

CAPÍTULO VIII

Una aplicación del análisis de series temporales funcionales a los precios horarios de la electricidad en el mercado MIBEL

Pedro Galeano*

Actualmente, muchas medidas se registran de manera prácticamente continua a lo largo del tiempo dando lugar a conjuntos de observaciones que tienen forma de funciones (curvas) relativamente suaves y que son observadas con alta frecuencia. El estudio de datos con tales características se puede realizar mediante técnicas para el análisis de series temporales funcionales, un área de la estadística que ha recibido gran atención en las dos últimas décadas. En este capítulo se realiza una aplicación del análisis de series temporales funcionales a las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad de España en el Mercado Ibérico de la Electricidad (MIBEL).

Palabras clave: análisis de datos funcionales, bandas de predicción, precios horarios de la electricidad, predicción, series temporales funcionales.

* El autor de este capítulo agradece el apoyo financiero de la Agencia Estatal de Investigación (PID2019-108311GB-I00/AEI/10.13039/501100011033).

1. INTRODUCCIÓN

La gran mayoría de métodos y procedimientos estadísticos tradicionales para el análisis de series temporales fueron diseñados para series con decenas, centenas o incluso algunos miles de observaciones. Estos métodos y procedimientos tratan de compensar la relativa falta de información con ciertos supuestos que permiten obtener predicciones relativamente fiables, lo que ha sido de gran ayuda para el desarrollo del conocimiento en un gran número de disciplinas. Si bien es cierto que el uso de modelos probabilísticos hace que el análisis de series temporales se pueda considerar un subcampo de la estadística, una parte fundamental de su desarrollo ha sido llevada a cabo por investigadores y científicos en otras áreas, muy especialmente en economía. Por ejemplo, los modelos de heterocedasticidad condicional, como el modelo GARCH de Engle (1982) y Bollerslev (1986) y el modelo de volatilidad estocástica de Taylor (1986), entre muchos otros modelos que permiten analizar la varianza condicionada de las series temporales, fueron desarrollados para entender y predecir la volatilidad de rendimientos financieros. También, las técnicas de cointegración de Granger (1981) y Engle y Granger (1987), entre otros, que permiten buscar correlaciones entre series temporales que ayuden a mejorar sus predicciones, fueron desarrolladas para entender la relación entre variables económicas, como el consumo y los ingresos, a largo plazo. Otro ejemplo son los contrastes de raíces unitarias, como los propuestos por Dickey y Fuller (1979), Sargan y Bhargava (1983), Said y Dickey (1984) y Kwiatkowski *et al.* (1992), que permiten determinar cuándo una serie temporal es estacionaria o no, fueron desarrollados en el contexto de variables económicas, como el PIB.

Durante los últimos veinte años han aparecido nuevas fuentes de información, como internet, los teléfonos inteligentes, las redes sociales y los sensores automáticos que producen grandes bases de datos estructuradas y no estructuradas con tamaños masivos y que pueden tener forma de textos, imágenes, vídeos y/o sonidos (véase Galeano y Peña, 2019). En particular, la monitorización hace que actualmente muchas medidas se registren prácticamente de manera continua a lo largo del tiempo, generando series temporales de gran longitud y con características especiales propias, como puede ser la presencia de ruido en forma de fluctuaciones que distorsionan las tendencias de la serie. Medidas de este estilo pueden ser, por ejemplo: (1) señales vitales humanas como la temperatura corporal, la presión arterial y las frecuencias cardíaca y respiratoria; (2) factores medioambientales como la temperatura, la velocidad del viento, la precipitación, y los niveles de contaminación de determinados contaminantes, como el monóxido de carbono o el óxido de azufre; y (3) el precio de activos o materias primas que se negocian en mercados financieros en tiempo real, como el precio de las acciones de una empresa o el precio de la energía eléctrica.

Debido a la gran longitud que pueden tener este tipo de series temporales y a que se pueden considerar varias de estas series de manera simultánea, los métodos y procedimientos más habituales para el análisis de la dependencia transversal y temporal pueden ser inadecuados y/o ineficientes. Por ello, están apareciendo nuevas metodologías entre las que se encuentra el análisis de series temporales funcionales. Este tipo de análisis surge cuando una serie temporal que se observa muy frecuentemente durante un extenso período de tiempo, se puede dividir en subseries de la misma longitud definidas en intervalos de tiempo conse-

cutivos definidos periódicamente, que pueden sufrir algún tipo de transformación posterior para alcanzar alguna propiedad como puede ser la estabilidad. Por ejemplo, una variable que se observa cada minuto durante varios meses o años se puede separar en subseries definidas en días o semanas consecutivas, dependiendo de sus características, que a su vez se pueden transformar en subseries estables o que comiencen con el mismo valor. Las subseries obtenidas de esta manera suelen mostrar dos características fundamentales. En primer lugar, tienen a tener un patrón común y formas relativamente suaves por lo que pueden ser estudiadas con las herramientas del análisis de datos funcionales (FDA, es su acrónimo en inglés). El FDA consiste en una colección de métodos y procedimientos para el análisis estadístico de muestras de funciones y es una de las áreas que mayor interés ha suscitado en la comunidad estadística en los últimos años. En la práctica, y aunque la frecuencia de toma de datos pueda ser muy alta, los procesos continuos se observan en un número discreto de puntos. Sin embargo, las técnicas de suavizado permiten reproducir las funciones que han sido observadas parcialmente obteniendo una única serie temporal funcional compuesta por una sucesión de funciones observadas durante el mismo período de tiempo, por ejemplo, en días, semanas, o meses sucesivos. Esta serie temporal funcional tiene una serie de características con dimensión subyacente infinita. Esto puede parecer problemático ya que es bien sabido que la maldición de la dimensión conlleva importantes problemas estadísticos. Sin embargo, los métodos de FDA permiten extraer información de la riqueza estructural inherente a los datos funcionales. En segundo lugar, las subseries obtenidas suelen ser dependientes debido a la manera en la que son definidas. Los procedimientos para el análisis de series temporales funcionales tratan de describir y entender dicha dependencia que puede ser utilizada, por ejemplo, para realizar predicciones de funciones futuras con bandas de predicción asociadas. De esta manera, las predicciones obtenidas corresponden a cada uno de los períodos en que se ha dividido la serie original. Por ejemplo, para subseries formadas por las observaciones correspondientes a los minutos de todo un día, la predicción funcional a un paso corresponderá a una predicción de todos los momentos del día siguiente al último observado, y no solamente para los minutos del día siguiente.

Existe una literatura creciente sobre el análisis de series temporales funcionales que, de alguna manera, extiende el análisis de series temporales univariantes y multivariantes al campo funcional. Aquí se incluyen: (1) las técnicas de suavizado propuestas por Ramsay y Silverman (2005), que permiten eliminar el ruido y suavizar los datos, lo que es beneficioso siempre que el proceso subyacente sea la principal preocupación; (2) los procedimientos de predicción basados en componentes principales funcionales propuestos por Hyndman y Ullah (2007), Hyndman y Shang (2009) y Aue, Norinho y Hörmann (2015), entre otros, que son extensiones de los procedimientos de reducción de la dimensión en el análisis de series temporales; (3) los modelos autorregresivos y de media móvil funcionales propuestos por Bosq (2000), Bosq y Blanke (2007), Klepsch, Klüppelberg y Wei (2017) y Li, Robinson y Shang (2020), entre otros, que son extensiones de los modelos autorregresivos y de media móvil del análisis de series temporales univariantes y multivariantes; (4) los procedimientos de estimación y predicción no paramétricos de Besse, Cardot y Stephenson (2000) y Antoniadis, Paparoditis y Sapatinas (2006), entre otros; (5) los modelos de heterocedasticidad condicional funcional propuestos por Hörmann, Horváth, y Reeder (2013) y Aue, Horváth y Pellatt (2017), entre otros; y (6) los modelos factoriales dinámicos funcionales propuestos por Hays,

Shen y Huang (2012), Kowal, Matteson y Ruppert (2012) y Kokoszka, Miao y Zhang (2015), entre otros. Estos y otros procedimientos han sido utilizados en un gran número de aplicaciones. Por ejemplo, Ramsay y Silverman (2005) detectaron diferentes tendencias estacionales y a largo plazo en el índice de fabricación de productos perecederos en USA mediante simples técnicas de suavizado. Hyndman y Ullah (2007) y Hyndman y Shang (2009), entre muchos otros, utilizaron los métodos de predicción basados en componentes principales funcionales para predecir curvas de mortalidad anual por edad de hombres y mujeres franceses, y para predecir tasas de fertilidad anual en Australia en función de la edad. Shen (2009) y Aue, Norinho y Hörmann (2015) también utilizaron componentes principales funcionales para predecir tasas de llegada de llamadas diarias a un servicio de atención al cliente de un banco y concentraciones de polución en Graz, Austria, respectivamente. Klepsch, Klüppelberg y Wei (2017) y Li, Robinson y Shang (2020) utilizaron modelos ARMA funcionales para predecir datos de tráfico en Alemania y tasas de fertilidad en varios países, respectivamente. Besse, Cardot y Stephenson (2000) y Antoniadis, Paparoditis y Sapatinas (2006) utilizaron métodos no paramétricos para predecir el efecto anual del fenómeno de *El Niño* y tasas de audiencias televisivas, respectivamente. Hörmann, Horváth, y Reeder (2013) y Aue, Horváth y Pellatt (2017) analizaron rendimientos diarios mediante modelos de heterocedasticidad condicional funcional. Por último, Hays, Shen y Huang (2012) y Kowal, Matteson y Ruppert (2012) utilizaron modelos factoriales dinámicos funcionales para predecir curvas de rendimientos financieros y señales cerebrales, respectivamente.

El objetivo principal de este capítulo es el de presentar una aplicación del análisis de series temporales funcionales de las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL). Para ello, la serie temporal de precios horarios, observada durante un período extenso de años, se transforma en una sucesión de subseries definidas mediante una estandarización de los precios en cada uno de los días de negociación. A continuación, cada subserie se transforma en un dato funcional genuino mediante técnicas de suavizado y todas las subseries suavizadas en forma de función dan lugar a una única serie temporal funcional. Mediante técnicas del análisis funcional de datos, se concluirá que dicha serie temporal funcional es predecible por lo que se obtendrán predicciones a un día vista de las curvas, junto con sus respectivas bandas de predicción. Para realizar los cálculos y gráficos mostrados en este artículo se han utilizado fundamentalmente las librerías *fda* y *ftsa* del *software* estadístico R. En el capítulo, se informará de las funciones utilizadas específicamente en cada uno de los cálculos realizados.

El resto de este capítulo se estructura como sigue. La sección segunda presenta la serie de precios horarios de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL) y cómo, a partir de esta serie, se obtienen sus curvas de rendimientos intradía acumulados mediante técnicas de suavizado, dando lugar a una serie temporal funcional. La sección tercera realiza un análisis descriptivo de las curvas de rendimiento y comprueba la estabilidad temporal de las mismas. La sección cuarta presenta los componentes principales funcionales de las curvas de rendimiento, que no solo son una herramienta muy útil para obtener las principales fuentes de variación de las curvas, si no que permiten obtener predicciones de las curvas futuras. De hecho, la sección quinta realiza predicción a un día

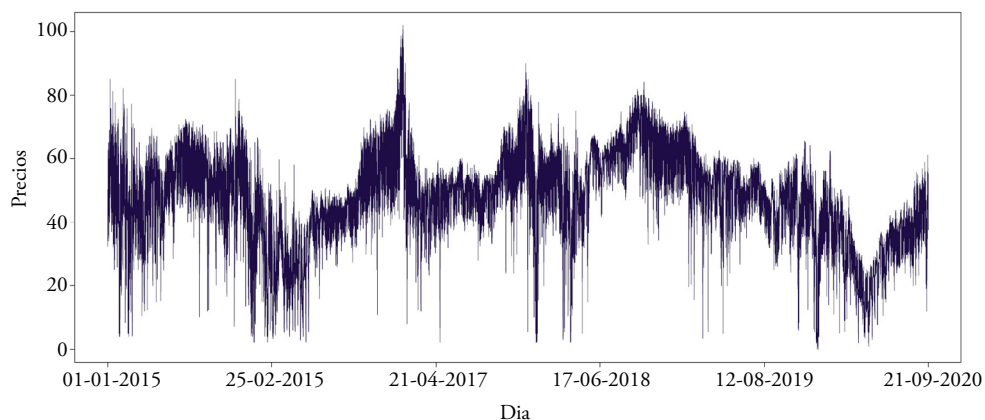
vista de las curvas observadas mediante componentes principales funcionales e introduce un procedimiento simple para la obtención de bandas de predicción, que son el equivalente de los intervalos y regiones de predicción en series temporales univariantes y multivariantes. Finalmente, la sección sexta presenta algunas conclusiones.

2. CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

La figura 1 muestra los precios de cierre cada hora de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL) en los 2100 días comprendidos entre el 1 de

Figura 1.

Precios horarios de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL)



Fuente: Elaboración propia.

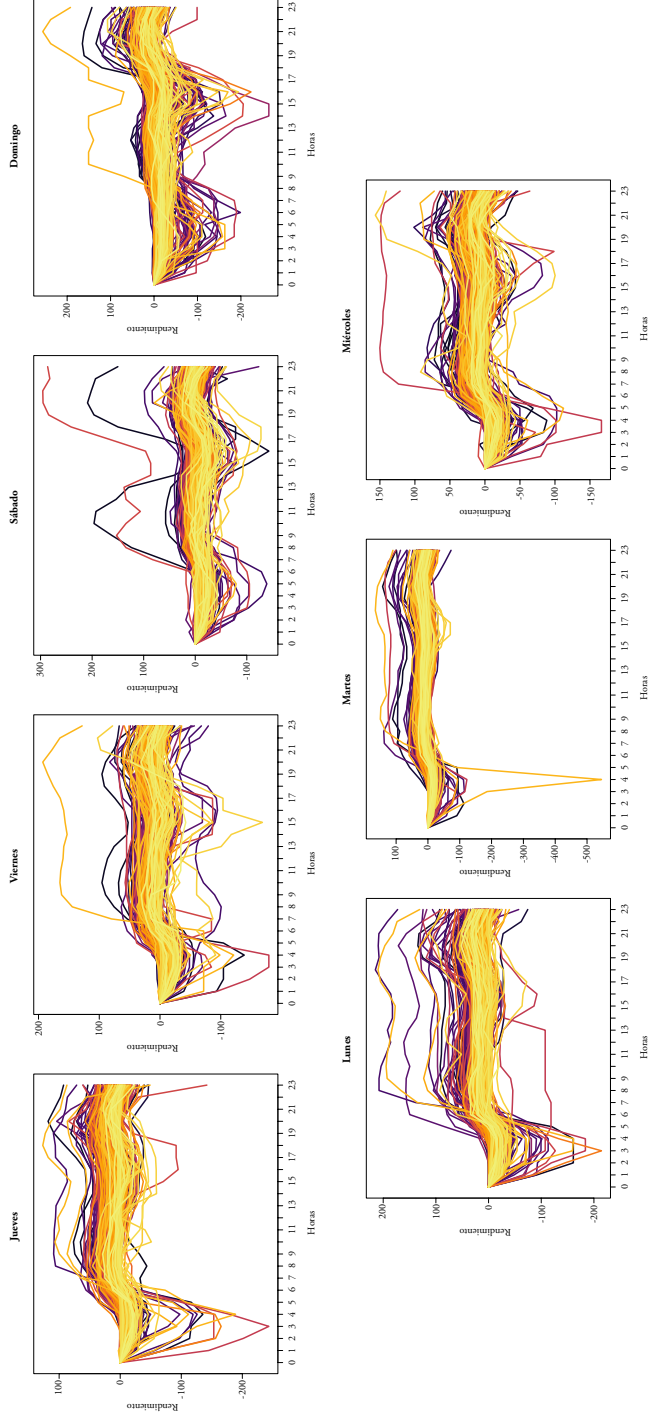
enero de 2015 y el 30 de septiembre de 2020, que han sido obtenidos de la página web de Red Eléctrica Española, <https://www.esios.ree.es/es>. Dado que el intervalo temporal entre observaciones es de una hora, la serie temporal de precios tiene $2.100 \times 24 = 50.400$ observaciones. Como se puede observar, la serie es claramente no estacionaria con los típicos crecimientos y decrecimientos habituales en series de precios y con la presencia de algunos valores extremos. Tal como hacen Gabrys, Horváth y Kokoszka (2010) con series de precios de gran longitud como esta, se definen los rendimientos intradía acumulados como:

$$r_i(t_j) = 100 \times [\ln p_i(t_j) - \ln p_i(t_1)], \quad [1]$$

donde $p_i(t_j)$ es el precio en la hora t_j del i -ésimo día de negociación, para $i = 1, \dots, 2100$ días, y t_1 y t_{24} son las 00:00 y las 23:00 horas del día, respectivamente. Dado que,

Figura 2.

Serie de rendimientos intradía acumulados de los precios horarios de la electricidad desde el 1 de enero de 2015 al 30 de septiembre de 2020 divididas en los siete días de la semana



Fuente: Elaboración propia.

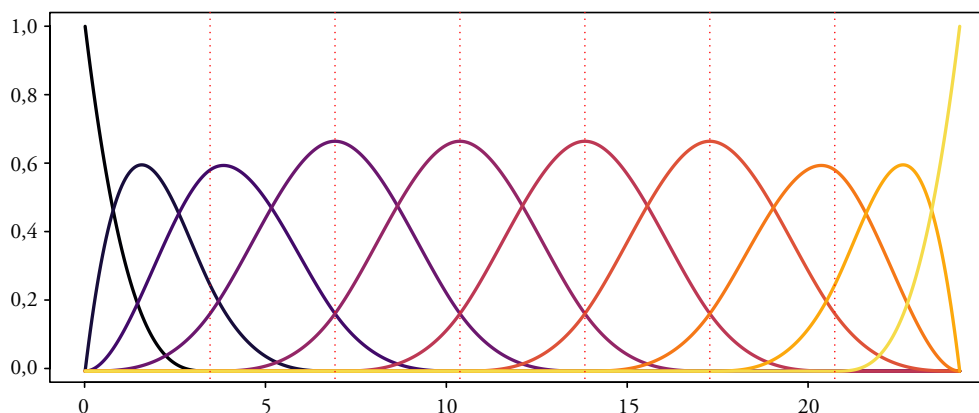
$$r_i(t_j) \simeq 100 \times \frac{p_i(t_j) - p_i(t_1)}{p_i(t_1)},$$

y teniendo en cuenta que $p_i(t_1)$ es un valor fijo para cada día i , los rendimientos intradía acumulados se pueden ver como una sucesión de series temporales diarias de 24 observaciones con el mismo origen, 0, y que resultan ser una especie de estandarización de los precios en cada uno de los días de negociación, ya que $p_i(\cdot)$ y $r_i(\cdot)$ tienen diferentes escalas y orígenes.

En particular, las 2.100 series de 24 rendimientos intradía acumulados obtenidas a partir de la serie de precios de la figura 1, se muestran en la figura 2. Las series se han dividido en los siete días de la semana, empezando por el jueves al ser el día correspondiente al 1 de enero de 2015, y se han dibujado con una escala de color de tal manera que las series de pasado más lejano tienen color oscuro y las series de pasado más cercano tienen color más claro. Nótese que muchas series de color claro ocultan a series de color más oscuros simplemente porque la función plot de R dibuja la curva del primer día y sucesivamente va añadiendo las curvas correspondientes al resto de días. Aunque es difícil extraer conclusiones específicas a partir de los gráficos, parece apreciarse que la mayoría de rendimientos intradía crecen en las primeras horas de la mañana y últimas de la tarde y decrecen en el resto de momentos del día. Este patrón común parece estar presente en todos los días de la semana. Además, existe un número determinado de series que parece tener un comportamiento diferente posiblemente debido a ser días festivos o de índole especial, e incluso algunas observaciones concretas de algunas series parecen mostrar también comportamientos extremos o atípicos. De hecho, la figura 2 sugiere que los rendimientos están afectados por fluctuaciones de precios o correcciones a corto plazo que pueden distorsionar las tendencias comunes.

Figura 3.

Funciones de una base de B-splines en el intervalo $[0,24)$



Fuente: Elaboración propia.

El proceso de suavizado que se describe a continuación permite, por un lado, suprimir la influencia del ruido más obvio en las series de rendimientos intradía y, por otro lado, transformar dichas series en genuinos datos funcionales. Para ello, nótese que si bien se disponen únicamente de los valores en [1], los rendimientos existen en cualquier momento del día. Por lo tanto, con los dos objetivos previos en mente, se supone que los rendimientos intradía acumulados en [1] se pueden escribir como

$$r_i(t_j) = \theta_i(t_j) + e_{ij},$$

donde θ_i es una función continua y suave definida en todo tiempo $t \in [0,24)$ y e_{ij} es un término de ruido. El conjunto de funciones:

$$\{\theta_i(t) : t \in [0,24), i = 1, \dots, 2100\}, \quad [2]$$

recibe el nombre de curvas de rendimientos intradía acumulados (CIDRs, es su acrónimo en inglés). Con el propósito de distinguir entre una CIDR como objeto funcional y el valor que toma dicha CIDR en un momento específico del día, se denota por θ_i a la i -ésima CIDR como objeto funcional y por $\theta_i(t)$ al valor puntual de la i -ésima CIDR en un tiempo $t \in [0,24)$ específico. Ahora, el paso clave para obtener las CIDR en [2] a partir de las series de rendimientos en [1] es expresar $\theta_i(t)$ mediante una expansión de funciones base, es decir, mediante una combinación lineal de funciones fijas determinadas que permiten aproximar cualquier función continua y suave. Los sistemas de funciones base más conocidos son la base de Fourier y las bases de B-splines (ver, por ejemplo, los capítulos 3 a 6 de Ramsay y Silverman, 2005). Por un lado, la base de Fourier es una sucesión de senos y cosenos con frecuencias enteras y está diseñada para el análisis de funciones periódicas, es decir, funciones que tienen el mismo valor al principio y al final del intervalo de observación. Por otro lado, las bases de B-splines están formadas por sucesiones de polinomios definidos en subintervalos consecutivos y están diseñadas para el análisis de cualquier tipo de función continua. Por lo tanto, para obtener las CIDRs se utiliza una base de B-splines con las 10 funciones definidas en el intervalo $[0,24)$ que aparecen en la figura 3 y que permiten escribir $\theta_i(t)$ como sigue:

$$\theta_i(t) = \sum_{m=1}^{10} c_{im} \beta_m(t) = c_i' \beta(t), \quad i = 1, \dots, 2100, \quad [3]$$

donde $c_i = (c_{i1}, \dots, c_{i10})'$ es el vector de coeficientes que minimizan la suma de cuadrados dada por,

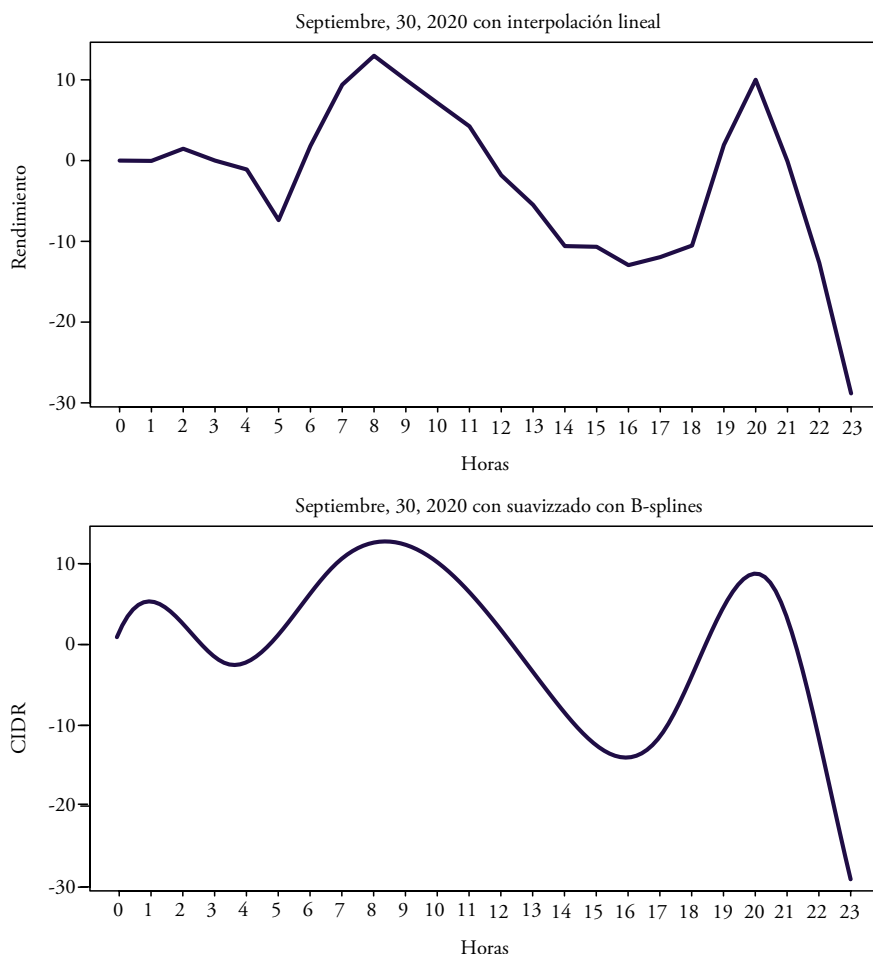
$$SC(c_{i1}, \dots, c_{i10}) = \sum_{j=1}^{24} \left(r_i(t_j) - \sum_{m=1}^{10} c_{im} \beta_m(t_j) \right)^2, \quad i = 1, \dots, 2100,$$

y $\beta(t) = (\beta_1(t), \dots, \beta_{10}(t))'$ es el vector de funciones base evaluadas en tiempo $t \in [0,24)$, respectivamente. Por ejemplo, la parte derecha de la figura 4 muestra los 24 rendimientos intradía acumulados correspondientes al 30 de septiembre de 2020 interpolados linealmente, mientras que la parte izquierda muestra la CIDR asociada suavizada con la base de 10 B-splines

de la figura 3. Como se puede apreciar, la CIDR preserva la forma de los rendimientos horarios observados. Para realizar el suavizado se han utilizado las funciones `create.bspline.basis` y `smooth.basis` de la librería `fda` de R.

Figura 4.

CIDR para los precios horarios de la electricidad del 30 de septiembre de 2020



Fuente: Elaboración propia.

Es muy importante resaltar el cambio de enfoque que supone el disponer de las CIDRs suaves en [3] frente a disponer de la serie temporal de precios original o de las series de rendimientos discretos en [1]. Nótese que inicialmente se ha transformado una serie temporal de 50.400 observaciones en las 2.100 series temporales de 24 observaciones en [1], que a su vez se han transformado en las 2.100 funciones suaves en [2] escritas en términos de los

B-splines como aparecen en [3]. Por lo tanto, de aquí en adelante el objeto de estudio son estas 2.100 funciones definidas en todo momento del día, y no los rendimientos horarios que han dado lugar a dichas funciones. Esto supone un cambio substancial, ya que, por ejemplo, en lugar de predecir los rendimientos acumulados en cada hora en punto del 1 de octubre de 2020, el objetivo es predecir la CIDR completa correspondiente al 1 de octubre de 2020, lo que permitiría conocer la predicción en cualquier momento del día. Para ello, como se detalla en la sección quinta, se hace uso de la expansión en [3], del vector de coeficientes $c_i = (c_{i1}, \dots, c_{i10})'$, para $i = 1, \dots, 2100$, y del vector de funciones base evaluadas en tiempo $t \in [0, 24)$, $\beta(t) = (\beta_1(t), \dots, \beta_{10}(t))'$.

Las CIDRs presentadas en esta sección dan lugar a muchas preguntas. Por ejemplo, es razonable preguntarse si las CIDRs de los precios de la electricidad son estacionarias bajo el punto de vista funcional. Aquí, como se describirá en la sección tercera, la estacionariedad funcional se refiere a la estabilidad de las CIDRs durante todo el período de observación. Conocer si las CIDRs son estacionarias o no es importante ya que ayuda a entender la dinámica de las funciones y a seleccionar un buen procedimiento para obtener predicciones de CIDR futuras. También parece adecuado preguntarse cuáles son las principales fuentes de variabilidad de las CIDRs que puedan ayudar a explicar las tendencias locales que parecen observarse en las figuras 2 y 4. Por último, también parece adecuado preguntarse si las CIDRs de los precios de la electricidad son funciones predecibles, es decir, si es posible predecir la CIDR de un día de negociación cualquiera con las CIDRs de los días de negociación previos. Para responder a estas y otras preguntas se realiza un análisis descriptivo funcional de las CIDRs en la sección tercera y se presentan los componentes principales funcionales en la sección cuarta.

3. ANÁLISIS DESCRIPTIVO FUNCIONAL DE LAS CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

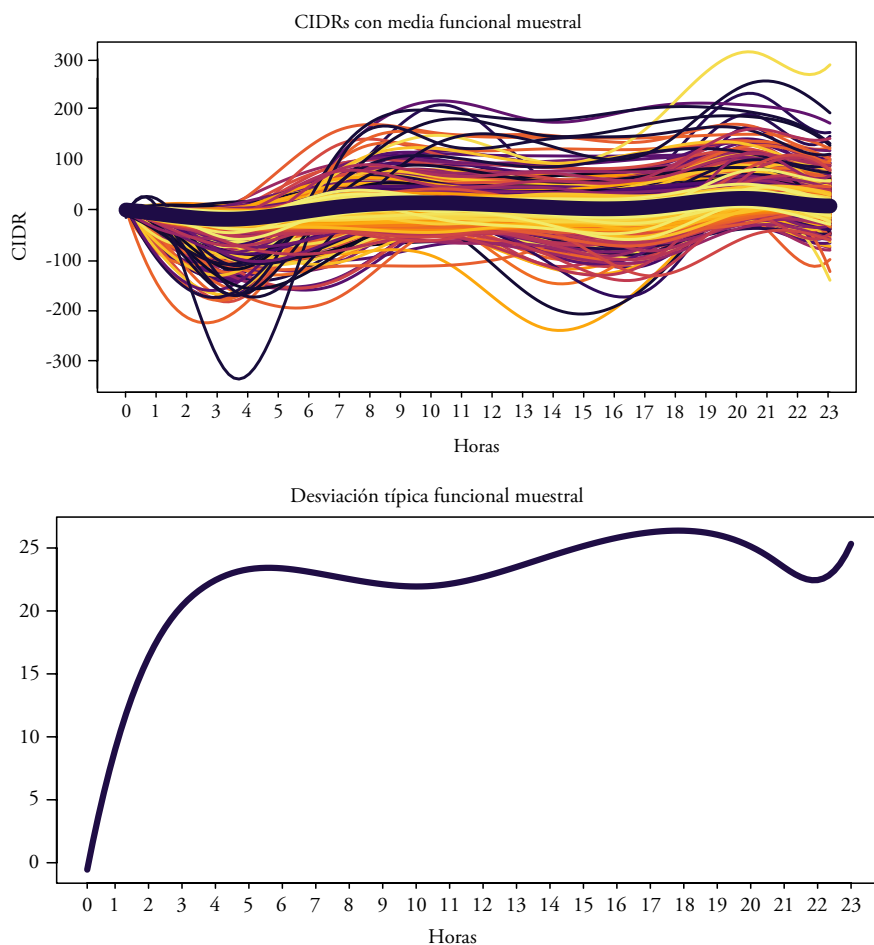
Las 2.100 CIDRs en [2] escritas mediante las expansiones en [3] forman una serie temporal funcional. Cada CIDR es una observación de la serie temporal funcional y se espera que las CIDRs de diferentes días tengan algún tipo de dependencia. Esta sección presenta un breve análisis descriptivo de la serie de CIDRs con el objetivo de entender sus principales características. Por sencillez en la exposición, dentro de lo posible se evita entrar en detalles técnicos complejos que los lectores interesados pueden consultar en referencias más generales sobre datos y series temporales funcionales, como pueden ser Ramsay y Silverman (2005), Ferraty y Vieu (2006), Hörmann y Kokoszka (2010), Horváth y Kokoszka (2012), Hsing y Eubank (2015), Srivastava y Klassen (2016) o Kokoszka y Reimherr (2017), entre otros.

En primer lugar, es importante conocer si las CIDRs son estacionarias bajo el punto de vista funcional. Aquí es necesario hacer hincapié que cuando se habla de estacionariedad desde un punto de vista funcional se refiere a la estabilidad entre las observaciones funcionales durante todo el período de observación y no a la estabilidad temporal de las observaciones puntuales, como es común en el análisis de series temporales univariantes y multivariantes.

Por ejemplo, tanto la figura 2 como la parte derecha de la figura 4 parecen indicar que las 24 observaciones discretas diarias tienen un comportamiento no estacionario. Sin embargo, la figura 2 parece indicar que las tendencias locales diarias parecen ser comunes, lo que sugiere que las CIDRs, que se muestran en el gráfico de la derecha de la figura 5, tienen un comportamiento estacionario. La estacionariedad de una serie temporal funcional se puede estudiar mediante un contraste de hipótesis propuesto por Horváth, Kokoszka y Rice (2014). La idea fundamental del estadístico de dicho contraste es comparar las sumas parciales de las funciones $\theta_1, \dots, \theta_{2100}$ con la suma total de todas las funciones, lo que permite conocer la

Figura 5.

Media funcional muestral y desviación típica funcional muestral para las CIDRs del precio de la electricidad



Fuente: Elaboración propia.

evolución de la serie temporal funcional. Bajo la hipótesis de estacionariedad, las diferencias entre las sumas parciales y la suma total no deben ser grandes, lo que sí debería ocurrir bajo la hipótesis alternativa de no estacionariedad. El estadístico del contraste proporciona un valor que cuantifica dichas diferencias que, en el caso concreto de las CIDRs, da lugar a un p -valor de 0.2347, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula de estacionariedad del conjunto de CIDRs. Este p -valor se ha obtenido gracias a la función `T_stationary` de la librería `fts` de R.

La condición de estacionariedad permite obtener un amplio conjunto de características de interés de las CIDRs. Las dos medidas más sencillas son la media funcional muestral, dada por:

$$\bar{\theta}(t) = \frac{1}{2100} \sum_{i=1}^{2100} \theta_i(t), \quad t \in [0, 24] \quad [4]$$

que proporciona el promedio de los valores de las CIDRs en cualquier momento $t \in [0, 24]$, y la varianza funcional muestral dada por:

$$\hat{\sigma}^2(t) = \frac{1}{2099} \sum_{i=1}^{2100} (\theta_i(t) - \bar{\theta}(t))^2, \quad t \in [0, 24] \quad [5]$$

que proporciona la varianza de los valores de las CIDRs en cualquier tiempo $t \in [0, 24]$. Estas medidas se pueden calcular muy fácilmente tras substituir $\theta_1, \dots, \theta_{2100}$ en las ecuaciones [4] y [5] con sus expansiones en términos de $\beta_1, \dots, \beta_{10}$ dadas en [3], dando lugar a:

$$\bar{\theta}(t) = \bar{c}' \beta(t), \quad t \in [0, 24] \quad [6]$$

y

$$\hat{\sigma}^2(t) = \beta(t)' S_0 \beta(t), \quad t \in [0, 24] \quad [7]$$

donde $\bar{c} = \frac{1}{2100} \sum_{i=1}^{2100} c_i$ y $S_0 = \frac{1}{2099} \sum_{i=1}^{2100} (c_i - \bar{c})(c_i - \bar{c})'$, respectivamente, es decir, el vector de medias muestral y la matriz de covarianzas muestral de los vectores de coeficientes $c_1, \dots, c_{2100}$. El gráfico de la derecha de la figura 5 muestra las 2.100 CIDRs junto con su media funcional muestral. Más concretamente, la media funcional muestral es la línea más gruesa que parece resumir razonablemente bien el comportamiento más común de las curvas, creciendo en las primeras horas de la mañana y últimas de la tarde y decreciendo en el resto de momentos del día. Por otro lado, el gráfico de la izquierda de la figura 5 muestra la desviación típica muestral de las CIDRs, es decir, la raíz cuadrada de la varianza funcional muestral en cada tiempo $t \in [0, 24]$. Como se puede observar, la variabilidad alcanza máximos alrededor de las 5:00 horas de la mañana, las 18:00 horas de la tarde y a última hora del día. El cálculo de la media funcional muestral y la varianza funcional muestral se ha realizado con las funciones `mean.fd` y `var.fd`, respectivamente, de la librería `fda` de R.

Más información sobre la relación de las CIDRs en diferentes momentos del día se puede obtener a partir de la covarianza funcional muestral dada por

$$\hat{\gamma}_0(t, s) = \frac{1}{2099} \sum_{i=1}^{2100} (\theta_i(t) - \bar{\theta}(t))(\theta_i(s) - \bar{\theta}(s)), \quad t, s \in [0, 24) \quad [8]$$

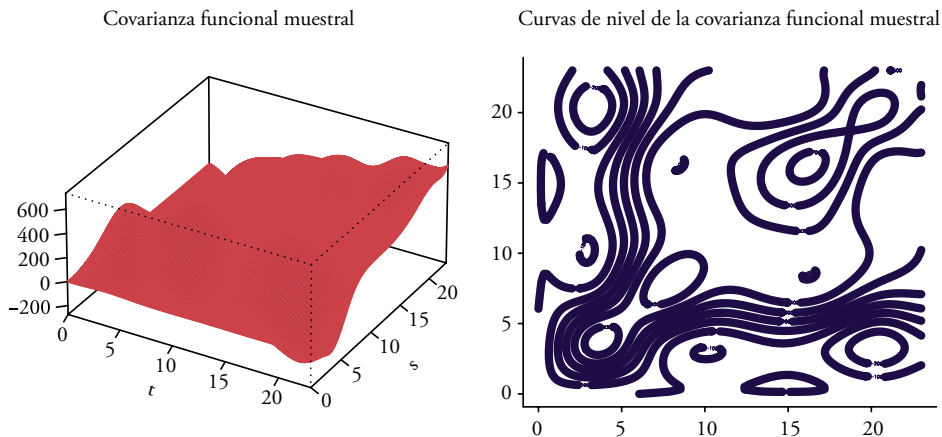
que proporciona la covarianza existente entre los valores que toman las CIDRs en dos tiempos cualesquiera $t, s \in [0, 24)$. Valores altos de $\hat{\gamma}_0(t, s)$ indican que los valores de las CIDR en los tiempos t y s tienden a estar simultáneamente por encima o por debajo de sus valores medios en dichos tiempos. En particular, nótese que $\gamma_0(t, t) = \hat{\sigma}^2(t)$. Tras substituir la expansión [3] y la expresión [6] en la expresión [8], es fácil comprobar que:

$$\hat{\gamma}_0(t, s) = \beta(t)' S_0 \beta(s), \quad t, s \in [0, 24)$$

lo que permite un cálculo muy rápido del valor de $\hat{\gamma}_0(t, s)$ para cualquier par de tiempos $t, s \in [0, 24)$. El gráfico de la derecha de la figura 6 muestra la covarianza funcional muestral de las CIDRs en el cuadrado $[0, 24) \times [0, 24)$, mientras que el gráfico de la izquierda de la figura 6 muestra las curvas de nivel asociadas a dichas covarianzas. Como se puede observar, la mayoría de las covarianzas son positivas. Si se excluyen las varianzas, las mayores covarianzas se localizan entre las 9:00 horas de la mañana y las 16:00 horas de la tarde. Además, parece haber covarianzas negativas entre las horas previas y posteriores al cambio de día. El cálculo de la covarianza funcional muestral se ha realizado con la función `var.fd` de la librería `fda` de R, mientras que el cálculo de las curvas de nivel asociadas se ha realizado con la función `contour` de la librería `graphics` de R.

Figura 6.

Covarianza funcional muestral para las CIDRs del precio de la electricidad



Fuente: Elaboración propia.

Por último, es de especial interés conocer la relación que existe entre las CIDRs de días diferentes. Se pueden utilizar dos medidas para ello. La primera de ellas es la autocovarianza funcional muestral dada por:

$$\hat{\gamma}_l(t, s) = \frac{1}{2099} \sum_{i=1}^{2100-l} (\theta_{i+l}(t) - \bar{\theta}(t))(\theta_i(s) - \bar{\theta}(s)), \quad t, s \in [0, 24] \quad [9]$$

para $l = 1, 2, \dots$, que proporciona la autocovarianza existente entre los valores que toman las CIDRs separadas l días en dos tiempos cualesquiera $t, s \in [0, 24]$. En este caso, valores altos de $\hat{\gamma}_l(t, s)$ implican que los valores de las CIDRs separadas l días en los tiempos t y s tienden a estar simultáneamente por encima o por debajo de sus valores medios en dichos tiempos. Tras substituir la expansión [3] y la expresión [6] en la expresión [9], es fácil comprobar que:

$$\hat{\gamma}_l(t, s) = \beta(t)' S_l \beta(s), \quad t, s \in [0, 24] \quad [10]$$

donde $S_l = \frac{1}{2099} \sum_{i=1}^{2100-l} (c_{i+l} - \bar{c})(c_i - \bar{c})'$ es la matriz de autocovarianzas muestral de orden l de los vectores de coeficientes $c_1, \dots, c_{2100}$. Si bien analizar las autocovarianzas muestrales puede resultar muy informativo, puede ser a su vez complicado determinar hasta que punto existe una dependencia fuerte o débil entre las CIDRs en días diferentes. Por ello, es conveniente definir autocorrelaciones funcionales muestrales a través de las autocovarianzas funcionales muestrales en [10]. La opción más popular es la propuesta por Horváth, Rice y Whipple (2016), quienes definieron el coeficiente de autocorrelación funcional dado por:

$$\hat{r}_l = \frac{\left(\int_0^{24} \int_0^{24} \hat{\gamma}_l(t, s)^2 ds dt \right)^{1/2}}{\int_0^{24} \hat{\sigma}^2(t) dt}, \quad l = 1, 2, \dots \quad [11]$$

tal que $\hat{r}_l \in [0, 1]$. Como se puede ver, el coeficiente $\hat{r}_l$ en [11] es la ratio entre las magnitudes de $\hat{\gamma}_l$ y $\hat{\sigma}^2$. Por lo tanto, $\hat{r}_l$ no está definido para un par de tiempos (t, s) , como la autocovarianza funcional en [9], si no que se trata de un valor entre 0 y 1 libre de escala y que mide la autocorrelación entre las CIDRs para el retardo l . Como en el caso de series temporales univariantes, valores de $\hat{r}_l$ cercanos a 1 indican autocorrelaciones fuertes y valores de r_l cercanos a 0 indican autocorrelaciones débiles. El conjunto de valores $\hat{r}_1, \hat{r}_2, \dots$, forma la llamada función de autocorrelación funcional, que es el equivalente a la función de autocorrelación en el análisis de series temporales univariantes, y que es útil para identificar qué retardos de la serie de CIDRs están asociados a autocorrelaciones elevadas. Substituyendo en [11] los valores de $\hat{\sigma}^2(t)$ y $\hat{\gamma}_l(t, s)$ en términos del vector $\beta(t)$ y las matrices S_0 y S_l en [7] y [10], respectivamente, es relativamente sencillo comprobar que:

$$\hat{r}_l = \frac{\text{Tr}(S_l B S_l' B)^{1/2}}{\text{Tr}(S_0 B)}, \quad [12]$$

donde $B = \int_0^{24} \beta(t) \beta(t)' dt$ es la matriz de dimensión 10×10 con elementos:

$$B_{mk} = \int_0^{24} \beta_m(t) \beta_k(t) dt = \langle \beta_m, \beta_k \rangle, \quad m, k = 1, \dots, 10 \quad [13]$$

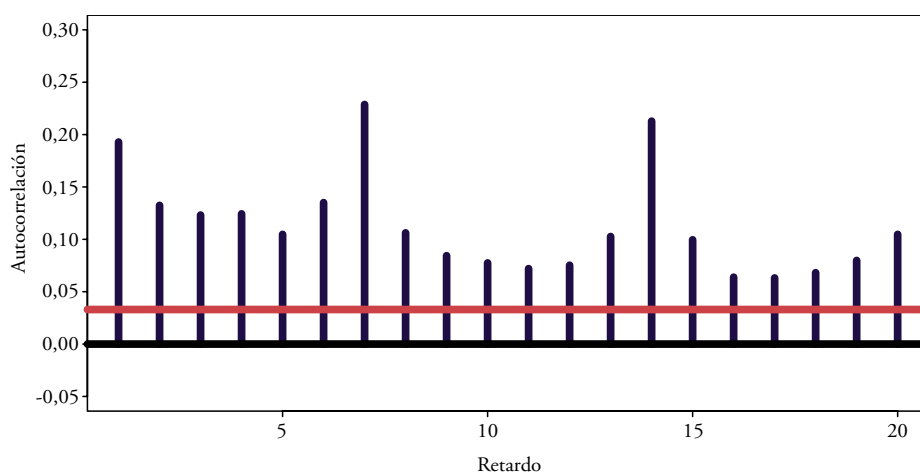
es decir, el producto interior entre las funciones base β_m y β_k . En particular, si $m = k$, entonces:

$$B_{mm} = \int_0^{24} \beta_m(t)^2 dt = \langle \beta_m, \beta_m \rangle = \|\beta_m\|^2, \quad m = 1, \dots, 10 \quad [14]$$

es decir, la norma al cuadrado de la función base β_m . La función `inprod` de la librería `fda` de R proporciona el valor de la matriz B tras realizar integración numérica de las integrales en [13] y [14] mediante la regla trapezoidal, por lo que el usuario no tiene que realizar el cálculo de ninguna integral. El proceso de cálculo es extremadamente rápido por lo que el coste computacional es muy bajo. En particular, la figura 7 muestra la función de autocorrelación funcional muestral en [12] para las CIDRs de los precios de la electricidad. Más concretamente el gráfico muestra los valores de r_l , para $l = 1, \dots, 20$, junto con una línea a la altura del intervalo de confianza al 95 % de dichas autocorrelaciones bajo la hipótesis de que la serie es ruido blanco funcional. El gráfico informa de dos aspectos interesantes. En primer lugar, existen autocorrelaciones significativas lo que parece indicar que las CIDRs para los precios de la electricidad pueden ser predecibles a partir de las CIDRs previas. En cualquier caso, las autocorrelaciones no son muy altas, por lo que las predicciones pueden no ser extremadamente buenas. En segundo lugar, las autocorrelaciones para $l = 7$ y $l = 14$ son significativas, indicando que existe un cierto efecto estacional debido a que las curvas representan los días de la semana. En cualquier caso, este efecto puede no parece ser lo suficientemente grande como para que el proceso generador sea no estacionario, ya que ninguna de estas autocorrelaciones está cerca de 1.

Figura 7.

Función de autocorrelación funcional muestral para las CIDRs del precio de la electricidad



Fuente: Elaboración propia.

4. COMPONENTES PRINCIPALES FUNCIONALES DE LAS CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

Todas las medidas definidas en la sección tercera resumen las características más importantes de la serie temporal funcional de CIDRs, $\theta_1, \dots, \theta_{2100}$. Sin embargo, la herramienta más útil para entender el comportamiento de las CIDRs y para realizar predicciones de futuras CIDRs es el análisis de componentes principales funcionales (FPCA, es su acrónimo en inglés). FPCA es el análogo funcional de la conocida técnica de reducción de dimensión en el análisis estadístico multivariante y aquí será útil para determinar las tendencias comunes que están presentes en la dinámica de las CIDRs. Más concretamente, los componentes principales funcionales (FPCs) son funciones que resumen las principales fuentes de variación de las CIDRs con respecto a la media funcional muestral en [6]. Más concretamente, el FPCA permite representar las CIDRs mediante una expansión de la forma,

$$\theta_i(t) = \bar{\theta}(t) + \sum_{m=1}^{10} s_{im} \psi_m(t), \quad t \in [0, 24)$$

donde los FPC, $\psi_1, \dots, \psi_{10}$, son funciones ortogonales suaves de norma uno, y las puntuaciones asociadas a los FPC, $s_{im} = \langle \theta_i - \bar{\theta}, \psi_m \rangle$, para $m = 1, \dots, 10$, son valores reales de media muestral 0, y de tal manera que cada CIDR se puede aproximar con mínima pérdida de información mediante los primeros $M < 10$ FPC como sigue:

$$\theta_i(t) \simeq \bar{\theta}(t) + \sum_{m=1}^M s_{im} \psi_m(t). \quad [15]$$

La expresión [15] muestra que $\psi_1, \dots, \psi_M$ representan las tendencias comunes que dominan el comportamiento diario de las CIDRs.

Para calcular los FPC se procede de manera similar a como se calculan sus homólogos multivariantes. En particular, el primer FPC se define como la función continua y suave ψ_1 definida en el intervalo $[0, 24)$ tal que la varianza de sus puntuaciones asociadas dada por,

$$\frac{1}{2099} \sum_{i=1}^{2100} s_{i1}^2$$

sea máxima sujeto a que $\|\psi_1\| = 1$. Para resolver el problema, nótese que, por un lado, dadas las expansiones [3] y [6], la i -ésima CIDR centrada $\theta_i - \bar{\theta}$ se puede escribir como:

$$\theta_i(t) - \bar{\theta}(t) = \tilde{c}_i \beta(t), \quad t \in [0, 24) \quad [16]$$

donde $\tilde{c}_i = c_i - \bar{c}$, para $i = 1, \dots, 2100$, mientras que, por otro lado, la primera FPC ψ_1 se puede escribir en términos de la base de B-splines mediante la expansión:

$$\psi_1(t) = \sum_{m=1}^{10} v_{1m} \beta_m(t) = v'_1 \beta(t), \quad t \in [0, 24) \quad [17]$$

donde $v_1 = (v_{11}, \dots, v_{1,10})'$ es un vector de coeficientes a determinar. Por lo tanto, substituyendo [16] y [17] en la definición de las puntuaciones de la primera FPC, se obtiene que,

$$s_{i1} = \langle \theta_i - \bar{\theta}, \psi_1 \rangle = \tilde{c}_i' B v_1,$$

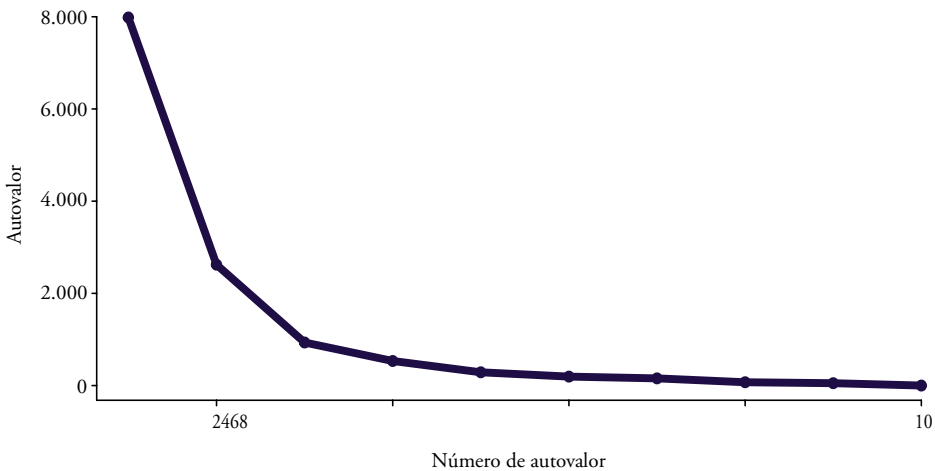
donde B es la matriz de productos interiores de los elementos de la base de B-splines $\beta_1, \dots, \beta_{10}$ introducida en la sección tercera, dando lugar a que la varianza de las puntuaciones se puedan escribir de manera sencilla mediante la expresión:

$$\frac{1}{2099} \sum_{i=1}^{2100} s_{i1}^2 = v_1' B S_0 B v_1.$$

Además, tras substituir [17] en la definición de $\|\psi_1\|$, se obtiene que la restricción se puede escribir como $\|\psi_1\|^2 = v_1' B v_1 = 1$. En resumen, el problema de optimización se reduce a obtener el vector v_1 tal que el valor de $v_1' B S_0 B v_1$ sea máximo sujeto a que $v_1' B v_1 = 1$, cuya solución viene dada por $v_1 = B^{1/2} d_1$, donde d_1 el autovector ligado al mayor autovalor a_1 de la matriz $B^{1/2} S_0 B^{1/2}$. De esta manera, se obtiene el vector de coeficientes v_1 en términos de la base de B-splines del primer FPC en [17]. Nótese además que las puntuaciones asociadas a la primera FPC tiene media muestral 0 y varianza muestral a_1 .

Figura 8.

Autovalores asociados a los componentes principales funcionales de las CIDRs para los precios de la electricidad



Fuente: Elaboración propia.

El segundo FPC se define como la función continua y suave ψ_2 definida en el intervalo $[0, 24)$ tal que la varianza de sus puntuaciones asociadas sea máxima sujeto a que $\|\psi_2\| = 1$ y a que ψ_2 sea ortogonal a ψ_1 , es decir, que $\langle \psi_1, \psi_2 \rangle = 0$. Como en el caso anterior, la segunda FPC ψ_2 se puede escribir en términos de la base de B-splines mediante la expansión:

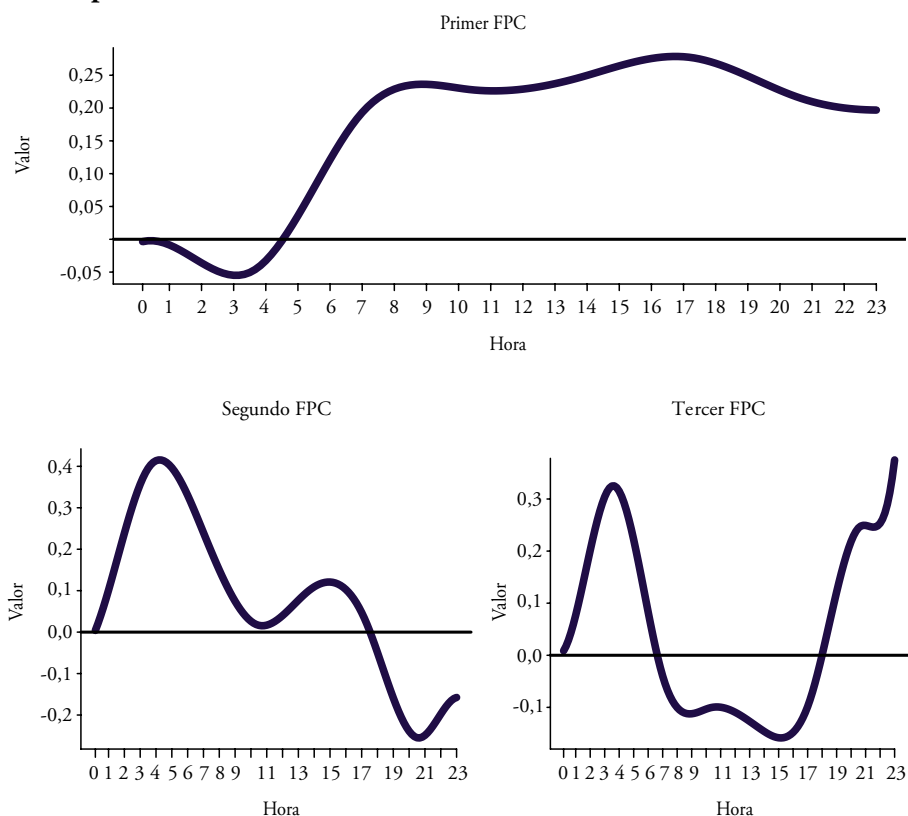
$$\psi_2(t) = \sum_{m=1}^{10} v_{2m} \beta_m(t) = v_2' \beta(t), \quad t \in [0, 24) \quad [18]$$

donde $v_2 = (v_{2,1}, \dots, v_{2,10})'$ es un vector de coeficientes a determinar. Procediendo de manera similar al cálculo del primer FPC, se llega a que la solución al problema de optimización viene dada por que $v_2 = B^{1/2}d_2$ es el vector de coeficientes del segundo FPC en [18], donde d_2 es el autovector ligado al segundo mayor autovalor a_2 de la matriz $B^{1/2}S_0B^{1/2}$. Como es previsible, el resto de FPC se obtiene de manera similar por lo que los FPC de las CIDRs son las funciones $\psi_1, \dots, \psi_{10}$ que tienen expansiones de B-splines con vectores de coeficientes $v_m = B^{-1/2}d_m$, para $m = 1, \dots, 10$, donde $d_1, \dots, d_{10}$ son los autovectores de la matriz $B^{1/2}S_0B^{1/2}$ asociados a los autovalores $a_1, \dots, a_{10}$. Además las puntuaciones del m-ésima FPC tienen media muestral 0 y varianza muestral a_m , lo que permite utilizar el criterio clásico de acumulación de varianzas para seleccionar un valor apropiado de M .

La figura 8 muestra los 10 autovalores $a_1, \dots, a_{10}$ asociados a los FPC de las CIDRs de tal manera que la acumulación de las varianzas del primero hasta el quinto FPC es capaz de

Figura 9.

Tres primeros componentes principales funcionales de las CIDRs para los precios de la electricidad



Fuente: Elaboración propia.

explicar el 62,21 %, 82,65 %, 89, 92 %, 94,07 % y 96,32 %, respectivamente, de la suma de las varianzas de todos los FPC. A la vista de la figura y de estos porcentajes, los primeros $M = 3$ FPC, que se muestran en la figura 9, proporcionan las principales fuentes de variación de las CIDRs. Los tres FPC parecen tener una relación evidente con varios momentos del día. Por ejemplo, el primer FPC está fundamentalmente relacionado con la variabilidad creada desde las 7:00 horas de la mañana hasta las 23:00 horas, es decir, el período de mayor actividad. A continuación, el segundo FPC contrapone la variabilidad creada durante la madrugada y la variabilidad creada a última hora de la tarde. De hecho, en la covarianza funcional muestral de las CIDRs que se muestra en la figura 6, se puede observar una relación negativa entre ambos momentos del día. Por último, el tercer FPC contrapone la variabilidad creada de madrugada y a última hora del día con la variabilidad creada desde la mañana hasta media tarde, que es el período ligado a la jornada laboral más habitual. El cálculo de los FPC de las CIDRs se ha llevado a cabo mediante la función `pca.fd` de la librería `fda` de R.

5. PREDICCIÓN DE LAS CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

Una vez descritas las características más importantes de las CIDRs como serie temporal funcional, el objetivo de esta sección es realizar predicción para las mismas. Existen dos opciones principales para la predicción de series temporales funcionales: (1) mediante los modelos autorregresivos y de media móvil funcional; y (2) mediante los procedimientos basados en componentes principales funcionales.

Los modelos autorregresivos y de media móvil funcionales son extensiones directas de los modelos autorregresivos y de media móvil univariantes y multivariantes. Dada la popularidad de este tipo de modelos para el análisis de series temporales univariantes y multivariantes, no es de extrañar que sus extensiones fuesen los primeros modelos que fueron utilizados para el análisis de series temporales funcionales. Más concretamente, Bosq (1991) propuso el modelo autorregresivo funcional de orden 1, denotado por $\text{FAR}(1)$, que fue extendido al modelo autorregresivo funcional de orden p , denotado por $\text{FAR}(p)$, por Bosq (2000) y Bosq y Blanke (2007). Más recientemente, Klepsch y Klüppelberg (2016), Klepsch, Klüppelberg y Wei (2017) y Li, Robinson y Shang (2020) han propuesto, respectivamente, el modelo de media móvil funcional, denotado por $\text{FMA}(q)$, el modelo autorregresivo y de media móvil funcional, denotado por $\text{FARMA}(p, q)$, y el modelo autorregresivo fraccionalmente integrado y de media móvil funcional, denotado por $\text{FFARIMA}(p, d, q)$. Expresiones, características y propiedades de todos estos modelos se pueden encontrar en los artículos mencionados y no se presentan aquí por su elevada complejidad técnica. De entre todos los modelos anteriores, el más popular en la práctica para el ajuste de series temporales funcionales es el modelo $\text{FAR}(1)$. Esto se debe a dos razones fundamentales. En primer lugar, los modelos $\text{FAR}(p)$ han sido analizados y utilizados desde hace alrededor de dos décadas (véanse las referencias mencionadas en la introducción), mientras que los modelos $\text{FMA}(q)$, $\text{FARMA}(p, q)$ y $\text{FFARIMA}(p, d, q)$ han sido propuestos muy recientemente. Por lo tanto, las ventajas del uso de estos modelos frente al uso de los modelos $\text{FAR}(p)$ no son claras todavía. Se menciona este punto porque, tal como

Kokoszka y Reimherr (2013) hicieron notar, al contrario que la autorregresión escalar y vectorial, solo valores muy pequeños de p , 1 o 2 como mucho, son relevantes en un modelo $\text{FAR}(p)$. Esto es debido a que cada una de las curvas se considera un único objeto funcional y, por lo tanto, la dependencia entre funciones estacionarias temporalmente alejadas debe ser escasa o nula, salvo que exista algún tipo de dependencia a muy largo plazo, quizás debida a algún tipo de efecto estacional. En segundo lugar, este tipo de modelos se define a partir de operadores complejos que requieren complicadas condiciones de existencia. Más aún, la estimación de dichos operadores está basada en formas regularizadas de las ecuaciones de Yule-Walker que son relativamente sencillas de manejar solamente en el caso del modelo $\text{FAR}(1)$, por lo que utilizar este modelo permite simplificar el procedimiento inferencial. Por estas y otras razones que se detallarán más adelante, los procedimientos basados en componentes principales funcionales son actualmente más populares para la obtención de predicciones de series temporales funcionales reales (véanse de nuevo las referencias mencionadas en la introducción). De hecho, para llevar a cabo predicciones de las CIDRs, que tienen un cierto efecto estacional, parece más conveniente el uso de este tipo de procedimientos que se presentan a continuación.

La predicción con componentes principales funcionales es bastante sencilla, estando a la vez también ligada, como los modelos autorregresivos y de media móvil funcionales, al uso de métodos para la predicción de series temporales univariantes y multivariantes. La idea de este método está basada en utilizar la expansión [15] de las CIDRs en términos de los primeros FPC. Más concretamente, dicha expansión permite escribir las CIDRs como:

$$\theta_i(t) = \bar{\theta}(t) + s_i' \psi(t) + \varepsilon_i(t), \quad [19]$$

donde $s_i = (s_{i1}, s_{i2}, s_{i3})'$ es el vector que contiene las puntuaciones de la i -ésima CIDR asociadas a las tres primeras FPC, $\psi(t) = (\psi_1(t), \psi_2(t), \psi_3(t))'$ es el vector de FPC evaluados en tiempo $t \in [0, 24)$, y ε_i es una función de media funcional muestral cero que no debe contener información útil de cara a la predicción de las CIDRs futuras. Ahora, dada la serie de CIDRs $\theta_1, \dots, \theta_{2100}$, el objetivo es obtener la predicción de la serie funcional de CIDRs a un día vista mediante la expansión [19]. En general, la CIDR correspondiente al día siguiente del último observado, es decir θ_{2100} , se puede predecir mediante:

$$\hat{\theta}_{2101}(t) = \bar{\theta}(t) + \hat{s}_{2101}' \psi(t), \quad [20]$$

donde $\hat{s}_{2101} = (\hat{s}_{2101,1}, \hat{s}_{2101,2}, \hat{s}_{2101,3})'$ es el vector de predicciones de las puntuaciones asociadas a los tres primeros FPC de las CIDRs. Por lo tanto, el problema de predecir la CIDR θ_{2100} se transforma en el problema de predecir el vector de puntuaciones dado por $s_{2101} = (s_{2101,1}, s_{2101,2}, s_{2101,3})'$. El resto de la sección se dedica a presentar dos maneras diferentes de obtener la predicción $\hat{s}_{2101}$ y, por tanto, de obtener la predicción $\hat{\theta}_{2101}$ en [20]. Además se presenta un procedimiento sencillo para obtener una banda de confianza para dicha predicción funcional.

En primer lugar, el procedimiento propuesto por Hyndman y Ullah (2007) y posteriormente analizado y utilizado en Hyndman y Shang (2009), Shen (2009) y Shang (2017), entre otros, consiste en obtener las predicciones $\hat{s}_{2101,1}$, $\hat{s}_{2101,2}$ y $\hat{s}_{2101,3}$, de manera

independiente. Es decir, las predicciones $\hat{s}_{2101,1}$, $\hat{s}_{2101,2}$ y $\hat{s}_{2101,3}$ resultan de predecir a un día vista las tres series temporales univariantes de puntuaciones $s_{1m}, \dots, s_{2100,m}$ para $m = 1, 2$ y 3 . Esta decisión es justificable en el sentido de que las puntuaciones de los FPC son incorreladas. En principio, para realizar predicciones univariantes, se puede emplear cualquier procedimiento apropiado para las características de las series de puntuaciones. Si las series de puntuaciones se pueden ajustar bien mediante modelos ARIMA, lo normal sería utilizar los procedimientos automáticos de predicción con modelos ARIMA univariantes donde el modelo se selecciona mediante un criterio de selección de modelos, como el criterio de Akaike (AIC), el criterio de Akaike corregido (AICc) o el criterio Bayesiano (BIC), ver Brockwell y Davis (2017), Shumway y Stoffer (2017) y Hyndman y Athanasopoulos (2018), para referencias sobre este tipo de procedimientos. Sin embargo, si las series de puntuaciones son relativamente complejas y los modelos ARIMA no son adecuados, se pueden utilizar otros procedimientos como pueden ser los diferentes métodos de alisado exponencial, modelos de series temporales no lineales, redes neuronales, o cualquier otra alternativa.

En segundo lugar, el procedimiento propuesto por Aue, Norinho y Hörmann (2015), consiste en obtener la predicción del vector $s_{2101} = (\hat{s}_{2101,1}, \hat{s}_{2101,2}, \hat{s}_{2101,3})'$, de manera conjunta. Es decir, se obtiene la predicción a un día vista de la serie temporal trivariante $s_i = (s_{i1}, s_{i2}, s_{i3})'$, donde $i = 1, \dots, 2100$. Esta decisión también es justificable ya que, si bien las puntuaciones de los FPC son incorreladas, pueden existir autocovarianzas cruzadas entre diferentes puntuaciones diferentes de 0. Para realizar las predicciones multivariantes, Aue, Norinho y Hörmann (2015) proponen utilizar modelos vectoriales autorregresivos (VAR) para los que también existen procedimientos automáticos de predicción basados en criterios de selección de modelos, véase, por ejemplo, Lütkepohl (2006) y Tsay (2017).

Una vez presentados los dos procedimientos basados en FPC, es importante resaltar dos ventajas adicionales que tienen estos métodos con respecto a los métodos basados en modelos ARMA funcionales. La primera ventaja está basada en que Aue, Norinho y Hörmann (2015) demostraron que las predicciones obtenidas con su procedimiento y utilizando un modelo FAR(p) son asintóticamente equivalentes. Por lo tanto, dado que la obtención de las predicciones con FPC es más sencilla, parece entonces razonable utilizar este método. La segunda ventaja está basada en que Liebl (2013) demostró que la descomposición en [15] es perfectamente válida incluso si la serie temporal funcional es no estacionaria. Por lo tanto, las predicciones con FPC se pueden utilizar sin mayor complicación en el caso no estacionario. Más detalles sobre este último punto se pueden encontrar en Lansangan y Barrios (2009) y Shen (2009).

En series temporales univariantes y multivariantes, los intervalos y las regiones de predicción, respectivamente, juegan un papel fundamental para medir la incertidumbre asociada a las predicciones puntuales. Además, son útiles para comparar predicciones de diferentes métodos y explorar diferentes escenarios basados en diferentes supuestos. En el caso de series temporales funcionales, los intervalos o regiones de predicción se substituyen con bandas de predicción que tienen en cuenta la estructura funcional de dicha predicción. Existen diferentes maneras de obtener estas bandas de predicción. Por un lado, Aue, Norinho y Hörmann (2015) propusieron un método simple basado en predicciones

dentro de la muestra. Por otro lado, Shang (2017) y Paparoditis y Shang (2020) propusieron procedimientos bootstrap para obtener dichas bandas. Por sencillez en la exposición y dado que los procedimientos bootstrap son computacionalmente más costosos, a continuación, se describe el procedimiento propuesto por Aue, Norinho y Hörmann (2015) adaptado a la obtención de una banda de predicción a un paso para las CIDRs de los precios de la electricidad en España. Para ello, se dispone de los FPC ψ_1, ψ_2, ψ_3 , de sus series de puntuaciones asociadas $s_{1m}, \dots, s_{2100,m}$, para $m = 1, 2, 3$, y de un procedimiento para las predicciones univariantes para las series de puntuaciones individuales o de un procedimiento para la predicción del vector de series de puntuaciones. El procedimiento funciona siguiendo los cinco pasos que se describen a continuación:

1. Utilizar el procedimiento de predicción univariante o multivariante para construir predicciones a un paso de las CIDRs $\theta_L, \dots, \theta_{2100}$, donde L es un cierto entero pequeño.
2. Obtener residuos funcionales basados en las predicciones anteriores, $\hat{\theta}_L, \dots, \hat{\theta}_{2100}$, dados por $\hat{\varepsilon}_i = \theta_i - \hat{\theta}_i$, para $i = L+1, \dots, 2100$.
3. Obtener la desviación típica funcional muestral de los residuos funcionales dada por:

$$\hat{\sigma}_{\varepsilon}(t) = \sqrt{\frac{1}{2100 - L - 1} \sum_{i=L}^{2100} \left(\hat{\varepsilon}_i(t) - \bar{\varepsilon}(t) \right)^2},$$

donde $\bar{\varepsilon}$ denota la media funcional muestral de $\hat{\varepsilon}_L, \dots, \hat{\varepsilon}_{2100}$.

4. Encontrar dos valores q_L y q_U tales que el $\alpha \times 100\%$ de los residuos funcionales verifiquen:

$$q_L \hat{\sigma}_{\varepsilon}(t) \leq \hat{\varepsilon}_i(t) \leq q_U \hat{\sigma}_{\varepsilon}(t), \quad t \in [0, 24].$$

5. La banda de confianza para la predicción $\hat{\theta}_{2100}$ está dada por:

$$\hat{\theta}_{2101}(t) + q_L \hat{\sigma}_{\varepsilon}(t) \leq \hat{\theta}_{2101}(t) \leq \hat{\theta}_{2101}(t) + q_U \hat{\sigma}_{\varepsilon}(t), \quad t \in [0, 24].$$

A continuación, se considera la predicción de las CIDRs a un paso utilizando los procedimientos basados en FPC descritos previamente. Para ello, se realizan dos ejercicios diferentes. En el primer ejercicio, se obtienen predicciones funcionales para las últimas 100 CIDRs observadas basadas en las CIDRs previas como sigue:

1. Para cada $i = 2001, \dots, 2100$, se utilizan las primeras $i-1$ CIDRs para obtener los primeros tres FPC y las series de puntuaciones asociadas.
2. Se aplican los procedimientos de Hyndman y Ullah (2007) y de Aue, Norinho y Hörmann (2015) para obtener sendas predicciones a un paso de la CIDR el día de negociación i utilizando la expresión [20].
3. Para cada predicción obtenida se calculan el error cuadrático integrado medio de predicción (MISFE) y el error absoluto integrado medio de predicción (MIAFE) dados por:

$$MISFE_i = \int_0^{24} \left(\theta_i(t) - \hat{\theta}_i(t) \right)^2 dt,$$

y

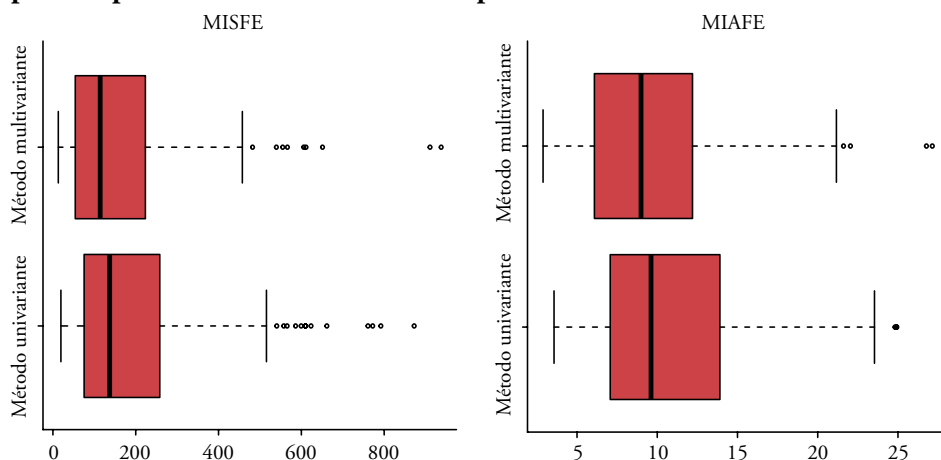
$$MIAFE_i = \int_0^{24} \left| \theta_i(t) - \hat{\theta}_i(t) \right| dt,$$

respectivamente, que permitirán determinar que método de predicción es más apropiada para cada CIDR $\theta_{2001}, \dots, \theta_{2100}$, y más en general para las 100 CIDRs en su conjunto.

Por un lado, para obtener las predicciones con el método de Hyndman y Ullah (2007), se utiliza la función `forecast.fsm` de la librería `fts` de R, que permite realizar las predicciones de las series de puntuaciones mediante una selección automática de modelos ARIMA, modelos de espacio de estado para suavizado exponencial o paseos aleatorios con deriva. Para las CIDRs, se ha seleccionado la predicción con modelos ARIMA. Por otro lado, para obtener las predicciones con el método de Aue, Norinho y Hörmann (2015), se utiliza la función `farforecast` de la misma librería, que solo permite realizar las predicciones del vector de series de puntuaciones mediante una selección automática de modelos VAR. Los 100 valores obtenidos de ambas medidas del error de predicción se pueden observar en la figura 10, que muestra que el procedimiento de Aue, Norinho y Hörmann (2015) es ligeramente superior al procedimiento de Hyndman y Ullah (2007), como se esperaba. De hecho, las medias muestrales de los valores del criterio MISFE son 216.7 y 178.6, respectivamente, para los procedimientos de Hyndman y Ullah (2007) y Aue, Norinho y Hörmann (2015), mientras que las medias respectivas para el criterio MIAFE son 11.129 y 9.867, respectivamente.

Figura 10.

Error cuadrático integrado medio de predicción (MISFE) y error absoluto integrado medio de predicción (MIAFE) para las 100 últimas CIDRs para los precios de la electricidad en España

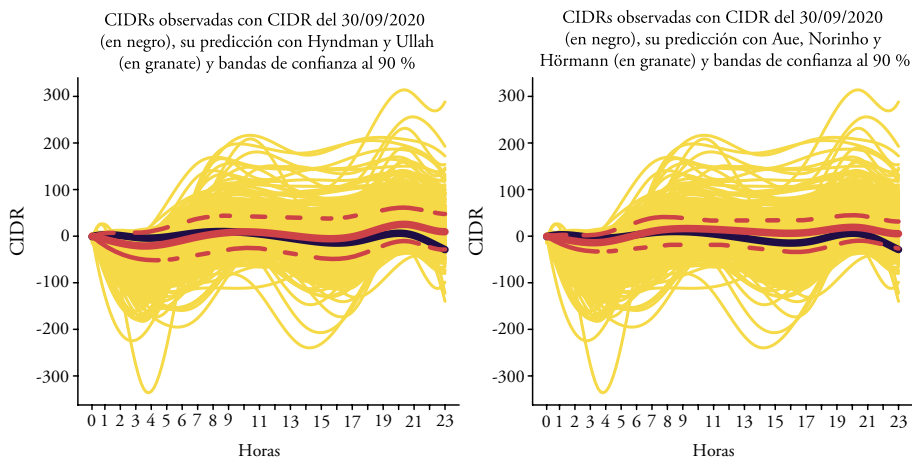


Fuente: Elaboración propia.

En el segundo ejercicio, se obtiene la predicción funcional y la banda de predicción asociada para la última CIDR observada correspondiente al 30 de septiembre de 2020 con los dos procedimientos considerados en el ejercicio anterior. Los valores del MIAFE para la predicción con el procedimiento de Hyndman y Ullah (2007) y con el procedimiento de Aue, Norinho y Hörmann (2015) son 12.220 y 12.513, respectivamente, que ocupan los lugares 65 y 78 de entre los valores del MIAFE obtenidos para los últimos 100 días. Por lo tanto, las predicciones obtenidas no son de las mejores de entre el conjunto de predicciones de los últimos 100 días observados. Más concretamente, los valores del criterio MISFE para esta CIDR son 228.7 y 221,5, respectivamente, para los procedimientos de Hyndman y Ullah (2007) y Aue, Norinho y Hörmann (2015), mientras que los valores respectivos para el criterio MIAFE son 12.22 y 12.51, respectivamente. Esto indica que ninguna de las dos predicciones para la última CIDR domina a la otra. La figura 11 muestra la verdadera CIDR correspondiente al 30 de septiembre de 2020 (en negro), las predicciones con los dos métodos utilizados (en granate), una en cada gráfico de la figura, sus correspondientes bandas de predicción asociadas al 90 % (en granate y línea discontinua), y el resto de CIDRs (en amarillo). Como se puede apreciar, las dos predicciones son bastante parecidas, por lo que no sorprende los valores parecidos obtenidos de los criterios MIAFE y MISFE, y bastante cercanas a la verdadera CIDR teniendo en cuenta los valores del resto de CIDRs. En ambos casos, la banda de predicción correspondiente contiene a toda la CIDR del 30 de septiembre de 2020 durante todo el día.

Figura 11.

A la derecha, predicción funcional para la CIDR del 30 de septiembre de 2020 junto con una banda de predicción asociada para el método de Hyndman y Ullah (2007), y a la izquierda, predicción funcional para la CIDR del 30 de septiembre de 2020 junto con una banda de predicción asociada para el método de Aue, Norinho y Hörmann (2015)



Fuente: Elaboración propia.

6. CONCLUSIONES

Este capítulo presenta una aplicación sencilla del análisis de series temporales funcionales a los precios horarios de la electricidad de España en el mercado MIBEL. Más concretamente, el análisis descrito analiza las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad. El objetivo fundamental del capítulo es el de llamar la atención sobre un conjunto de técnicas que están apareciendo en el área del análisis de datos funcionales y que están despertando el interés de un número importante de estadísticos y de profesionales en otras áreas. Evidentemente, todavía hay un gran espacio para la mejora y para el desarrollo de técnicas más completas y que cubran aspectos relevantes como, por ejemplo, el análisis conjunto de varias series temporales funcionales. Además, el análisis de las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad puede ser más completo teniendo en cuenta aspectos que no se describen en este artículo meramente introductorio.

Referencias

- ANTONIADIS, A., PAPARADOTIS, E. y SAPATINAS, T. (2006). A functional wavelet-kernel approach for time series prediction. *Journal of the Royal Statistical Society, Series B*, 68, pp. 837–857.
- AUE, A., HORVÁTH, L. y PELLATT, D. F. (2017). Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis*, 38, pp. 3–21.
- AUE, A., NORINHO, D. D. y HÖRMANN, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110, pp. 378–392.
- BESSE, P. C., CARDOT, H. y STEPHENSON, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27, pp. 673–687.
- BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, pp. 307–327.
- BOSQ, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. En G. ROUSSAS (eds.), *Nonparametric Functional Estimation and Related Topics*. NATO ASI Series (Series C: Mathematical and Physical Sciences), 335, pp. 509–529.
- BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Berlin: Springer.
- BOSQ, D. y BLANKE, D. (2007). *Inference and Prediction in Large Dimensions*. Chichester: Wiley.
- BROCKWELL, P. J. y DAVIS, R. A. (2017). *Introduction to Time Series and Forecasting* (third edition). New York: Springer.
- DICKEY, D. A. y FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, pp. 427–431.
- ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, pp. 987–1007.
- ENGLE, R. F. y GRANGER, C. W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica*, 55, pp. 251–276.
- FERRATY, F. y VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.
- GABRYS, R., HORVÁTH, L. y KOKOSZKA, P. (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, 105, pp. 1113–1125.

- GALEANO, P. y PEÑA, D. (2019). Data science, big data and statistics. *Test*, 28, pp. 289–329.
- GRANGER, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 23, pp. 121–130.
- HAYS, S., SHEN, H. y HUANG, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *Annals of Applied Statistics*, 6, pp. 870–894.
- HÖRMANN, S., HORVÁTH, L. y REEDER, R. (2013). A functional version of the ARCH model. *Econometric Theory*, 29, pp. 267–288.
- HÖRMANN, S. y KOKOSZKA, P. (2010). Weakly dependent functional data. *Annals of Statistics*, 38, pp. 1845–1884.
- HORVÁTH, L. y KOKOSZKA, P. (2012). *Inference with Functional Data with Applications*. New York: Springer.
- HORVÁTH, L., KOKOSZKA, P. y RICE, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, 179, pp. 66–82.
- HORVÁTH L., RICE, G. y WHIPPLE, S. (2016). Adaptive bandwidth selection in the estimation of the long run covariance of functional time series. *Computational Statistics & Data Analysis*, 100, pp. 676–693.
- HSING, T. y EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. West Sussex: John Wiley & Sons.
- HYNDMAN, R. J. y ATHANASOPOULOS, G. (2018). *Forecasting: Principles and Practice* (second edition). Melbourne: OTexts.
- HYNDMAN, R. J. y SHANG, H. L. (2009). Forecasting functional time series (with discussions). *Journal of the Korean Statistical Society*, 38, pp. 199–221.
- HYNDMAN, R. J. y ULLAH, S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51, pp. 4942–4956.
- KLEPSCH, J. y KLÜPPELBERG, C. (2016). An innovations algorithm for the prediction of functional linear processes. *Journal of Multivariate Analysis*, 155, pp. 252–271.
- KLEPSCH, J., KLÜPPELBERG, C. y WEI, T. (2017). Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics*, 1, pp. 128–149.
- KOKOSZKA, P., MIAO, H. y ZHANG, X. (2015). Functional dynamic factor model for intraday price curves. *Journal of Financial Econometrics*, 13, pp. 456–477.
- KOKOSZKA, P. y REIMHERR, M. (2013). Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34, pp. 116–129.
- KOKOSZKA, P. y REIMHERR, M. (2017). *Introduction to Functional Data Analysis*. CRC press, Boca Ratón.
- KOWAL, D. R., MATTESON, D. S. y RUPPERT, D. (2017). A Bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, 112, pp. 733–744.
- KWIATKOWSKI, D., PHILIPS, P. C. B., SCHMIDT, P. y SHIN, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54, pp. 159–178.
- LANSANGAN, J. R. G. y BARRIOS, E. B. (2009). Principal components analysis of nonstationary time series data. *Statistics and Computing*, 19, pp. 173–187.
- LI, D., ROBINSON, P. M. y SHANG, H. L. (2020). Long-range dependent curve time series. *Journal of the American Statistical Association*, 115, pp. 957–971.
- LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *The Annals of Applied Statistics*, 7, pp. 1562–1592.
- LÜTKEPOHL, H. (2006). *New Introduction to Multiple Time Series Analysis*. New York: Springer.
- PAPARODITIS, E. y SHANG, H. L. (2020). Bootstrap prediction bands for functional time series. <https://arxiv.org/abs/2004.03971v1>

- RAMSAY, J. O. y SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd edition. New York: Springer.
- SAID, E. S. y DICKEY, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71, pp. 599–607.
- SARGAN, J. D. y BHARGAVA, A. (1983). Testing residuals from the least squares regression for being generated by the Gaussian random walk. *Econometrica*, 51, pp. 153–174.
- SHANG, H. L. (2017). Forecasting intraday S&P 500 index returns: A functional time series approach. *Journal of Forecasting*, 36, pp. 741–755.
- SHEN, H. (2009). On modeling and forecasting time series of smooth curves. *Technometrics*, 51, pp. 227–238.
- SHUMWAY, R. H. y STOFFER, D. S. (2017). *Time Series Analysis and its Applications* (fourth edition). New York: Springer.
- SRIVASTAVA, A. y KLASSEN, E. (2016). *Functional and Shape Data Analysis*. New York: Springer.
- TAYLOR, S. J. (1986). *Modeling Financial Time Series*. Chichester: Wiley.
- TSAY, R. S. (2017). *Multivariate Time Series Analysis: With R and Financial Applications*. New Jersey: Wiley.