

DETECCIÓN DE ATÍPICOS EN DATOS FUNCIONALES DEPENDIENTES

Paula Raña¹, Juan Manuel Vilar¹ y Germán Aneiros¹

¹Departamento de Matemáticas, Universidade da Coruña.

RESUMEN

En este trabajo se proponen diferentes enfoques para detectar atípicos en datos funcionales, teniendo en cuenta la dependencia en las observaciones. En los procedimientos se utilizan tanto la profundidad funcional como herramientas exploratorias y contrastes de hipótesis. Nuestra propuesta principal sigue el procedimiento de Febrero et al. (2008). Este método asume la hipótesis de independencia entre los datos. En datos obtenidos de una serie de tiempo funcional, debe haber alguna dependencia entre las curvas que debe tenerse en cuenta. Nuestro enfoque modifica a este de dos formas, cambiando algunos pasos para introducir otras técnicas y teniendo en cuenta la dependencia de los datos, con el uso de diferentes herramientas disponibles para datos dependientes. Otros procedimientos para detectar atípicos en datos funcionales bajo dependencia están en estudio. Están basados en realizar una transformación de la serie de tiempo funcional en una serie de tiempo univariante. Se muestran aplicaciones a los datos eléctricos horarios de la demanda y el precio en España para ilustrar el comportamiento de los métodos.

Palabras y frases clave: Datos funcionales, Atípicos, Dependencia, Profundidad funcional, Bootstrap.

1. INTRODUCCIÓN

En algunos contextos los datos deben ser tratados como funciones, en lugar de una cadena de valores. Este tipo de datos, los llamados datos funcionales, aparecen en muchas áreas como meteorología, economía o medicina. El estudio de datos funcionales es un campo reciente en la investigación estadística, en desarrollo durante los últimos años. Para una revisión acerca de este tema se pueden ver las monografías de Ramsay y Silverman (2005) y Ferraty y Vieu (2006).

El estudio de atípicos es importante en cualquier análisis estadístico de datos. Se han estudiado algunas técnicas para detectar atípicos en datos funcionales utilizando la profundidad funcional, basándose en herramientas exploratorias, como Hyndman y Shang (2010), o también en contrastes de hipótesis como Febrero et al. (2008). La profundidad funcional permite ordenar una muestra de curvas desde el cento hacia fuera. La curva con mayor profundidad se corresponde con la mediana, mientras que los atípicos se definen como las curvas de menor profundidad.

Nuestro principal objetivo es detectar atípicos en un conjunto de datos funcionales, considerando la dependencia entre las observaciones. Los métodos presentados se aplican a datos eléctricos, en los que la demanda y el precio de la electricidad forman series de tiempo funcionales. Cada dato funcional se corresponde con un día construido a partir de 24 observaciones horarias. Debido a la dependencia temporal entre las observaciones, no se puede asumir la independencia en los datos.

El trabajo se estructura de la siguiente forma: en la Sección 2 se presentan los métodos propuestos para detectar atípicos funcionales. En la Sección 3 se muestran algunas aplicaciones de los mismos a datos reales del contexto eléctrico, tanto a la demanda como al precio de la electricidad en distintos grupos según el día de la semana. La Sección 4 son las conclusiones del trabajo.

2. MÉTODOS DE DETECCIÓN DE ATÍPICOS FUNCIONALES

En esta sección se presentan los métodos utilizados para detectar atípicos en datos funcionales. Se trata de tres métodos, clasificados en dos grupos. El primero de ellos, método basado en contraste de hipótesis, está basado en el método propuesto por Febrero et al. (2008). Los otros dos métodos, métodos basados en reducción de la dimensión, se caracterizan por transformar la serie de tiempo funcional que componen los datos en series de tiempo univariantes que puedan ser tratadas con las técnicas habituales.

2.1. Método basado en contraste de hipótesis

El método propuesto por Febrero et al. (2008) está basado en la profundidad de los datos funcionales, asumiendo la independencia de los mismos y obteniendo un contraste de hipótesis para saber si cada una de las curvas de un conjunto de datos es atípica o no. Una de las variaciones propuestas será no asumir la hipótesis de independencia por las características de los datos utilizados, adaptando pues el método a datos dependientes. Se entiende que los atípicos se encuentran entre las curvas con poca profundidad. De esta forma, el método tiene como objetivo calcular un percentil, punto de corte, que determina que los datos funcionales con profundidad asociada menor que dicho percentil son atípicos.

Partiendo de un conjunto de curvas formado por n datos funcionales: $x_1, \dots, x_n$, se procede como sigue:

1. Calcular la profundidad asociada (D) a cada curva del conjunto de datos, $D(x_1), \dots, D(x_n)$.
2. Sean $x_{i_1}, \dots, x_{i_k}$ las k curvas cuya profundidad es menor o igual que el punto de corte C , es decir, $D(x_{i_k}) \leq C$. Las k curvas son los atípicos detectados y se eliminan del conjunto de datos.
3. Se vuelve al paso 1 y se repite el proceso hasta que no se detecten más atípicos. Este paso es necesario ya que, al detectar las curvas que son atípicas y eliminarlas de la muestra, el conjunto de datos se reduce. Se tiene entonces un nuevo conjunto de datos cuyas profundidades deben ser recalculadas y, en función de esas nuevas profundidades, se pueden detectar o no nuevos atípicos que hasta el momento estaban “enmascarados” por los demás.

El punto C se calcula eligiendo el error de tipo I en el contraste que fija como hipótesis nula que un dato no es atípico. Se fija dicho error en el 1 %, de forma que, en ausencia de atípicos, el porcentaje de curvas mal clasificadas como atípicas sea del 1 %. Por lo tanto el C seleccionado sería:

$$\mathbb{P}(D(x) \leq C \mid x \text{ no atípico}) = 0,01.$$

Sin embargo, la distribución de las profundidades funcionales es desconocida, por lo que no se puede calcular C como el primer percentil de dicha distribución. Se presenta entonces un procedimiento para obtener C basado en el remuestreo de las curvas del conjunto de datos. En concreto, se utilizan procedimientos bootstrap para encontrar un estimador robusto del percentil.

Los datos de demanda y precio de la electricidad proceden, como se ha indicado, de una serie de tiempo funcional que se transformará en una serie estacionaria a partir de la que realizar el procedimiento. Como serie de tiempo funcional, se espera que exista una cierta dependencia entre los datos, que debe tenerse en cuenta a la hora de buscar atípicos y calcular el punto de corte C . Una opción será utilizar remuestreos bootstrap que contemplen la dependencia de los datos.

El proceso mediante el que se calcula el punto de corte C es el siguiente:

1. Calcular la profundidad asociada a cada curva del conjunto de datos, $D(x_1), \dots, D(x_n)$. Se propone el uso de dos profundidades:

- Profundidad modal de Cuevas et al. (2006).
 - Profundidad de banda modificada (MBD) de López–Pintado y Romo (2009).
2. Limpiar la muestra de posibles atípicos. Este paso es necesario para obtener una estimación robusta. Para ello se utiliza el método gráfico del boxplot funcional para datos dependientes propuesto por Sun y Genton (2012). Se eliminan de la muestra los atípicos detectados por dicho método, utilizando la misma profundidad elegida entre las dos propuestas. En este método, el gráfico se construye siguiendo el procedimiento de un gráfico de cajas punto a punto, pero considerando toda la curva, lo que supone una extensión del boxplot clásico de Tukey (1970) al caso funcional realizada por Sun y Genton (2011). Se agrupan en una banda el 50 % de curvas más profundas y se representan los “bigotes” del gráfico inflando la envoltura de dicha banda según un factor. De esta forma, las curvas que en algún momento salen de los límites del gráfico son consideradas candidatas a atípicos y son las que eliminamos del conjunto de datos. El factor indicado es 1.5 en el caso de datos independientes, justificándolo igual que en el caso univariante, donde se obtiene de una distribución normal. Para el caso de datos dependientes es necesario modificar dicho factor de forma que, la probabilidad de no encontrar atípicos con el boxplot funcional y el factor considerado, en un conjunto de datos sin atípicos sea de 0.993. Para cada conjunto de datos y para cada profundidad es necesario calcular el factor apropiado.
 3. Obtener B remuestras bootstrap de tamaño n del conjunto de datos. Las curvas remuestreadas se denotan por y_i^b , con $i = 1, \dots, n$ y $b = 1, \dots, B$. Los métodos bootstrap propuestos son:
 - Bootstrap suavizado: obtener las remuestras bootstrap suavizadas, $y_i^b = x_i^b + z_i^b$, donde x_i^b se obtienen mediante bootstrap naïve y se suaviza sumándole z_i^b , normalmente distribuida con media cero y matriz de covarianzas $\gamma \Sigma_x$, siendo γ el parámetro de suavizado y Σ_x la matriz de covarianzas del conjunto de datos del que previamente se han limpiado los atípicos.
 - Bootstrap por bloques: Se emplea el bootstrap por bloques móviles (MBB), propuesto en Künsch (1989) para datos dependientes, adaptándolo al caso funcional. Depende de la elección del tamaño de los bloques.
 - Bootstrap estacionario: Se emplea el bootstrap estacionario de Politis y Romano (1994) para datos dependientes, adaptándolo también al caso funcional. Este método remedia la falta de estacionariedad del MBB. Depende de un parámetro p , inverso del tamaño medio del bloque en el MBB.
 - Remuestreo de los residuos de un modelo autorregresivo funcional: Se estima la función de regresión de un modelo autorregresivo no paramétrico de orden p , con variables funcionales. A partir del modelo, se obtienen los residuos, que se espera sean independientes. Se tiene entonces, si q es el número de atípicos detectados en el paso 2 para limpiar el conjunto de datos, $n - q$ datos funcionales de los que se obtienen $n - q - p$ residuos. Por cada residuo se tiene el correspondiente valor ajustado. Se remuestran dichos residuos utilizando el bootstrap suavizado de la misma forma que en el primero de los métodos bootstrap, obteniendo n réplicas. A partir de las remuestras de los residuos se reconstruyen las remuestras de los datos, y_i^b .
 4. Para cada conjunto bootstrap $b = 1, \dots, B$, obtener C^b como el percentil 1 % de la distribución de las profundidades $D(y_i^b)$.
 5. Obtener C como la mediana de los B valores de C^b .

Se utilizará en la práctica 200 remuestras bootstrap. En el bootstrap suavizado se considera el parámetro de suavizado $\gamma = 0,05$. El cálculo de C se realiza solamente una vez, manteniendo el mismo valor en las distintas iteraciones del proceso. El cálculo de dicho punto de corte podría actualizarse en cada iteración, pero esta opción se rechaza debido al incremento del coste computacional requerido para ello y al planteamiento de C como un estimador robusto del percentil señalado.

Se tienen por lo tanto cuatro formas de detectar atípicos, atendiendo al tipo de remuestreo utilizado. Además, en cada una se comparará la eficacia de las dos profundidades funcionales propuestas.

2.2. Métodos basados en reducción de la dimensión

Esta clase de métodos se basa en transformar la serie de tiempo funcional en una serie de tiempo univariante. Una vez realizada la transformación, se modeliza la serie de tiempo univariante obtenida mediante un modelo ARIMA, ajustando los atípicos detectados mediante las técnicas habituales. Cada uno de los atípicos detectados en la serie de tiempo univariante se corresponde con un dato funcional, resultando ser el atípico funcional buscado.

El segundo de los métodos propuestos utiliza para la transformación proyecciones aleatorias, igual que en Cuesta y Febrero (2010). Las proyecciones aleatorias permiten, por lo tanto, reducir la dimensión. Se puede obtener un determinado número de proyecciones, modelizando y ajustando los atípicos univariantes en cada una de las series de tiempo univariantes asociadas a cada proyección. Este método proporciona una herramienta exploratoria para ver qué datos funcionales son atípicos, sin embargo en este trabajo se omiten los resultados obtenidos por su falta de estabilidad en las aplicaciones.

El tercero de los métodos propuestos tiene el mismo fundamento que el anterior. En esta ocasión, para realizar la transformación, se utiliza una descomposición en componentes principales funcionales robustas, siguiendo la propuesta de Hyndman y Ullah (2007). Se toman las componentes principales necesarias para lograr una proporción de la varianza explicada de los datos suficiente para el análisis. Con cada una de las componentes principales se crea una serie de tiempo univariante, formada por los coeficientes asociados. Las series de tiempo univariantes que resultan se estudian mediante un modelo ARIMA sobre el que se buscan los atípicos. Cada atípico detectado se asocia a un dato funcional, resultando el atípico funcional buscado.

3. APLICACIONES

En esta sección se muestran aplicaciones de dos de los métodos presentados, el método basado en contrastes de hipótesis y el de reducción de la dimensión mediante componentes principales. Cada uno de ellos se aplica tanto a la demanda como al precio de la electricidad en España a lo largo del año 2012. Para el primero de los métodos se compara el uso de las dos profundidades funcionales citadas anteriormente y los cuatro procedimientos bootstrap.

En el mercado eléctrico los datos de demanda y precio son datos horarios que se agrupan por días, a lo largo de todo el año. De esta forma, cada día se corresponde con un dato funcional obtenido a partir de 24 observaciones. Para el análisis de datos atípicos se establecen los siguientes grupos: Días de semana (de lunes a viernes a lo largo de todo el año), sábados y domingos.

Esta separación en tres grupos se debe a la diferencia de comportamiento entre los tipos de días. Los días de semana, de lunes a viernes, son días laborables en los que la demanda y precios son generalmente altos. Los domingos son días festivos, con lo cual se reduce la actividad empresarial e industrial y, por consiguiente se reduce el consumo eléctrico, con lo cual son días de demanda eléctrica baja. Los sábados se encuentran en una situación intermedia. Son días laborables pero también se sitúan en el fin de semana, por lo que no se asemejan ni a los días de semana ni a los domingos, de ahí que se clasifiquen en un grupo independiente.

Se debe tener en cuenta que los datos funcionales de demanda y precio eléctrico conforman una serie de tiempo funcional. Como serie de tiempo es necesario estudiar su tendencia, ya que se pretende trabajar con datos estacionarios, por lo que habrá que sustraer la tendencia estimada antes de analizar los atípicos funcionales que pueda contener. La tendencia se estimará mediante suavización tipo núcleo, utilizando el núcleo de Epanechnikov, siendo la serie estacionaria obtenida

la diferencia entre la serie original y la tendencia estimada. Esta transformación de la serie se realiza para los datos dentro de cada uno de los grupos de días considerados, no sobre el total del año.

A lo largo de los resultados que se muestran a continuación aparecen días particulares detectados como atípicos. Es el caso de algunos días festivos o días especiales que no son festivos, como puede ser nochevieja. Además, algunos de los atípicos se señalan como días de precio cero por la fluctuación del precio en el mercado eléctrico debido a la procedencia de la energía.

El precio de la electricidad depende entre otros factores de la demanda y de la producción de energías renovables en general, y de energía eólica en particular. Según la información recogida en la prensa y en el informe anual de la Red Eléctrica de España (REE), tanto en abril como en septiembre se batieron récords de producción eólica. Esto repercute en una bajada de los precios, llegando a ser cero en algunas horas. En concreto, durante el 2012 el precio de la electricidad bajó hasta cero durante varias horas de los siguientes días:

- Enero: día 2.
- Abril: días 15, 16, 19 y 25.
- Septiembre: día 24.
- Octubre: día 28.
- Noviembre: días 1 y 2.
- Diciembre: días 16 y 24.

Destaca el día 1 de noviembre con precio cero durante 8 horas consecutivas y el 15 de abril, en el que el precio fue cero durante 8 horas no consecutivas y con precios igualmente muy bajos la mayor parte del día en ambos casos. En otros días también se ha producido una notable bajada de precios, aunque no llegue a cero. Es el caso, por ejemplo, del 13 de febrero. En ese día, durante tres horas, el precio fue de 0.1 Cent/kWh.

3.1. Método 1 aplicado a demanda eléctrica

En esta sección se muestran los resultados de la aplicación del primero de los métodos (basado en contrastes) a la demanda eléctrica en España en 2012. En la Figura 1 se muestra el conjunto de datos funcionales sobre el que se quieren detectar atípicos, junto con la serie de tiempo funcional correspondiente. Es esa figura el conjunto de datos abarca todo el 2012, sin dividirlo en los tres grupos de días mencionados. A continuación se muestra la aplicación a cada uno de esos tres grupos.

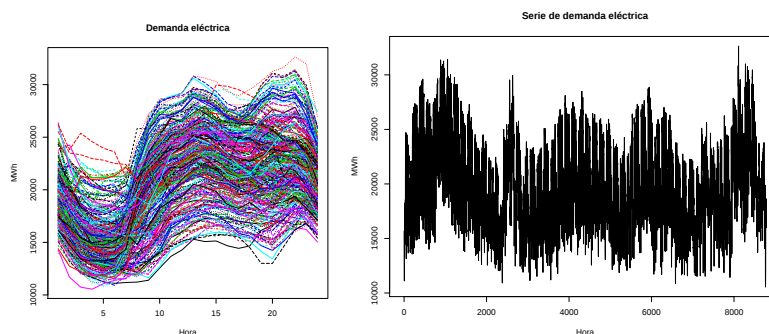


Figura 1: Demanda eléctrica en España, año 2012.

3.1.1. Días de semana

En la Figura 2 se muestra el conjunto de datos de la demanda eléctrica en días de semana sin tendencia. A la derecha del gráfico se puede ver la serie de tiempo funcional estacionaria correspondiente.

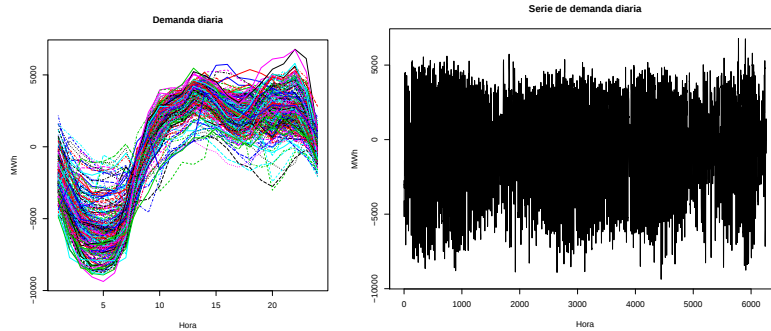


Figura 2: Demanda eléctrica sin tendencia en días de semana.

El análisis de atípicos se realiza por separado con cada una de las profundidades funcionales consideradas. En primer lugar se utiliza la profundidad modal. El primer paso del análisis consiste en eliminar los atípicos detectados por el boxplot funcional, adaptando el factor a la dependencia de los datos con lo que resulta un factor 1.34. En esta situación se detecta como atípico el día 1 de mayo, festivo. Este día se elimina del conjunto de datos para calcular el punto de corte C .

En la Tabla 1 se muestran los atípicos obtenidos para este conjunto de datos, con profundidad modal, comparando los cuatro bootstrap considerados. El bootstrap de residuos es el que más atípicos detecta, siendo muchos de ellos días especiales. Destacan los festivos 6 de abril, 1 de noviembre y 6 de diciembre que los otros métodos de remuestreo no detectan. Así como el día de huelga del 14 de noviembre. El 1 de noviembre tiene la peculiaridad de ser festivo y además ser un día de precio cero. Todos los métodos coinciden en señalar como atípicos los festivos 6 de enero, 1 de mayo, 15 de agosto, 12 de octubre y 25 de diciembre.

Día	Boot. suavizado	Boot. bloques/estacionario	Boot. residuos
06.01.2012 Festivo	X	X	X
06.04.2012 Festivo			X
19.04.2012 Precio cero			X
24.04.2012			X
01.05.2012 Festivo	X	X	X
15.08.2012 Festivo	X	X	X
12.10.2012 Festivo	X	X	X
18.10.2012			X
01.11.2012 Festivo y precio			X
02.11.2012 Precio cero			X
07.11.2012	X		X
14.11.2012 Huelga general			X
06.12.2012 Festivo			X
21.12.2012			X
25.12.2012 Festivo	X	X	X
Nº atípicos	6	5	15

Tabla 1: Atípicos en demanda diaria con profundidad modal.

En la Figura 3 se muestran los atípicos detectados con el bootstrap suavizado frente a los

obtenidos con el bootstrap de residuos.

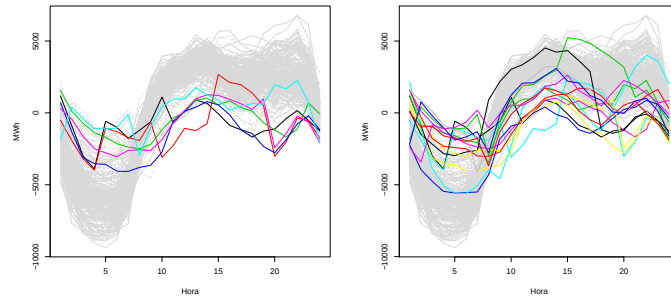


Figura 3: Atípicos detectados en demanda diaria con profundidad modal por bootstrap suavizado (izquierda) y por bootstrap de residuos (derecha).

Cambiando la profundidad modal por la profundidad de banda modificada se obtiene que el factor del boxplot funcional adaptado a la dependencia es de 1.225. Sin embargo, el resultado es el mismo, ya que detecta el día festivo 1 de mayo como único atípico.

Con la profundidad de banda el número de atípicos detectado es menor (Tabla 2). Los principales días festivos del año también se consideran atípicos, como con la profundidad modal. El bootstrap de residuos se puede considerar la mejor opción porque es capaz de detectar el día 1 de noviembre, que es festivo y tiene precio cero.

Día	Boot. suavizado/bloques/estacionario	Boot. residuos
06.01.2012 Festivo	X	X
15.08.2012 Festivo	X	X
12.10.2012 Festivo	X	X
01.11.2012 Festivo y precio		X
02.11.2012 Precio cero	X	X
25.12.2012 Festivo	X	X
Nº atípicos	5	6

Tabla 2: Atípicos en demanda diaria con profundidad de banda

En la Figura 4 se muestran los atípicos detectados en los dos casos sobre el conjunto de datos.

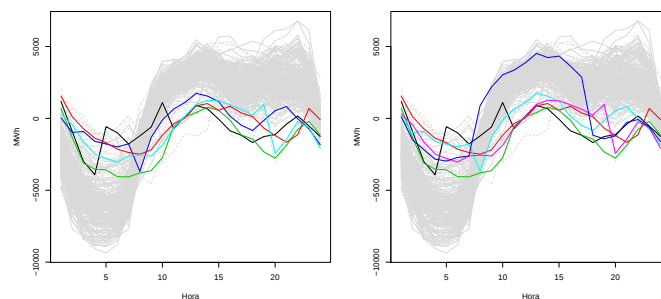


Figura 4: Atípicos en demanda diaria con profundidad de banda. Izquierda: bootstrap suavizado/bloques y estacionario, derecha: bootstrap de residuos.

3.1.2. Sábados

En este apartado se presentan los resultados de aplicar el primero de los métodos en la demanda eléctrica de los sábados. En la Figura 5 se puede ver el conjunto de datos funcionales sin tendencia a estudiar, así como la serie de tiempo funcional estacionaria correspondiente.

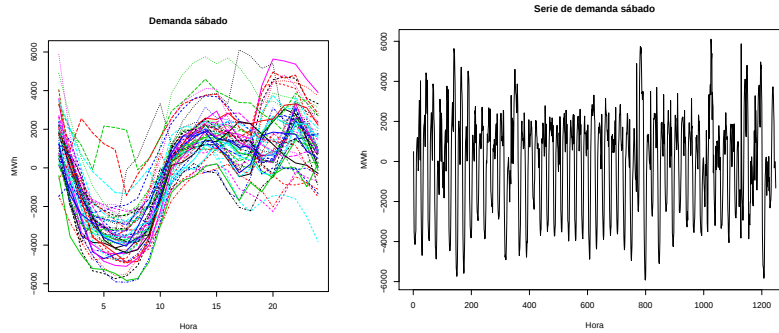


Figura 5: Demanda eléctrica.

En primer lugar se realiza el análisis de datos atípicos con la profundidad modal. En este caso el factor de boxplot funcional ajustado es 1.42 y se detectan tres días como atípicos en esta primera fase: 14.04.2012, 27.10.2012 y 24.11.2012.

De nuevo el bootstrap de residuos es el que más atípicos detecta (Tabla 3), siendo además el único en detectar como atípico el festivo del 8 de diciembre. Los días 7 de abril y 18 de agosto se sitúan en semanas en las que hay festivos. El bootstrap suavizado coincide con los resultados obtenidos con bootstrap por bloques o estacionario. En la Figura 6 se representan los atípicos detectados por los tres métodos bootstrap frente al bootstrap de residuos.

Día	Boot. suavizado/bloques/estacionario	Boot. residuos
04.02.2012		X
11.02.2012		X
07.04.2012 Especial		X
14.04.2012	X	X
18.08.2012 Especial	X	X
20.10.2012		X
27.10.2012	X	X
24.11.2012		X
01.12.2012		X
08.12.2012 Festivo		X
15.12.2012	X	X
22.12.2012		X
29.12.2012		X
Nº atípicos	4	13

Tabla 3: Atípicos en demanda en sábados con profundidad modal.

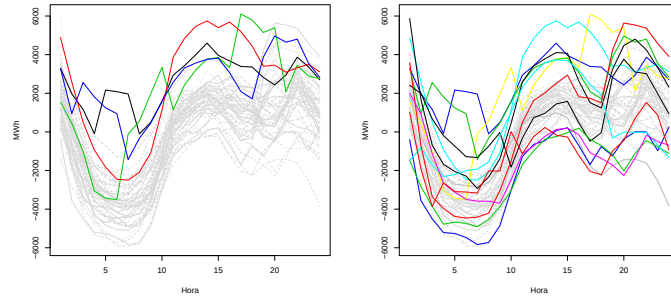


Figura 6: Atípicos en demanda en sábados con profundidad modal. Izquierda: bootstrap suavizado/bloques/estacionario, derecha: bootstrap de residuos.

Al realizar el análisis con la profundidad de banda modificada se obtiene un factor para el boxplot ajustado de 1.29, detectando cuatro días como atípicos: 14.04.2012, 18.08.2012, 27.10.2012 y 24.11.2012.

En esta ocasión el bootstrap suavizado, por bloques y estacionario con la profundidad de banda no detectan ningún atípico. El bootstrap de residuos sí lo hace, siendo algunos de ellos días cercanos a festivos (Tabla 4). En la Figura 7 se muestran los atípicos detectados por el bootstrap de residuos.

Día	Boot. suavizado/bloques/estacionario	Boot. residuos
07.04.2012 Especial		X
14.04.2012		X
18.08.2012 Especial		X
15.12.2012		X
22.12.2012		X
Nº atípicos	0	5

Tabla 4: Atípicos en demanda en sábados con profundidad de banda.

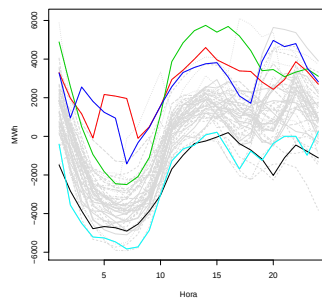


Figura 7: Atípicos en demanda en sábados con profundidad de banda y bootstrap de residuos.

3.1.3. Domingos

En la Figura 8 se muestra el conjunto de datos funcionales correspondiente a la demanda eléctrica en domingo, junto con la serie de tiempo funcional estacionaria correspondiente. Se trata de un conjunto de datos similar al de los sábados, pero con valores de demanda generalmente menores.

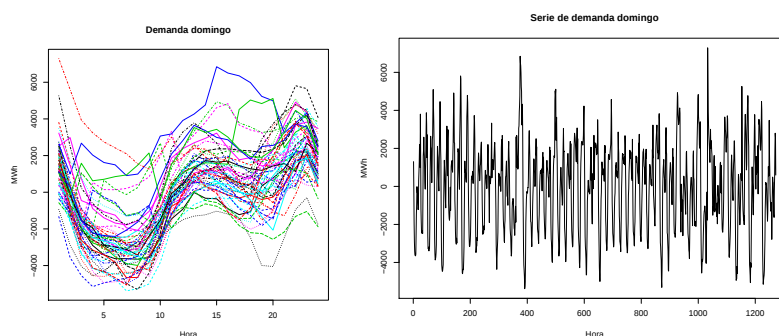


Figura 8: Demanda eléctrica en domingos.

Con la profundidad modal se obtiene un factor para el boxplot de 1.37, detectando los siguientes atípicos: 15.04.2012, 01.07.2012, 28.10.2012 y 16.12.2012. Los atípicos son días de precio cero, salvo el 01.07.2012.

En esta ocasión el método de bootstrap por bloques es el único que proporciona un resultado distinto, detectando solo dos atípicos (Tabla 5). El resto de métodos bootstrap detecta cuatro atípicos. Destaca el hecho de que 3 de los 4 atípicos detectados por los otros métodos coinciden con días de precio cero.

Día	Boot. suavizado/estacionario/residuos	Boot. bloques
15.04.2012 Precio cero	X	X
21.10.2012	X	
28.10.2012 Precio cero	X	X
16.12.2012 Precio cero	X	
Nº atípicos	4	2

Tabla 5: Atípicos en demanda en domingos con profundidad modal.

En la Figura 9 se representan los atípicos detectados comparando los tipos de bootstrap.

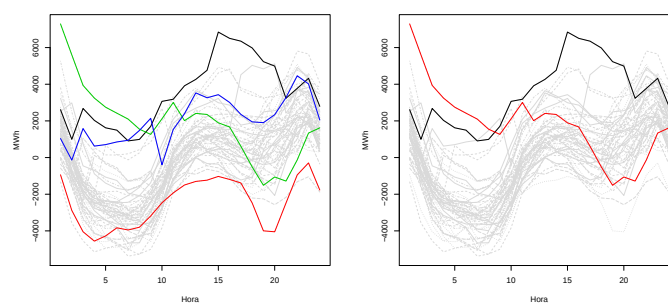


Figura 9: Atípicos en demanda en domingos con profundidad modal. Izquierda: bootstrap suavizado/estacionario/residuos, derecha: bootstrap por bloques.

En el caso de la profundidad de banda modificada el factor ajustado es 1.3 y los atípicos detectados por el boxplot son: 15.04.2012 (precio cero), 28.10.2012 y 16.12.2012 (precio cero).

En este caso el bootstrap suavizado, por bloques, estacionario y de residuos coinciden, detectando solo dos días atípicos (Tabla 6). El resultado se representa gráficamente en la Figura 10.

Día	Boot. suavizado/bloques/estacionario/residuos
15.04.2012 Precio cero	X
21.10.2012	X
Nº atípicos	2

Tabla 6: Atípicos en demanda en domingos con profundidad de banda.

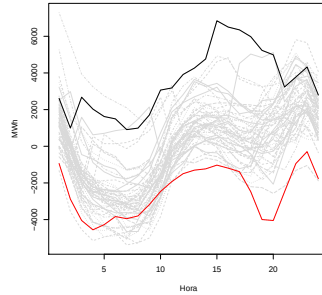


Figura 10: Atípicos en demanda en domingos con profundidad de banda para todos los bootstrap.

3.2. Método 1 aplicado a precio eléctrico

En este apartado se realiza un análisis de atípicos funcionales sobre el precio eléctrico en España a lo largo de todo el año 2012 utilizando el método 1, basado en contraste. Se representan en la Figura 11 los datos funcionales y la serie de tiempo funcional correspondiente. Esta serie no presenta tendencia, por lo que no es necesaria ninguna transformación previa en los datos.

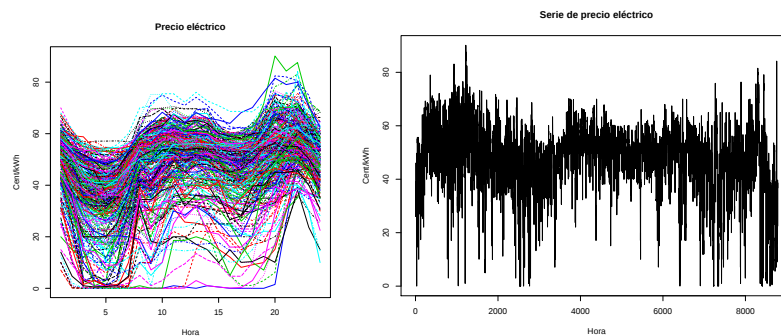


Figura 11: Precio eléctrico.

3.2.1. Días de semana

Se representan en la Figura 12 los datos funcionales y la serie de tiempo funcional correspondiente del precio eléctrico para el grupo de días de semana. Los resultados se presentan de forma análoga al caso de la demanda eléctrica.

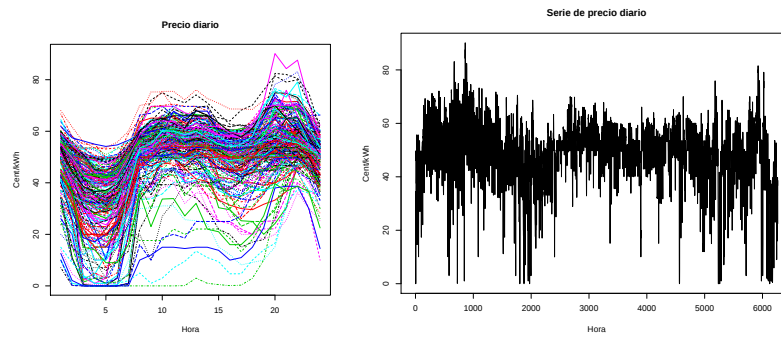


Figura 12: Precio eléctrico en días de semana.

Utilizando la profundidad modal se obtiene un factor para el boxplot ajustado de 1.8, con el que se obtienen los siguientes 6 días como atípicos: 25.04.2012, 15.08.2012, 01.11.2012, 02.11.2012, 24.12.2012 y 25.12.2012. Todos los atípicos son días festivos o días en los que la alta producción de energía eólica ha bajado notablemente los precios, llegando a cero.

Día	Boot. suavizado	Boot. bloques/estacionario	Boot. residuos
02.01.2012 Precio cero			X
13.02.2012 Precio cero			X
20.02.2012 Especial			X
10.04.2012			X
18.04.2012	X		X
19.04.2012 Precio cero	X	X	X
23.04.2012			X
25.04.2012 Precio cero	X	X	X
01.05.2012 Festivo	X	X	X
04.05.2012			X
15.08.2012 Festivo	X		X
24.09.2012 Precio cero			X
01.11.2012 Festivo y precio	X	X	X
02.11.2012 Precio cero	X	X	X
21.12.2012			X
24.12.2012 Precio cero	X	X	X
25.12.2012 Festivo	X	X	X
26.12.2012			X
31.12.2012 Especial	X	X	X
Nº atípicos	10	8	19

Tabla 7: Atípicos en precio en días de semana con profundidad modal.

De los métodos bootstrap utilizados, el más acertado parece ser el de los residuos (Tabla 7). Todos los métodos detectan algunos de los atípicos más destacados del año, coincidentes con días festivos y días de precio cero. El 31 de diciembre, por ejemplo, a pesar de no ser festivo es un día especial por lo que es normal que aparezca también como atípico. Algunos de los atípicos son días que por sí mismos no son especiales, pero que son el víspera o el día siguiente a algunos de los días destacados.

Entre el bootstrap suavizado y el bootstrap por bloques o estacionario, parece mejor el suavizado. La diferencia principal está en que este último incluye como atípico el día festivo 15 de agosto. Sin embargo en general es mejor el bootstrap de residuos, ya que es el único en incluir los días 2 de enero, 13 de febrero y 24 de septiembre que son de los destacados por tener precio cero. Entre

otros de los días detectados destaca el 20 de febrero, coincidente con las fiestas de carnaval, día festivo en muchos lugares de España. Los resultados del bootstrap suavizado, frente al bootstrap de residuos se representan en la Figura 13.

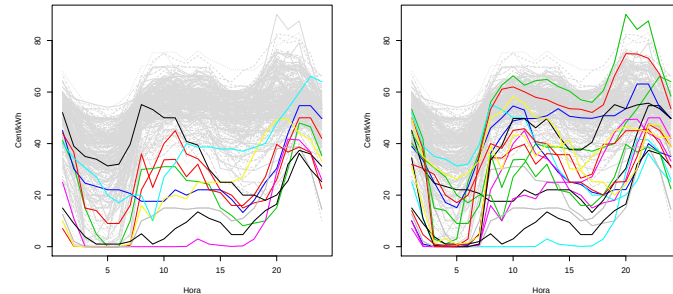


Figura 13: Atípicos en precio en días de semana con profundidad modal. Izquierda: bootstrap suavizado, derecha: bootstrap de residuos.

Utilizando la profundidad de banda modificada se obtiene un factor 1.61 para el boxplot funcional ajustado, que detecta los siguientes atípicos: 25.04.2012, 01.05.2012, 15.08.2012, 01.11.2012, 02.11.2012, 24.12.2012 y 25.12.2012. El boxplot funcional con esta nueva medida de profundidad detecta un atípico más que con la profundidad modal, el día 1 de mayo, festivo nacional.

En la Tabla 8 se muestran los atípicos detectados por cada procedimiento bootstrap, que en este caso proporcionan resultados diferentes en cada caso.

Día	Boot. suav.	Boot. bloques	Boot. estac.	Boot. resid.
21.02.2012 Especial			X	X
19.04.2012 Precio cero		X	X	X
25.04.2012 Precio cero		X	X	X
01.11.2012 Festivo y precio	X	X	X	X
02.11.2012 Precio cero	X	X	X	X
11.12.2012				X
12.12.2012	X	X	X	X
24.12.2012 Precio cero	X	X	X	X
25.12.2012 Festivo	X	X	X	X
26.12.2012		X	X	X
31.12.2012 Especial				X
Nº atípicos	5	8	9	11

Tabla 8: Atípicos en precio en días de semana con profundidad de banda.

El bootstrap suavizado es el único que no contempla la dependencia de los datos y en este caso es notablemente peor que los que sí la contemplan, porque no detecta algunos de los días más destacados por sus bajos precios. Entre los métodos de remuestreo para datos dependientes hay pocas diferencias, siendo la principal que el de residuos detecta el 31 de diciembre. La comparación entre el bootstrap suavizado y el de residuos, que son el que menos atípicos detecta y el que más, respectivamente, se representa en la Figura 14.

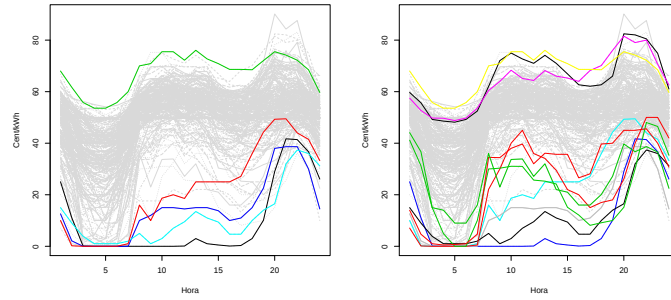


Figura 14: Atípicos en precio en días de semana con profundidad de banda. Izquierda: bootstrap suavizado, derecha: bootstrap de residuos.

3.2.2. Sábados

Se representan en la Figura 15 los datos funcionales y la serie de tiempo funcional correspondiente del precio eléctrico para el grupo de sábados.

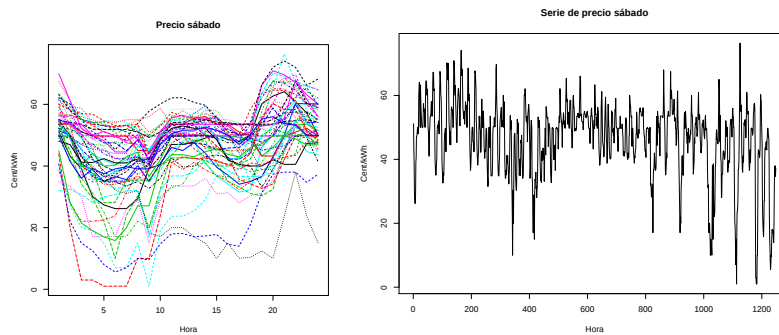


Figura 15: Precio eléctrico en sábados.

Utilizando la profundidad modal el factor para el boxplot funcional ajustado es 1.6. Los atípicos detectados en el paso previo del método son: 05.05.2012, 29.09.2012, 27.10.2012, 24.11.2012, 15.12.2012 y 29.12.2012.

En la Tabla 9 se muestran los atípicos detectados. De nuevo el bootstrap de residuos es el que más atípicos detecta, aunque el resultado en este caso es similar al bootstrap suavizado. Dentro del conjunto de sábados del año 2012 apenas se encuentran festivos o días especiales. Es por eso que los días más destacados son aquellos que pertenecen a semanas en las que sí se encuentran festivos. El 3 de noviembre viene precedido de un festivo y un día de precio cero mientras que, por ejemplo, el 14 de abril y el 27 de octubre son víspera de un día con precio cero.

Día	Boot. suavizado	Boot. bloques/estacionario	Boot. residuos
14.04.2012	X		X
05.05.2012	X	X	X
01.09.2012	X		X
29.09.2012	X		X
27.10.2012	X	X	X
03.11.2012 Especial			X
24.11.2012	X	X	X
15.12.2012	X	X	X
22.12.2012	X		X
29.12.2012	X	X	X
Nº atípicos	9	5	10

Tabla 9: Atípicos en precio en sábados con profundidad modal.

En la Figura 16 se representan los atípicos detectados por el bootstrap por bloques o estacionario frente a los detectados con el bootstrap de residuos.

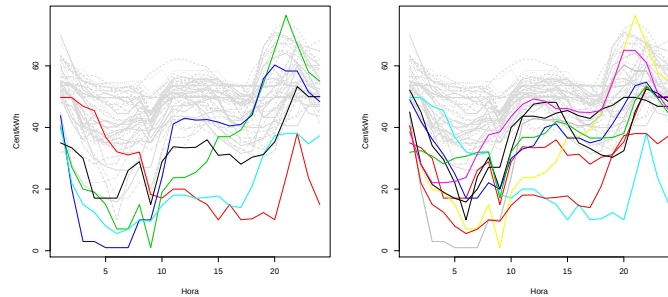


Figura 16: Atípicos en precio en sábados con profundidad modal. Izquierda: bootstrap por bloques/estacionario, derecha: bootstrap de residuos.

Con la profundidad de banda modificada el factor ajustado para el boxplot es 1.51, con el que se obtienen los siguientes atípicos: 27.10.2012, 24.11.2012, 15.12.2012 y 29.12.2012.

En esta ocasión todos los métodos de remuestreo obtienen los mismos resultados (Tabla 10). El único atípico es el 29 de diciembre, que se encuentra en medio de los días de navidad, en los que sí hay festivos y además algún día de precio bajo.

Día	Boot. suavizado/bloques/estacionario/residuos
29.12.2012	X
Nº atípicos	1

Tabla 10: Atípicos en precio en sábados con profundidad de banda.

En la Figura 7 se representa el único atípico detectado por los métodos bootstrap.

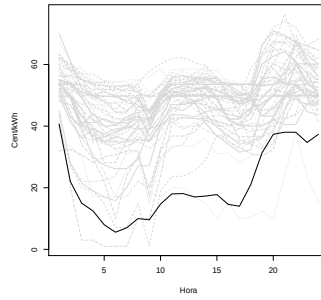


Figura 17: Atípicos en precio en sábados con profundidad de banda para todos los tipos de bootstrap.

3.2.3. Domingos

Se representan en la Figura 18 los datos funcionales y la serie de tiempo funcional correspondiente del precio eléctrico para el grupo de domingos.

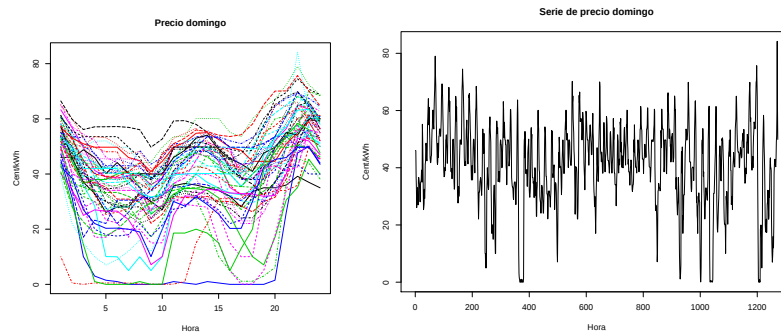


Figura 18: Precio eléctrico en domingos.

Utilizando la profundidad modal, el factor ajustado para el boxplot en este caso es 1.52, con el que se detectan tres atípicos que son días de precio cero: 15.04.2012, 28.10.2012 y 16.12.2012.

Los domingos detectados como atípicos son todos días de precio cero o bien vísperas de días especiales. Apenas hay diferencias entre los métodos de remuestreo, siendo los resultados del bootstrap suavizado iguales a los de bootstrap de residuos y los de bootstrap por bloques iguales a los del estacionario (Tabla 11).

Día	Boot. suavizado/residuos	Boot. bloques/estacionario
15.04.2012 Precio cero	X	X
23.09.2012	X	X
14.10.2012 Especial	X	X
28.10.2012 Precio cero	X	X
16.12.2012 Precio cero	X	X
30.12.2012	X	
Nº atípicos	6	5

Tabla 11: Atípicos en precio en domingos con profundidad modal.

Los atípicos detectados se representan en la Figura 19. En el gráfico de la izquierda se representa el resultado del bootstrap suavizado y de residuos, frente al de bloques y estacionario.

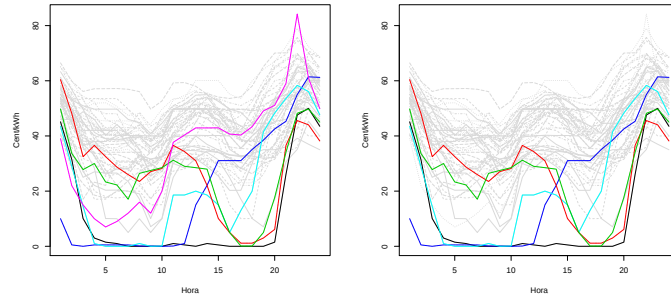


Figura 19: Atípicos en precio en domingos con profundidad modal. Izquierda: bootstrap suavizado/residuos, derecha: bootstrap por bloques/estacionario.

Cambiando la profundidad modal por la profundidad de banda modificada se obtiene un factor 1.47, con el que se detectan dos atípicos en días de precio cero: 15.04.2012 y 28.10.2012.

Con la profundidad de banda se detectan dos atípicos con los bootstrap por bloques y estacionario y uno solo en los demás (Tabla 12). El 15 de abril es uno de los días de precio cero y es el atípico detectado por todos los métodos. Estos resultados se representan en la Figura 20

Día	Boot. suavizado/residuos	Boot. bloques/estacionario
15.01.2012		X
15.04.2012 Precio cero	X	X
Nº atípicos	1	2

Tabla 12: Atípicos en precio en domingos con profundidad de banda.

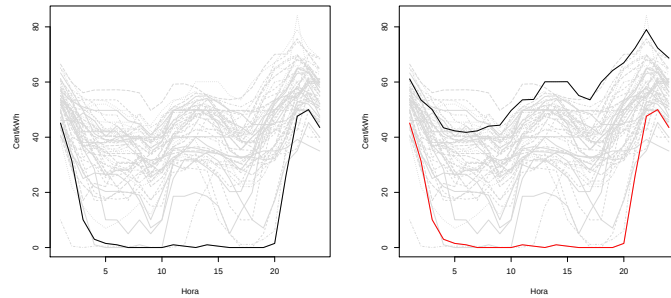


Figura 20: Atípicos en precio en domingos con profundidad de banda. Izquierda: bootstrap suavizado/residuos, derecha: bootstrap por bloques/estacionario.

3.3. Método 3 aplicado a demanda eléctrica

En este apartado se muestran los resultados de la aplicación a los datos eléctricos del tercero de los métodos, el que utiliza la descomposición en componentes principales de la serie de tiempo funcional. En todos los casos presentados se obtiene el número de componentes necesario, por el método robusto propuesto por Hyndman y Ullah (2007), de forma que la proporción de varianza explicada entre todas sea suficiente.

Se muestran en primer lugar los atípicos detectados en el grupo de días de semana a los que se les ha sustraído la tendencia estimada para obtener datos estacionarios.

Se encuentran once días atípicos, que se muestran en la Tabla 13 y que se representan gráficamente en la Figura 21. Muchos de los atípicos coinciden con los detectados por el método anterior, siendo la mayoría de ellos festivos, días de precio cero o día de huelga.

10.04.2012	16.04.2012 (Precio cero)	19.04.2012 (Precio cero)	24.04.2012
01.05.2012 (Festivo)	15.08.2012 (Festivo)	12.10.2012 (Festivo)	07.11.2012
14.11.2012 (Huelga)	06.12.2012 (Festivo)	25.12.2012 (Festivo)	

Tabla 13: Atípicos en demanda diaria.

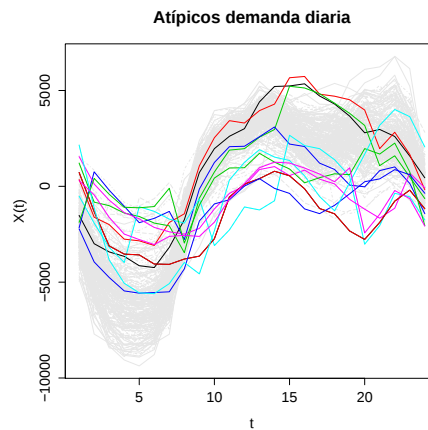


Figura 21: Atípicos en demanda diaria.

En el grupo de sábados se detecta un único día atípico, el 12 de mayo. En el de domingos se encuentran dos atípicos, el 19 de agosto y el 30 de septiembre. En estos dos grupos de días los resultados difieren de los obtenidos por el método 1.

3.4. Método 3 aplicado al precio eléctrico

En el estudio de atípicos en el precio eléctrico con el método de componentes principales, en el grupo de días de semana, se encuentran siete atípicos que se muestran en la Tabla 14.

15.08.2012 (Fest)	24.09.2012 (Precio 0)	01.11.2012 (Fest y precio)	02.11.2012 (Precio 0)
06.11.2012	24.12.2012 (Precio cero)	25.12.2012 (Festivo)	

Tabla 14: Atípicos en precio diario.

Entre los atípicos detectados se mezclan festivos, días de precio cero y vísperas de festivo, la mayoría de ellos han sido también detectados por el método 1 y se representan en la Figura 22.

En el grupo de sábados se encuentran dos atípicos, los días 30 de junio y 11 de agosto, mientras que en los domingos no se encuentra ningún atípico.

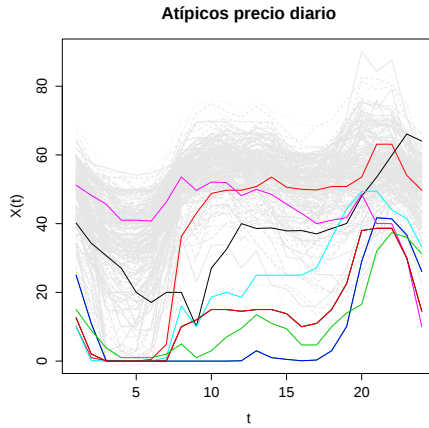


Figura 22: Atípicos en precio diario.

4. CONCLUSIONES

En este trabajo se han presentado tres propuestas de métodos para detectar atípicos en datos funcionales bajo dependencia. De ellos, el primero (basado en contrastes) y el tercero (componentes principales) proporcionan resultados bastante adecuados. La gran mayoría de días detectados como atípicos se corresponden con festivos, vísperas de festivo, huelgas o días de precio cero. Estos últimos son un caso especial del mercado eléctrico español que influye no solo en el precio, si no también en la demanda.

El método 2, tal y como se presenta en este informe, proporciona resultados inestables que se han omitido en este trabajo. El estudio detallado de este método con datos simulados permitirá comprobar su eficacia real. El método 3 proporciona resultados correctos, en el sentido de que detecta días especiales a lo largo del año, que también son señalados como atípicos por el método 1. Sin embargo, en los grupos pequeños de sábados y domingos, se encuentran menos atípicos y son diferentes de los obtenidos con el método 1.

El método 1 parece ser el más fiable y el que proporciona una mayor variedad de resultados. Dentro del mismo, en la mayoría de los casos el remuestreo de los residuos de un modelo autorregresivo es el que más atípicos detecta, siendo la mayoría de ellos justificables atendiendo a días especiales. Esto pone de manifiesto la importancia de considerar un método de remuestreo que contemple la dependencia de los datos. Sin embargo, en muchas ocasiones, los métodos de bootstrap por bloques y estacionario proporcionan los mismos resultados que el bootstrap suavizado.

Entre las dos profundidades consideradas, la profundidad modal es la que más atípicos detecta, siendo la profundidad de banda más conservadora. En concreto, en los grupos pequeños de sábado y domingo, especialmente en el análisis de los precios, detecta solo uno o dos atípicos con poca justificación. El caso contrario se produce en los grupos grandes, de días de semana, en los que el número de atípicos detectados con la profundidad modal incluso dobla a los obtenidos con la profundidad de banda. En estos casos quizá la profundidad modal sea demasiado permisiva a la hora de clasificar un dato como atípico.

Un estudio de simulación permitirá comparar la eficacia real de cada uno de los métodos presentados así como estudiar en detalle el comportamiento de cada profundidad funcional y de cada método de remuestreo propuesto.

REFERENCIAS

- Cuesta, J. A. y Febrero, M. (2010) A Simple Multiway ANOVA for Functional Data. *TEST*. Vol. 19, 537–557.
- Cuevas, A., Febrero, M. y Fraiman, R. (2006) On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*. Vol. 51, 1063–1074.
- Febrero, M., Galeano, P. y González-Manteiga, W. (2008) Outlier Detection in Functional Data by Depth Measures, With Application to Identify Abnormal NOx Leveles. *Environmetrics*. Vol. 19 (4), 331–345.
- Ferraty F. y Vieu P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer.
- Hyndman, R. J. y Ullah, S. (2007) Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach. *Computational Statistics & Data Analysis*. Vol. 51, 4942–4956.
- Hyndman, R. J. y Shang, H. L. (2010) Rainbow Plots, Bagplots, and Boxplots for Functional Data. *Journal of Computational and Graphical Statistics*. Vol. 19, 29–45.
- Künsch, H. R. (1989) The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*. Vol. 17, 1217–1241.
- López-Pintado, S. y Romo, J. (2009) On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*. Vol. 104, 718–734.
- Politis, D. N. y Romano, J. R. (1994) The Stationary Bootstrap. *Journal of American Statistical Association*. Vol. 89, 1303–1313.
- Ramsay, J. O. y Silverman, B. W. (2005) *Functional Data Analysis (2nd ed.)*. Springer.
- Sun, Y. y Genton M. G. (2011) Functional Boxplots. *Journal of Computational and Graphical Statistic*. Vol. 20, 316–334.
- Sun, Y. y Genton, M. G. (2012) Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*. Vol. 23, 54–64.
- Tukey, J. W. (1970) *Exploratory Data Analysis. (Limited Preliminary Edition)*, Vol. 1, Ch. 5. Addison-Wesley. Reading, MA.