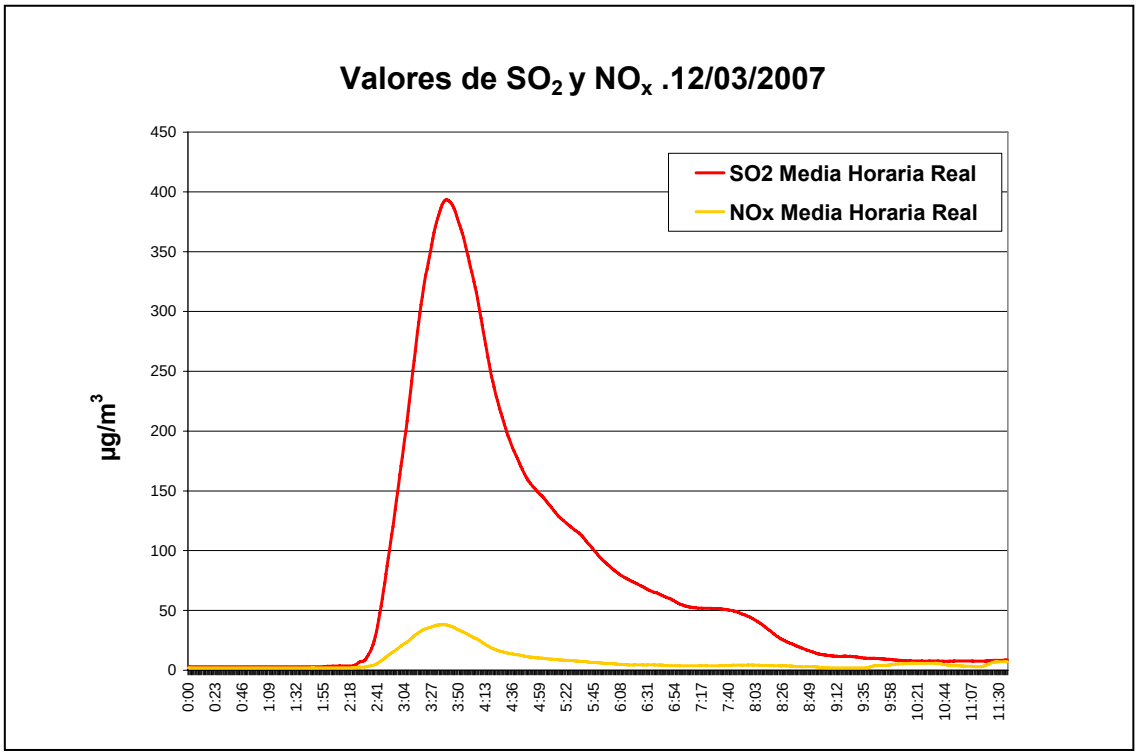

Figura 1.3: Episodio de alteración de la calidad del aire ocurrido en una de las estaciones de medida el 12 de marzo de 2007.

Capítulo 2. Revisión de los modelos de predicción

Desde 1992, el Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela ha venido desarrollando diversos modelos de predicción de niveles de SO_2, enfocados a funcionar como herramientas de apoyo en la toma de decisiones en la Central Térmica de As Pontes. Dichos desarrollos se han llevado a cabo en el marco de diversos proyectos de investigación en colaboración con la Sección de Medio Ambiente de la Central Térmica, propiedad de Endesa Generación S. A.

La evolución de los distintos modelos ha dado lugar a un Sistema de Predicción Estadística de Inmisión (SIPEI) que proporciona predicciones puntuales y de probabilidad, para cada una de las 17 estaciones de la Red de Vigilancia y Control de la Calidad Atmosférica, además de mapas de predicción espacial. Este sistema funciona en continuo, recibiendo los datos medidos en las distintas estaciones y proporcionando las predicciones en tiempo real.

En este capítulo vamos a hacer una revisión de algunos de los modelos de predicción puntual que han sido desarrollados.

2.1 Modelos Semiparamétricos

A lo largo de los años, los modelos estadísticos desarrollados se han ido adaptando a la normativa vigente en cada momento y a las necesidades de la Central Térmica.

En los primeros años de desarrollo, la frecuencia de envío de datos al Sistema de Predicción Estadística de Inmisión era pentaminutal. Desde las estaciones de medida se transmitían al laboratorio central, cada diez segundos, las concentraciones de distintas sustancias. Las medidas recibidas se promediaban cada cinco minutos y se creaba un nuevo fichero de datos.

Además, la normativa vigente en ese momento, anterior a la entrada en vigor de la Directiva Europea 1999/CE/30 y el posterior Real Decreto 1073/2002, establecía los valores límite para la media bihoraria de las medidas del SO_2.

Por esta razón los modelos de predicción puntuales de niveles SO_2 trabajaban inicialmente con la serie temporal de medias bihorarias arrastradas, x_t, donde t representa un instante pentaminutal, y cada valor se obtiene como promedio de los valores correspondientes a las 2 horas anteriores:

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

donde $SO_2(t)$ representa la concentración de SO_2 en el instante t, medida en μg/m^3.

El objetivo era obtener la predicción puntual, con horizonte de media hora, para esta serie temporal. Por tanto, cada vez que se recibe una nueva observación, x_t, se ha de predecir el valor a seis instantes de tiempo, x_{t+6}.

Como ya se ha comentado en el capítulo anterior, dicha serie tiene un comportamiento peculiar ya que puede incrementarse de forma brusca y acusada. Estos cambios generalmente están bastante espaciados en el tiempo. Además, no verifica la estabilidad de la varianza. Dichas peculiaridades hacen pensar que la metodología clásica de Box – Jenkins no es muy apropiada para este problema.

En la Figura 2.1 se muestran las predicciones (a media hora) utilizando esta metodología para un episodio real ocurrido el 18 de Marzo de 1993 en una de las estaciones de medida. La línea roja representa la serie real y la línea azul las predicciones obtenidas con Box – Jenkins.

En la figura se puede apreciar que la predicción tiene una serie de defectos. Uno de estos defectos son los valores excesivamente altos que toma la predicción con respecto a la serie real. Además, se predicen valores negativos de una variable que no puede ser negativa y existen instantes en los que no se puede realizar la predicción dado que no es posible encontrar un modelo ARIMA adecuado con un número de parámetros moderado (no se permiten más dos diferencias y más de cinco parámetros en las partes AR y MA). Los resultados no son muy satisfactorios. Hay que tener en cuenta que, en cada instante, se utilizan los datos recogidos las últimas seis horas (72 observaciones).

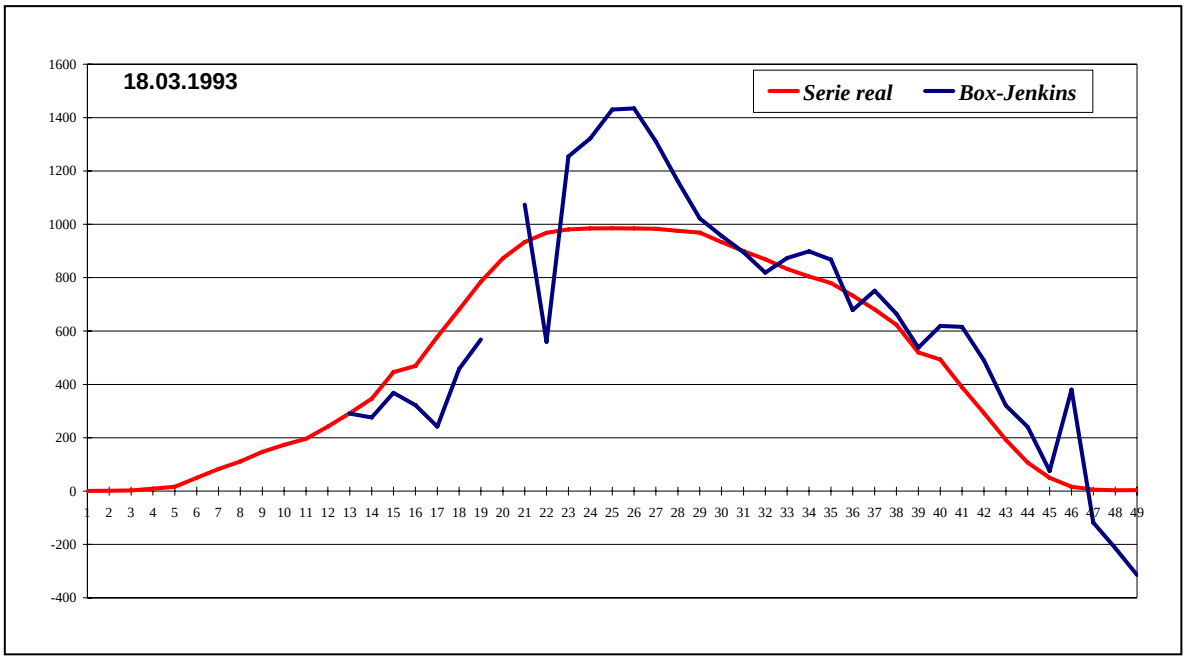

Figura 2.1: Episodio de alteración de la calidad del aire ocurrido en una de las estaciones de medida el 18 de Marzo de 1993. Predicción dada por la metodología Box-Jenkins (Prada-Sánchez, 1997).

En vista de los inconvenientes de la metodología Box – Jenkins, se optó por un planteamiento no paramétrico.

En cada instante (pentaminutal) t, se estima $E\left(X_{t+6} / X_t, X_{t-1}\right)$ mediante el estimador de Nadaraya-Watson para la regresión con una ventana óptima de tipo de Validación Cruzada local y kernel Gaussiano (ver Hastie & Tibshirani, 1990).

Hay que tener en cuenta que, con la regresión no paramétrica lo que se hace, en definitiva, es ponderar los datos existentes para realizar una predicción en un punto. Por tanto si la serie tiene un máximo en un determinado valor, cualquier predicción con este sistema siempre será menor o igual que dicho valor. Notemos, por otra parte que, como ya se ha comentado para el caso Box – Jenkins, puede ocurrir que el último episodio producido en una estación ocurriera un mes antes. Estas dificultades podrían superarse si se contase con más datos para la estimación y se pudiera además introducir en la modelización alguna componente capaz de aportar capacidad de improvisación a la misma.

Añadir datos para poder predecir no supone una dificultad en este enfoque no paramétrico ya que se están tratando los datos en un contexto puro de regresión. En el

enfoque Box – Jenkins la estimación depende fuertemente del orden temporal de los datos. Esta dependencia temporal provoca que, si se quiere estimar en un punto determinado con información que se produjo en un pasado lejano, se deben suministrar todos los datos entre esta información pasada y el momento actual en el orden en que se produjeron. En cambio, en un contexto puro de regresión, para predecir en ese punto, basta considerar registros procedentes de fragmentaciones independientes de la serie, conteniendo información de tipo causa-efecto.

Para responder a esta demanda de información (distinta para cada estación de muestreo) se diseñó un tipo de memoria denominado *matriz histórica* (González Manteiga, *et al*., 1993, García Jurado, *et al*., 1995 y Prada-Sánchez, *et al*., 1997). Esta matriz se compone de un número grande de registros (1000 en este caso) de la forma $\left(x_{t-1}, x_t, x_{t+6}\right)'$, ternas de datos reales de medias bihorarias de SO_2, elegidos de forma que cubran todo el rango de la variable y que harán el papel de memoria histórica de ésta. Para asegurar que cubren todo el rango de la variable en cuestión, se divide la matriz en bloques atendiendo al nivel de la variable respuesta, x_{t+6}. Para actualizar la memoria, cada vez que llega un nuevo dato se construye el registro correspondiente al que pertenece. Dicho registro entra en ese bloque sustituyendo al registro más antiguo del mismo.

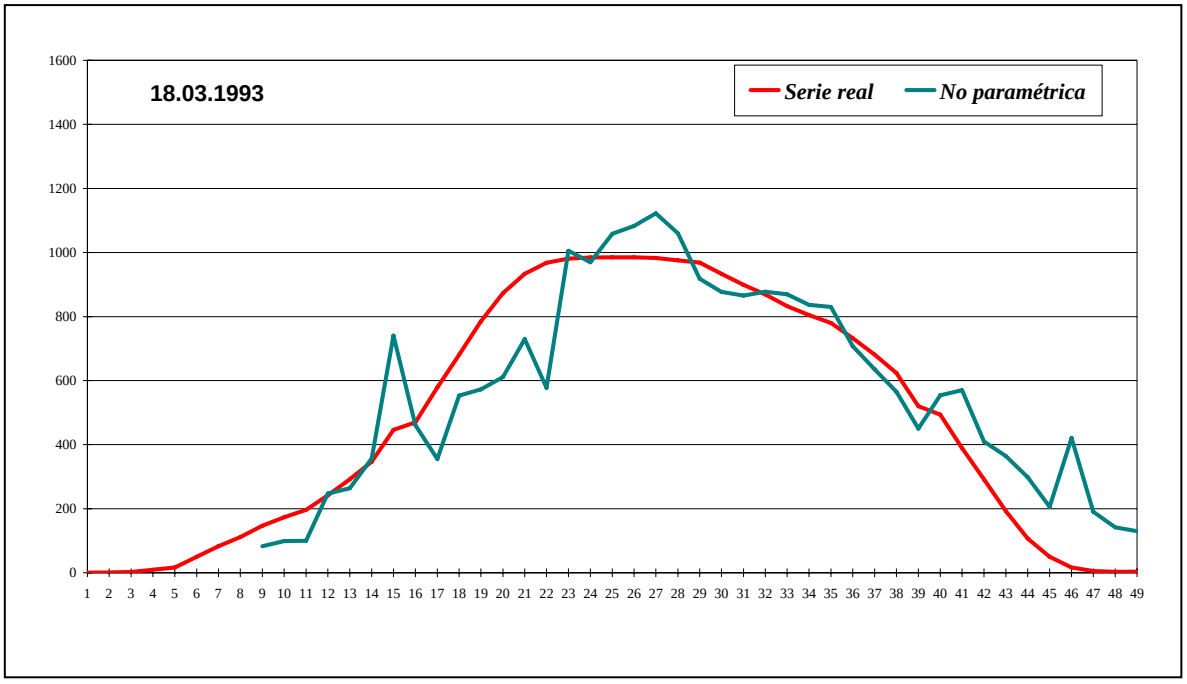

Figura 2.2: Episodio de alteración de la calidad del aire ocurrido en una de las estaciones de medida el 18 de Marzo de 1993. Predicción dada por la metodología no paramétrica (Prada-Sánchez JM, 1997).

Con una muestra así construida, se asegura que en todo momento se dispone de información actualizada sobre todo el rango de variación de la variable de interés, en este caso, la concentración de SO_2 en media bihoraria.

Si utilizamos la matriz histórica para calcular el estimador de Nadaraya – Watson de $E\left(X_{t+6} / X_t, X_{t-1}\right)$, denotado por $\hat{E}\left(X_{t+6} / X_t, X_{t-1}\right)$, la predicción no paramétrica mejora considerablemente.

En la Figura 2.2 puede observarse el buen comportamiento de la metodología no paramétrica con *memoria histórica*.

A la vista de los dos últimos gráficos, se puede decir que la metodología no paramétrica se comporta mejor que los métodos Box – Jenkins. Sin embargo, este enfoque no es completamente satisfactorio ya que, si en cada instante t (en el cual la serie con la que se trabaja es $X_{t-71},\ldots,X_t$) se realiza un test de adecuación del modelo (por ejemplo, el test de Ljung – Box; Ljung & Box, 1978) a la serie $\hat{Z}_{t-64},...,\hat{Z}_t$,donde $\hat{Z}_i := X_i - \hat{E}(X_i / X_{i-6}, X_{i-7})$ para cada i, en algunas ocasiones no se puede decir que dicha serie de residuos sea ruido blanco.

Por esto, y con el objetivo de mejorar la predicción, se propone el siguiente enfoque semiparamétrico: en cada instante t, primero se estima $E\left(X_{t+6} / X_t, X_{t-1}\right)$ con Nadaraya – Watson y la matriz histórica. En un segundo paso, se calcula la serie $\hat{Z}_{t-64},...,\hat{Z}_t$ relativa a las últimas 6 horas y se ajusta un modelo ARIMA adecuado para la misma. Por último se obtiene la predicción de Box – Jenkins de $\hat{Z}_{t+6}$ (que será cero si la correspondiente serie es ruido blanco). La predicción puntual final propuesta es:

$$\hat{E}\left(X_{t+6} / X_t, X_{t-1}\right) + \hat{Z}_{t+6} \qquad (2.1)$$

La Figura 2.3 muestra las predicciones (a media hora) utilizando el modelo semiparamétrico (línea discontinua) y la serie real observada (línea continua) durante un episodio real de alteración de calidad de aire.

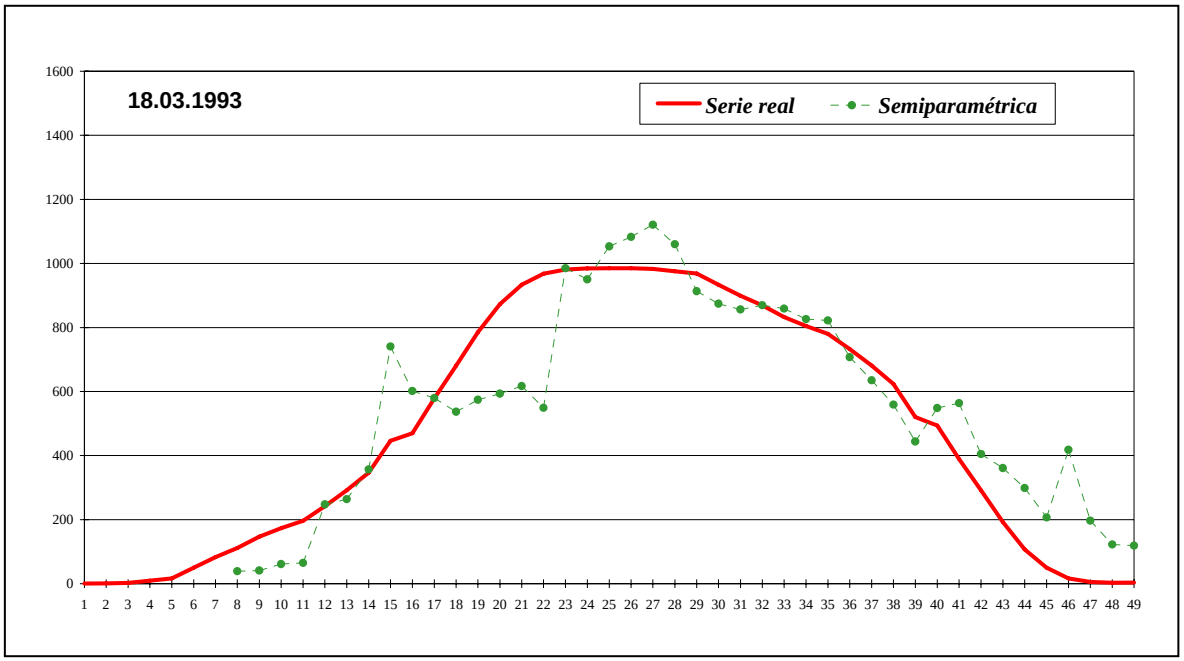

Figura 2.3: Episodio de alteración de la calidad del aire ocurrido en una de las estaciones de medida el 18 de Marzo de 1993. Predicción dada por la metodología semiparamétrica (Prada-Sánchez JM, 1997).

Comparando las Figuras 2.1, 2.2 y 2.3 se puede ver que el modelo semiparamétrico supera al enfoque Box – Jenkins puro y parece ligeramente mejor que el no paramétrico. Estas consideraciones se confirman en la Tabla 2.1. Dicha tabla contiene dos medidas de precisión para cada uno de los tres predictores. Si se denota por e_t el error de predicción observado en el instante t ($t = 1,\ldots,T;\ T = 49$ en este caso), el error cuadrado medio (MSE) viene dada por

$$MSE = \frac{1}{T}\sum_{t=1}^{T} e_t^2$$

y el error absoluto medio (MAE) por

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |e_t|$$

Ambas medidas confirman que, durante el episodio seleccionado, el modelo semiparamétrico es el que mejor comportamiento tiene, seguido del enfoque no paramétrico y por último el Box – Jenkins puro.

Modelo	MSE	MAE
Box – Jenkins puro	59614.10	179.38
No paramétrico	19861.26	103.56
Semiparamétrico	17947.96	100.56

Tabla 2.1: Errores de predicción.

En general, se puede decir que la componente no paramétrica del modelo de predicción captura la tendencia histórica de la serie, mientras que la modelización ARIMA de la serie reciente de residuos correspondiente (componente paramétrica del modelo), mejora la predicción aportando al sistema capacidad de improvisación.

Una descripción más detallada de estos modelos se puede ver en González Manteiga, *et al*. (1993), García Jurado, *et al*. (1995), y en Prada-Sánchez, *et al*. (1997).

2.2 Modelos Parcialmente Lineales

La información utilizada por los modelos semiparamétricos para obtener las predicciones, es el pasado de la propia serie; sin embargo, se podría pensar en introducir información adicional con el objetivo de mejorar dichas predicciones.

Concretamente, se han utilizado variables meteorológicas y de emisión con los modelos parcialmente lineales (Prada-Sánchez, *et al.*, 2000) para obtener predicciones de los niveles de SO_2, en media bihoraria, con horizonte de predicción de una hora.

2.2.1 Planteamiento general

Sea $\{(Z_i, Y_i)\}_{i=1}^{k}$ una muestra aleatoria de una serie de tiempo $\{(Z_l, Y_l)\}$, $l = 0, \pm 1, \pm 2, \ldots$, donde Z_l es una serie r-dimensional e Y_l una serie respuesta

unidimensional. Para algún $n = k + \kappa \ \ (\kappa > 0)$, se pretende predecir Y_n a través de Z_n, en base a la muestra dada. Por ejemplo, la predicción podría ser $\hat{Y}_n = \hat{\varphi}_k (Z_n)$, donde $\hat{\varphi}_k$ es una estimación de la función de regresión

$$\varphi(Z_n) = E(Y_n / Z_n)$$

que ha sido calculada de la muestra de tamaño k (como es bien sabido, φ es un predictor de error cuadrático medio mínimo). Una situación muy común es cuando $Z_l = (X_l, \ldots, X_{l-r+1})$ e $Y_l = X_{l+\kappa}$, donde $\{X_l\}$, $l = 0, \pm 1, \pm 2, \ldots$, es una serie unidimensional. En tal caso, el problema es predecir X_n estimando $E(X_n / X_k, \ldots, X_{k-r+1})$ en base a $\{X_{l-r+1}, \ldots, X_k\}$, una muestra de $\{X_l\}$, $l = 0, \pm 1, \pm 2, \ldots$, de tamaño $k + r - 1$.

Una posible estimación de φ es un estimador no paramétrico

$$\hat{\varphi}_k (Z_n) = \sum_{i=1}^{k} w_i^{H,k} (Z_n, (Z_1, \ldots Z_k)) Y_i, \qquad (2.2)$$

donde $\{w_i^{H,k}\}$ es un conjunto de pesos generados por un kernel y H es una matriz $r \times r$ simétrica definida positiva (matriz de ventanas o parámetro de suavizado).

Otra posibilidad es un estimador semiparamétrico

$$\tilde{\varphi}_k (Z_n) = \hat{\varphi}_k (Z_n) + \dot{e}_n, \qquad (2.3)$$

donde $\hat{\varphi}_k (Z_n)$ es el estimador no paramétrico anterior y $\dot{e}_n$ es la predicción a κ instantes del residuo $\hat{e}_n = Y_n - \hat{\varphi}_k (Z_n)$, el cual se obtiene de la serie de residuos $\hat{e}_i = Y_i - \hat{\varphi}_k (Z_i), \ \ i = 1, \ldots, k$.

Este enfoque semiparamétrico se ha explicado con detalle en la sección anterior, así como las peculiaridades de la serie de interés que hacen necesaria la utilización de las matrices históricas. El horizonte de predicción para estos modelos es de 30

minutos ($\kappa = 6$) y como se vio, los resultados son satisfactorios para el predictor $\tilde{\varphi}_k(Z_n)$, con $r = 2$ y las matrices históricas.

Si se amplia el horizonte de predicción a una hora ($\kappa = 12$), este último predictor no obtiene resultados tan satisfactorios. Por esto, y con el objetivo de ampliar dicho horizonte a una hora, se va a reforzar el modelo con una combinación lineal de variables exógenas, las cuales serán estimadas como parte del proceso de predicción.

En lugar de las series $\{(Z_l, Y_l)\}$ se consideran las series $\{(V_l, Z_l, Y_l)\}$, $l = 0, \pm 1, \pm 2, \ldots$, donde V_l es un vector q-dimensional de variables explicativas exógenas, y se asume que las series se ajustan al siguiente modelo parcialmente lineal

$$Y_l = V_l'\beta + \varphi(Z_l) + \varepsilon_l \qquad (2.4)$$

donde β es un vector q-dimensional de coeficientes desconocidos y ε_l es un término de error de media cero. Las variables exógenas, de las que se hablará más adelante, se introducen de forma lineal en el modelo por simplicidad computacional, por la interpretabilidad de β, y por la, bien conocida, "maldición de la dimensión".

Las propiedades estadísticas de los modelos definidos por (2.4) despertaron un gran interés en las últimas décadas del siglo pasado. Uno de los primeros estudios fue llevado a cabo por Engle, *et al.* (1986) en relación a un modelos de ventas de electricidad. Otros trabajos han investigado los errores cuadráticos medios en la estimación de β y φ (Speckman, 1988; Robinson, 1988 y 1995) o en la estimación de $\sigma^2 = Var(\varepsilon_l^2)$ (Liang, 1994; Gao 1995), y la root-k consistencia del estimador de β (Schich, 1996; Linton, 1995). Sin embargo, no se había estudiado el comportamiento predictivo de este tipo de modelos hasta Prada-Sánchez, *et al.* (2000). En lo que sigue se mostrará la discusión hecha por estos autores para cuatro modelos de predicción, que se definirán más adelante en las ecuaciones (2.5), (2.7), (2.9) y (2.11). Todos estos modelos son de la clase definida en (2.4), pero utilizan diferentes métodos para estimar β y φ. Además, se verá la aplicación de estos modelos al problema de predicción de contaminación de SO_2, que se está tratando a lo largo de este trabajo.

Primero se consideran las predicciones basadas en un análisis de regresión lineal en dos pasos. Este modelo se define como un caso particular de (2.4)

$$Y_l = V_l'\beta + Z_l'\gamma + \varepsilon_l,$$

donde γ es un vector r-dimensional de coeficientes. Para muestras de tamaño k se puede escribir

$$Y = V\beta + Z\gamma + \varepsilon,$$

donde Y y ε son vectores k–dimensionales y V y Z son $k\times q$ y $k\times r$ matrices cuyas filas i-ésimas son V_i' y Z_i', respectivamente. Las estimaciones de mínimos cuadrados a dos pasos de β y γ (denotadas por $\hat{\beta}_0$ y $\hat{\gamma}_0$, respectivamente) vienen dadas por

$$\hat{\beta}_0 = \left(V'\left(I - P_Z\right)V\right)^{-1} V'\left(I - P_Z\right)Y$$
$$Z\hat{\gamma}_0 = P_Z\left(Y - V\hat{\beta}_0\right),$$

donde I es la matriz identidad y $P_Z = Z\left(Z'Z\right)^{-1}Z'$ es la matriz de proyección en el espacio generado por las columnas de Z. Por consiguiente, una predicción natural de Y_n a partir de $\left(V_n, Z_n\right)$ en base a la muestra $\left\{\left(V_i, Z_i, Y_i\right)\right\}_{i=1}^{k}$ es

$$p_k^0\left(V_n, Z_n\right) = V_n'\hat{\beta}_0 + Z_n'\hat{\gamma}_0 \qquad (2.5)$$

Una generalización natural de (2.5) se obtiene reempleando P_Z por la correspondiente matriz de suavizado no paramétrica, Z_H, dada por $\left(Z_H\right)_{ij} = w_j^{H,k}\left(Z_i, \left(Z_1, \ldots, Z_k\right)\right)$. Así, la estimación de β viene dada por

$$\hat{\beta}_1 = \left(V'\left(I - Z_H\right)V\right)^{-1} V'\left(I - Z_H\right)Y \qquad (2.6)$$

y la predicción de Y_n por

$$p_k^1\left(V_n, Z_n\right) = V_n'\hat{\beta}_1 + \sum_{j=1}^{k} w_j^{H,k}\left(Z_n, \left(Z_1, \ldots, Z_k\right)\right)\left(Y_j - V_j'\hat{\beta}_1\right) \qquad (2.7)$$

Una alternativa, propuesta por Speckman (1988), surge al restar $E\{Y_l / Z_l\}$ a cada lado de (2.4)

$$Y_l - E\{Y_l / Z_l\} = \left(V_l - E\{Y_l / Z_l\}\right)' \beta + \varepsilon_l$$

Si se denota $\tilde{Y} = (I - Z_H)Y$ y $\tilde{V} = (I - Z_H)V$, entonces se trata de estimar β por

$$\hat{\beta}_2 = \left(\tilde{V}'\tilde{V}\right)^{-1} \tilde{V}'\tilde{Y} \qquad (2.8)$$

y el predictor será

$$p_k^2(V_n, Z_n) = V_n'\hat{\beta}_2 + \sum_{j=1}^{k} w_j^{H,k}\left(Z_n, (Z_1, \ldots, Z_k)\right)\left(Y_j - V_j'\hat{\beta}_2\right) \qquad (2.9)$$

Para datos independientes, las propiedades de $\hat{\beta}_1$ han sido estudiadas por Green, *et al.* (1985) y Speckman (1988), y las de $\hat{\beta}_2$ por Speckman (1988). Este último demostró, para el caso univariante, que el error cuadrático medio de $\hat{\beta}_2$ es menor que el de $\hat{\beta}_1$. Se espera un comportamiento similar para el caso multivariante.

Cuando V_l es independiente de Z_l en el modelo dado por (2.4), una tercera posibilidad es hacer una regresión de Y_l en V_l, después de haber ajustado Y a través de su valor esperado, dado Z_l. Esto conduce al siguiente estimador para β

$$\hat{\beta}_3 = \left(V'V\right)^{-1} V'\tilde{Y} \qquad (2.10)$$

y para el predictor

$$p_k^3(V_n, Z_n) = V_n'\hat{\beta}_3 + \sum_{j=1}^{k} w_j^{H,k}\left(Z_n, (Z_1, \ldots, Z_k)\right) Y_j \qquad (2.11)$$

Finalmente, cada uno de los predictores p^1, p^2 y p^3 puede ser corregido añadiendo el predictor de Box-Jenkins para el residuo correspondiente $Y_n - p_k^j(V_n, Z_n)$, $j = 1, 2, 3$, como en (2.3).

2.2.2 Aplicación al problema medioambiental

Antes de ver los resultados que proporcionan estos estimadores en el problema de predicción de los niveles de SO_2, se va a detallar la relación de variables disponible.

La Estación Meteorológica de A Mourela mide las variables que se presentan en la Tabla 2.2. Dichas variables se reciben a través de un sistema de comunicación mediante archivos cada cinco minutos, como ya se ha comentado en el capítulo anterior.

Variables meteorológicas	
T_t^0	Temperatura a nivel del suelo, en ºC
T_t^1	Temperatura a 10 m sobre el nivel del suelo, en ºC
$\Delta_{10}^{80}T_t$	Diferencia entre las temperaturas a 80 y 10 m sobre el nivel del suelo, en ºC
$\Delta_{10}^{30}T_t$	Diferencia entre las temperaturas a 30 y 10 m sobre el nivel del suelo, en ºC
V_t^1	Velocidad de viento a 10 m sobre el nivel del suelo, en m/s
V_t^8	Velocidad de viento a 80 m sobre el nivel del suelo, en m/s
D_t^1	Dirección de viento a 10 m sobre el nivel del suelo, en grados a partir del Norte
D_t^8	Dirección de viento a 80 m sobre el nivel del suelo, en grados a partir del Norte
R_t	Radiación solar, en cal/(cm^2min)
P_t	Precipitaciones, en l/m^2
H_t	Humedad relativa, en %

Tabla 2.2: Relación de las variables medidas en la Estación Meteorológica de A Mourela.

Además, el Sistema de Control de Emisiones muestrea, entre otras, las series que aparecen en la siguiente tabla.

Variables de emisión	
E_t	SO_2 emitido por la Central Térmica, en µg/m^3
W_t	Potencia generada por la Central Térmica, en MW

Tabla 2.3: Alguna de las variables medidas por el Sistema de Control de Emisiones.

Por razones de relevancia, fiabilidad, estabilidad y redundancia (por ejemplo, T^1 está fuertemente correlacionada con T^0), las variables elegidas para ser incluidas en los modelos son $\Delta_{10}^{80}T$, T^1, V^8, R, H y E. Además, con la finalidad de modelizar tanto el efecto instantáneo como los retardos de estas variables, y ya que la predicción requiere un horizonte de una hora $(\kappa = 12)$, se van a considerar los valores en tiempo l y $l-12$. De hecho, se van a utilizar dos conjuntos de variables exógenas en los diferentes modelos.

$$V_l^1 = \left(\Delta_{10}^{80}T_{l-12}, \Delta_{10}^{80}T_l, T_{l-12}^1, T_l^1, V_{l-12}^8, V_l^8, R_{l-12}, R_l, H_{l-12}, H_l, E_{l-12}, E_l\right)$$

$$V_l^2 = \left(\Delta_{10}^{80}T_l - \Delta_{10}^{80}T_{l-12}, T_l^1 - T_{l-12}^1, V_l^8 - V_{l-12}^8, R_l - R_{l-12}, H_l - H_{l-12}, E_l - E_{l-12}\right)$$

En V_l^2, las variables son tratadas a través de su porcentaje de cambio, medido en una escala de una hora.

Además, se han construido matrices históricas, cada una con mil filas, de la forma descrita en la sección anterior. Para ello, se han seleccionado datos de la forma $\left(V_l^i, Z_l, Y_l\right)$, donde V_l^i es el descrito arriba, $i = 1,2$; $Z_l = \left(X_l, X_{l-3}\right)$ e $Y_l = X_{l+12}$, siendo $\{X_t\}$ la serie de medias bihorarias de valores de SO_2. En el modelo semiparamétrico, Z_l tenía dos componentes adyacentes (X_{l-1} y X_l) y el predictor era capaz de distinguir si un episodio de contaminación estaba en fase de ascenso o por el contrario, en fase de descenso. Sin embargo, ahora la predicción es a una hora, no a 30 minutos, por lo que no resulta adecuado que esas dos componentes sean consecutivas; la experiencia ha demostrado que el intervalo óptimo, en este caso, es de 15 minutos (3 unidades de tiempo).

Estas matrices históricas se utilizan en la elaboración de los cinco estimadores: uno de los cuales no tendrá en cuenta las variables exógenas (ecuación (2.2)), y los otros cuatro serán construidos a través de las ecuaciones (2.5), (2.7), (2.9) y (2.11). Los pesos utilizados para la estimación de la parte no paramétrica de los predictores son de la forma

$$w_j^{H,k}\left(\left(X_n, X_{n-3}\right),\left(M_n\right)\right)=\frac{K\left(H^{-1/2}\left(X_n - X_j, X_{n-3} - X_{j-3}\right)\right)}{\sum_{i\in I} K\left(H^{-1/2}\left(X_n - X_i, X_{n-3} - X_{i-3}\right)\right)}$$

donde I es el conjunto del número de filas de la matriz histórica M_n (de la cual j es una de ellas), K es el kernel Gaussiano bidimensional y $H = h^2 S$, siendo S la matriz de varianzas-covarianzas construida con $\left\{\left(X_i, X_{i-3}\right)\right\}_{i\in I}$ y h un parámetro de suavización determinado por el método de Validación Cruzada. También se ha utilizado una modificación de este método de selección, con el objetivo de reducir la carga computacional; pero en este trabajo sólo se mostrarán los resultados obtenidos para el primer método.

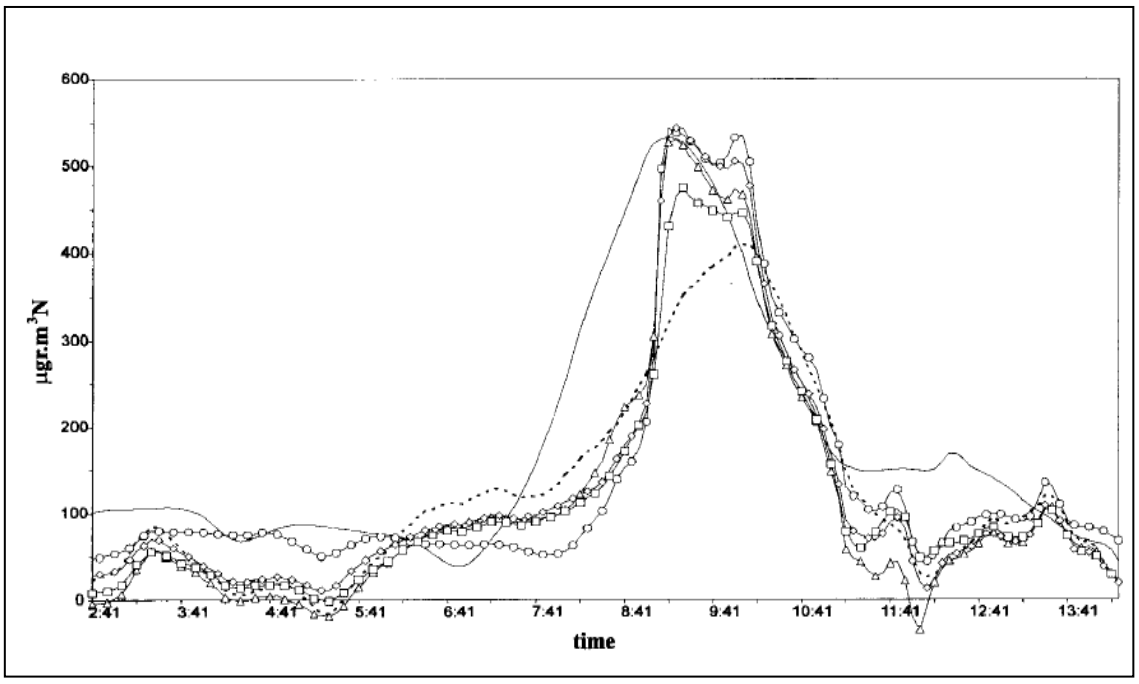

Figura 2.4: Episodio de alteración de la calidad del aire ocurrido el 23 de Agosto 1995 (-). Predicciones dadas por los predictores de ecuaciones (2.2) (círculos), (2.5) (línea discontinua), (2.7) (triángulos), (2.9) (cuadrados) y (2.11) (rombos). Todos ellos utilizan el vector de variables exógenas V^1 (Prada-Sánchez, *et. al.*, 2000).

El comportamiento de los predictores definidos, se muestra evaluando los resultados de su aplicación a la predicción de un episodio real de contaminación ocurrido el 23 de Agosto de 1995. La evolución de este episodio se puede ver en la

Figura 2.4 (curva continua) junto con las predicciones dadas por los estimadores $\hat{\varphi}_k$ y $p_k^j \ \ (j = 0,1,2,3)$ con el vector V^1 de variables exógenas. El predictor no paramétrico puro $\hat{\varphi}_k$ trabaja bien en el pico del episodio y durante la siguiente caída a cero, pero el estimador de regresión múltiple p_k^0, el cual utiliza la información adicional, es mejor al comienzo del episodio; p_k^1, p_k^2 y p_k^3 muestran un comportamiento intermedio.

Para todo el período de tiempo que aparece en la Figura 2.4 y para el período que cubre el episodio propiamente dicho (de 07:00 a 11:30 horas) la Tabla 2.4 compara la precisión de estos predictores en términos del error cuadrático y del error absoluto.

		Período completo		**07:00 a 11:30 horas**	
		SE	***AE***	***SE***	***AE***
$\hat{\varphi}_k(Z_n)$	Media	7408.4	52.11	18933.6	96.19
	DT	18188.1	68.51	27566.3	98.39
$p_k^0(V_n, Z_n)$	Media	7660.8	68.41	15037.1	97.13
	DT	11981.8	54.59	17643.2	74.86
$p_k^1(V_n, Z_n)$	Media	7968.0	70.61	10815.0	75.19
	DT	11392.1	54.60	16068.2	71.84
$p_k^2(V_n, Z_n)$	Media	7937.7	67.40	14478.0	86.98
	DT	15123.1	58.27	22423.9	83.14
$p_k^3(V_n, Z_n)$	Media	7269.9	63.02	11673.8	74.31
	DT	12489.8	57.43	18025.2	78.73

Tabla 2.4: Errores de predicción para los cinco estimadores y el conjunto de variables exógenas V^1. (Prada-Sánchez, *et. al*., 2000).

Si se compara la Figura 2.5 con la 2.4 se observa que para p_k^1, p_k^2 y p_k^3, el vector de variables exógenas V^2 proporciona mejores predicciones que V^1. La Tabla 2.5, la cual evalúa estas predicciones en términos de los mismos estadísticos que la Tabla 2.4, muestra que cuando se utiliza V^2, p_k^1 y p_k^3 son estimadores más precisos durante el episodio propiamente dicho, pero que p_k^2 es mejor sobre el período completo que aparece en las figuras.

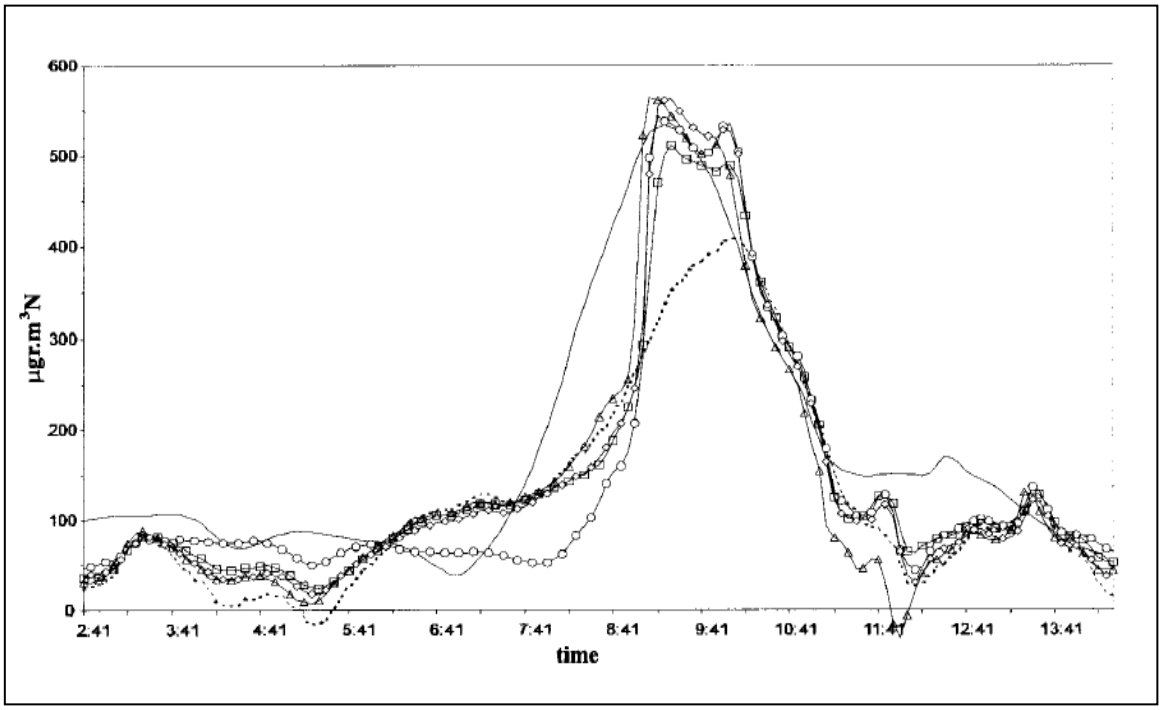

Figura 2.5: Episodio de alteración de la calidad del aire ocurrido el 23 de Agosto 1993. Predicciones dadas por los predictores de ecuaciones (2.2), (2.5), (2.7), (2.9) y (2.11). Todos ellos utilizan el vector de variables exógenas V^2 (Prada-Sánchez, *et. al*., 2000).

		Período completo		**07:00 a 11:30 horas**	
		SE	***AE***	***SE***	***AE***
$\hat{\varphi}_k(Z_n)$	Media	7408.4	52.11	18933.6	96.19
	DT	18188.1	68.51	27566.3	98.39
$p_k^0(V_n, Z_n)$	Media	7660.8	68.41	15037.1	97.13
	DT	11981.8	54.59	17643.2	74.86
$p_k^1(V_n, Z_n)$	Media	6073.7	58.48	9210.1	69.33
	DT	10450.7	51.51	14512.3	66.36
$p_k^2(V_n, Z_n)$	Media	5600.8	52.99	11210.8	75.17
	DT	12564.1	52.85	18622.1	74.57
$p_k^3(V_n, Z_n)$	Media	6137.5	56.25	10644.2	73.37
	DT	11786.3	54.52	16942.2	72.54

Tabla 2.5: Errores de predicción para los estadísticos de ecuación (2.7), (2.9) y (2.11) con el conjunto de variables exógenas V^2; para los estadísticos de ecuaciones (2.2) y (2.5) con el conjunto de variables exógenas V^1 (Prada-Sánchez, *et. al*., 2000).

2.3 Modelos de Redes Neuronales

El cambio en la serie de interés que establece la Directiva Europea 1999/CE/30, de medias bihorarias a medias horarias, provoca que la serie objetivo sea menos suave. En un principio se adaptó el modelo semiparamétrico diseñado para trabajar sobre la nueva serie de medias horarias. En los resultados se observó un considerable aumento de la variabilidad de las predicciones dadas, respecto a los resultados habitualmente obtenidos para la serie de medias bihorarias.

La excesiva variabilidad de las predicciones es una característica poco deseable a la hora de incluir las mismas en un sistema de apoyo a la toma de decisiones en general, y en particular, en el diseñado para la Central Térmica de As Pontes. La variabilidad de las predicciones hace saltar intermitentemente las alertas del sistema. La inestabilidad de las alertas provoca desconfianza, con lo que se reduce, de forma considerable, la eficacia del sistema.

En un intento por mejorar la respuesta del Sistema de Predicción Estadística de Inmisión, y en particular, sus predicciones puntuales con horizonte de media hora, se desarrollaron modelos de redes neuronales (Fernández de Castro, *et al*., 2003).

2.3.1 Planteamiento general

Las redes neuronales son métodos utilizados para el aprendizaje de datos (Hastie, *et al*., 2001). Dos características notables de los modelos de redes neuronales son la flexibilidad y la capacidad de adaptación. Además de sus conocidas aplicaciones en clasificación y reconocimiento de patrones, las redes neuronales son utilizadas en diferentes áreas, como son la Estadística (Kay, *et al*., 1999) y la contaminación ambiental (Pérez, *et al*., 2000).

Este grupo de técnicas surgen como un mecanismo para simular los procesos del cerebro humano. Las redes neuronales artificiales están formadas por pequeños elementos: neuronas (o nodos), agrupadas en capas y conectadas entre ellas. La red debe adquirir el suficiente conocimiento para reaccionar correctamente ante un estímulo. Dicho conocimiento se obtiene trabajando con ejemplos o patrones.

Cada conexión entre los nodos tiene un peso asociado, y cada nodo de la red neuronal, una función de activación. A través de esta función, cada nodo procesa la información que obtiene a través de las conexiones, ponderada por cada conexión. El resultado de este proceso es enviado a los nodos de la siguiente capa a través de las conexiones salientes. Estos pasos se repiten en cada nodo de la red, desde la primera capa (denominada *capa de entrada*) hasta la última (denominada *capa de salida*). Los resultados obtenidos en los nodos de la capa de salida son las salidas de la red neuronal.

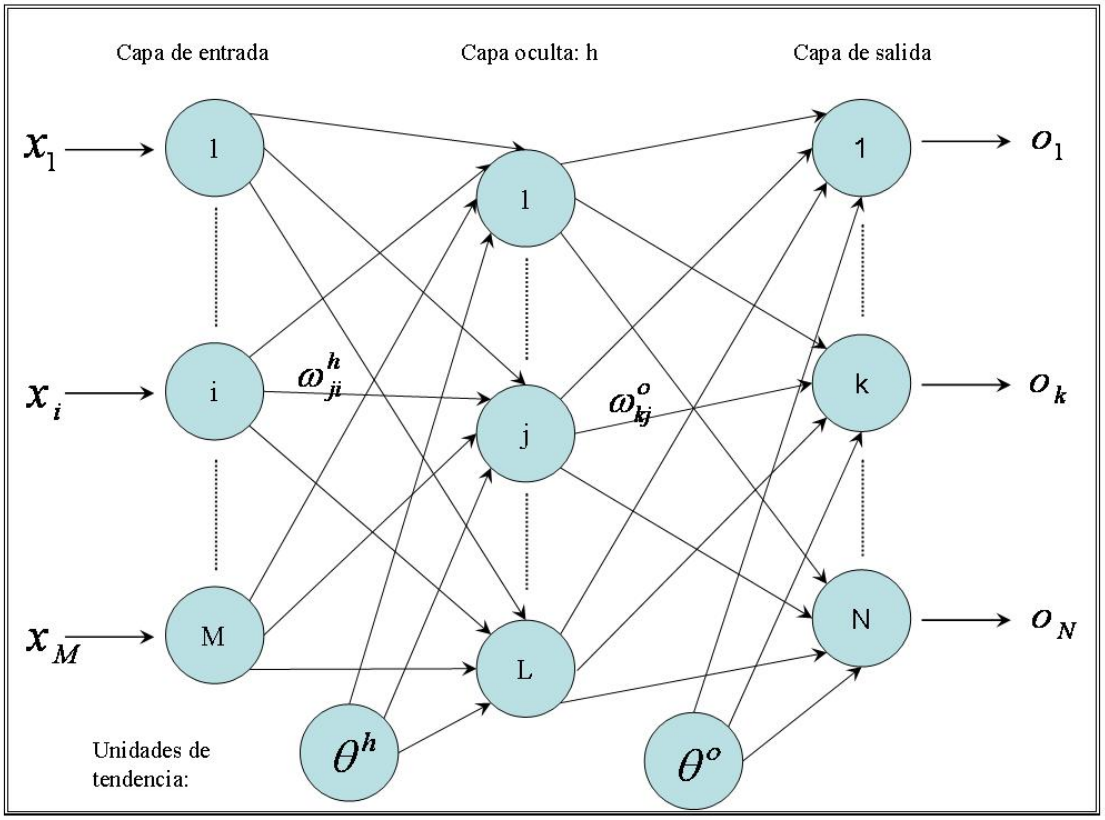

Figura 2.6: Esquema general de una red neuronal multicapa (MLP).

El modelo de red neuronal que se va a considerar, es la denominada "Backpropagation Network" o también, red neuronal perceptrón multicapa (MLP). Esta red tiene una capa de entrada con M nodos, una capa oculta con L nodos y una capa de salida con N nodos. Además, es posible introducir términos de tendencia en las capas oculta y de salida. En la Figura 2.6 aparece representada una red de estas características.

Un planteamiento más específico de este tipo de modelos es el que sigue: sea $X = (x_1, x_2, \ldots, x_M)'$ un vector de variables explicativas de entrada e $Y = (y_1, y_2, \ldots, y_N)'$ su vector de respuestas.

Hay que diseñar una topología de red neuronal cuyo vector de salidas $\mathrm{O} = (o_1, o_2, \ldots, o_N)'$ reproduzca, lo mejor posible, la respuesta real Y para cada X.

Se denota por h a los elementos de la capa oculta y por o a los de la capa de salida. Así, se define ω_{ji}^h como el peso asociado al arco que llega al nodo j de la capa h desde el nodo i de la capa anterior; ω_{kj}^o como el peso asociado al arco que llega al nodo k de la capa de salida desde el nodo j de la capa previa; f_j^h como la función de activación del nodo j de la capa h; f_k^o como la función de activación del nodo k de la capa de salida; θ_j^h como el término de tendencia del nodo j de la capa h (este término es opcional); θ_k^o como el término de tendencia del nodo k de la capa de salida (también es opcional); o_j^h como la salida del nodo j de la capa h y o_k como la salida del nodo k de la capa de salida.

De esta forma, la salida de cada nodo de la capa o puede escribirse

$$o_k = f_k^o\left(\theta_k^o + \sum_{j=1}^{L} \omega_{kj}^o f_j^h\left(\theta_j^h + \sum_{i=1}^{M} \omega_{ji}^h x_i\right)\right) \quad k = 1, \ldots, N$$

Una vez que se ha adaptado la red neuronal al problema concreto para el que se quiere utilizar (elección del número de capas, nodos en cada capa, y funciones de activación), se somete a un proceso de aprendizaje en el que se fijarán todos los pesos de la red.

Este proceso de entrenamiento consiste en proporcionar a la red un conjunto de vectores de entrada con sus respuestas conocidas: $(X, Y)' = (x_1, x_2, \ldots, x_M, y_1, y_2, \ldots, y_N)'$, denominado conjunto de entrenamiento; comparar las salidas obtenidas con las respuestas reales; y, modificar los pesos de la red, según el error cometido. Este es un proceso largo, ya que ha de repetirse hasta

que el error cometido sea admisible para todos los elementos de la muestra de entrenamiento.

Durante el proceso de entrenamiento, la red aprenderá los patrones incluidos en el conjunto de entrenamiento. Por tanto, es recomendable que esta muestra sea representativa de toda la población.

2.3.2 Aplicación al problema medioambiental

La red neuronal que se ha diseñado para poder dar predicciones, con media hora de antelación, de los valores de inmisión de SO_2 en media horaria, consta de una capa de entrada, una única capa oculta y una capa de salida. El número de nodos de la capa de salida viene determinado por la dimensión de la respuesta que se quiere obtener de la red; en este caso interesa una predicción puntual para x_{t+6}.

La experiencia adquirida en la utilización de los anteriores modelos estadísticos, diseñados para series de tiempo de calidad de aire, indica que la información más relevante para este tipo de predicciones se encuentra en el pasado de la propia serie de tiempo de valores medios horarios de SO_2.

Por esta razón, se ha tomado como entrada de la red neuronal el vector bidimensional $(x_{t-3}, x_t)'$, que representa, para cada instante de tiempo t, el nivel de SO_2 hace 15 minutos y el actual, en media horaria. La red neuronal resultante tiene, por tanto, $M = 2$ nodos en la capa de entrada. La topología de la red diseñada es la que se representa en la Figura 2.7.

En los nodos de la capa oculta se ha tomado como función de activación la función logística, y en la capa de salida la función identidad.

Se ha construido, por tanto, una red neuronal que predice, en cada instante t, el nivel de SO_2 en media horaria, media hora después (instante $t+6$), utilizando la información presente (t) y de hace 15 minutos ($t-3$). El predictor dado por la red neuronal tiene la siguiente expresión:

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

con $f_j^{h}(z) = \dfrac{1}{1+e^{-z}}$ para cada nodo j de la capa oculta.

El conjunto de pesos $\left\{ \omega_{j1}^{h}, \omega_{j2}^{h}, \omega_{1j}^{o};\ \ j = 1,...,L \right\}$ y las tendencias $\left\{ \theta_j^{h};\ \ j = 1,...,L \right\}$ serán determinados durante el proceso de entrenamiento.

El número final L de nodos de la capa oculta, se elegirá también durante el entrenamiento como el valor cuya red neuronal proporcione mejores resultados, tras haber entrenado redes con idéntica arquitectura y distintos valores de L.

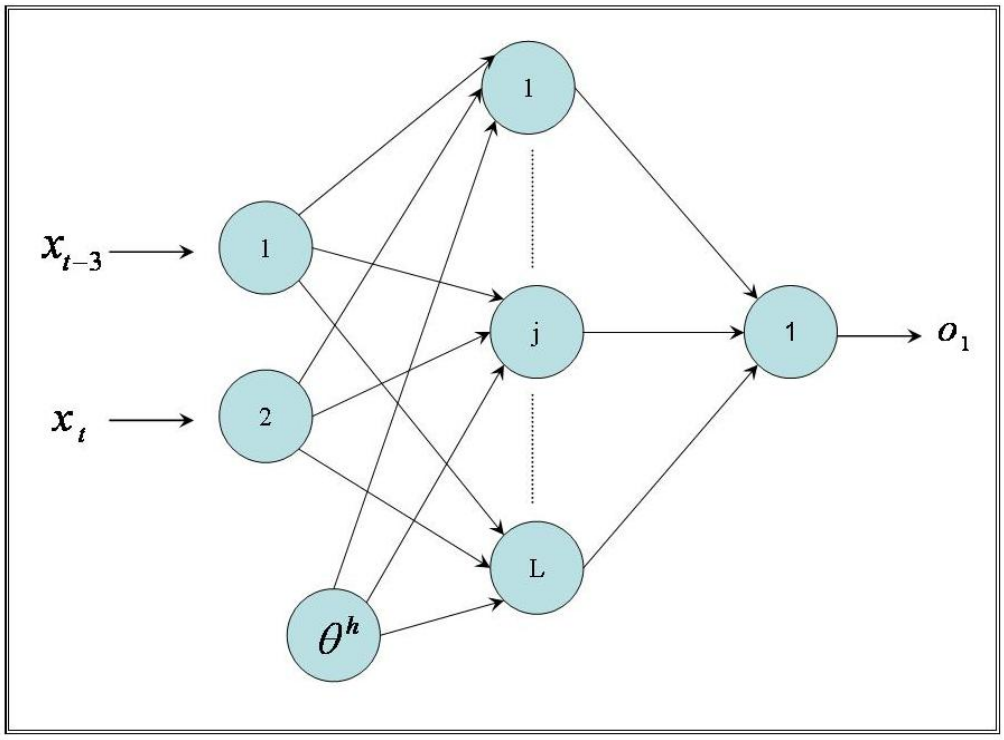

Figura 2.7: Esquema de una red neuronal diseñada para predecir el nivel de inmisión de SO_2 en media horaria, con media hora de antelación.

Para diseñar el conjunto de entrenamiento de la red neuronal se han considerado las matrices históricas, introducidas en la sección anterior, convenientemente adaptadas. El conjunto de entrenamiento para la red diseñada está construido con vectores de la forma $\left(x_{t-3}, x_t, x_{t+6} \right)'$ constituidos por datos reales de medias horarias de

SO_2. Concretamente se han diseñado matrices de 2000 registros, divididas en 10 estratos. Cada estrato tiene un rango de valores de x_{t+6}, de manera que cada nuevo vector será incluido en el estrato correspondiente a su x_{t+6}, reemplazando al vector más antiguo de dicho estrato, como ya se ha explicado en anteriores secciones.

Las ternas seleccionadas para rellenar la matriz histórica provienen de datos reales correspondientes al año 1999. No siempre es posible llenar por completo en todas las estaciones los 2000 registros de la matriz, ya que no se puede garantizar la existencia de tantos vectores en todos los estratos. Los registros vacíos serán eliminados para entrenar la red neuronal.

Una vez construido el conjunto de entrenamiento, se han entrenado redes neuronales con la arquitectura reflejada en la Figura 2.7 variando el número, L, de nodos de la capa oculta. Se han probado distintos valores para L entre 40 y 60.

Se han entrenado todas las redes neuronales utilizando el algoritmo backpropagation y se han comparado los errores cometidos con cada una. Finalmente se han seleccionado los pesos correspondientes a la red con menor error.

Para determinar el buen funcionamiento de la red neuronal elegida, se ha evaluado su funcionamiento sobre un episodio de alteración de la calidad de aire, cuya información no había sido incluida en la matriz histórica. Así, es posible tener una evaluación del funcionamiento real de la red neuronal, al poder comprobar cómo aplica a situaciones reales lo aprendido durante el proceso de entrenamiento.

Esta comprobación es esencial, ya que las redes se diseñan como herramienta de predicción del Sistema de Control Suplementario de la Contaminación Atmosférica de la Central Térmica de As Pontes. Dentro de este sistema, las redes neuronales trabajan en continuo, recibiendo datos de la Red de Control de Calidad de Aire, y devolviendo al sistema una predicción para cada una de las 17 estaciones de la red, cada 5 minutos. De esta manera se dispone de predicciones que se actualizan con la misma frecuencia que los datos reales medidos en la Red de Inmisión. Para que la predicción no sufra retrasos se incorporan al sistema ya entrenadas.

La Figura 2.8 muestra las predicciones dadas, con media hora de antelación, por la red neuronal con 50 nodos en su capa oculta, para un episodio de alteración de la

calidad de aire ocurrido el 20 de mayo de 2000 en una de las estaciones de medida. Se puede observar el buen comportamiento de las predicciones (línea azul), persiguiendo los valores reales de la media horaria de concentraciones de SO_2 (línea roja).

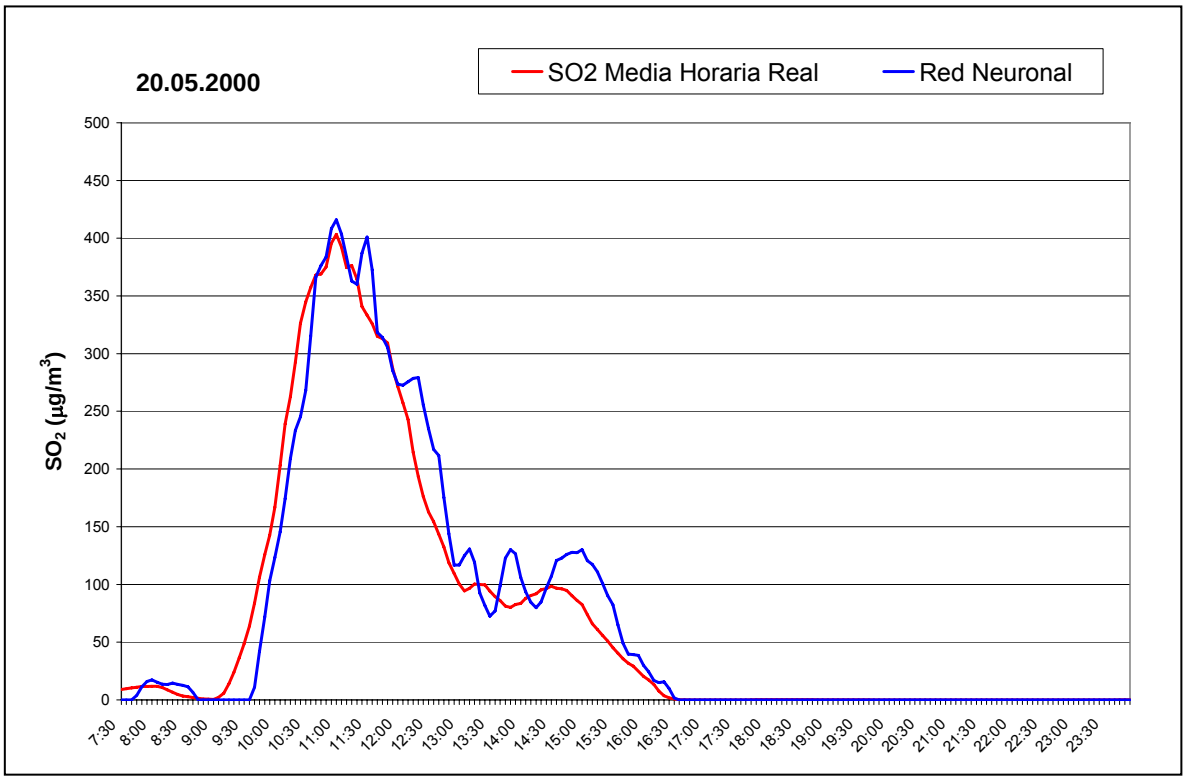

Figura 2.8: Episodio de alteración de la calidad del aire ocurrido el 20 de Mayo de 2000. Predicción dada por red neuronal (Fernández de Castro, et al., 2003).

En los últimos años se ha desarrollado un nuevo Sistema de Seguimiento y Control de la Calidad de Aire Atmosférica para la sección de Medio Ambiente de la U. P. T. de As Pontes. Este sistema, denominado MEDAS, se ha puesto en marcha durante el año 2002, y es ahora el encargado de enviar los datos reales de inmisión, emisión y meteorología, al SIPEI.

La frecuencia de envío de datos se ha mantenido como la anterior: se crea un nuevo fichero cada 5 minutos, donde el dato de cada variable es la media de todos los recogidos en esa variable en los últimos 5 minutos. Además, se ha incorporado la creación de un nuevo fichero con una frecuencia de un minuto. En este fichero cada dato representa la media de los datos recogidos para esa variable en el último minuto.

Por esto, se han adaptado los modelos de redes neuronales, inicialmente construidos para datos pentaminutales, a los datos con frecuencia minutal a los que ahora se tiene acceso.

Al trabajar con datos minutales, la serie de interés será $\{x_t,\ t=0,1,2,...\}$ donde, cada t, corresponde a un instante minutal y cada valor x_t se obtiene como el promedio de las concentraciones de SO_2, medidas en µg/m^3, correspondientes a la última hora. Esto es, si $SO_2(t)$ representa la concentración de SO_2 en el instante t medida en µg/m^3:

$$x_t = \frac{1}{60}\sum_{i=0}^{59} SO_2(t-i)$$

Se ha tomado como entrada de la red el vector bidimensional $(x_{t-15}, x_t)'$ que representa, exactamente, el nivel medio horario de SO_2 en el instante actual t (ahora instante minutal) y en el $t-15$, es decir, 15 minutos antes. Además, se han construido matrices históricas apropiadas a partir de los datos minutales correspondientes a 2003.

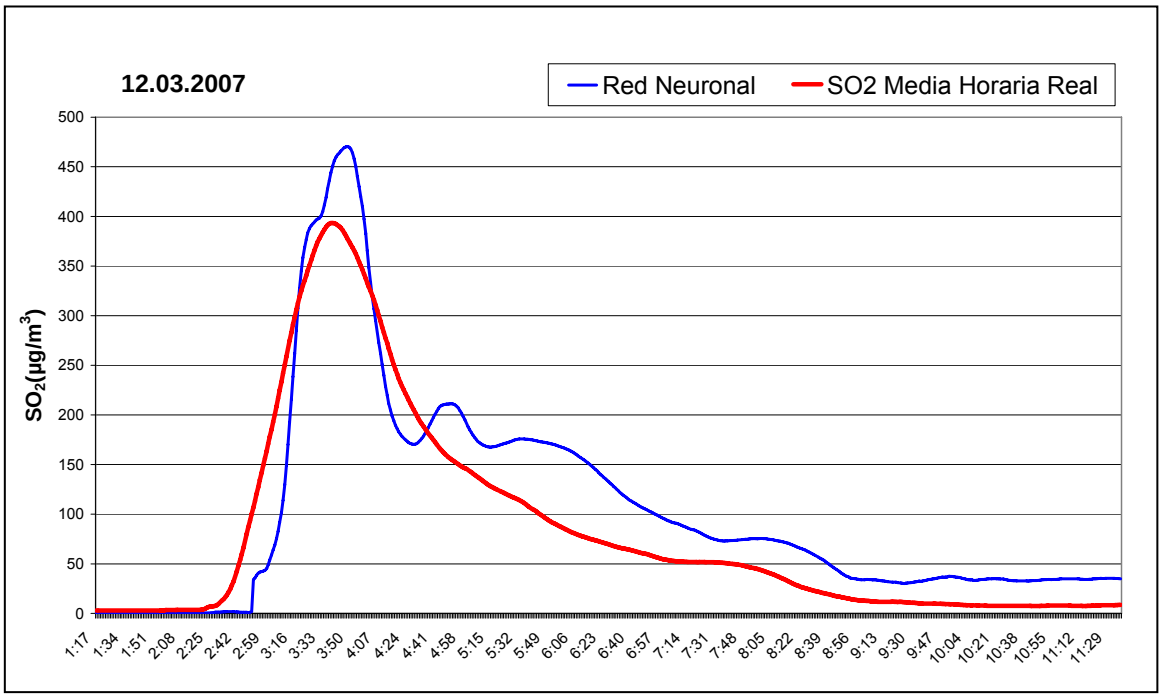

Figura 2.9: Episodio de alteración de la calidad del aire ocurrido en el 12 de Marzo de 2003. Predicción dada por red neuronal

La Figura 2.9 muestra las predicciones dadas, con media hora de antelación, por la red neuronal con 50 nodos en su capa oculta, para un episodio de alteración de la calidad de aire ocurrido en una de las estaciones el 12 de marzo de 2007.

2.3.3 Comparación con otros modelos

Es posible comparar la predicción dada por la red neuronal, descrita con anterioridad, con la predicción obtenida mediante el modelo semiparamétrico. Para ello, se ha adaptado el modelo descrito en la sección 2.1, a la predicción de la serie de medias horarias de SO_2.

Para hacer la comparación de manera apropiada se ha construido una matriz histórica adecuada para este modelo semiparamétrico a partir de los mismos datos correspondientes a 1999.

La Figura 2.10 muestra las predicciones dadas por el modelo semiparamétrico (línea punteada verde) así como las de la red neuronal (línea azul) para el episodio ocurrido el 20 de mayo de 2000. Como se puede observar, la predicción dada por el modelo semiparamétrico es más inestable que la predicción de la red neuronal.

Para evaluar los errores cometidos por las predicciones de cada uno de los modelos considerados, se han utilizado dos medidas:

- Error cuadrático: $SE = \left(x_{t+6} - \hat{x}_{t+6} \right)^2$
- Error absoluto: $AE = \left| x_{t+6} - \hat{x}_{t+6} \right|$

Se han calculado la media (M), mediana (Md) y desviación típica (DT) de las dos medidas consideradas. Se ha evaluado también una medida relativa: el error cuadrático medio normalizado (NMSE), que es el cociente entre el error cuadrático medio (MSE) y la varianza de los datos observados. Los resultados se muestran en la Tabla 2.6, para las medidas de error absolutas, y en la Tabla 2.7, para la medida de error relativa.

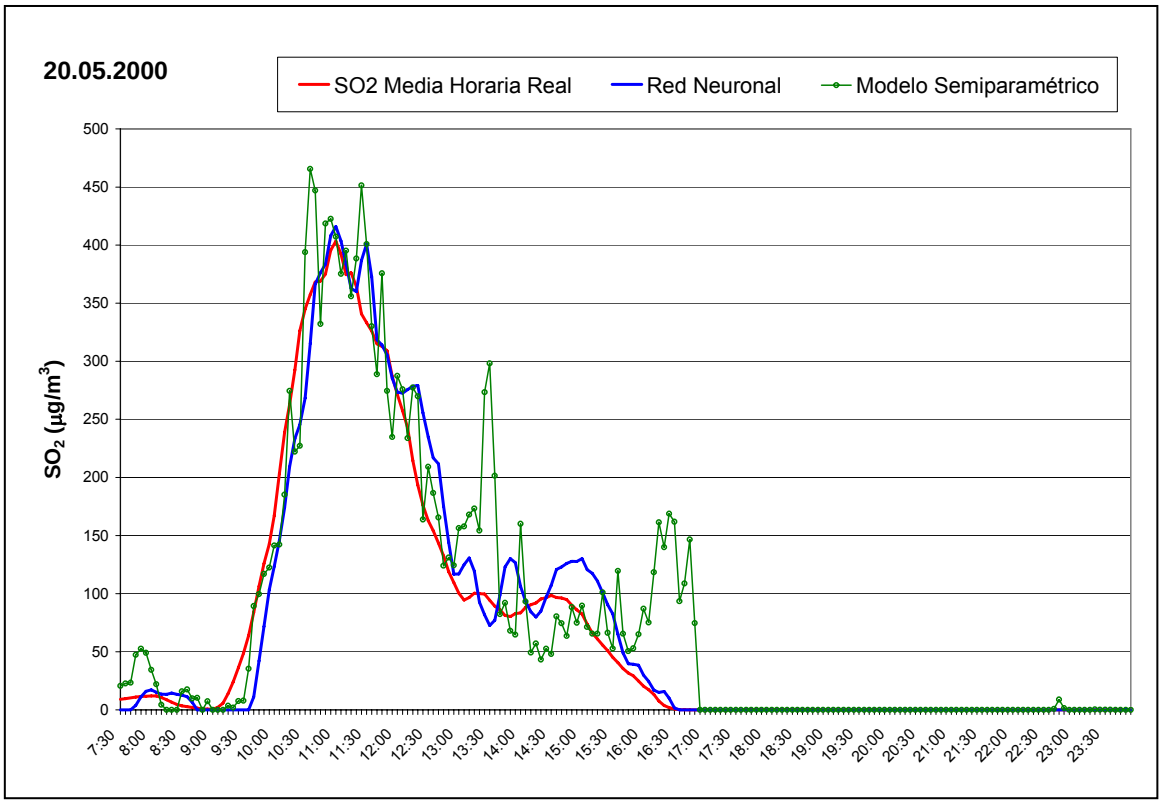

Figura 2.7: Episodio de alteración de la calidad del aire ocurrido el 20 de Mayo de 2000. Predicciones dadas por red neuronal y modelos semiparamétrico (Fernández de Castro, et al., 2003).

Las tablas de errores confirman los resultados de la Figura 2.7: los errores cuadrático y absoluto cometidos por la red neuronal son mejores en media, mediana y desviación típica, que los cometidos por el modelo semiparamétrico. Más aún, los errores cometidos por la red neuronal tienen menor variabilidad que los del modelo semiparamétrico, tal y como indican las desviaciones típicas evaluadas. Una alta variabilidad en las predicciones no es deseable si se pretenden utilizar como herramienta de ayuda en la operación de la instalación, ya que provocaría diversos altibajos en las alertas del sistema, generando desconfianza.

Modelo	**Red Neuronal**		**Modelo Semiparamétrico**	
Error	**SE**	**AE**	**SE**	**AE**
M	1525.18	31.09	2736.36	37.66
Md	628.50	25.07	689.06	26.25
DT	1827.57	23.78	5935.60	36.52

Tabla 2.6: Errores de predicción en el episodio de alteración de la calidad de aire ocurrido el 20 de mayo de 2000 de 9:00 a 16:00 horas. Medidas absolutas. (Fernández de Castro, et al., 2003)

Modelo	Red Neuronal	Modelo Semiparamétrico
ECMN	0,0536	0.1626

Tabla 2.7: Errores de predicción en el episodio de alteración de la calidad de aire ocurrido el 20 de mayo de 2000 de 9:00 a 16:00 horas. Medida relativa. (Fernández de Castro, et al., 2003)

2.4 Modelos Funcionales

Las técnicas de análisis de datos funcionales se han hecho cada vez más populares en la comunidad estadística debido, principalmente, a la habilidad que presentan a la hora de comprender la evolución global de un proceso estocástico.

Podemos encontrar en la literatura aplicaciones al tráfico (Besse, *et al.*, 1996) y al conocido fenómeno climático El Niño (Besse, *et al.*, 2000), en los que se comparan los resultados obtenidos mediante modelos autorregresivos de Hilbert y modelos autorregresivos estacionales (SARIMA). Las técnicas funcionales obtienen mejores resultados en las predicciones de las series estudiadas. Damon, *et al.*, (2002) presentan una aplicación de modelos funcionales a la predicción de valores máximos diarios de ozono.

Se van a aplicar los modelos de núcleo funcional y los modelos autorregresivos de Hilbert de orden 1, al problema de predicción que se ha tratado hasta ahora (Fernández de Castro, et al., 2005).

2.4.1 Planteamiento general

Dado un proceso estocástico en tiempo continuo $x(u)$, $u \in \mathbb{R}$, se trata de predecir valores futuros del proceso $\left(x(u), u \geq T\right)$ a partir de la información contenida en las infinitas variables del pasado $\left(x(u), u \leq T\right)$, considerando proporciones del proceso estocástico en tiempo continuo como curvas.

Sea H el espacio de Hilbert definido como $H = L^2\left([0,\delta]\right)$. Se consideran variables aleatorias X_n evaluadas en H, construidas de la siguiente forma:

$$X_n(u) = x(\delta n + u), \quad 0 \le u \le \delta, \quad n \in \mathbb{R}$$

así $X = \left(X_n, n \in \mathbb{R}\right)$ es un proceso infinito-dimensional en tiempo discreto.

Se define un *ruido blanco fuerte de Hilbert* (Bosq, 2000) como una sucesión $\left(\varepsilon_n, n \in \mathbb{R}\right)$ de variables aleatorias independientes e idénticamente distribuidas (iid) evaluadas en H, verificando:

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

Se considera el análisis del siguiente modelo estadístico:

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

en donde $\rho : H \to H$ es el operador a estimar.

Bosq (2000) proporciona un completo estudio teórico sobre procesos lineales con valores en espacios de funciones.

A partir de este momento el estudio se va a restringir a modelos autorregresivos de orden 1, aunque es posible encontrar análisis para órdenes mayores.

El modelo Autorregresivo de Hilbert de orden 1 (ARH(1)) asume la estimación de ρ en el espacio de operadores lineales acotados en un espacio de Hilbert separable, como por ejemplo el considerado $H = L^2\left([0,\delta]\right)$.

Dada una sucesión de variables aleatorias evaluadas en H $\left(X_n, n \in \mathbb{Z}\right)$ con media $\mu = 0$, se define el *operador covarianza* C como:

$$C(x) = E\left[< X_0, x > X_0\right], \quad x \in H$$

y el operador *covarianza cruzada* de orden 1 como:

$$D(x) = C_{X_0, X_1}(x) = E\left[< X_0, x > X_1\right], \quad x \in H$$

Es posible demostrar que, bajo ciertas condiciones (Bosq, 2000), un proceso ARH(1) de media $\mu = 0$ satisface la siguiente relación:

$$D = \rho C$$

La estimación de ρ se lleva a cabo en varios pasos, teniendo en cuenta la relación anterior y las versiones empíricas de los operadores C y D:

$$C_n(x) = \frac{1}{n}\sum_{i=1}^{n} < X_i, x > X_i, \quad x \in H$$

$$D_n(x) = \frac{1}{n-1}\sum_{i=1}^{n-1} < X_i, x > X_{i+1}, \quad x \in H$$

respectivamente, para una muestra de datos funcionales $X_1, ..., X_n$.

La matriz C_n no es invertible en general, por lo que para obtener la estimación de ρ será necesario hacer una proyección sobre un espacio de dimensión finita. Para ello se elige el espacio generado por los k_n autovectores asociados a los k_n mayores autovalores de C_n.

Así, los pasos a seguir en la estimación de ρ son los siguientes:

Paso 0: Eliminar la media del proceso.

Paso 1: Mediante análisis de componentes principales, calcular estimadores empíricos de los autovalores y autovectores del operador de covarianzas empírico C_n.

Paso 2: Proyectar la relación $D = \rho C$ en el subespacio generado por los k_n autovectores asociados a los k_n mayores autovalores de C_n.

Paso 3: Calcular un estimador $\hat{\rho}_{k_n}$ consistente de ρ utilizando la relación proyectada.

En algunos casos puede resultar restrictivo considerar únicamente operadores lineales, para el análisis de muestras de curvas dependientes. Besse, *et al.* (2000) proponen una extensión del estimador clásico de Nadaraya Watson tipo núcleo de regresión al contexto funcional.

Siguiendo este planteamiento, el operador ρ definido como la esperanza condicionada:

$$\rho(x) = E\left[X_i / X_{i-1} = x\right], \quad x \in H$$

se puede estimar con el estimador funcional tipo núcleo:

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

donde $x \in H$, K denota una función núcleo, n es el tamaño muestral y h_n es la ventana o parámetro de suavización.

Como función núcleo se ha considerado el núcleo Gaussiano usual:

$$K(x) = \left(\sqrt{2\pi}\right)^{-1} e^{-\frac{x^2}{2}}$$

Para la selección de la ventana se han considerado criterios globales y locales, utilizando técnicas de validación cruzada sobre un subconjunto de datos funcionales de la muestra original.

2.4.2 Aplicación al problema medioambiental

Se van a tratar los datos en media horaria de los niveles de SO_2 como observaciones de un proceso estocástico en tiempo continuo que modeliza los niveles de SO_2. El interés es, como ya se ha comentado en repetidas ocasiones, predecir con un horizonte de media hora, por lo que cada una de las curvas consideradas representa, precisamente, media hora, y se obtiene al considerar seis observaciones pentaminutales consecutivas, como puntos de muestreo de cada dato funcional.

Algunos autores utilizan técnicas de interpolación o splines para obtener curvas a partir de los datos observados. Se pueden encontrar ejemplos de estas aplicaciones en Besse, P., *et al.* (1996) y en Besse, P., *et al.* (2000). En este caso, los autores no han considerado la aplicación de estas técnicas para evitar perder información relevante en el proceso de suavización. Precisamente el mayor interés es predecir las rápidas y bruscas subidas en la serie analizada; suavizar los datos iniciales podría ocultar una valiosa información.

Se considera el análisis de variables aleatorias en el espacio de Hilbert $H = L^2\left([0,6]\right)$. Las variables aleatorias se expresan como:

$$X_n(u) = x(6n+u), u \in [0,6], \quad n \in \mathbb{Z}$$

Se trata de predecir con un horizonte de media hora, por ello nuestro análisis se ha restringido al caso de modelos de orden $p=1$:

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

Se ha considerado para la estimación de la metodología ARH(1) y el núcleo funcional.

Para evaluar los distintos procedimientos planteados se utilizan los errores-L^p empíricos, para los enteros $p=1,2$:

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

para una muestra de curvas de media hora, o seis datos pentaminutales, de tamaño n.

Además, se considera el error L^∞ evaluado como sigue:

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

Por otro lado, se ha adaptado el concepto de matriz histórica al caso en que los datos son curvas, en vez de valores reales. La muestra original de datos considerada, contiene todos los datos disponibles correspondientes al año 2001. Los modelos

funcionales considerados tratan de predecir la curva X_{n+1} a partir de X_n, por lo que se han diseñado matrices de vectores funcionales de la forma $\left(X_n, X_{n+1}\right)$, en donde ahora cada dato X_n es una curva, representada por seis medias consecutivas de SO_2, en media horaria.

La matriz histórica contiene 2000 registros que, como en el caso de datos reales, se han dividido en estratos. Para clasificar cada dato funcional dentro de la matriz histórica funcional se han considerado dos alternativas:

- *Clasificación "clásica":* una primera aproximación que replica la clasificación de la matriz histórica real. La matriz histórica funcional se divide en 10 clases. Cada clase tiene asociado un rango de valores reales de SO_2. Cada vector funcional $\left(X_n, X_{n+1}\right)$ se incorpora a la clase a la que corresponde el último valor real de la respuesta funcional X_{n+1}, reemplazando al vector funcional más antiguo de cada clase.

- *Clasificación "funcional":* un segundo criterio atendiendo a características funcionales de los datos. La matriz histórica funcional se divide en cinco clases. Cada clase tiene asociado un tipo de curva. Distinguimos cuatro tipos de curvas atendiendo a su forma: curva creciente, curva decreciente, curva plana y curva cambiante. Construimos una última clase en la matriz histórica funcional, en la que se introducen todas las curvas que no pertenecen estrictamente a ninguno de los tipos anteriores: curvas libres. Para determinar la clase a la que pertenece un dato funcional, $X_n = \left(X_n^1, \ldots, X_n^6\right)$, evaluamos las diferencias entre sus componentes $\left(X_n^2 - X_n^1, \ldots, X_n^6 - X_n^5\right)$. Cuando el valor absoluto de una diferencia es estrictamente menor que 5, se sustituye por un 0 para evitar pequeñas inestabilidades producidas realmente por las técnicas de medida. Cuando una diferencia es mayor que 5, se sustituye por un signo "+". Cuando una diferencia es menor que -5 se sustituye por un signo "-". Con estas indicaciones, se asocia a cada dato funcional un vector de signos +, - y ceros. Las cinco clases descritas se definen como sigue: cinco signos "+" serán una curva creciente, cinco signos "-" serán una curva decreciente, cinco 0 una curva plana, una curva cambiante será la que tenga asociado un vector de

diferencias con al menos un signo "+" y un "-", y sin ceros. La clase de curvas libres contendrá los datos funcionales con vectores de signos asociados que no verifiquen ninguna de las cuatro condiciones dadas. La categoría de curvas cambiantes resulta especialmente difícil de llenar, ya que representan una pequeña parte de los datos históricos. Así, la matriz histórica se compone de cinco clases de pares funcionales de la forma (X_n, X_{n+1}), clasificadas según la forma de la curva respuesta X_{n+1}.

En lo sucesivo, se utilizarán las denominaciones *matriz de niveles* y *matriz de formas*, para hacer referencia a las matrices históricas funcionales construidas según los criterios "clásico" y "funcional", respectivamente.

Los resultados de los estudios realizados con los datos reales de SO_2 revelan que la utilización de las matrices históricas proporciona mejores resultados en la predicción. Más adelante se compararán las predicciones obtenidas, con distintos modelos funcionales, sin matrices históricas y haciendo uso de ellas.

Se han utilizado las dos metodologías de estimación descritas: ARH(1) y núcleo funcional. Esta última, con dos versiones: ventana global y ventana local.

Además, se han utilizado en la estimación de todos los modelos las dos matrices históricas funcionales que se acaban de describir: matriz de niveles y matriz de formas, construidas a partir de datos reales correspondientes a 2001.

Se presentan a continuación los resultados obtenidos en las predicciones sobre el episodio de alteración de la calidad de aire, ocurrido en una de las estaciones de medida, el día 22 de abril de 2002. Las distintas predicciones se muestran en las Figuras 2.11, 2.12 y 2.13 junto con los datos reales del episodio.

Se puede observar que ante bajos niveles de concentración de SO_2 la respuesta de los modelos es una subida, que es realmente una respuesta incorrecta. Este error, sin embargo, no es excesivamente grave, ya que el nivel alcanzado por esas subidas no se puede considerar crítico al no alcanzar los márgenes de alerta del Sistema de Control Suplementario de la Contaminación Atmosférica.

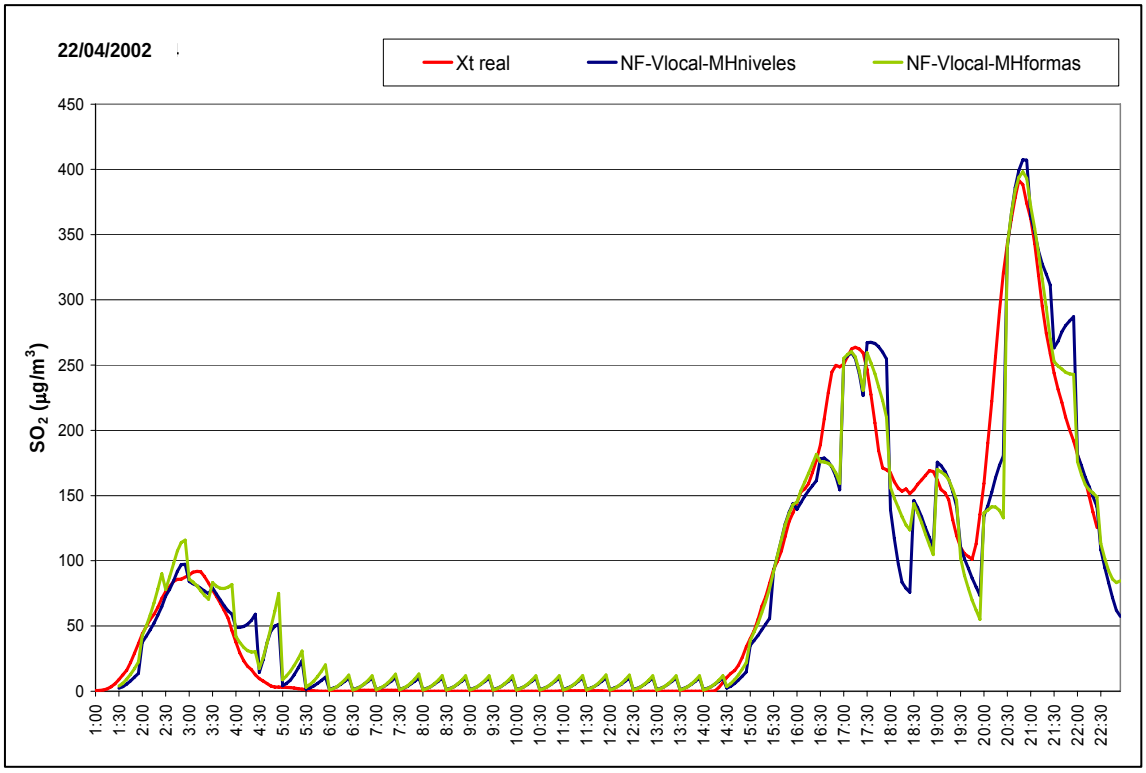

Figura 2.11: Episodio correspondiente al día 22 de abril de 2002. Predicciones de Núcleo Funcional con ventana local, utilizando matriz histórica de niveles y de formas (Fernández de Castro, et al., 2005).

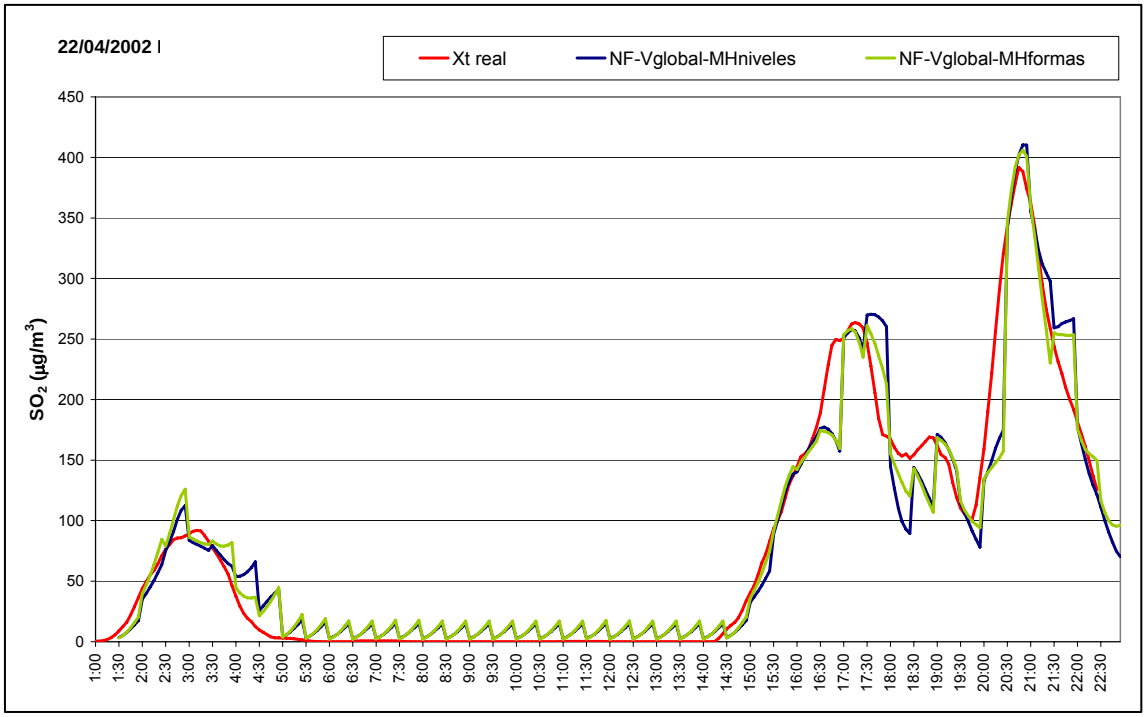

Figura 2.12: Episodio correspondiente al día 22 de abril de 2002. Predicciones de Núcleo Funcional con ventana global, utilizando matriz histórica de niveles y de formas (Fernández de Castro, et al., 2005).

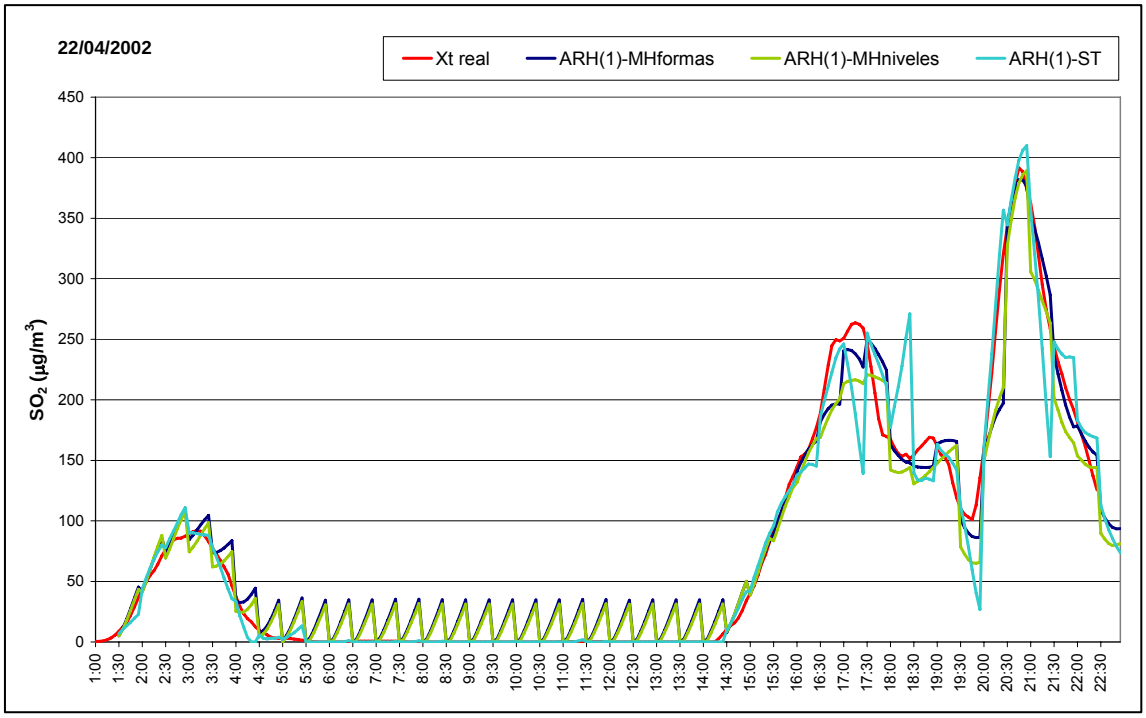

Figura 2.13: Episodio correspondiente al día 22 de abril de 2002. Predicciones de ARH(1), utilizando matriz histórica de formas, de niveles y sin matriz histórica (ST) (Fernández de Castro, et al., 2005).

Por otro lado, la parte interesante de la evolución de la serie, comienza cuando la concentración de SO_2, en media horaria, supera los 100-150 $\mu g/m^3$, ya que esos niveles suponen la activación de los sistemas de reducción de emisiones de la Central Térmica. En esta parte de los episodios, las predicciones de los distintos modelos son bastante buenas. Los modelos captan con gran exactitud la subida inicial de los episodios, lo que representa un excelente resultado de cara a la prevención. Las predicciones acumulan mayores errores en la evolución del episodio, sobre todo cuando éste presenta diversos repuntes.

Si se comparan los resultados obtenidos con las distintas matrices históricas consideradas, se observa que los episodios se predicen mejor si el modelo dispone de la información contenida en la matriz histórica de formas.

En la Tabla 2.8 se muestran los errores de predicción de los distintos modelos en el episodio del día 22 de abril de 2002. Observando estas medidas, parece que la mejor elección para predecir los niveles de SO_2 en media horaria resulta ser el modelo ARH(1) que no utiliza matriz histórica, se estima con los datos correspondientes a tres días del pasado. Pero, en realidad esta medida resulta engañosa para los objetivos

perseguidos. Este modelo tiene errores tan bajos debido a que su predicción es 0 cuando la realidad es 0, mientras que los demás modelos predicen una subida. Pero como ya se ha comentado, esta parte resulta de menor interés frente a la predicción de altos valores de SO_2.

Modelo	Error		
	L^1	L^2	L^∞
NF ventana local, MH-niveles	16.14	18.27	28.12
NF ventana global, MH-niveles	16.66	18.65	28.52
NF ventana local, MH-formas	14.61	16.78	26.96
NF ventana global, MH-formas	15.26	17.36	27.60
ARH(1), sin MH	10.84	12.88	21.31
ARH(1), MH-niveles	16.57	19.65	31.67
ARH(1), MH-formas	15.24	18.75	31.74

Tabla 2.8: Errores de predicción correspondientes al episodio del día 22 de abril de 2002 (Fernández de Castro, et al., 2005).

Modelo	Error		
	L^1	L^2	L^∞
NF ventana local, MH-niveles	29,15	32.88	49,91
NF ventana global, MH-niveles	26,76	30.03	45,38
NF ventana local, MH-formas	23,63	26.70	41,76
NF ventana global, MH-formas	23,13	26.04	40,17
ARH(1), sin MH	24,11	28.61	47,43
ARH(1), MH-niveles	23,49	25.85	37,20
ARH(1), MH-formas	17,55	20.42	32,00

Tabla 2.9: Errores de predicción correspondientes al episodio del día 22 de abril de 2002, entre las 14:00 y las 22:30 (Fernández de Castro, et al., 2005).

En la Tabla 2.9 se muestran los errores para ese mismo episodio, teniendo en cuenta únicamente los valores correspondientes al periodo crítico, entre las 14:00 y las 22:30. En este periodo, el modelo ARH(1) estimado sin matriz histórica ya no se encuentra entre los mejores.

Descartado este modelo, los mejores resultados evaluando el día completo se obtienen con los modelos de núcleo funcional que utilizan la matriz de formas. Aunque durante el periodo crítico, sus resultados se ven mejorados por los del modelo ARH(1), de nuevo haciendo uso de la matriz histórica de formas.

2.4.3 Comparación con otros modelos

Es posible establecer comparaciones entre los resultados obtenidos con los modelos funcionales desarrollados y los modelos para series temporales reales utilizados anteriormente en la predicción de niveles de SO_2 en media horaria.

Se han comparado las predicciones dadas por los modelos funcionales con las predicciones obtenidas mediante el modelo semiparamétrico y la red neuronal, descritos en las anteriores secciones. Recuérdese que, estos modelos han sido diseñados para obtener predicciones, a seis retardos, de la serie temporal de valores reales $\{x_t\}$. Es decir, en cada instante pentaminutal t, se obtiene la predicción $\hat{x}_{t+6}$.

Para poder llevar a cabo esta comparación, se han calculado las predicciones dadas por los modelos funcionales cada 5 minutos. Para ello se procede del siguiente modo: dado un instante pentaminutal t, se considera el dato funcional $X_{n,t} = (x_{t-5},...,x_t)$, se obtiene la predicción $\hat{X}_{n+1,t} = (\hat{x}_{t+1},...,\hat{x}_{t+6})$ para la curva correspondiente a la siguiente media hora, y se extrae de ella su último valor real. Iterando para cada instante t, es posible reconstruir una serie de predicciones reales a partir de los valores reales $\hat{x}_{t+6}$ de cada predicción $\hat{X}_{n+1,t}$.

En la Figura 2.14 se representa la serie real correspondiente al día 22 de abril de 2002 junto con las predicciones obtenidas mediante redes neuronales, modelo semiparamétrico, y las obtenidas con modelos funcionales. Para hacer más legible el gráfico se representan las predicciones del modelo ARH(1) con matriz histórica de

formas, y del núcleo funcional con ventana global y matriz histórica de formas. Se puede observar una gran variabilidad de las predicciones dadas por el modelo semiparamétrico, frente a los modelos funcionales, especialmente durante el episodio de alteración de la calidad del aire.

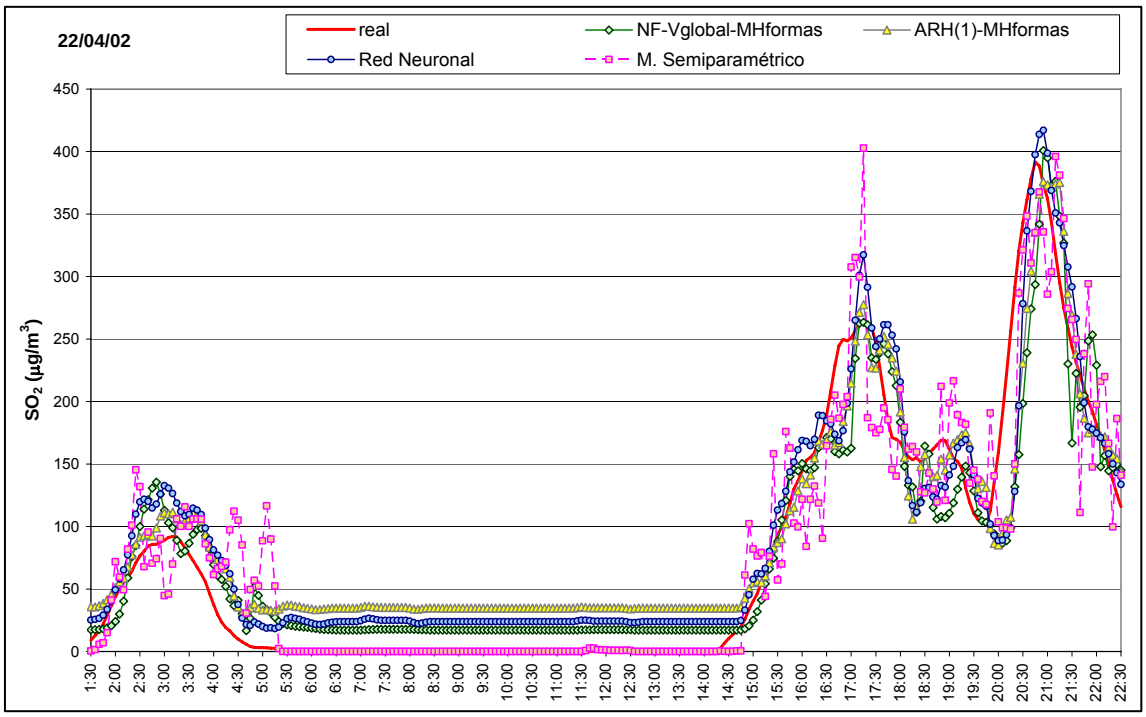

Figura 2.14: Episodio correspondiente al día 22 de abril de 2002. Comparación de predicciones de modelos funcionales: ARH(1) y Núcleo Funcional con ventana global, ambos con matriz histórica de formas, y modelos reales: red neuronal y modelo semiparamétrico (Fernández de Castro, et al., 2004).

Modelo	Error	
	ECM	***EAM***
NF ventana global, MH-formas	1341,16	25,66
ARH(1), MH-formas	1420,56	31,57
Red Neuronal	1194,70	27,84
Modelo Semiparamétrico	1504,92	23,84

Tabla 2.10: Errores de predicción correspondientes al episodio del día 22 de abril de 2002. Comparativa entre modelos funcionales y modelos reales (Fernández de Castro, et al., 2004).

En la Tabla 2.10 se resumen el error cuadrático medio (*ECM*) y absoluto medio (*EAM*) para cada uno de los métodos considerados. Como se puede observar, los métodos funcionales desarrollados son muy competitivos. De hecho, los mejores resultados de predicción se obtienen con el núcleo funcional y la red neuronal, para la que ya se han observado mejoras frente al modelo semiparamétrico.

Capítulo 3. Nuevas aportaciones a la predicción

La puesta en marcha de la nueva Central de Ciclo Combinado y la transformación de todos los grupos de la Central Térmica traen consigo la necesidad de predecir valores futuros de los óxidos de nitrógeno, además de los de dióxido de azufre, como ya se ha comentado en el primer capítulo.

Las redes neuronales fueron diseñadas para la predicción de los niveles de SO_2. Podríamos pensar en adaptar dichos modelos para obtener simultáneamente las predicciones de NO_x, ya que se espera que los episodios de alteración de la calidad de aire provocados por este contaminante tengan un comportamiento similar a los causados por el SO_2, aunque a menor escala.

Por otro lado, además de predecir también nos interesa diseñar una herramienta que permita decidir cuál es el origen de los episodios de alteración de la calidad de aire: la nueva Central de Ciclo Combinado, la actual Central Térmica transformada, ambos potenciales emisores simultáneamente, o bien, ninguno de ellos (otros posibles focos). Las redes neuronales tienen una gran capacidad predictiva pero no nos permiten visualizar el proceso hasta la predicción, lo que supone un inconveniente ante esta nueva necesidad. Debido a esto se van utilizar nuevos modelos de predicción tanto para el NO_x como para el SO_2: modelos aditivos (AM).

3.1 Nuevos modelos de predicción: Modelos Aditivos

Dentro de las distintas técnicas estadísticas que podemos utilizar para obtener las predicciones, tan mencionadas a lo largo de este trabajo, cabe destacar los modelos de regresión.

El objetivo de cualquier estudio de regresión es encontrar un modelo matemático que se ajuste a los datos y que permita una interpretación razonable de la relación entre una variable respuesta y un conjunto de variables explicativas independientes (o covariables). En este sentido los modelos de regresión lineal (Seber, 1997) suponen que el efecto de las covariables en la respuesta media es lineal y aditivo. Buja, *et al*. (1989) y Hastie & Tibshirani (1986) proponen extender esta idea a una forma más flexible conocida como *modelo aditivo*. La idea es reemplazar la función lineal de una

covariable con una función suave desconocida. El modelo aditivo consiste en la suma de estas funciones. Este modelo es no paramétrico en el sentido de que no hay que imponer una forma paramétrica a las funciones, pero en cambio hay que estimarlas de una forma iterativa a través de la utilización de mecanismos de suavización no paramétricos.

3.1.1 Planteamiento general

Sea Y una variable respuesta y sean $\left(X_1,\ldots,X_p\right)$ p variables explicativas. Un modelo aditivo se define por

$$Y=\alpha+\sum_{j=1}^{p} f_j\left(X_j\right)+\varepsilon \qquad (3.1)$$

donde los errores ε son independientes de los X_j, $E\left(\varepsilon\right)=0$ y $\mathrm{var}\left(\varepsilon\right)=\sigma^2$. En este modelo, las f_j son funciones arbitrarias unidimensionales, una para cada predictor.

Hay que destacar que este modelo retiene uno de los rasgos interpretativos más importantes del modelo lineal: la variación de la superficie respuesta fijando los valores de todos los predictores excepto de uno solo dependerá de los valores del predictor que queda por fijar, y en ningún caso dependerá de los valores que hayan tomado los otros.

Para no imponer restricciones demasiado severas en la formulación del modelo, en la ecuación (3.1) las funciones $f_j(\cdot)$ suelen ser funciones suaves de las covariables. La ventaja de utilizar estos modelos es que no se está imponiendo ninguna estructura paramétrica rígida para relacionar las covariables con la variable respuesta, sino que se deja a los datos que den forma a esta relación.

El método más general para estimar modelos aditivos consiste en estimar cada una de las funciones f_j mediante un suavizador arbitrario. Los suavizadores más comunes son los *smoothing splines ponderados* y los *kernel smoothers*. Existen varias aproximaciones para estimar estas funciones suaves. Las más utilizadas son la estimación por el método *backfitting* y técnicas basadas en la integración.

El método backfitting (Buja, *et al*., 1989; Opsomer, 2000) es un algoritmo iterativo que permite ajustar un modelo aditivo utilizando mecanismos de ajuste similares a los de regresión. La idea es estimar cada componente del modelo aditivo suavizando iterativamente los residuos parciales. Los residuos parciales correspondientes al término suave j-ésimo son calculados eliminando del modelo el efecto del resto de covariables.

El esquema del algoritmo es el siguiente:

Paso 1: Inicializar : $\alpha = \overline{Y}$, $f_j = f_j^0 \quad \forall j = 1,\ldots, p$

Paso 2: Iterar: Para $j = 1,\ldots, p, 1,\ldots, p,\ldots$

$$f_j = S_j\left(y - \alpha - \sum_{k \neq j} f_k \mid x_j \right)$$

siendo $S_j\left(y \mid x_j\right)$ un suavizador de la respuesta y en función del predictor x_j.

Paso 3: Repetir el paso 2 hasta que las funciones f_j no cambien.

En el problema de predicción que se está tratando a lo largo de este trabajo se va a utilizar este método para la estimación de los términos suaves del modelo aditivo, planteado más adelante en la ecuación (3.2).

3.1.2 Muestra de trabajo

Para construir la muestra de trabajo hemos utilizado nuevamente la idea de matriz histórica con alguna modificación.

Las matrices históricas utilizadas en los modelos anteriores estaban divididas en 10 estratos según el nivel de la variable respuesta, todos exactamente iguales para cada una de las estaciones de medida. Sin embargo el rango de valores de la variable de interés varía en función de la estación que se considere. Así, por ejemplo, en la estación F4 se alcanzaron valores de más de 700 µg/m^3 en 2006, mientras que en G4 no se superaron los 300 µg/m^3 durante ese año. En la Figura 3.1 se puede observar la

estimación de la densidad de los datos de SO_2 (medidos en $\mu g/m^3$) registrados en 2006 para las estaciones F4 y G4. Esto provoca que no siempre sea posible llenar por completo todos los registros de las matrices de manera eficiente, ya que no se puede garantizar la existencia de datos reales actualizados para ello en las circunstancias actuales, y menos aun en las futuras (con niveles más bajos de emisiones contaminantes). SO2

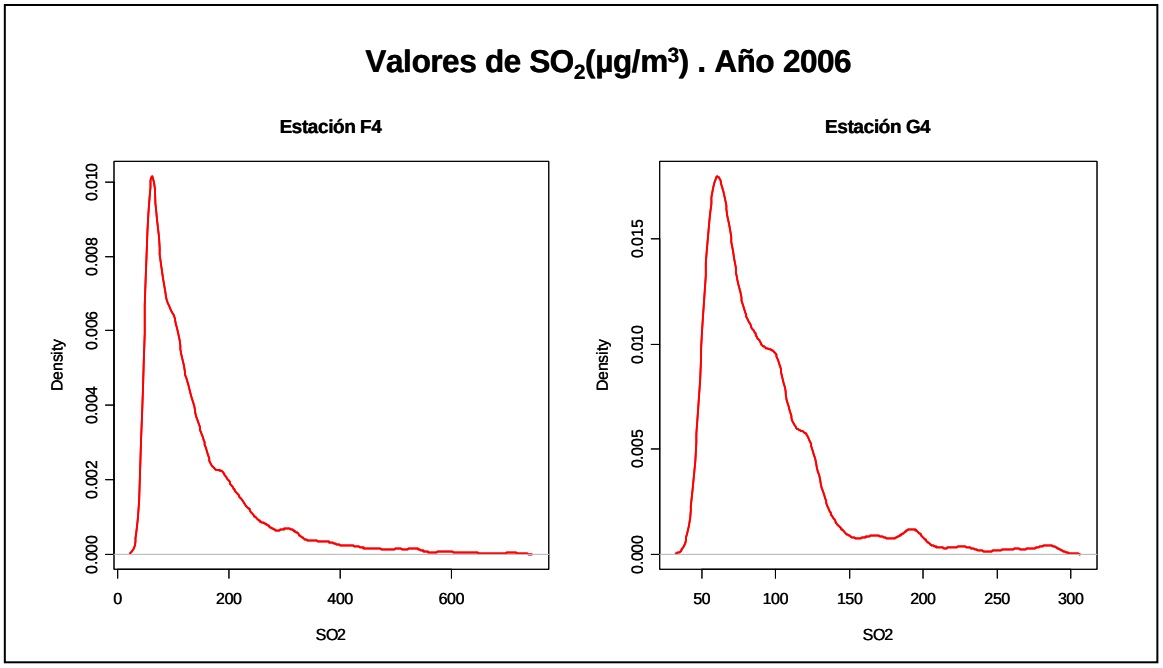

Figura 3.1: Estimación de la densidad de los datos reales de SO_2 (medidos en $\mu g/m^3$) registrados en 2006 para las estaciones F4 y G4.

Mediante un análisis de las densidades de las emisiones en 2006 se han construido estratos específicos para cada una de las estaciones de medida y para cada una de las variables de interés, SO_2 y NO_x, evitando reservar espacio para valores que sabemos que no se producen, y disponiendo así de un número mayor de registros.

El número de estratos considerados varía entre 3 y 6 dependiendo de la estación y de la variable en cuestión. Se reduce considerablemente con respecto al número de estratos considerados por los modelos anteriores, ya que al finalizar el 2006 los grupos III y IV ya estaban transformados, lo que implica una reducción considerable en las

emisiones de la Central Térmica, es decir, en 2006 no se alcanzaron valores tan altos como en otros años.

Además, al construir las nuevas matrices se han filtrado los datos de forma que no se introduzcan valores atípicos debidos a fallos en los aparatos de medida o a fallos eléctricos.

Por tanto, las matrices históricas se construyen con vectores $\left(x_{t-5}, x_t, x_{t+30}\right)'$ e $\left(y_{t-5}, y_t, y_{t+30}\right)'$, formados por datos reales de medias horarias de SO_2 y NO_x, respectivamente, donde t representa el instante minutal.

Concretamente se han diseñado matrices de 8000 registros para el SO_2 y de 4000 para el NO_x, repartidos en un cierto número de estratos variable para cada estación. Cada estrato tiene un rango de valores de x_{t+6}, de manera que cada nuevo vector será incluido en el estrato correspondiente a su x_{t+6}, reemplazando al vector más antiguo de dicho estrato.

Los vectores seleccionados para rellenar la matriz histórica provienen de datos reales correspondientes a los años comprendidos entre 2003 y 2006 (ambos incluidos) para el SO_2 y a los años 2005 y 2006 para el NO_x. Como se puede observar hemos aumentado notablemente el tamaño de las matrices históricas.

3.1.3 Aplicación al problema medioambiental

Un campo donde han sido muy utilizados los modelos aditivos es en el estudio de las series de tiempo medioambientales. Por este motivo vamos a utilizarlos para obtener las predicciones, a media hora, de los niveles de SO_2 y NO_x en el entorno de la Central Térmica. Estamos planteando así nuestro problema de predicción desde el punto de vista de la regresión, donde la variable respuesta es aquella que queremos predecir, x_{t+30} (t representa instante minutal).

Por razones ya expuestas, vamos a utilizar un vector de covariables bidimensional conteniendo información sobre el pasado de la serie temporal $\left(x_t - x_{t-5}, x_t\right)'$, que

representa el nivel medio horario del contaminante en cuestión (SO_2 ó NO_x) en el instante actual t y en el $t-5$, es decir, 5 minutos antes. Nótese que utilizamos el gradiente en lugar de considerar directamente la variable en el instante $t-5$. Esto se hace con el objetivo de evitar la *concurvity* (análogo no paramétrico de la colinealidad) que provoca una infraestimación de los errores típicos.

El modelo que planteamos es

$$\hat{X}_{t+30} = \beta_0 + f_1\left(X_t\right) + f_2\left(X_t - X_{t-5}\right) \qquad (3.2)$$

donde X_i representa el nivel medio horario del contaminante en el instante i (instante minutal), β_0 es la constante desconocida y, f_1 y f_2 son funciones suaves desconocidas. Estimaremos el modelo de forma no paramétrica utilizando splines con penalizaciones y las matrices históricas correspondientes. El parámetro de suavizado, presente en cualquier metodología no paramétrica, se va a estimar utilizando el método de validación cruzada generalizado.

Hay que tener en cuenta que la estimación de los modelos se hace de forma independiente para cada uno de los dos contaminantes, es decir, por un lado utilizamos un modelo de la forma expuesta arriba y las correspondientes matrices históricas para obtener las predicciones de SO_2,y por otro, utilizaremos otro modelo similar para obtener las del NO_x.

Actualmente esta es la predicción puntual que está funcionando en el Sistema de Predicción Estadística de Inmisión instalado en la Central Térmica de As Pontes.

Para poder observar el comportamiento del modelo aditivo seleccionado hemos evaluado su funcionamiento sobre un episodio de alteración de la calidad de aire, cuya información no ha sido incluida en las matrices históricas. Así veremos si se comporta de forma adecuada ante situaciones reales a la hora de predecir tanto los valores futuros de SO_2 como los de NO_x.

Las Figuras 3.2 y 3.3 muestran las predicciones (a media hora) utilizando el modelo propuesto y la serie real observada para un episodio de alteración de la calidad de aire ocurrido el 12 de Marzo de 2007, para el SO_2 y el NO_x,

respectivamente. En ambas se puede apreciar el buen comportamiento de las predicciones obtenidas por el modelo propuesto.

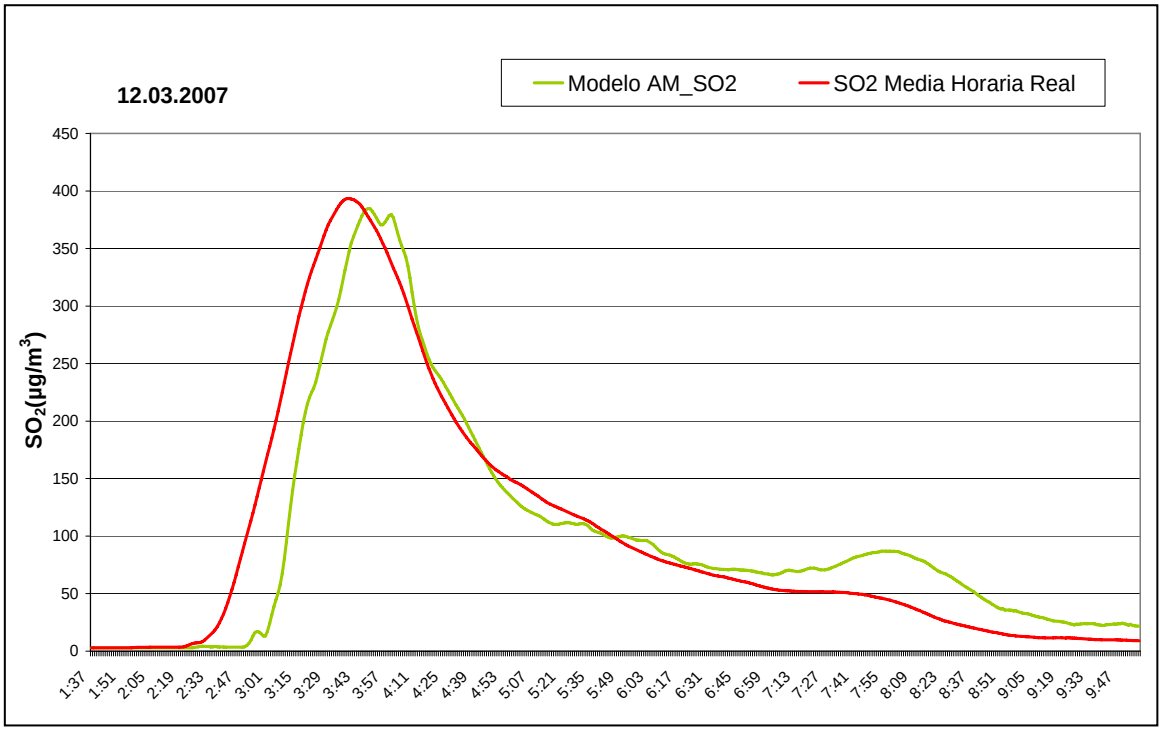

Figura 3.2: Episodio de alteración de la calidad del aire ocurrido el 12 de Marzo de 2007. Predicción dada por modelo aditivo.

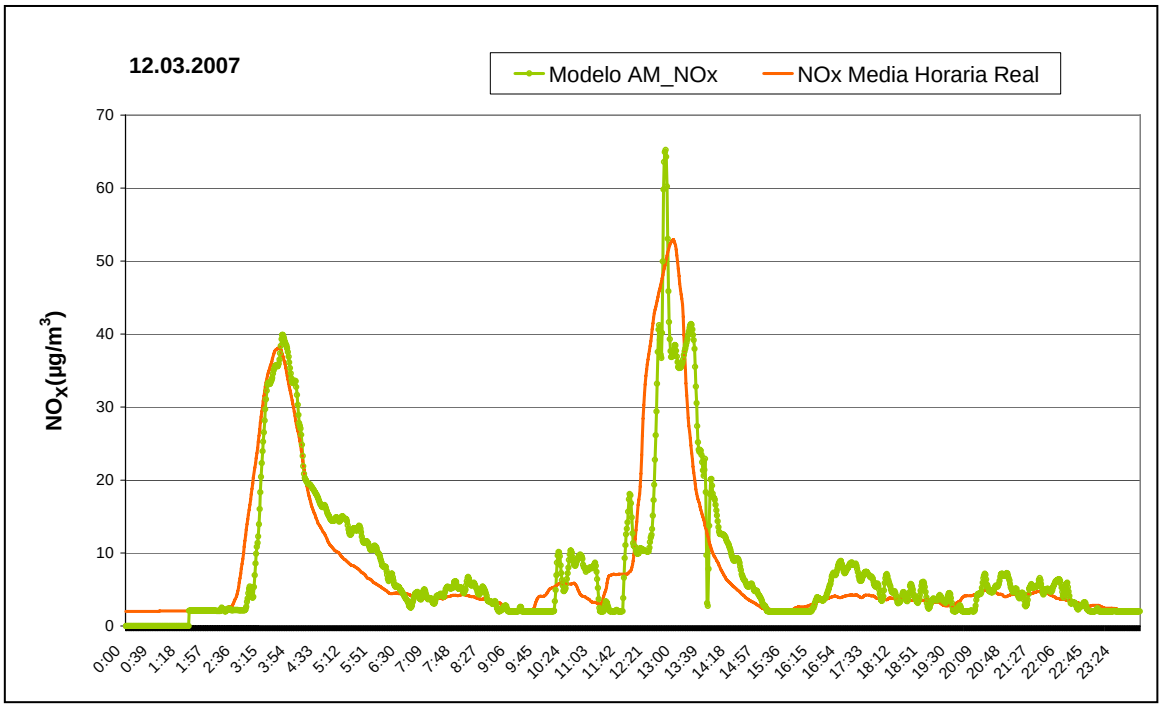

Figura 3.3: Episodio de alteración de la calidad del aire ocurrido el 12 de Marzo de 2007. Predicción dada por modelo aditivo.

De todas formas, las predicciones para el NO_x no podrán ser validadas hasta dentro de un tiempo, cuando ya se disponga de datos suficientes del Ciclo Combinado; hasta el momento, los valores medidos en la inmisión no son representativos de la situación en un futuro próximo, cuando el Ciclo esté en pleno funcionamiento.

3.1.4 Comparación con otros modelos

Se va a realizar una comparación entre el modelo aditivo propuesto y la red neuronal, descrita en la sección 2.3 del capítulo anterior y que ha estado en funcionamiento hasta hace unos meses, para la predicción de los valores medios horarios de SO_2.

Para hacer la comparación de manera apropiada se ha ajustado el modelo aditivo utilizando las matrices históricas con las que se habían entrenado la red neuronal. Recuérdese que estas matrices están construidas con datos minutales de 2003.

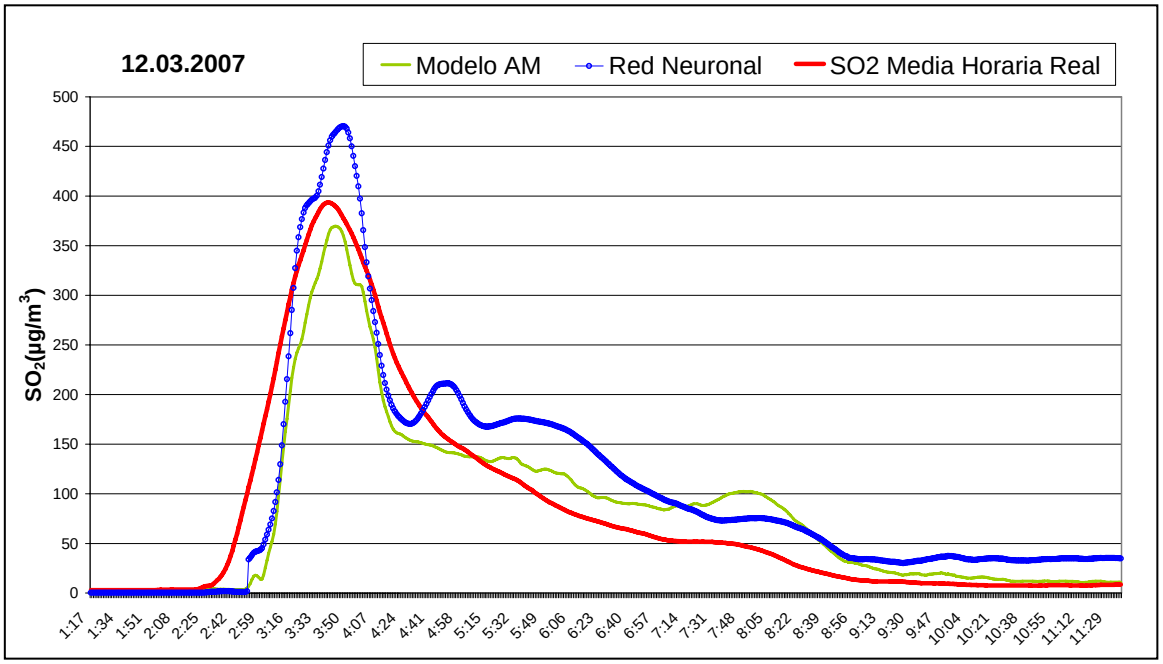

Figura 3.4: Episodio de alteración de la calidad del aire ocurrido el 12 de Marzo de 2007. Predicciones dadas por modelo aditivo y por la Red Neuronal.

La Figura 3.4 muestra las predicciones dadas por el modelo aditivo así como las de la red neuronal para el episodio ocurrido en una de las estaciones de medida el 12

de Marzo de 2007. Como se puede observar la predicción dada por el modelo aditivo persigue mejor a la serie real y además, no es tan sensible a los cambios como la red neuronal.

Para evaluar los errores cometidos por las predicciones de cada uno de los modelos considerados se han utilizado las dos medidas de error absolutas, ya utilizadas en anteriores comparaciones. Los resultados se muestran en la Tabla 3.1.

La tabla de errores confirma los resultados de la última figura: los errores cuadrático y absoluto cometidos por el modelo aditivo son mejores, en media y en mediana, que los cometidos por la red neuronal. Sin embargo, los errores cometidos por la red neuronal tienen menor variabilidad que los del modelo aditivo, tal y como indican las desviaciones típicas evaluadas.

Modelo	**Red Neuronal**		**Modelo AM**	
Error	**EC**	**EA**	**EC**	**EA**
M	3660.72	54.32	3024.55	43.17
Md	2556.82	50.56	1016.66	31.89
DT	3568.78	26.70	5135.46	34.12

Tabla 3.1: Errores de predicción en el episodio de alteración de la calidad de aire ocurrido el 12 de Marzo de 2007. Medidas absolutas.

3.2 Estructura de correlación: relación de cointegración

Los modelos aditivos que actualmente están integrados en el SIPEI obtienen la predicción puntual de los valores de SO_2 y NO_x de forma independiente, por tanto no tienen en cuenta la posible relación de dependencia que existe entre las series de dichos contaminantes. En esta sección se va a realizar un estudio de esta relación de dependencia con el objetivo futuro de plantear un modelo de predicción bidimensional.

Para ello se hará una pequeña revisión de conceptos básicos de series temporales unidimensionales y multidimensionales, y se introducirá el concepto de cointegración. Por último, se analizará si las series medioambientales objeto de este estudio siguen una relación de cointegración, y se sugerirá un nuevo modelo de predicción.

3.2.1 Conceptos básicos

Una *serie de tiempo* (univariante) es un conjunto de observaciones $\{x_t\}_{t\in T}$ de una variable aleatoria X, cada una de ellas tomada en un instante específico de tiempo t.

Se dice que una serie de tiempo es *estacionaria en el sentido estricto* si la distribución conjunta de cualquier conjunto de variables no se modifica si trasladamos las variables en el tiempo, es decir: $F(x_{t_1},\ldots,x_{t_n}) = F(x_{t_1+h},\ldots,x_{t_n+h})$ para cualesquiera $t_1,\ldots,t_n$ y h.

La estacionaridad estricta es una condición muy fuerte, ya que para contrastarla es necesario disponer de las distribuciones conjuntas para cualquier selección de variables del proceso. Una propiedad más débil, pero más fácil de contrastar en la práctica, es la *estacionaridad en sentido débil*, que implica la estabilidad de la media, la varianza y la estructura de covarianzas a lo largo del tiempo. Una serie es estacionaria en sentido débil, y para abreviar, se denominará estacionaria a partir de este momento, si para todo t:

1. $E(X_t) = \mu = cte,$
2. $E\left(X_t - \mu\right) = \sigma^2 = cte,$
3. $\operatorname{cov}\left(X_t, X_{t-k}\right) = E\left[\left(X_t - \mu\right)\left(X_{t-k} - \mu\right)\right] = \gamma_k, \quad k = 0, \pm 1, \pm 2, \ldots$

Existen dos representaciones clásicas para expresar una serie de tiempo. Una es escribir x_t como una combinación lineal de una sucesión de variables aleatorias incorreladas:

$$X_t = \mu + a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + \ldots = \mu + \sum_{j=0}^{\infty} \Psi_j a_{t-j} \qquad (3.3)$$

donde $\Psi_0 = 1$, $\{a_t\}$ es un proceso de ruido blanco de media cero, y $\sum_{j=0}^{\infty} \Psi_j^2 < \infty$.

Si se considera $\tilde{X}_t = X_t - \mu$ se puede escribir la ecuación anterior utilizando la notación del operador retardo, B definido por $BX_t = X_{t-1}$, es decir, la ecuación (3.3) se puede escribir como:

$$\tilde{X}_t = \Psi(B) a_t \qquad (3.4)$$

donde $\Psi(B) = \sum_{j=0}^{\infty} \Psi_j B^j$.

La expresión (3.4) se denomina *representación en medias móviles de un proceso*. Wold (1938) probó que un proceso estacionario puede ser siempre escrito en la forma indicada en (3.4). Una expresión del tipo de la ecuación (3.4) se dice que es un *proceso de medias móviles de orden* q (MA (q)) si viene dado por:

$$\tilde{X}_t = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

o equivalentemente, $\tilde{X}_t = \theta_q(B) a_t$, donde $\theta_q(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$.

Dado que $1 + \theta_1^2 + \ldots + \theta_q^2 < \infty$, un proceso finito de medias móviles es siempre estacionario. Se dirá que es invertible si las raíces de $\theta_q(B) = 0$ caen fuera del círculo unidad.

La otra importante clase de series de tiempo lineales es el conjunto de modelos autorregresivos. Un *modelo autorregresivo de orden* p (AR(p)) viene dado por

$$\tilde{X}_t = \phi_1 \tilde{X}_{t-1} + \ldots + \phi_p \tilde{X}_{t-p} + a_t$$

o más brevemente $\phi_p(B) \tilde{X}_t = a_t$, donde $\phi_p(B) = (1 - \phi_1 B - \ldots - \phi_p B^p)$.

El término del error a_t es un proceso de ruido blanco verificando: $\operatorname{cov}(a_t, X_{t-k}) = 0, \quad k = 1, 2, \ldots$

Para ser estacionario, las raíces de $\phi_p(B)=0$ deben estar fuera del círculo unidad. Sin embargo, la mayoría de las series reales no son estacionarias y su nivel varía con el tiempo.

Se dice que una serie es *integrada de orden* d y se denota por $I(d)$, si el número de diferencias necesarias para obtener un proceso estacionario es igual a d, es decir, si la serie $\Delta^d X_t$ es estacionaria y todas las diferencias de orden j, con $j<d$, no lo son. Las series estacionarias se van a denominar *integradas de orden cero* $I(0)$.

3.2.2 Modelos Autorregresivos Vectoriales

El modelo vectorial autorregresivo (VAR) es uno de los modelos con más éxito, flexible y fácil de utilizar para el análisis de series de tiempo multivariantes. Es una extensión natural del modelo autorregresivo univariante a las series de tiempo multivariantes. El modelo VAR ha demostrado ser especialmente útil para describir el comportamiento dinámico de las series de tiempo económicas y financieras, así como en predicción. A menudo proporciona mejores predicciones que los modelos univariantes y hace posible el desarrollo de una teoría basada en modelos de ecuaciones simultáneas. Las predicciones de los modelos VAR son bastante flexibles porque se pueden condicionar a los posibles caminos futuros de variables especificadas en el modelo.

En primer lugar se va a hacer una pequeña introducción de los modelos de series de tiempo vectoriales en general, y después se centrará en los modelos VAR.

- *Representaciones autorregresivas y en medias móviles de procesos vectoriales*:

Se dice que un proceso vectorial estacionario n-dimensional Z_t es un proceso lineal si se puede escribir como una combinación lineal de una sucesión de vectores aleatorios de ruido blanco n-dimensionales:

$$Z_t = \mu + a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + \ldots = \mu + \sum_{s=0}^{\infty} \Psi_s a_{t-s} \qquad (3.5)$$

donde $\Psi_0 = \mathrm{I}_{n\times n}$, Ψ_j son matrices $n \times n$ de coeficientes y a_t son vectores aleatorios n-dimensionales de ruido blanco con media cero y estructura de matriz de covarianzas

$$E\left[a_t a'_{t+k}\right] = \begin{cases} \Sigma, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

siendo Σ una matriz arbitraria $n \times n$ simétrica y definida positiva. Por tanto, aunque los elementos a_t son incorrelados en diferentes instantes, pueden ser correlados en instantes iguales. Equivalentemente, utilizando la notación del operador retardo la ecuación (3.5) se puede escribir:

$$\tilde{Z}_t = \Psi(B) a_t$$

donde $\tilde{Z}_t = Z_t - \mu$ y $\Psi(B) = \sum_{s=0}^{\infty} \Psi_s B^s$. Esta representación se denomina *representación en medias móviles* o *representación de Wold*.

Otra forma práctica de expresar un proceso vectorial es la representación autorregresiva, como ocurría en el caso de los procesos univariantes. Dicha representación devuelve el valor de Z en un instante t a través de los propios valores pasados más un vector de efectos aleatorios, es decir:

$$\tilde{Z}_t = \Phi_1 \tilde{Z}_{t-1} + \Phi_2 \tilde{Z}_{t-2} + \ldots + a_t = \sum_{s=1}^{\infty} \Phi_s \tilde{Z}_{t-s} + a_t \qquad (3.6)$$

o en términos del operador retardo

$$\Phi(B) \tilde{Z}_t = a_t$$

donde $\Phi(B) = \mathrm{I} - \sum_{s=1}^{\infty} \Phi_s B^s$ y Φ_s son matrices $n \times n$ de coeficientes autorregresivos.

- *Procesos vectoriales ARMA:*

Una clase útil de modelos parsimoniosos es el *proceso en medias móviles autorregresivo vectorial* ARMA (p , q):

$$\Pi_p(B)\tilde{Z}_t = \Theta_q(B)a_t$$

donde

$$\Pi_p(B) = \Pi_0 - \Pi_1 B - \Pi_2 B^2 - \ldots - \Pi_p B^p$$

y

$$\Theta_q(B) = \Theta_0 - \Theta_1 B - \Theta_2 B^2 - \ldots - \Theta_q B^q$$

son los polinomios matriciales autorregresivos y en medias móviles de órdenes p y q, respectivamente y, Π_0 y Θ_0 son matrices no singulares $n \times n$. Para cualquier caso no degenerado donde la matriz de covarianza Σ de a_t es definida positiva, se asume en lo que sigue, sin pérdida de generalidad, que $\Theta_0 = \Pi_0 = \mathrm{I}_{n\times n}$.

Si $p = 0$, el proceso seguirá un modelo vectorial MA (q):

$$\tilde{Z}_t = a_t - \Theta_1 a_{t-1} - \ldots - \Theta_q a_{t-q} \qquad (3.7)$$

Si $q = 0$, el proceso seguirá un modelo vectorial AR (p):

$$\tilde{Z}_t = \Pi_1 \tilde{Z}_{t-1} + \ldots + \Pi_p \tilde{Z}_{t-p} + a_t \qquad (3.8)$$

El proceso es estacionario si los ceros del polinomio característico $\left|\Pi_p(B)\right|$ están fuera del círculo unidad y es invertible si los ceros del polinomio característico $\left|\Theta_q(B)\right|$ están fuera del círculo unidad.

- *Modelos vectoriales AR*

El modelo vectorial AR(p) viene dado por la ecuación (3.8). Por ejemplo, la ecuación de un modelo bidimensional VAR(2) tiene la siguiente expresión

$$\begin{pmatrix} z_{1t} \\ z_{2t} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} z_{1t-1} \\ z_{2t-1} \end{pmatrix} + \begin{pmatrix} \pi_{11}^2 & \pi_{12}^2 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix} \begin{pmatrix} z_{1t-2} \\ z_{2t-2} \end{pmatrix} + \begin{pmatrix} a_{1t} \\ a_{2t} \end{pmatrix}$$

o bien,

$$z_{1t} = \mu_1 + \pi_{11}^1 z_{1t-1} + \pi_{12}^1 z_{2t-1} + \pi_{11}^2 z_{1t-2} + \pi_{12}^2 z_{2t-2} + a_{1t}$$
$$z_{2t} = \mu_2 + \pi_{21}^1 z_{1t-1} + \pi_{22}^1 z_{2t-1} + \pi_{21}^2 z_{1t-2} + \pi_{22}^2 z_{2t-2} + a_{2t}$$

donde $\operatorname{cov}(a_{1t}, a_{2s}) = \begin{cases} \sigma_{12}, & para \quad t = s \\ 0, & en \quad otro \quad caso \end{cases}$. Nótese que cada ecuación tiene las mismas variables retardadas de z_{1t} e z_{2t}. Por tanto, el modelo VAR(p) es un modelo de regresión sin relación aparente (SUR) con variables retardadas y términos determinísticos como regresoras comunes.

Una característica importante de un proceso VAR(p) es su estabilidad. Esto significa que genera series de tiempo estacionarias con medias, varianzas y covarianzas invariantes en el tiempo, con los suficientes valores iniciales. Una forma de chequear esta característica es evaluando las raíces del polinomio característico, como ya se ha comentado:

$$\det\left(\mathrm{I}_n - \Pi_1 b - \ldots - \Pi_p b^p\right) = 0$$

Si las raíces de dicho polinomio caen fuera del círculo complejo unidad (es decir, las raíces tienen módulo menor que 1) entonces el VAR(p) es estable.

Sin embargo, si la solución de la ecuación anterior tiene una raíz para $b = 1$, entonces alguna o todas las variables del proceso VAR(p) son integradas de orden 1. También podría ocurrir que exista una relación de cointegración entre las variables. En este caso podrían ser analizadas en el contexto de un modelo de corrección de errores vectorial (VECM), del que se hablará en la próxima sección.

El modelo VAR(p) básico propuesto en (3.8) puede ser demasiado restrictivo para representar las principales características de los datos. En particular, otros términos determinísticos tales como una tendencia temporal lineal o variables estacionales pueden ser necesarios, para representar los datos adecuadamente. Además, también se pueden requerir otras variables exógenas. La forma general de un modelo VAR(p) con términos determinísticos y variables exógenas viene dada por

$$Z_t = \Pi_1 Z_{t-1} + \Pi_2 Z_{t-2} + \ldots + \Pi_p Z_{t-p} + \Phi D_t + G X_t + a_t$$

donde D_t representa una matriz $(l \times 1)$ de componentes determinísticas, X_t una matriz $(n \times 1)$ de variables exógenas, y Φ y G son matrices de parámetros.

Considérese el modelo VAR(p) dado por (3.8). Se asume que este modelo es estacionario en covarianza, y que no hay restricciones en los parámetros del modelo. En la notación SUR, cada ecuación en el VAR(p) puede ser escrita como

$$z_i = Y\pi_i + e_i, \quad i = 1, \ldots, n$$

donde z_i es un vector de T observaciones en la i-ésima ecuación, Y es una matriz $(T \times k)$ con la filia t-ésima dada por $Y_t' = (1, Z_{t-1}', \ldots, Z_{t-p}')$, $k = np + 1$, π_i es un vector de parámetros con k componentes y e_i es el término de error con T y matriz de covarianzas $\sigma_i^2 \mathrm{I}_T$.

Así, el VAR(p) tiene la forma de un modelo SUR, donde cada ecuación tiene las mismas variables explicativas, y entonces, cada una de ellas se puede estimar por separado utilizando mínimos cuadrados ordinarios sin pérdida de eficiencia relativa con respecto a mínimos cuadrados generalizados.

Se puede determinar el orden del modelo VAR(p) utilizando criterios de selección de modelos. Se ajustan modelos VAR(p) con órdenes $p = 0, \ldots, p_{\max}$ y se elige el valor de p que minimiza algún criterio de selección.

Los criterios de selección para modelos VAR(p) son de la forma

$$IC(p) = \ln \left| \tilde{\Sigma}(p) \right| + c_T \cdot \varphi(n, p)$$

donde $\tilde{\Sigma}(p) = T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$ es la matriz de covarianzas residual sin una corrección de grados de libertad de un modelo VAR(p), c_T es una sucesión indexada para la muestra de tamaño T, y $\varphi(n, p)$ es una función penalizada. Los tres criterios de información más comunes son el Akaike (AIC), Schwarz-Bayesian (BIC) y Hannan-Quinn (HQ):

$$AIC(p) = \ln\left|\tilde{\Sigma}(p)\right| + \frac{2}{T} pn^2$$

$$BIC(p) = \ln\left|\tilde{\Sigma}(p)\right| + \frac{\ln T}{T} pn^2$$

$$HQ(p) = \ln\left|\tilde{\Sigma}(p)\right| + \frac{2\ln\ln T}{T} pn^2$$

El criterio AIC asintóticamente sobreestima el orden con probabilidad positiva, mientras que los criterios BIC y HQ estiman el orden consistentemente bajo condiciones generales si el verdadero orden p es menor o igual que $p_{\max}$. Para más información sobre los criterios de información ver Lütkepohl (1991).

3.2.3 Cointegración

La noción de cointegración ha sido uno de los conceptos más importantes en series de tiempo desde que Granger (1983) y Engle & Granger (1987) lo desarrollaron formalmente. Sin embargo, este concepto ya estaba implícito en los modelos de corrección de errores propuestos por Dabison, *et al*., 1978. El tema de la cointegración ha tenido aplicaciones generales en el análisis de datos económicos así como diversas publicaciones en la literatura económica. El libro de Engle & Granger (1991) contiene una colección de artículos que han sido importantes en el desarrollo de este concepto.

- *Concepto de cointegración*

Sea $Z_t = (z_{1t}, \ldots, z_{nt})'$ un vector de n series temporales $\mathrm{I}(1)$. Se dice que Z_t está *cointegrada* si existe un vector $\beta = (\beta_1, \ldots, \beta_n)'$ tal que

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

En otras palabras, las series no estacionarias de Z_t están cointegradas si existe una combinación lineal de ellas que sí es estacionaria. Si alguno de los elementos de

β es igual a cero, entonces sólo el subconjunto de series en Z_t con coeficientes no cero estará cointegrado.

El vector β recibe el nombre de *vector de cointegración*. Dicho vector no es único ya que para cualquier escalar c la combinación lineal $c\beta' Z_t = \beta^{*\prime} Z_t \sim \mathrm{I}(0)$. Por esto, se suele asumir una normalización para poder identificar β de forma única. Una normalización típica es $\beta = (1, -\beta_2, \ldots, -\beta_n)'$ y así la relación de cointegración se puede expresar como sigue

$$\beta' Z_t = z_{1t} - \beta_2 z_{2t} - \ldots - \beta_n z_{nt} \sim \mathrm{I}(0)$$

o bien

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

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

- *Relaciones de cointegración múltiples*

Si el vector Z_t está cointegrado puede haber $0 < r < n$ vectores de cointegración. Por ejemplo, si $n = 3$ y se supone que hay $r = 2$ vectores de cointegración $\beta_1 = (\beta_{11}, \beta_{12}, \beta_{13})'$ y $\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23})'$. Así se verifica que

$$\beta_1' Z_t = \beta_{11} z_{1t} + \beta_{12} z_{2t} + \beta_{13} z_{3t} \sim \mathrm{I}(0)$$

$$\beta_2' Z_t = \beta_{21} z_{1t} + \beta_{22} z_{2t} + \beta_{23} z_{3t} \sim \mathrm{I}(0)$$

y la matriz

$$B' = \begin{pmatrix} \beta_1' \\ \beta_2' \end{pmatrix} = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}$$

forma una base del espacio de vectores de cointegración. Los vectores linealmente independientes β_1 y β_2 de la base de cointegración B no son únicos a menos que se

asuma algún tipo de normalización. Es más, cualquier combinación lineal de estos dos vectores, por ejemplo $\beta_3 = c_1\beta_1 + c_2\beta_2$ donde c_1 y c_2 son constantes, es también un vector de cointegración.

- *Cointegración y modelo de corrección de errores*

Se considera un vector bidimensional $Z_t = (z_{1t}, z_{2t})'$ de series $\mathrm{I}(1)$ y se asume que Z_t está cointegrado con vector de cointegración $\beta = (1, -\beta_2)'$ y así $\beta' Z_t = z_{1t} - \beta_2 z_{2t}$ es $\mathrm{I}(0)$. En un importante artículo, Engle & Granger (1987) demostraron que la cointegración implica la existencia de un *modelo de corrección de errores* (ECM) de la forma

$$\Delta z_{1t} = c_1 + \alpha_1 \left(z_{1t-1} - \beta_2 z_{2t-1} \right) + \sum_j \psi_{11}^j \Delta z_{1t-j} + \sum_j \psi_{12}^j \Delta z_{2t-j} + \varepsilon_{1t}$$

$$\Delta z_{2t} = c_2 + \alpha_2 \left(z_{1t-1} - \beta_2 z_{2t-1} \right) + \sum_j \psi_{21}^j \Delta z_{1t-j} + \sum_j \psi_{22}^j \Delta z_{2t-j} + \varepsilon_{2t}$$

que describe el comportamiento dinámico de z_{1t} y z_{2t}. El ECM vincula las relaciones de equilibrio a largo plazo, implicadas por la cointegración, con los mecanismos de ajustes dinámicos a corto plazo, que describen cómo las variables reaccionan cuando se mueven hacia el equilibrio a largo plazo. Este ECM hace que el concepto de cointegración sea útil para la modelización de series de tiempo que verifican esta propiedad.

- *Test de cointegración*

Sea Z_t un vector de n series $\mathrm{I}(1)$. Recuérdese que Z_t es cointegrado con $0 < r < n$ vectores de cointegración si existe B' una matriz $r \times n$ tal que

$$B' Z_t = \begin{pmatrix} \beta_1' Z_t \\ \vdots \\ \beta_r' Z_r \end{pmatrix} = \begin{pmatrix} u_{1t} \\ \vdots \\ u_{rt} \end{pmatrix} \sim \mathrm{I}(0)$$

Los test de cointegración pueden plantearse como tests de existencia de relaciones de equilibrio a largo plazo entre los elementos de Z_t. Los tests de cointegración cubren dos situaciones:

1. Hay más de un vector de cointegración.
2. Posiblemente hay $0 \leq r < n$ vectores de cointegración.

El primer caso fue originalmente considerado por Engle & Granger (1987) que desarrollaron un simple procedimiento en dos pasos basado en los residuos de cointegración y utiliza técnicas de regresión. El segundo tipo de test fue propuesto por Johansen (1988) que desarrolló un sofisticado procedimiento secuencial para determinar la existencia y el número de relaciones de cointegración, basado en técnicas de máxima verosimilitud.

El procedimiento propuesto por Engle & Granger consiste en determinar si un cierto vector β es vector de cointegración. Para ello primero se construye el residuo de cointegración $\beta' Z_t = u_t$ y una vez construido se realiza un test de raíces unitarias para determinar si u_t es $\mathrm{I}(0)$.

La hipótesis nula en este procedimiento es la de no cointegración y la alternativa es la cointegración. Hay que considerar dos casos. En el primero el vector de cointegración propuesto es preespecificado (no estimado). En el segundo caso el vector se estima a partir de los datos por mínimos cuadrados y se construye el residuo de cointegración estimado $\hat{\beta}' Z_t = \hat{u}_t$. Los test que utilizan el vector de cointegración preespecificado son generalmente más potentes que los que utilizan el vector estimado.

Una vez construido el residuo de cointegración, ya sea directamente o estimándolo, se plantea el siguiente test de hipótesis

$$\begin{cases} \mathrm{H}_0: \ u_t \sim \mathrm{I}(1) & \text{(no cointegración)} \\ \mathrm{H}_1: \ u_t \sim \mathrm{I}(0) & \text{(cointegración)} \end{cases}$$

Para evaluar estas hipótesis se puede utilizar cualquier test de raíces unitarias. Los más utilizados son el de Dickey-Fuller aumentado (ADF) y el de Phillips-Perron (PP).

Los residuos de cointegración pueden incluir términos determinísticos (constante o tendencia), en cuyo caso los test de raíces unitarias deben plantearse teniendo en cuenta esta cuestión.

Se había comentado que existen dos grandes grupos de test para determinar las relaciones de cointegración, el propuesto por Engle & Granger que se acaba de explicar y el propuesto por Johansen, del que se va a dar una idea general a continuación.

- *Procedimiento de Johansen: Modelos VAR y cointegración*

Johansen trata el tema de la cointegración y el modelo de corrección de errores en el marco de los modelos VAR.

Considérese el modelo VAR(p) para el vector de n series temporales Z_t

$$Z_t = \Phi D_t + \Pi_1 Z_{t-1} + \ldots + \Pi_p Z_{t-p} + \varepsilon_t, \quad t = 1, \ldots, T \qquad (3.9)$$

donde D_t contiene los términos determinísticos (constante, tendencia, …). Recuérdese que el modelo VAR es estable si

$$\det\left(\mathrm{I}_n - \Pi_1 b - \ldots - \Pi_p b^p\right) = 0$$

tiene todas las raíces fuera del círculo unidad. Si existe alguna raíz en el círculo unidad entonces alguna o todas las variables son $\mathrm{I}(1)$ y por tanto, pueden estar cointegradas.

Supóngase que Z_t es $\mathrm{I}(1)$ y posiblemente cointegrada. Entonces la representación VAR (3.9) no es la más adecuada para el análisis, ya que las relaciones de cointegración no aparecen explícitamente. Para tener en cuenta dichas relaciones se va a transformar el modelo VAR en un modelo de corrección de errores vectorial (VECM)

$$\Delta Z_t = \Phi D_t + \Pi Z_{t-1} + \Gamma_1 \Delta Z_{t-1} + \ldots + \Gamma_{p-1} \Delta Z_{t-p+1} + \varepsilon_t \qquad (3.10)$$

donde $\Pi = \Pi_1 + \ldots + \Pi_p - \mathrm{I}_n$ y $\Gamma_k = -\sum_{j=k+1}^{p} \Pi_j, \quad k = 1, \ldots, p-1$. La matriz Π se denomina *matriz de impactos a largo plazo* y Γ_k son las *matrices de impactos a corto plazo*. Nótese que los parámetros Π_i del VAR pueden recuperarse a partir de los parámetros Π y Γ_k del VECM:

$$\begin{cases} \Pi_1 = \Gamma_1 + \Pi + \mathrm{I}_n, \\ \Pi_k = \Gamma_k - \Gamma_{k-1}, \quad k = 2, \ldots, p \end{cases}$$

En el VECM (3.10) ΔZ_t y sus retardos son $\mathrm{I}(0)$. El término ΠZ_{t-1} es el único que incluye las posibles variables $\mathrm{I}(1)$ y para que ΔZ_t sea $\mathrm{I}(0)$ debe ocurrir que ΠZ_{t-1} sea también $\mathrm{I}(0)$. Por tanto, ΠZ_{t-1} contendrá las relaciones de cointegración en caso de que existan.

Bajo la suposición de que el modelo VAR(p) tiene raíces unitarias, es claro que Π es una matriz singular, y en consecuencia de rango reducido. Supóngase que $rang(\Pi) = r < n$. Hay que considerar dos casos:

1. $rang(\Pi) = 0$. Esto implica que $\Pi = 0$ y Z_t es $\mathrm{I}(1)$ pero no cointegrada. El VECM (3.10) se reduce a un VAR($p-1$) en primeras diferencias:

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

2. $0 < rang(\Pi) = r < n$. Esto implica que Z_t es $\mathrm{I}(1)$ con r vectores de cointegración linealmente independientes y $n-r$ tendencias estocásticas (raíces unitarias). Ya que Π tiene rango r se puede escribir como el producto

$$\Pi = \alpha \beta'$$

donde α y β son matrices $n \times r$ con $rang(\alpha) = rang(\beta) = r$. Las filas de β' forman una base de r vectores de cointegración y los elementos de α distribuyen el impacto de los vectores de cointegración en la evolución de ΔZ_t. Así el VECM (3.10) resulta

$$\Delta Z_t = \Phi D_t + \alpha\beta' Z_{t-1} + \Gamma_1 \Delta Z_{t-1} + \ldots + \Gamma_{p-1} \Delta Z_{t-p+1} + \varepsilon_t \quad (3.11)$$

con $\beta' Z_{t-1} \sim \mathrm{I}(0)$ ya que β' es una matriz de vectores de cointegración. Hay que tener en cuenta que esta descomposición de la matriz Π no es única, para que lo fuese habría que imponer restricciones al modelo.

La metodología propuesta por Johansen para modelizar la cointegración consta de los siguientes pasos:

i. Especificar y estimar un modelo VAR(p) para Z_t .

ii. Construir test de razón de verosimilitudes para el rango de Π y poder determinar el número de vectores de cointegración.

iii. Si fuese necesario, imponer restricciones de normalización e identificación para los vectores de cointegración.

iv. Una vez obtenidos los vectores de cointegración normalizados estimar el VECM cointegrado resultante por máxima verosimilitud.

Para ver como se construye el test de razón de verosimilitud para determinar el número de vectores de cointegración, se denota el VECM propuesto en (3.11) por $\mathrm{H}(r)$. Dicho modelo puede ser formulado con la condición de que el rango de Π es menor o igual que r . Esto crea un conjunto de modelos anidados

$$\mathrm{H}(0) \subset \ldots \subset \mathrm{H}(r) \subset \ldots \subset \mathrm{H}(n)$$

donde $\mathrm{H}(0)$ representa el modelo VAR no cointegrado con $\Pi = 0$ y $\mathrm{H}(n)$ un modelo VAR(p) estacionario sin restricciones. Esta formulación anidada es útil para desarrollar un procedimiento secuencial para contrastar el número de relaciones de cointegración.

Como el rango de la matriz Π proporciona el número de relaciones de cointegración en Z_t , Johansen formula estadísticos de razón de verosimilitudes (LR) para el número de dichas relaciones como estadísticos LR para determinar el rango

de Π. Estos tests están basados en la estimación de los autovalores $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_n$ de esta matriz. Estos autovalores coinciden con las correlaciones canónicas cuadradas entre ΔZ_t y Z_{t-1} corregida por el retardo ΔZ_t y D_t, así que están entre 0 y 1. Recuérdese que el rango de Π es igual al número de autovalores no cero.

Johansen propone el siguiente test estadístico de hipótesis anidadas:

$$\begin{cases} \mathrm{H}_0(r_0): & r = r_0 \\ \mathrm{H}_1(r_0): & r > r_0 \end{cases}$$

El estadístico de LR, denominado *estadístico de la traza*, viene dado por:

$$LR_{trace}(r_0) = -T \sum_{i=r_0+1}^{n} \ln\left(1-\hat{\lambda}_i\right)$$

Si $rang(\Pi) = r_0$ entonces $\hat{\lambda}_{r_0+1},\ldots,\ \hat{\lambda}_n$ deberían ser próximos a cero y $LR_{trace}(r_0)$ debería ser pequeño. En cambio, si $rang(\Pi) > r_0$ entonces, alguno de los $\hat{\lambda}_{r_0+1},\ldots,\ \hat{\lambda}_n$ será distinto de cero (pero menor que 1) y $LR_{trace}(r_0)$ debería ser grande. La distribución asintótica bajo la hipótesis nula del estadístico no es chi-cuadrado sino que es una versión multivariante de la distribución de Dickey-Fuller que depende de la dimensión $n - r_0$ y la especificación de los términos determinísticos. Los valores críticos para esta distribución están tabulados en Osterwald-Lenum (1992).

Johansen propone un procedimiento secuencial para determinar el número de vectores de cointegración. Primero se contrasta $\mathrm{H}_0(r_0 = 0)$ frente a $\mathrm{H}_1(r_0 > 0)$. Si no se rechaza la hipótesis nula, se concluye que no hay vectores de cointegración entre las n variables de Z_t. Si se rechaza se concluye que existe al menos un vector de cointegración y se plantearía el siguiente test $\mathrm{H}_0(r_0 = 1)$ frente a $\mathrm{H}_0(r_0 > 1)$. Si no se rechaza H_0 entonces hay un único vector de cointegración. En cambio, si la hipótesis nula se rechazada existen al menos dos vectores de cointegración y se plantea es siguiente test. El procedimiento secuencial continúa hasta que la hipótesis nula no es rechazada.

Una vez que se ha determinado que $rang(\Pi) = r$, $0 < r < n$, para estimar el VECM se utiliza regresión multivariante de rango reducido. Se estiman por máxima verosimilitud los vectores de cointegración y se obtiene que

$$\hat{\beta}_{mle} = (\hat{v}_1, \ldots, \hat{v}_r)$$

donde $\hat{v}_i$ son los autovectores asociados a los autovalores $\hat{\lambda}_i$. El resto de los parámetros del modelo se estiman por mínimos cuadrados multivariante reemplazando β por $\hat{\beta}_{mle}$.

3.2.4 Aplicación al problema medioambiental

En el primer capítulo se ha comentado que la evolución de las series de emisión de la Central Térmica para el SO_2 y el NO_x tienen un comportamiento parecido aunque a escalas distintas y por tanto, con las series de inmisión ocurre algo similar (véase Figura 1.3). Esto hace que se plantee la pregunta de sí estas dos series están cointegradas.

En cada instante t se considera la serie bidimensional $X_t = (x_{1t}, x_{2t})$ relativa a las últimas cuatro horas, donde $x_{1t} = (x_{1t}, x_{1t-1}, \ldots, x_{1t-239})$ representa la serie de valores medios horarios de SO_2 y $x_{2t} = (x_{2t}, x_{2t-1}, \ldots, x_{2t-239})$ representa la serie de valores medios horarios de NO_x. Se va a comprobar si dichas series están cointegradas utilizando el procedimiento secuencial de Johansen descrito en la sección anterior.

Esta comprobación se va a realizar durante el episodio de alteración de la calidad de aire ocurrido en una de las estaciones de medida el día 12 de Marzo de 2007, que ya se ha analizado a lo largo de este trabajo.

Para ilustrar el ejemplo se va a considerar un instante en concreto: las 05:45 horas. En la Figura 3.5 se representan los valores en media horaria de los dos contaminantes durante las cuatro horas anteriores a dicho instante. Como se puede observar ambas series evolucionan de forma muy similar aunque a distinta escala, por lo que cabe esperar que sí estén cointegradas. Si se utiliza el procedimiento de

Johansen para contrastar si las dos series verifican la relación de cointegración, la respuesta es que sí lo están y el vector de cointegración estimado es $\hat{\beta} = (1, -13.51)$, considerando la normalización expuesta en la sección 3.2.3. Esto quiere decir que a cada unidad de NO_x le corresponden aproximadamente 13 de SO_2.

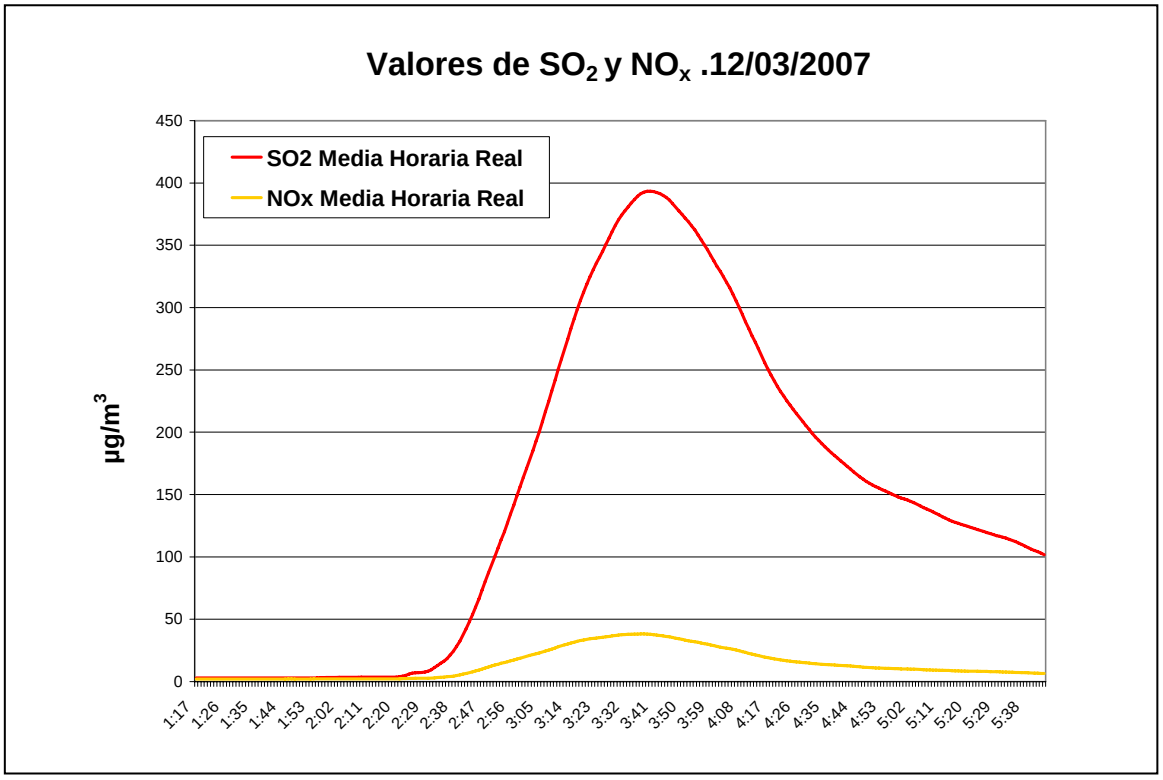

Figura 3.5: Episodio de alteración de la calidad del aire ocurrido el 12 de Marzo de 2007.

En la Tabla 3.2 se muestra en número de veces en que el criterio de Johansen indica si las series están cointegradas, si no lo están y alguna ocasión en que este criterio no es capaz de decidir, tanto para el día completo como para el período en que ocurrió el episodio propiamente dicho. Como se puede observar el número de veces que las series están cointegradas es casi la mitad de instantes en el período completo 46.46%. Si sólo se consideran los instantes comprendidos entre las 2:00 y las 10:00 horas, que es cuando realmente ocurre el episodio, el número de veces que las series están cointegradas es casi el triple que el número de veces que no lo están. Por tanto, parece que va a ser útil considerar esta estructura de correlación a la hora de realizar las predicciones.

Criterio de Johansen	Período completo	De 02:00 a 10:00
No cointegradas	642 (44,58%)	121 (25,20%)
Cointegradas	669 (46.46%)	306 (63.75%)
No decide	129 (8.95%)	54 (11.25%)

Tabla 3.2: Número de veces que las series de SO_2 y NO_x ocurridas en una de las estaciones de medida el 12 de Marzo de 2007 están cointegradas o no cointegradas según el procedimiento de Johansen.

Los nuevos modelos, sobre los que actualmente se está trabajando, van tener en cuenta esta relación de dependencia pero también van utilizar el mecanismo de memoria con el fin de capturar la tendencia histórica de las series. La idea es utilizar un enfoque semiparamétrico en la línea del propuesto por García Jurado, *et al*. (1995), ya comentado en este trabajo.

En cada instante t se va a predecir x_{1t+30} y x_{2t+30} de forma independiente utilizando los modelos aditivos propuestos en la primera sección de este capítulo. El siguiente paso será construir la serie de residuos bidimensional $\hat{Z}_{t-205},...,\hat{Z}_t$ donde cada $\hat{Z}_i = X_i - \hat{X}_i$ y $\hat{X}_i = (\hat{x}_{1i}, \hat{x}_{2i})$ es la predicción dada por el modelo aditivo (3.2). Una vez construida la serie bidimensional de residuos se utilizará el procedimiento de Johansen para comprobar si existe cointegración en ese instante. En caso de existir dicha relación se obtendrá la predicción de $\hat{Z}_{t+30}$ con un modelo de corrección de errores de la forma del (3.11); si no existe relación de cointegración se calculará la predicción de $\hat{Z}_{t+30}$ a través de un modelo VAR como el de la ecuación (3.8). La predicción final dada será:

$$\hat{\hat{X}}_{t+30} = \hat{X}_{t+30} + \hat{\hat{Z}}_{t+30}$$

donde

$$\hat{X}_{t+30} = \begin{pmatrix} \beta_1 + f_{11}(x_{1t}) + f_{12}(x_{1t} - x_{1t-5}) \\ \beta_2 + f_{21}(x_{2t}) + f_{22}(x_{2t} - x_{2t-5}) \end{pmatrix}$$

y

$$\hat{\hat{Z}}_{t+30} = \begin{cases} VECM(p), \text{ si hay cointegración} \\ VAR(p), \text{ si no hay cointegración} \end{cases} \quad , \; para \;\; \hat{Z}_i = X_i - \hat{X}_i$$

Se espera que estas predicciones mejoren las dadas por los modelos aditivos, ya que la componente no paramétrica del modelo de predicción capturará la tendencia histórica de la serie, mientras que la modelización de la serie bidimensional residual más reciente (componente paramétrica del modelo) mejorará la predicción aportando al sistema capacidad de improvisación, al igual que ocurría con el anterior modelo semiparamétrico.

Bibliografía

1. Besse P and Cardot H. *Spline Approximation of the Prediction of a First-Order Autoregressive Functional Process*. Canadian Journal of Statistics 1996; 24: 467-487.

2. Besse P, Cardot H and Stephenson D. *Autoregressive Forecasting of Some Functional Climatic Variations*. Scandinavian Journal of Statistics 2000; 27: 673-687.

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

4. Buja A, Hastie TJ and Tibshirani RJ. *Linear smoothers and additive models*. Annals of Statistics 1989; 17: 453-555.

5. Dabison JEH., David F Hendry, Frank Srba and Stephen Yeo. *Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom*. Economical Journal 1978; 88: 661-692.

6. Damon J and Guillas S. *The Inclusion of Exogenous Variables in Functional Autoregressive Ozone Forecasting*. Environmetrics 2002; 13: 759-774

7. Engle RF, Granger WJ, Rice J and Weiss A. *Semiparametric estimates of the relationship between weather and electricity sales*. Journal of the American Statistical Association 1986; 81: 310-320.

8. Engle RF and Granger CWJ. *Co-Integration and Error Correction: Representation, Estimation and Testing*. Econometrica 1987; 55: 251-276.

9. Engle RF and Granger CWJ. *Long-run economic relationships: readings in cointegration*. Oxford University Press, Oxford 1991.

10. Eric Zivot and Jiahui Wang. *Modeling Financial Time Series with S-Plus*. Springer, 2006.

11. Febrero-Bande M. *Modelización para la predicción en series de tiempo: aspectos computacionales con nuevas aportaciones y aplicaciones*. Tesis Doctoral. Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela 1995.

12. Fernández de Castro BM, Prada-Sánchez JM, González-Manteiga W, Febrero-Bande, M, Bermúdez-Cela JL and Hernández Fernández JJ. *Prediction of SO_2*

levels using neural networks. Journal of the Air and Waste Management Association 2003; 53: 532-538.

13. Fernández de Castro BM. *Modelos de predicción con redes neuronales y modelos funcionales: una aplicación a un problema medioambiental*. Tesis Doctoral. Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela 2004.

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

15. Gao J. *The laws of the iterated logarithm of some estimates in Partly Linear Models*. Statistics & Probability Letters 1995; 25: 153-162.

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

17. González-Manteiga W, Prada-Sánchez JM, Cao R, García-Jurado I, Febrero-Bande M and Lucas-Domínguez T. *Time-series analysis for ambient concentrations*. Atmospheric Environment 1993; 27A: 153-158.

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

19. Green P, Jennison C and Seheult A. *Analysis of field experiments of least squares smoothing*. Journal of the Royal Statistical Society 1985; 47: 299-315.

20. Hamilton JD. *Time Series Analysis*. Princeton University Press, Princeton 1994.

21. Hastie T and Tibshirani R. *Generalized additive models (with discussion)*. Statistical Science 1986; 1: 297–318.

22. Hastie TJ and Tibshirani RJ. *Generalized Additive Models*. Chapman and Hall, Londres 1990.

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

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

25. Kay JW and Titterington DM. *Statistics and Neural Networks: Advances at the Interface*. Oxford University Press, Oxford 1999.

26. Liang H. *The Berry-Essen bounds of error variance estimation in semiparametric regression model*. Communications in Statistics-Theory Methods 1994; 23:3439-3451.

27. Linton O. *Second order approximation in the partially linear regression model*. Econometrica 1995; 63: 1079-1112.

28. Ljung GM and Box GEP. *On the Measure of Lack of Fit in Time Series Models*. Biometrika 1978; 65: 297-303.

29. Lütkepohl H. *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin 1991.

30. Opsomer J. *Asymptotic properties of backfitting estimators*. Journal of Multivariate Analysis, 2000; 73: 166-179.

31. Osterwald-Lenum M. *A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Statistics*. Oxford Bulletin of Economics and Statistics 1992; 54: 461-472.

32. Peña D. *Análisis de Series Temporales*. Alianza Editorial S.A., Madrid 2005.

33. Pérez P, Trier A and Reyes J. *Prediction of $PM_{2.5}$ Concentrations Several Hours in Advance Using Neural Networks in Santiago, Chile*. Atmospheric Environments 2000; 34: 1189-1196.

34. Prada-Sánchez JM 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.

35. Prada-Sánchez JM, Febrero-Bande M, Cotos-Yáñez T, González-Manteiga W, Bermúdez-Cela JL 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.

36. Robinson PM. *Root n-consistent semiparametric regression*. Econometrica 1988; 56: 931-954.

37. Robinson PM. *Nearest-neighbour estimation in partly linear models*. Nonparametric Statistics 1995; 5: 33-41.

38. Schick A. *Root n consistent estimation in partly linear regression models*. Statistics & Probability Letters 1996; 28: 353-358.

39. Seber GAF. *Linear Regression Analysis*. John Wiley, Nueva York 1997.

40. Speckman P. *Kernel smoothing in partial linear models*. Journal of the Royal Statistics Society 1988; 50: 413-436.

41. Wei WWS. *Time Series Analysis: Univariate and Multivariate Methods*. Addison-Wesley, California 1990.

42. Wold HOA. *A study in the Analysis of Stationary Time Series*. Almquist and Wiksell, Uppsala 1938.