Trabajo Fin de Máster

Regiones de predicción con datos funcionales. Aplicación a mercados eléctricos.

José Graña Colubi

Máster en Técnicas Estadísticas

Curso 2020-2021

Propuesta de Trabajo Fin de Máster

Título en galego: Rexións de predición con datos funcionais. Aplicación a mercados eléctricos.
Título en español: Regiones de predicción con datos funcionales. Aplicación a mercados eléctricos.
English title: Prediction regions with functional data. Aplication to electricity markets.
Modalidad: Modalidad A
Autor/a: José Graña Colubi, Universidad de Santiago de Compostela
Director/a: Juan Manuel Vilar Fernández, Universidade da Coruña; Germán Aneiros Pérez, Universidade da Coruña
Breve resumen del trabajo: Utilizando métodos de predicción funcional (regresiones no paramétrica y parcialmente lineal) se desarrolla un algoritmo bootstrap para el cálculo de regiones de predicción. Dicho algoritmo se aplica a curvas diarias de demanda y precio del mercado eléctrico español. Además, se estudia la influencia de la norma seleccionada en el cálculo de las regiones de predicción. Finalmente, los resultados obtenidos se comparan con unos intervalos de predicción puntual obtenidos utilizando los mismos métodos de predicción.
Recomendaciones: Conocimientos básicos de regresión no paramétrica, datos funcionales y métodos de remuestreo. Manejo de R.
Otras observaciones: Conocimientos básicos de series de tiempo.

Índice general

Índice de figuras

Resumen

Resumen en español

Se utilizan herramientas pertenecientes al campo de los datos funcionales con el fin de construir regiones de predicción para las curvas diarias del precio y la demanda en el mercado eléctrico de España. Se utilizarán dos métodos de predicción puntual funcional: uno no paramétrico puramente autorregresivo, basado en un estimador kernel, y otro semifuncional parcialmente lineal que incluye en el modelo covariables funcionales y escalares. A continuación se construirá un algoritmo basado en la aproximación bootstrap de la distribución del error en la predicción para calcular las regiones de predicción. Este algoritmo se utilizará para calcular las regiones tanto del precio como de la demanda eléctrica de todos los días del año 2012. Además, se considerarán tres normas y tres niveles de confianza, por lo que se calcularán un total de treinta y seis regiones para cada día. A continuación, se procederá a comparar los resultados de cada combinación a través de la cobertura y la amplitud media de la región resultante. Además, se incluye un método benchmark de predicción que se utilizará como medida comparativa.

English abstract

Statistical tools from the field of functional data are used in order to build prediction regions for daily curves of electricity demand and price in the Spanish Market. Two methods of pointwise functional prediction are proposed: one using a functional nonparametric predictor based on kernel estimation and one semifunctional partial linear method that includes linear scalar covariates. Then, an algorithm based on a bootstrap approximation to the distribution of the error in prediction is built. Using this algorithm, the prediction regions for both demand and price are calculated for each day of the year 2012. Three norms and three confidence levels are considered, which means that a total of thrty-six prediction regions are calculated for each day. Then, a comparative study among each combination of parameters is performed, considering the coverage and mean length of the regions, and also featuring a benchmark method in order to study the efficiency of the previous.

Capítulo 1

Introducción

1.1. Contexto matemático

En este trabajo se realizarán predicciones en un espacio de Hilbert separable. Sea H un espacio de Hilbert, se sabe que está dotado de un producto interno $\langle \cdot, \cdot \rangle$ y una correspondiente norma asociada $\| \cdot \|_H$ verificando que $\|\zeta\|_H^2 = \langle \zeta, \zeta \rangle \ \forall \zeta \in H$. Además, si H es separable, significa que cuenta con una base ortonormal $\{e_k\}_{k \in \mathbb{N}}$. Así pues, se considerará que los datos de estudio se encuentran en un espacio de estas condiciones. En particular, los datos serán curvas vistas como datos funcionales.

1.2. Los datos

Se dispone de un archivo con los datos de demanda y precio de la electricidad en España durante cada hora entre el 1 de Enero de 2011 y el 31 de Diciembre de 2012. Estos datos son públicos y pueden encontrarse en la web de OMIE (Operador de Mercado Ibérico de España). Se cuenta también con los datos de producción eólica horaria durante el mismo periodo de tiempo (extraídos de la web de REE), así como la temperatura máxima diaria (proporcionada por AEMET).

1.2.1. Trabajando con datos funcionales

Un dato funcional se puede definir como una observación de una variable aleatoria que toma valores en un espacio de dimensión infinita. Esto permite trabajar directamente con curvas, superficies, imágenes, etc. En el caso particular de este trabajo, los datos serán curvas diarias de demanda y de precio del mercado eléctrico español. Así, cada dato funcional será de la forma

$$\chi = \{\chi(t) : t \in T\}$$

donde $T \subset \mathbb{R}$ es un intervalo real.

Medidas de proximidad

Para trabajar con datos funcionales es necesaria una forma de medir la proximidad entre los mismos. Una posibilidad sería utilizar una norma como medida de la cercanía entre dos elementos. Es bien sabido que en un espacio de dimensión finita todas las normas son equivalentes; sin embargo, esta propiedad no se cumple en espacios de dimensión infinita, es decir, donde se encuentran los datos funcionales, por lo que la elección de la norma se vuelve determinante. En este trabajo se utilizarán semimétricas como medidas de proximidad entre datos, dado que el uso de métricas es más restrictivo y puede revelar menos estructura dentro de los datos. A continuación se muestran dos ejemplos de semimétricas.

La semnimétrica $d_v^{deriv}(\cdot,\cdot)$ se basa en la $v-$ésima derivada de las curvas y se define como

$$d_v^{deriv}(\chi_i,\chi_j) = \sqrt{\int (\chi_i^{(v)}(t), \chi_j^{(v)}(t))^2 dt}.$$

Esta semimétrica es recomendable cuando se trabaja con curvas suaves. En otro caso es preferible utilizar $d_s^{proj}(\cdot,\cdot)$, una semimétrica basada en la proyección en los s primeros autovectores, $v_1(\cdot),...,v_s(\cdot)$, asociados a los s mayores autovalores del operador covarianza del predictor funcional $\boldsymbol{\chi}$, que se define como

$$d_s^{proj}(\chi_i,\chi_j) = \sqrt{\sum_{k=1}^{s} \left(\int (\chi_i(t) - \chi_j(t)) v_k(t) dt \right)^2}.$$

Medidas de centralización

Dado un conjunto de curvas $S = \{\chi_1,...,\chi_n\}$, la noción de media se puede extender fácilmente de la conocida para datos unidimensionales definiendo

$$\chi_{media,S} = \frac{1}{n} \sum_{i=1}^{n} \chi_i(t),\ \forall t \in T.$$

Sin embargo otras medidas como la moda o la mediana no son tan sencillas de definir. Para ello se utilizarán las profundidades, unas funciones definidas inicialmente dentro de la estadística multivariante que, extendidas a dimensión infinita, sirven para generalizar la noción de centralización a los datos funcionales. Así, un dato con una profundidad alta se interpretará como un dato central, mientras que datos con profundidades inusualmente bajas serán candidatos a ser outliers.

1.2.2. Análisis exploratorio

Se cuenta con los datos horarios de demanda y precio de la electricidad en España de cada día de los años 2011 y 2012. Más concretamente, los conjuntos de datos reúnen, respecticamente, la cantidad total de energía intercambiada en el mercado diario (medida eb MWh) y el precio marginal del mercado español en cada hora (medido en Cent/kWh). Es decir, para cada día entre el 1 de Enero de 2011 y el 31 de Diciembre de 2012 se tienen veinticuatro datos de demanda y otros veinticuatro de precio. En este trabajo se considerarán como datos las curvas diarias de demanda y de precio, entendiendo los valores con los que se cuenta como veinticuatro observaciones de cada una de estas curvas.

El modo de empleo de los datos será el siguiente: se buscará construir una región de predicción para cada curva de cada día del año 2012 utilizando la información de los 365 días previos, aprovechando que se tienen los datos reales de dicho año y utilizándolos para la evaluación de la predicción.
Por otro lado, debido a la similitud de los datos de 2011 con los de 2012, solo se tendrán en cuenta estos últimos para el análisis exploratorio.

Demanda

La siguiente gráfica muestra la demanda de electricidad (en MWh) en España en el año 2012. Las líneas verticales rojas delimitan los meses del año, con el fin de facilitar la observación del comportamiento a lo largo del año.
A simple vista se puede apreciar cómo cambia la demanda en función del mes. En los primeros y últimos meses esta tiende a ser mayor, mientras que durante el resto del año es, en general, menor. Eso sí, sufre fluctuaciones durante todo año y experimenta grandes subidas en momentos concretos.

En el siguiente gráfico se muestran los datos correspondientes únicamente al mes de Enero de 2012, con el fin de poder apreciar mejor el comportamiento de los datos de demanda. Se pueden apreciar dos características que, como se verá más adelante, resultan clave a la hora de comprender este conjunto de datos. La primera es la tendencia de la serie a mantenerse en valores parecidos durante cinco días para a continuación disminuir durante dos. Esto no es casual: ahondando más se descubre que la racha de cinco días coincide con los días laborables de la semana, mientras que los dos siguientes se corresponden con el fin de semana. En relación a esto aparece la otra característica: el día 6, festivo, tiene unos datos de demanda más similares a los de un fin de semana que a los de un día laborable. Esto invita a pensar que los días festivos que coincidieran entre semana a lo largo del año mostrarán un comportamiento atípico.

Cambiando de perspectiva, a continuación se presentan los datos como curvas diarias superpuestas

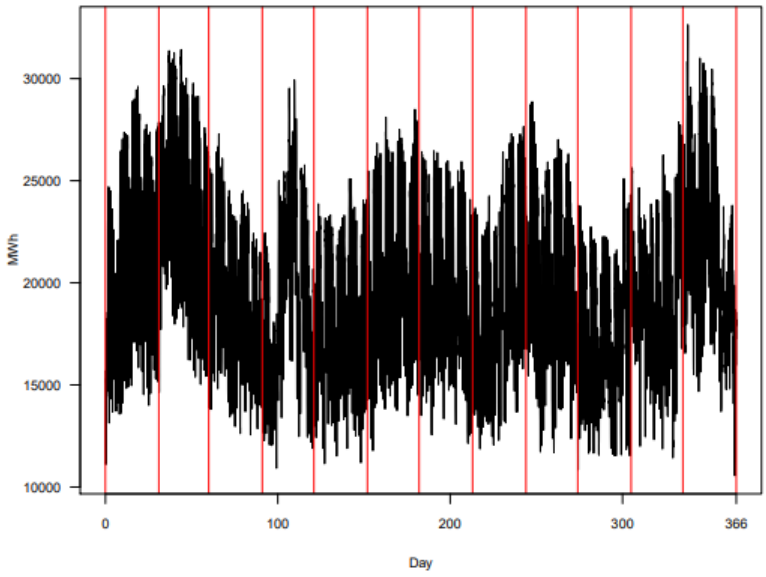

Figura 1.1: Serie de tiempo funcional de demanda de electricidad en España en 2012.

(demanda frente a hora del día). Así, cada curva puede ser vista como un dato funcional, de forma que la muestra representada en el gráfico sean 366 datos correspondientes con las curvas de demanda de cada uno de los 366 días de 2012 (año bisiesto, cabe recordar).

Los valores de la demanda oscilan entre 10000 y 30000 MWh. Los valores mínimos se alcanzan aproximadamente a las 5 am, mientras que los máximos aparecen a la 1 pm y las 10pm. No se aprecia gran cambio en la forma de las curvas.
Sí se aprecia una diferencia en la altura de la curva en función del color de la misma, es decir, del día de la semana al que pertenece. Más concretamente, llama la atención la escasa altura de las curvas correspondientes a los domingos (representadas en azul).

A raíz de estas observaciones, se plantea la idea de separar los datos según el día de estudio, a fin de comprobar mejor las diferencias entre cada uno. Las siguientes gráficas muestran , cada una, las curvas correspondientes a un día de la semana en concreto en color y el resto de curvas en gris. Además, se incluye en trazo discontinuo la media funcional de las curvas coloreadas. Así, se puede concluir que las curvas de demanda correspondientes a los cinco primeros días de la semana (es decir, los días laborables) se comportan de forma más o menos similar entre sí, mientras que las relativas a los sábados adoptan un comportamiento significativamente distinto y las de los domingos un tercer comportamiento diferente.

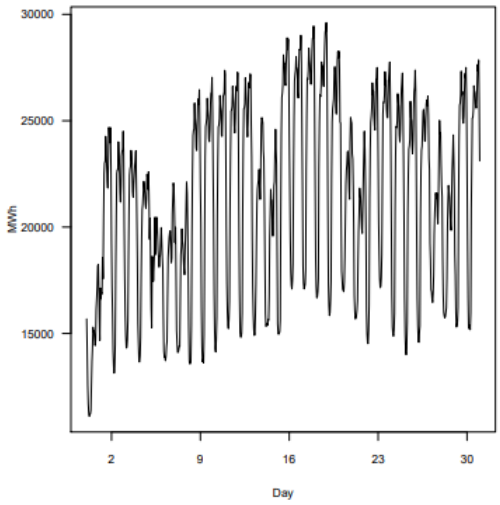

Figura 1.2: Serie de tiempo fundional de la demanda de electricidad en Enero de 2012.

En general, las curvas de los sábados son más bajas (menor demanda total) y no tienen picos máximos tan pronunciados como las de los días laborables. Este comporamiento se acrecenta más aún en los domingos. Desde el punto de vista interpretativo, este fenómeno es bastante coherente. Los domingos casi todas las fábricas y demás negocios descansan, mientras que los sábados se mantienen en un punto medio, por lo que es de esperar que la demanda de electricidad, siendo menor que en los días laborables, no sea tan baja como en los domingos.

A consecuencia de las conclusiones extraídas, el conjunto de curvas de demanda se divide en tres subgrupos: uno correspondiente a los días laborables, otro recogiendo los datos de los sábados y otro los de los domingos. Los conjuntos resultantes, junto a su media funcional, se pueden observar en la siguiente gráfica.

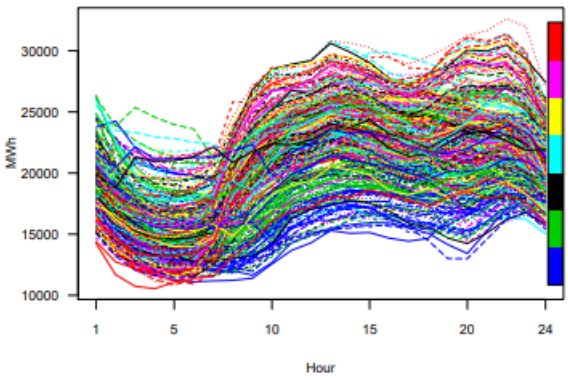

Figura 1.3: Curvas diarias de demanda. El color indica el día de la semana.

Cabría también preguntarse si existen diferencias significativas entre las curvas diarias de demanda en función del trimestre del año al que pertenezca el día. Como se ha visto, la tendencia de la serie iba cambiando a lo largo del día, observándose en general mayores valores de demanda en los primeros y los últimos meses del año, es decir, en el primer y el cuarto trimestre. La siguiente gráfica recoge las curvas separadas por semestres, indicando además dentro de cada gráfico qué curvas se corresponden con un día laborable (línea continua negra) y cuáles con un día de fin de semana (línea discontinua roja).

Debido a las características del consumo energético y su relación con el clima y las actividades económicas e industriales, es esperable un cambio de comportamiento de las curvas en función del trimestre del año al que pertenezcan. Observando el gráfico queda patente que estas diferencias existen, aunque, como se aprecia en esta misma imagen, no son tan pronunciadas como las existentes según el tipo de día.

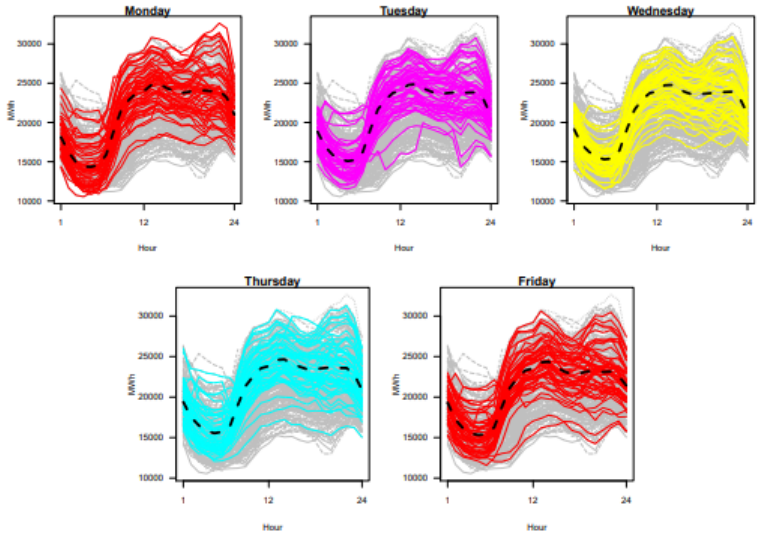

Figura 1.4: Curvas diarias de demanda de los días laborables.

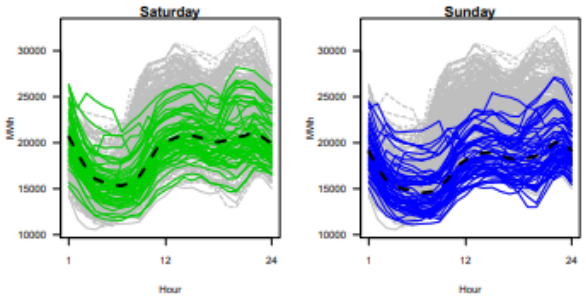

Figura 1.5: Curvas diarias de demanda de los fines de semana.

A continuación se incluye una comparativa de las curvas diarias de demanda según el mes del año. Una vez más, las curvas continuas negras se corresponden con datos de días laborables y las punteadas rojas con datos de fin de semana.

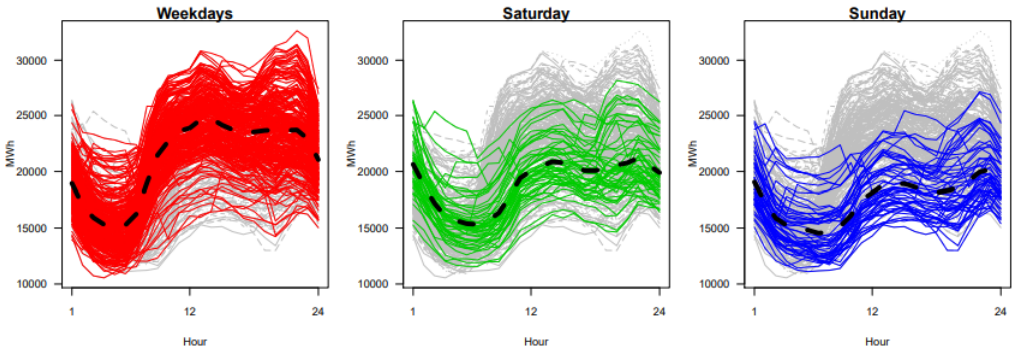

Figura 1.6: Curvas diarias de demanda separadas por el tipo de día. En discontinuo, la media funcional de cada grupo.

Precio

Los datos de precio de la electricidad comparten algunas características con los vistos de demanda, pero también tienen sus particularidades propias. Para analizarlos se seguirá el mismo procedimiento que con los del apartado anterior, echando primero un vistazo a la serie de tiempo completa para, a continuación, extraer los 366 datos funcionales.

Al observar la gráfica 1.9 (expresada en Cents/kWh frente a hora) lo primero que llama la atención del precio de la electricidad es que periódicamente experimenta fuertes disminuciones, llegando incluso a alcanzar en varias ocasiones el valor 0. Esto se debe a que en el mercado eléctrico español la energía eólica entra en la oferta diaria a un precio de 0Cents/kWh. Esto significa que en las horas en las que la energía proveniente de parques eólicos es suficiente como para abastecer toda la demanda requerida en el mercado el precio de la electricidad será de 0Cents/kWh. Este fenómeno explica el comportamiento de la serie de tiempo del precio horario en ese sentido.

Por otro lado, otra diferencia crucial es el cambio en la escala. Mientras que en la gráfica de la demanda nos encontrábamos con valores de entre 10000 y 30000 (MWh) en esta ocasión los valores del precio suelen estar entre 0 y 80 (Cents/kWh). Más allá de esto, la forma general de las curvas sí es similar, siendo la zona más baja la situada en la madrugada y con dos "picos.ªltos a mediodía y por la noche. Por otro lado, también se aprecia dependencia con el tipo del día de la semana, aunque no tan notable como en el caso de la demanda.

Observando el comportamiento del precio de la electricidad a lo largo de un solo mes (concretamente, Enero) quedan aún más patentes estas afirmaciones. Se aprecia de forma más clara la forma de las curvas de cada día, la ligera dependencia con el tipo de día y las bajadas abruptas que sufre el precio (en este caso la más reseñable ocurre el día 2).

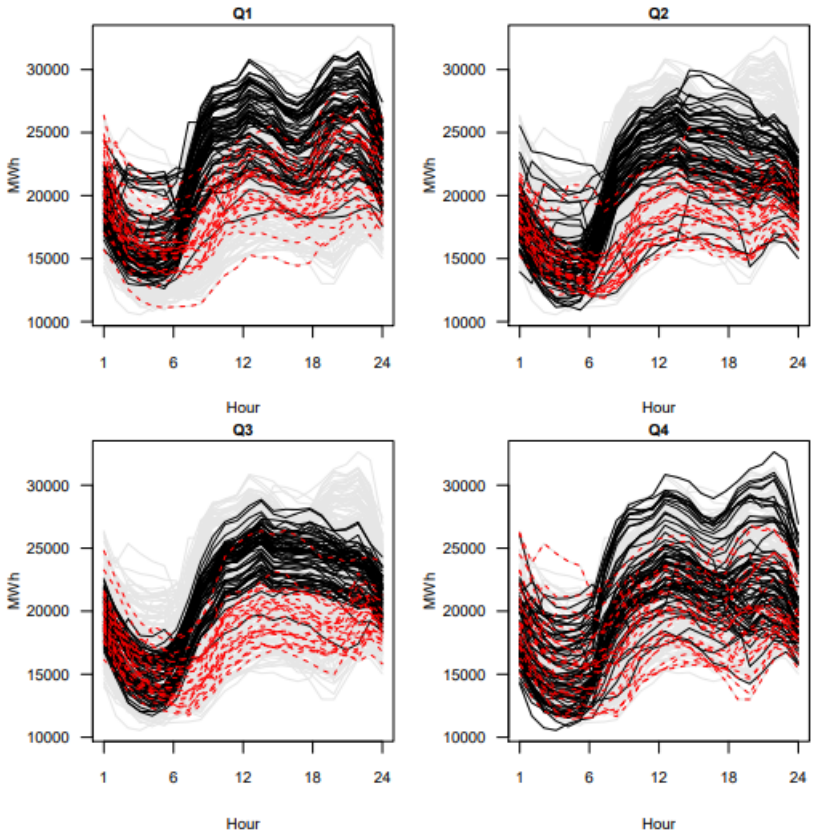

Figura 1.7: Curvas diarias de demanda separadas por trimestres.

En el gráfico 1.11 se vuelve a apreciar la similitud entre la forma de las curvas y la forma que tenían las curvas de demanda.

Se incluye también un gráfico comparativa del comportamiento de las curvas diarias de precio en el mercado eléctrico español separando los días en siete grupos en función del día de la semana al que pertenezcan.

Tras esta comparativa sí se observan diferencias significativas entre los días laborables, los sábados y los domingos. Los sábados tienden a seguir un patrón más suave, manteniéndose el precio relativamente estable a lo largo de las horas del día. Por otro lado, los domingos parecen distribuirse de forma distinta al resto de días, presentando en general precios sensiblemente más bajos. En conclusión, para trabajar con estos datos también los dividirmos en tres subconjuntos atendiendo al tipo de día al que pertenezcan.

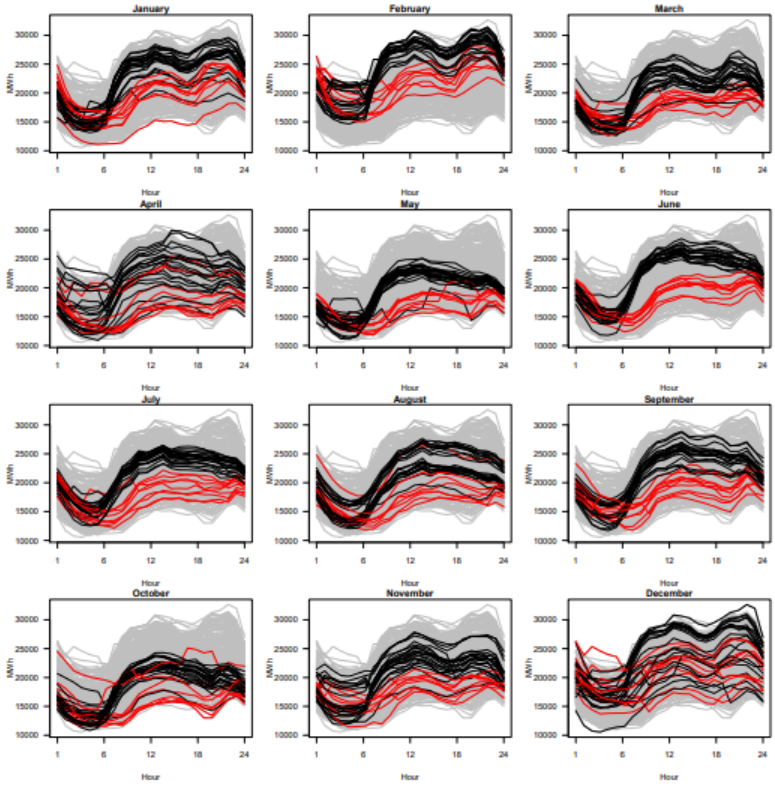

Figura 1.8: Curvas diarias de demanda separadas por meses.

Por último, siguiendo el mismo esquema que en el análisis de las curvas diarias de demanda, se presentan dos gráficas más que compararán las curvas diarias de precio en distintos momentos del año, a fin de encontrar alguna tendencia o relación. Así, en 1.15 se comparan las curvas atendiendo al trimestre del año al que pertenecen y en 1.16 se muestra lo propio pero dividiendo esta vez el año en sus meses.

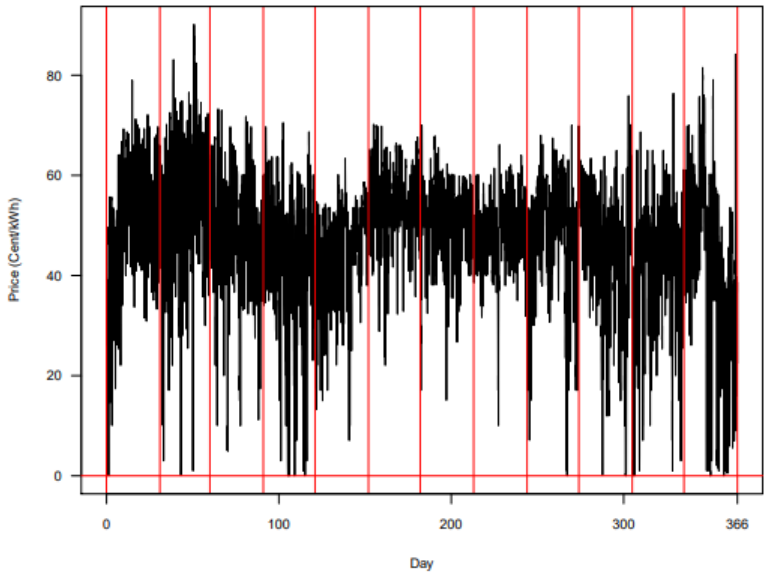

Figura 1.9: Serie de tiempo funcional de precio de electricidad en España en 2012.

Datos adicionales

Existen muchas variables que pueden influir en el valor de la demanda y/o el precio de la electricidad en el mercado diario español. Durante el propio análisis exploratorio de los datos se han visto cambios notables en el comportamiento de las curvas a lo largo del año que sugerían la influencia de factores climáticos en los valores de demanda y precio.
Hay que tener en cuenta que los datos sobre el clima no siempre son sencillos de obtener. Pueden ser complejos de recopilar, no ser gratuitos o presentar dificultades a la hora de incluirlos en el modelo. Por este motivo, se tendrán en cuenta únicamente dos fuentes externas de información: la temperatura y la producción eólica.

La temperatura influye en la cantidad de demanda de electricidad. Si es muy baja, la demanda aumentará a consecuencia del gasto energético que suponen los calefactores. De igual modo, si la temperatura es muy alta, la demanda se verá también incrementada por el gasto en aires acondicionados y demás. De este modo, en este trabajo se considerará la temperatura máxima diaria. Los datos se extraen de la web de AEMET.

La otra variable externa que se tendrá en cuenta es la energía eólica. Como se ha visto en la sección anterior, la producción eólica juega un papel clave en el precio de la electricidad. Debido a que esta

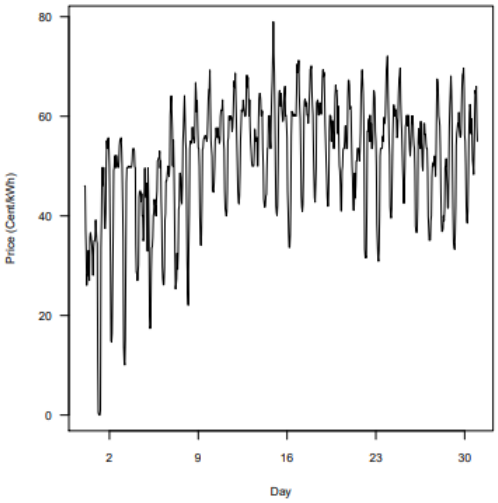

Figura 1.10: Serie de tiempo fundional del precio de la electricidad en Enero de 2012.

entra en el mercado ofertando a 0Cents/kWh, las curvas de precio de los días con mayor producción eólica serán más bajas, pudiendo incluso llegar a pasar por el 0. Por ello, se considerarán los datos de cantidad de demanda cubierta por la producción eólica cada día. Estos datos se pueden extraer de la web de REE.

1.2.3. Outliers

En cualquier análisis estadístico, y en especial si este implica predicción, es muy importante tener en cuenta la presencia de outliers. Un outlier es un dato anómalo, que se aleja del resto de datos o no sigue los mismos patrones. Estas observaciones pueden estropear el análisis de los datos, haciendo que se extraigan conclusiones contrarias a la realidad. Por ello es importante identificarlos.

Durante el análsis exploratorio de los datos de curvas diarias quedó patente la presencia de outliers. El caso más notable sería el de los días con precio nulo. Así pues, antes de empezar con la predicción se identificarán los outliers, tanto entre las curvas de demanda como en los de precio, y estos no serán tenidos en cuenta en lo sucesivo.

Una forma clásica de detectar outliers en la estadística univariante es mediante un diagrama de cajas. Este método no se puede aplicar directamente sobre los datos de este trabajo, pues son datos funcionales, pero sí se podrían discretizar las curvas y realizar un diagrama de caja para cada hora.

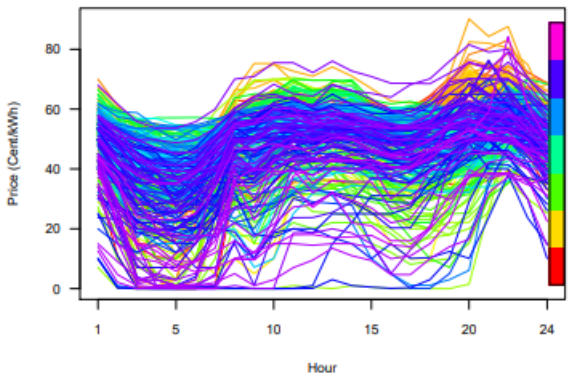

Figura 1.11: Curvas diarias de precio. El color indica el día de la semana.

Este método puede proporcionar una idea general de los outliers del conujnto de datos (por ejemplo, que entre las curvas de precio haya outliers por ser demasiado bajas), pero no se debe catalogar una curva entera como outlier porque uno de sus valores horarios sea outlier marginalmente (entre los valores de dicha hora). Será necesario recurrir a herramientas específicas de detección de outliers para datos funcionales.

Cabe destacar que, debido al poco tiempo que tiene la estadística con datos funcionales, no existe aún consenso en la definición formal de outlier al trabajar con curvas. En este trabajo se considerará la definición dada en Febrero y González-Manteiga (2008), que dicta que üna curva es outlier si ha sido generada por un proceso estocástico con distribución diferente al del resto de curvas".

Al trabajar con curvas, pueden aparecer dos tipos de outliers: outliers de magnitud y outliers de forma. Los outliers de magnitud son aquellas curvas que se encuentran fuera del rango habitual del resto de curvas. Los outliers de forma son curvas que, aun estando en el mismo rango que la mayoría, tienen una forma diferente. Además de estos dos tipos de outliers también podría aparecer, por supuesto, uno que combinase ambas condiciones.

En el caso particular de este trabajo, conviene recordar que no se está trabajando con unos datos funcionales cualesquiera, sino que se trata de series de tiempo, es decir, existe una dependencia temporal. Esto significa que cada curva tiene dependencia con la anterior y, además, dentro de la curva cada dato horario también depende del anterior. Esta dependencia también ha de ser tenida en cuenta a la hora de detectar los outliers.

Así, en este trabajo se utilizará el método propuesto en Raña, Aneiros, y Vilar (2015), a su vez

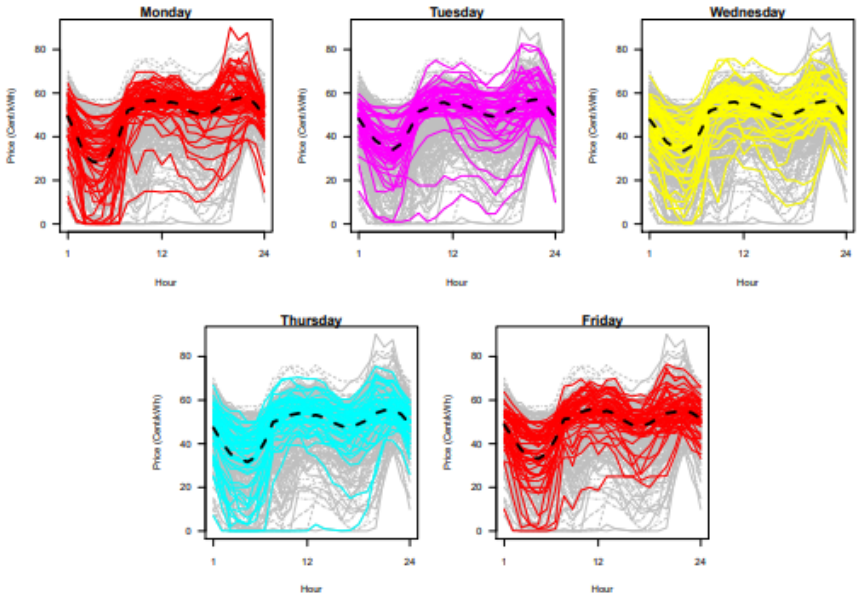

Figura 1.12: Curvas diarias de precio de los días laborables.

una modificación del propuesto en Febrero y González-Manteiga (2008) para identificar outliers con datos funcionales. Este método se basa en el uso de una función de profundidad funcional. Fijada la profundidad y una cota C, se calculan todas las profundidades del conjunto de curvas, se consideran outliers curvas cuya profundidad esté por debajo de C y se repite el proceso hasta que ninguna curva sea catalogada como outlier.

Es claro que la elección de C resulta clave para este método. En Raña y cols. (2015) se propone el siguiente procedimiento para su elección: Sea $S = \{\chi_i\}_{i=1}^n$ el conjunto de curvas donde se quieren detecar los outliers.

1. Detectar outliers en S a través de un método gráfico y definir S_1 el subconjunto de S sin los outliers.

2. Utilizar un procedimiento bootstrap aplicable a datos dependientes para obtener B muestras bootstrap, S^b, de tamaño n de S_1.

3. Obtener C^b como el cuantil empírico de orden α_1 de la distribución de profundidades $\{D_{S^b}(\chi_i^b) : \chi_i^b \in S^b\}$.

4. Tomar C como la mediana de $\{C^b\}_{b=1}^B$

De este modo la elección de C dependerá de la profundida utilizada, $D(\cdot)$, y del procedimiento utilizado para el remuestreo. En el presente trabajo se considerarán outliers los así considerados en

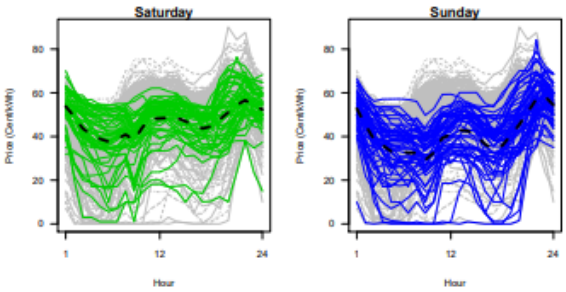

Figura 1.13: Curvas diarias de precio de los fines de semana.

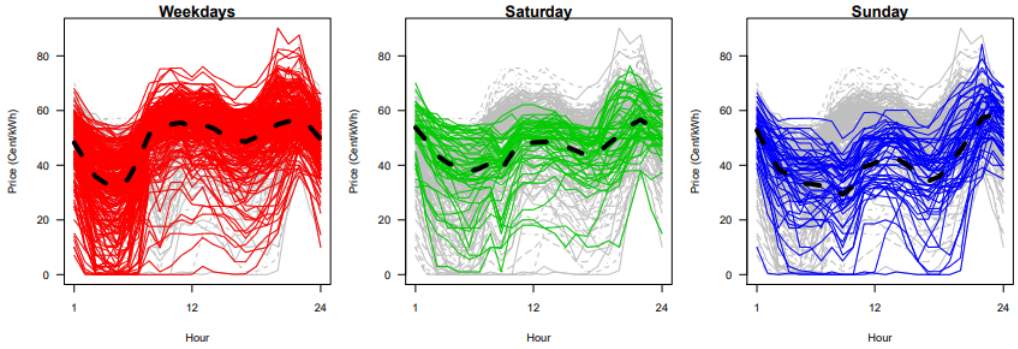

Figura 1.14: Curvas diarias de precio separadas por el tipo de día. En discontinuo, la media funcional de cada grupo.

Raña (2016) tras realizar un estudio considerando el procedimiento descrito con tres profundidades distintas (Modal Depth, Band Depth y Modified Modal Depth) y cuatro técnicas bootstrap (Standard Smoothed Bootstrap on Data, Moving Blocks Bootstrap, Stationary Bootstrap y Standard Smoothed Bootstrap on Residuals). Como resultado se obtienen 19 outliers, 10 en los datos de 2011 y 9 en los 2012. Para no interferir con la estructura de los datos, en lugar de eliminarse cada uno se sustituye por una ponderación de sus cuatro datos más cercanos en el tiempo, de forma que un dato outlier χ_i se sustituirá en el conjunto de datos por $0.2\chi_{i-2} + 0.3\chi_{i-1} + 0.3\chi_{i+1} + 0.2\chi_{i+2}$.

1.3. Predicción con datos funcionales

Los datos de demanda y precio se supondrán procesos estocásticos en tiempo continuo. Ya que son dos problemas análogos, se utilizará una misma notación para referirse a cualquiera de ellos indistintamente. De esta forma, contamos con un proceso estocástico $\{\zeta(t)\}_{t\in\mathbb{R}}$ estacionario de periodo $\tau = 24$. Sea N el número de días de los que se tienen datos, se puede definir el intervalo de observación del

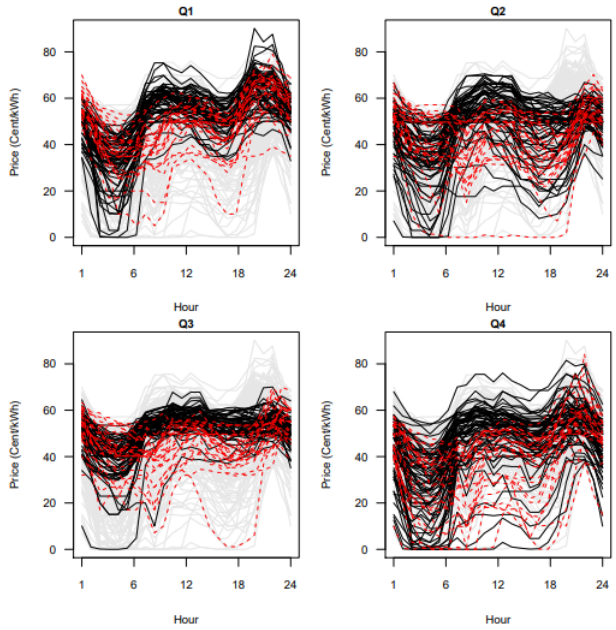

Figura 1.15: Curvas diarias de precio separadas por trimestres.

proceso como $(a,\ a + N\tau]$ para cierto $a \in \mathbb{R}$ y con ello definir

$$\zeta_i(t) = \zeta(a + (i-1)\tau + t), \quad t \in (0, \tau]. \tag{1.1}$$

Se obtiene así el conjunto de datos funcionales $\{\zeta_i\}_{1 \leq i \leq N}$ donde cada dato recoge la información (curva de precio o de demanda) de uno de los días observados. Se puede ahora plantear el problema de predecir ζ_{N+1} utilizando la información dada por las N curvas anteriores.

Se plantea el siguiente modelo de regresión:

$$\zeta_{i+1} = r(\chi_{i+1}) + \varepsilon_{i+1} = r(\zeta_i, \mathbf{x}_{i+1}, \xi_{i+1}) + \varepsilon_{i+1}\ 1 \leq i < N \tag{1.2}$$

donde

- ζ_{i+1} es la respuesta funcional.
- $r(\cdot)$ es la función de regresión desconocida (aunque supuesta de algún tipo: lineal, parcialmente lineal, autorregresiva...)
- χ_{i+1} es el vector de variables explicativas compuesto por
  - ζ_i: parte autorregresiva (supuesta de orden 1 pero generalizable).

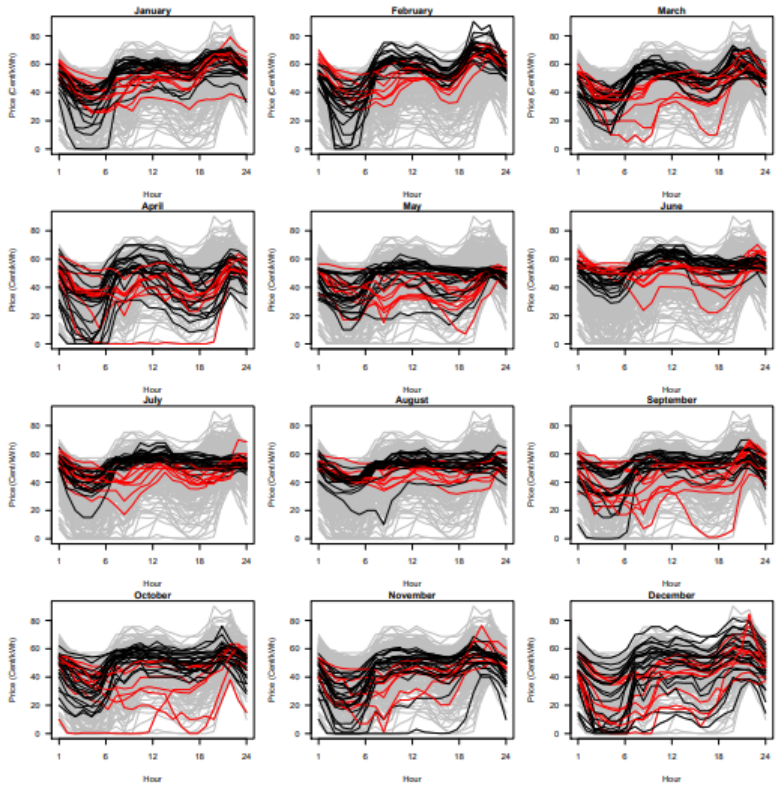

Figura 1.16: Curvas diarias de precio separadas por meses.

- $\mathbf{x}_{i+1}$: vector de explicativas escalares.
- ξ_{i+1}: vector de explicativas funcionales.

- ε_{i+1} es el error aleatorio funcional de media 0.

Así, para obtener una predicción $\widehat{\zeta}_{N+1}$ de la curva ζ_{N+1} bastará con encontrar una función $\widehat{r}(\cdot)$ que estime $r(\cdot)$, de forma que $\widehat{\zeta}_{N+1} = \widehat{r}(\chi_{N+1})$. Por tanto, la predicción de la curva dependerá del predictor de la función de regresión del modelo, el cual a su vez depende de las suposiciones que se hayan fijado sobre dicha función. En este trabajo se trabajará con los siguientes modelos:

- **Modelo autorregresivo (FNP):** El modelo se supone puramente autorregresivo, es decir, $\chi_{i+1} = \zeta_i$. De esta forma, $\zeta_{i+1} = m(\zeta_i) + \varepsilon_{i+1}$, donde $m(\cdot)$ es una función suave desconocida. Para calcular $\widehat{\zeta}_{N+1}$ se estima $m(\zeta_N)$ de forma no paramétrica.

- **Modelo parcialmente lineal (SFPL):** En este caso, se supone que el modelo tiene una parte autorregresiva y una parte lineal. Esto se consigue añadiendo al modelo anterior una componente lineal que depende de p variables escalares. Así, $\chi_{i+1} = (\zeta_i, \mathbf{x}_{i+1})$ y $\zeta_{i+1} = m(\zeta_i) + \mathbf{x}_{i+1}^T\boldsymbol{\beta} + \varepsilon_{i+1}$,

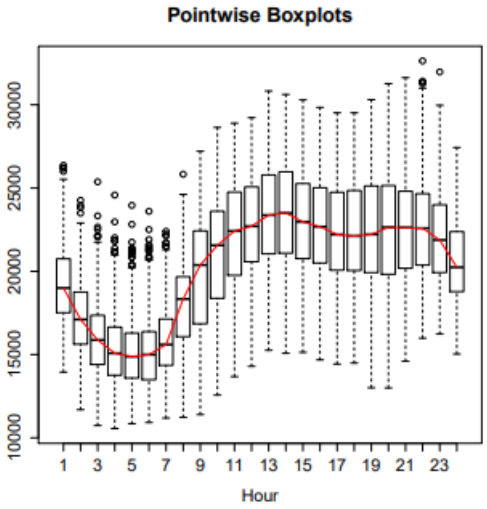

Figura 1.17: Diagramas de caja para demanda energética de cada hora.

siendo $\mathbf{x}_{i+1}$ el vector de las p variables escalares y $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ un vector de parámetros funcionales desconocidos. En esta ocasión, para calcular $\widehat{\zeta}_{N+1}$ será necesario encontrar $\widehat{\boldsymbol{\beta}}$ y $\widehat{m}(\zeta_N)$, estimaciones de β y $m(\zeta_N)$ respectivamente.

Estos dos modelos son los planteados en Aneiros, Vilar, y Raña (2016), donde se estudia a fondo su eficacia para la estimación puntual de curvas.

Nótese que el modelo está planteado para un caso general, con unas curvas cualesquiera. En el caso particular de este trabajo, los días se dividen en laborables, sábados y domingos. Para la predicción de cada día se utilizará la información de los 365 anteriores, pero únicamente la de los días de su misma categoría. Se especifica a continuación la notación a seguir.

Notación: Tomando las curvas $\{\zeta_i\}_{1\leq i\leq N}$ definidas en 1.1 se consideran los conjuntos

- $I_0 = \{N - 364,\ N - 363, ...,\ N - 1,\ N\}$
- $I_l = \{i \in I_0 : \zeta_i\ es\ laborable\}$
- $I_s = \{i \in I_0 : \zeta_i\ es\ sábado\}$
- $I_d = \{i \in I_0 : \zeta_i\ es\ domingo\}$

Y se define $\zeta_i^{I_l} = \zeta_{(i)}$ donde (i) es el i-ésimo elemento de I_l. Análogamente se definen $\zeta_i^{I_s}$ y $\zeta_i^{I_d}$.

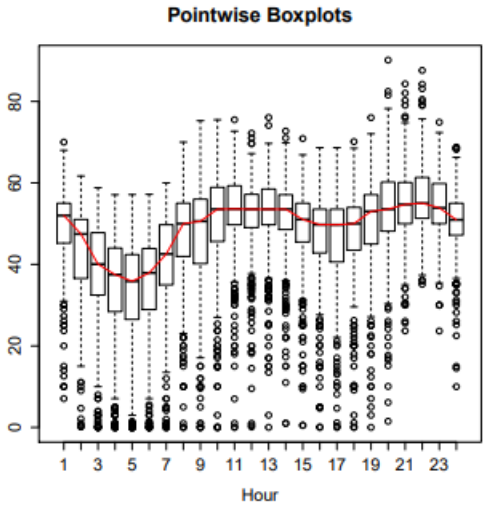

Figura 1.18: Diagramas de caja para precio de la electricidad de cada hora.

En conclusión, se tendrán seis conjuntos diferentes de curvas correspondientes respectivamente a la demanda y el precio de los días laborables, los sábados y los domingos. Además, cada uno de estos conjuntos se estudiará aplicando tanto el modelo FNP como el SFPL. Así, sea $S \in \{I_l,\ I_s,\ I_d\}$ según la categoría del día a predecir, tendremos estos dos modelos:

- **Modelo FNP:** $\widehat{\zeta}_{N+1} = \widehat{m}(\zeta_N^S)$.
- **Modelo SFPL:** $\widehat{\zeta}_{N+1} = \mathbf{x}_{N+1}^T \widehat{\boldsymbol{\beta}} + \widehat{m}(\zeta_N^S)$.

1.4. Regiones de predicción con datos funcionales

Sea ψ_{i+1} un elemento aleatorio de H, una región de predicción (PR) de ψ_{i+1} a un nivel de confianza $1-\alpha$ es un subconjunto aleatorio $R \subset H$ tal que

$$P(\psi_{i+1} \in R) = 1 - \alpha. \tag{1.3}$$

Para construir la PR de ψ_{i+1} se utilizará una predicción $\widehat{\psi}_{i+1}$ para así, fijada una norma $\|\cdot\|$ de H, hallar $q(\alpha) \in \mathbb{R}$ tal que la bola

$$B(\widehat{\psi}_{i+1}, q(\alpha)) = \left\{\xi : \|\xi - \widehat{\psi}_{i+1}\| < q(\alpha)\right\}$$

verifique 1.3.

La PR dependerá, por tanto, de

- el modelo de regresión en base al cual se obtiene la predicción,
- el método que se utilice para estimar la regresión,
- la norma con la que se trabaje,
- el nivel de significación fijado.

En este trabajo se considerarán dos modelos de regresión (FNP y SFPL) y tres niveles de significación ($(1-\alpha) \in \{0.95, 0.9, 0.8\}$). Además, se estudiará la influencia de la norma utilizando las siguientes:

- Norma $\|\cdot\|_1$, definida como $\|(x_1, ..., x_n)\|_1 = \sum_{i=1}^{n} |x_i|$
- Norma $\|\cdot\|_2$, definida como $\|(x_1, ..., x_n)\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$
- Norma $\|\cdot\|_\infty$, definida como $\|(x_1, ..., x_n)\|_\infty = \text{máx}_i\, |x_i|$

Para comparar distintas PR obtenidas al modificar uno o varios de esos elementos se puede calcular la cobertura real de cada una. Además, fijada una norma, también se pueden comparar las amplitudes de cada PR. Por la construcción de la PR, el $q(\alpha)$ nos indica el tamaño de la región, por lo que bastará compararlos; sin embargo, carecería de sentido utilizar este método para comparar regiones obtenidas utilizando diferentes normas.

En este trabajo se calcularán las PR basadas en los dos métodos de predicción expuestos con anterioridad, utilizando a su vez diversas normas y niveles de significación, y se compararán las coberturas (y si se acercan a la teórica) y, para cada norma, las amplitudes (a igualdad de cobertura se buscará menor amplitud).

Capítulo 2

Modelos de predicción

En este trabajo se calcularán las PR basadas en los modelos de regresión definidos en 1.2. Así, para cada día del año 2012 se construirán, tanto para los datos de demanda como para los de precio, regiones utilizando el modelo FNP y el modelo SFPL. En ambos casos, para la estimación de la función de regresión se utilizará, como ha sido mencionado, la información de los días de la misma categoría que el día a predecir transcurridos en el año natural previo. Además, se considerarán tres normas distintas ($\|\cdot\|_1,\ \|\cdot\|_2,\ \|\cdot\|_\infty$) y tres niveles de significación distintos (0.95, 0.9, 0.8)). Es decir, un total de 36 regiones por día (18 por cada variable).

2.1. Modelo funcional no paramétrico

Basándose en la propuesta de Ferraty, Laksaci, Tadj, y Vieu (2011) para regresión no paramétrica con variable y respuesta funcional para curvas independientes, en Aneiros, Vilar, Cao, y Muñoz San-Roque (2013) se propone un modelo de regresión para predecir curvas de demanda residual en el mercado eléctrico. Dado que es un modelo pensado para series de tiempo funcionales, también se puede aplicar para predecir las curvas de demanda y de precio.

La idea es construir tres modelos de regresión, uno para cada tipo de día a predecir. De esta forma, en función de si se desea predecir la curva correspondiente a un día laborable, un sábado o un domingo, se utilizará la información proporcionada por el anterior día laborable, el anterior sábado o el anterior domingo, respectivamente.
Fijado $S \in \{I_l,\ I_s,\ I_d\}$ en función del tipo de día que se quiera predecir, el modelo puramente

autorregresivo es el siguiente:

$$\zeta_{i+1}^S = m(\zeta_i^S) + \varepsilon_{i+1} \;\; 1 \leq i < \#S \tag{2.1}$$

donde $m(\cdot)$ es una función suave y ε_{i+1} un error aleatorio funcional de media 0. Por tanto, para predecir ζ_{N+1}^S basta con estimar $m(\zeta_N^S)$, lo cual se puede hacer mediante un estimador tipo Nadaraya-Watson para datos funcionales:

$$\widehat{m}_h^{FNP}(\zeta_N^S) = \sum_{1 \leq i < \#S} w_h(\zeta_N^S, \zeta_i^S)\zeta_{i+1}^S$$

donde los pesos $w_h(\cdot,\cdot)$ se construyen como sigue

$$w_h(\zeta_N^S, \zeta_i^S) = \frac{K(d(\zeta_N^S, \zeta_i^S)/h)}{\sum_{1 \leq i \leq \#S} K(d(\zeta_N^S, \zeta_i^S)/h)}$$

siendo $K : [0,\infty) \to [0,\infty)$ una función kernel, h un parámetro de suavizado y $d(\cdot,\cdot)$ una semimétrica. En Ferraty y cols. (2011) se pueden encontrar propiedades asintóticas de este estimador bajo el supuesto de independencia.

Este modelo depende de las elecciones que se hagan sobre la ventana, h, la semimetrica, $d(\cdot,\cdot)$ y la función kernel, $K(\cdot)$.

2.2. Modelo semifuncional parcialmente lineal

Como se observó en 1.2.2, existen otras variables que se pueden introducir en el modelo de predicción de curvas diarias de demanda y precio de la electricidad en España que podrían mejorar su precisión frente a modelos que solo consideren la autorregresión. En Aneiros y cols. (2013) se plantea una extensión del modelo anterior, añadiendo a este p covariables escalares que aportan información extra sobre la variable respuesta. De este modo, fijado $S \in \{I_l,\ I_s,\ I_d\}$ en función del tipo de día que se quiera predecir, el modelo resulta como sigue.

$$\zeta_{i+1}^S = r(\chi_{i+1}) = r(\zeta_i^S, \mathbf{x}_{i+1}) = m(\zeta_i^S) + \boldsymbol{\beta}\mathbf{x}_{i+1} + \varepsilon_{i+1} \;\; 1 \leq i < \#S \tag{2.2}$$

donde $\mathbf{x}_{i+1} = (x_1^{i+1}, ..., x_p^{i+1})^T$ denota un vector de p covariables escalares, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ es un vector de parámetros funcionales desconocidos, $m(\cdot)$ una función suave desconocida y ε_{i+1} un error aleatorio funcional de media 0. Por tanto, para predecir ζ_{N+1}^S se necesitará estimar $\boldsymbol{\beta}$ y $m(\zeta_N^S)$.

Se buscarán $\widehat{\boldsymbol{\beta}}$ y $\widehat{m}(\zeta_N^S)$ estimadores de $\boldsymbol{\beta}$ y $m(\zeta_N^S)$ respectivamente, para así a continuación definir

$$\widehat{\zeta}_{N+1}^S = \widehat{m}(\zeta_N^S) + \mathbf{x}_{N+1}^T \widehat{\boldsymbol{\beta}}. \tag{2.3}$$

En este trabajo se utilizarán los estimadores propuestos en Aneiros y cols. (2013), basados respectivamente en mínimos cuadrados y estimación kernel, ie

$$\widehat{\beta}_h = (\tilde{\mathbf{X}}_h^T \tilde{\mathbf{X}}_h)^{-1} \tilde{\mathbf{X}}_h^T \tilde{\boldsymbol{\zeta}}_h^S$$

donde por $\tilde{\mathbf{A}}_h$ se denota a $(\mathbf{I} - \mathbf{W}_h)\mathbf{A}$, con $\mathbf{W}_h = (w_h(\zeta_i^S, \zeta_j^S))_{1 \leq i,j \leq \#S}$ y $\boldsymbol{\zeta} = (\zeta_j^S)_{1 \leq j \leq \#S}$, y

$$\widehat{m}_h^{SFPL}(\zeta_N^S) = \sum_{1 \leq i < \#S} w_h(\zeta_N^S, \zeta_i^S) \left(\zeta_{i+1}^S - \mathbf{x}_{i+1}^T \widehat{\beta}_h \right).$$

2.2.1. Covariables

La demanda de electricidad se ve influida por factores meteorológicos (Hyde y Hodnett, 2015), (Taylor y Buizza, 2003), (Taylor, de Menezes, y McSharry, 2006). Se pueden tener en cuenta varios factores, como la humedad, las horas de sol, la presión o la cantidad de lluvia; sin embargo, la temperatura resume bien los cambios que experimenta la demanda debido al clima. Así pues, consideraremos la temperatura máxima de cada día como covariable.

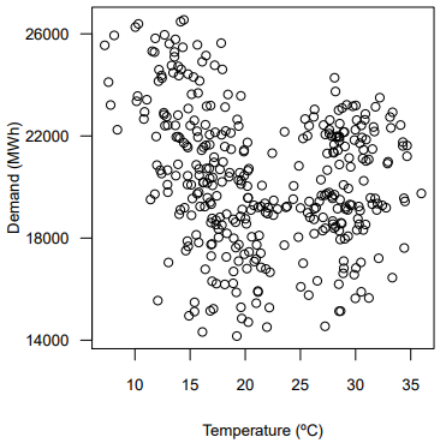

Figura 2.1: Temperatura máxima diaria frente a demanda eléctrica diaria.

Se puede apreciar que el efecto de la temperatura sobre la demanda no es lineal, sino que esta aumenta cuando la temperatura es especialmente baja o especialmente alta. Esto se debe a que a bajas temperaturas se tiende a utilizar la calefacción, mientras que bajo altas temperaturas aumenta el uso de los ventiladores y aires acondicionados. Dado que en el modelo planteado solo se contemplan

covariables con efecto lineal, dividiremos la variable de temperatura en dos, representando los días calurosos y los fríos. En el gráfico se observa que en el intervalo entre 20 y 24 grados centígrados la demanda se ve menos afectada, por lo que las nuevas variables se definirán como sigue:

$$HDD = max\{20 - T(t), 0\},$$

$$HDD = min\{T(t) - 24, 0\}$$

donde $T(t)$ representa la temperatura máxima diaria en el día t.
Obtenemos así las variables HDD y CDD:

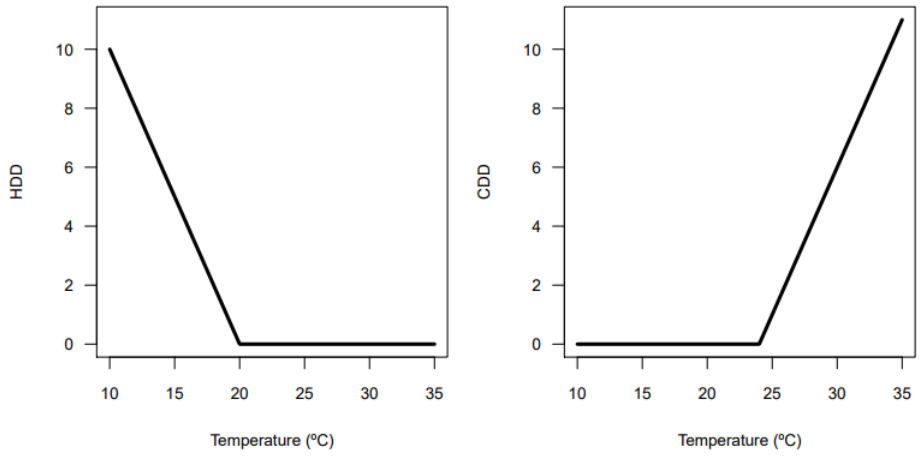

Figura 2.2: Covariables HDD y CDD.

Por otro lado, en lo referido a las curvas de precio se considerarán como variables externas la demanda diaria y la producción eólica. La primera variable representa la cantidad total de energía eléctrica (medida en MWh) que se demandó a lo largo de cada día, mientras que la segunda contiene la información de la cantidad de esa demanda diaria que fue cubierta por la energía eólica ese día.
A continuación se comparan ambas variables con el precio medio diario.

Se observa que en ambos casos la relación se puede suponer lineal. Debido a la estructura del mercado eléctrico español, ambas correlaciones son esperables, dado que al ser un sistema de oferta y demanda el precio depende de la demanda y, además, la cantidad de energía procedente del viento reduce el precio y es con diferencia la energía renovable con mayor presencia en el mercado.

Desde el punto de vista práctico, puede parecer poco útil la elección de covariables que se ha realizado, debido a que todos son datos desconocidos hasta el propio día del que queremos hacer la

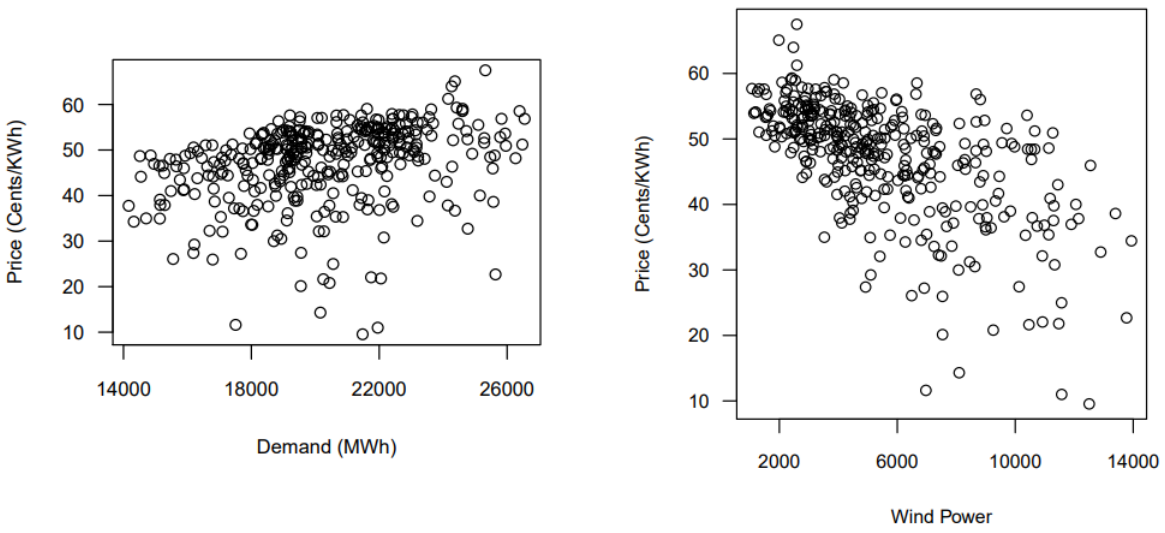

Figura 2.3: Demanda diaria y producción eólica frente a precio medio diario.

estimación; sin embargo, es bien sabido que existen métodos muy precisos, aunque no públicos, para estimar tanto la temperatura máxima como la producción eólica del día siguiente, por lo que incluir los datos reales no desvirtuará de forma notable el resultado. Por otro lado, para el caso de la demanda basta con utilizar como estimación la predicción realizada en este mismo trabajo.

2.3. Algoritmo de predicción

Se tiene que el predictor de $\zeta^S_{N+1}|\chi_{N+1}$ es $\widehat{r}(\chi_{N+1})$, y podemos hacer la siguiente descomposición:

$$\zeta^S_{N+1}|\chi_{N+1} = r(\chi_{N+1}) - (\widehat{r}(\chi_{N+1}) - \widehat{r}(\chi_{N+1})) + (\varepsilon_{N+1}|\chi_{N+1})$$
$$\zeta^S_{N+1}|\chi_{N+1} - \widehat{r}(\chi_{N+1}) = (r(\chi_{N+1}) - \widehat{r}(\chi_{N+1})) + (\varepsilon_{N+1}|\chi_{N+1})$$

Desconocemos el valor real de $r(\chi_{N+1})$, así como de $\varepsilon_{N+1}|\chi_{N+1}$, por lo que debemos aproximarlos. Para ello, siguiendo la filosofía expuesta en Zhu y Politis (2017) y Vilar, Raña, y Aneiros (2018), utilizaremos procedimientos bootstrap.

Antes de elaborar el algoritmo debemos definir una nueva función $\widehat{r}^*_{hb}(\cdot)$, que dependerá de las suposiciones que hagamos sobre $r(\cdot)$. Partiendo de una muestra $\{(X_k, Y_k) : k \in K\}$ y queriendo aproximar $Y_N = r(X_N) + \varepsilon_N$, definimos:

- Siguiendo el modelo FNP

$$\widehat{r}^*_{hb}(X_N) = \widehat{m}^{*FNP}_h(X_N)$$

con

$$\widehat{m}_h^{*FNP}(X_N) = \sum_{k \in K} w_h(X_N, X_k) Y_k.$$

- Siguiendo el modelo SFPL

$$\widehat{r}_{hb}^{*}(X_N) = \widehat{m}_h^{*SFPL}(Y_N) + \widehat{\beta}_b^{*} \mathbf{x}_N$$

con

$$\widehat{\beta}_b^{*} = (\tilde{\mathbf{X}}_b^T \tilde{\mathbf{X}}_b)^{-1} \tilde{\mathbf{X}}_b^T \tilde{\mathbf{Y}}_b \;\; con \;\; \mathbf{Y} = (Y_k)_{k \in K}$$

y

$$\widehat{m}_h^{*SFPL}(Y_N) = \sum_{k \in K} w_h(X_N, X_k) \left(Y_k - \mathbf{x}_k^T \widehat{\beta}_b \right).$$

Establecido esto, se propone una versión del algoritmo utilizado en Vilar y cols. (2018) para la construcción de intervalos de predicción utilizando datos funcionales, pero adaptándolo a un modelo de respuesta funcional siguiendo la propuesta de Zhu y Politis (2017).

ALGORITMO BOOTSTRAP

1. Calcular el predictor de ζ_{N+1}^S dado por $\widehat{r}_h(\chi_{N+1})$, donde h representa la ventana de suavizado requerida en los modelos planteados.

2. Computar $\widehat{r}_b(\chi_i)$ para $1 < i \leq \#S$, es decir, las predicciones de cada día de la muestra utilizando la función estimada para el caso $N+1$ pero con una ventana b (asintóticamente mayor que h).

3. Calcular los residuos $\widehat{\varepsilon}_{i,b} = \zeta_i^S - \widehat{r}_b(\chi_i)$ para $1 < i \leq \#S$.

4. Centrar los residuos: $\widehat{\epsilon}_{i,b} = \widehat{\varepsilon}_{i,b} - \bar{\widehat{\varepsilon}}_{i,b}$ donde $\bar{\widehat{\varepsilon}}_{i,b} = (\#S - 1)^{-1} \sum_{1<i\leq \#S} = \widehat{\varepsilon}_{i,b}$.

5. Denotando por $F_{\widehat{\epsilon}}$ la distribución empírica de $\widehat{\epsilon}_{i,b}$, generar para $1 < i \leq \#S$ los pseudo-residuos $\widehat{\varepsilon}_i^*$ i.i.d de $F_{\widehat{\epsilon}}$, y con ellos construir la muestra bootstrap

$$\zeta_i^* = \widehat{r}_b(\chi_i) + \widehat{\varepsilon}_i^* \;\; 1 < i \leq \#S.$$

6. Con la muestra bootstrap $\{(\chi_i, \zeta_i^*) : 1 < i \leq \#S\}$ construir el predictor

$$\widehat{\zeta}_{N+1}^* = \widehat{r}_{hb}^*(\chi_{N+1}).$$

7. Repetir B veces los pasos 5-6 para obtener las B predicciones $\{\widehat{\zeta}_{N+1}^{*j}\}_{j=1}^B$.

8. Generar $\{\varepsilon_{j,N+1}^*\}_{j=1}^B$ m.a.s de tamaño B de la distribución $F_{\widehat{\epsilon}}$.

9. Almacenar el conjunto de errores bootstrap

$$Errors.Boot = \{\widehat{r}_b(\chi_{N+1}) - \widehat{r}_{hb}^*(\chi_{N+1}) + \varepsilon_{j,N+1}^*\}_{j=1}^B.$$

10. Fijada la norma $\|\cdot\|$ se calcula

$$\delta_j^* = \|\widehat{r}_b(\chi_{N+1}) - \widehat{r}_{hb}^*(\chi_{N+1}) + \varepsilon_{j,N+1}^*\| \ , \ j = 1, ..., B.$$

11. Finalmente, fijado un nivel de confianza $1-\alpha$, se extrae el cuantil $q(\alpha) = q_{1-\alpha}(\delta^*)$ y la PR buscada será

$$\{\xi \in H : \|\xi - \widehat{r}_b(\chi_{N+1})\| \leq q(\alpha)\}.$$

De esta forma, estamos estimando el error real

$$\zeta_{N+1}^S|\chi_{N+1} - \widehat{r}(\chi_{N+1}) = (r(\chi_{N+1}) - \widehat{r}(\chi_{N+1})) + (\varepsilon_{N+1}|\chi_{N+1})$$

obteniendo B errores bootstrap

$$\widehat{r}_b(\chi_{N+1}) - \widehat{r}_{hb}^*(\chi_{N+1}) + \varepsilon_{j,N+1}^*$$

aproximando así tanto la variación debida a la estimación de r $((r(\chi_{N+1}) - \widehat{r}(\chi_{N+1})))$ como la variación no explicada debida al error aleatorio del modelo $(\varepsilon_{N+1}|\chi_{N+1})$.

Capítulo 3

Resultados

El algoritmo propuesto se utiliza para calcular las PR tanto de la demanda como del precio de la electricidad de cada día del año 2012 (siempre utilizando la información de los días de su misma categoría de entre los 365 previos). Para cada día se calculan un total de 18 regiones para cada variable de estudio, combinando los dos métodos propuestos (FNP y SFPL) con las tres normas ($\|\cdot\|_1,\ \|\cdot\|_2,\ \|\cdot\|_\infty$) y los tres niveles de significación (0.95, 0.9, 0.8) que se había propuesto utilizar. Para los parámetros de los modelos se siguen las recomendaciones incluidas en Aneiros y cols. (2013), es decir, se utiliza una semimétrica basada en componentes principales (pues los datos no son suaves), se calcula el número óptimo de componentes y la ventana óptima mediante validación cruzada y se utiliza el kernel Epanechnikov ($K(u) = 0.75(1-u^2)1_{0,1}$). Para la ventana de remuestreo se toma $b = h$ en el caso FNP y $b = 2h$ en el SFLP. Por otro lado, el número de remuestreos se fija en $B = 1000$ excepto para el método SFPL, en el cual, debido a limitaciones temporales producidas por el alto gasto computacional, se ha tenido que ver reducido a $B = 700$ para los fines de semana y $B = 200$ para los laborables.

3.1. Método Benchmark

A fin de estudiar la eficacia del algoritmo propuesto se confecciona un método Benchmark para calcular la PR de forma menos elaborada. Así, fijados variable de estudio, tipo de día, método de predicción, norma y nivel de confianza, se calcularán dos regiones de predicción y sus correspondientes amplitudes y coberturas, lo que permitirá relativizar mejor los resultados. El método que se propone como Benchmark es el siguiente:

- Calcular el predictor de ζ_{N+1}^S dado por $\widehat{r}_h(\chi_{N+1})$

- Calcular los residuos $\widehat{\varepsilon}_{i,h} = \zeta_i - \widehat{r}_h(\chi_i)$ para $1 < i \leq \#S$.
- Centrar los residuos: $\widehat{\epsilon}_{i,h} = \widehat{\varepsilon}_{i,h} - \bar{\widehat{\varepsilon}}_{i,h}$ donde $\bar{\widehat{\varepsilon}}_{i,h} = (\#S-1)^{-1} \sum_{1<i\leq \#S} = \widehat{\varepsilon}_{i,h}$.
- Fijada la norma $\|\cdot\|$ se calcula
$$\eta_i = \|\widehat{\epsilon}_{i,h}\| \; 1 < i \leq \#S$$
- Finalmente, fijado un nivel de confianza $1-\alpha$, se extrae el cuantil $q_{BM}(\alpha) = q_{1-\alpha}(\boldsymbol{\eta})$ y la PR buscada será
$$\{\xi \in H : \|\xi - \widehat{r}_b(\chi_{N+1})\| \leq q_{BM}(\alpha)\}.$$

3.2. Tablas de coberturas

A continuación se procede construir las PR objetivo del trabajo y a evaluar la calidad de la predicción. Para ello, aprovechando que se dispone de los datos del año 2012, se hará uso de las coberturas reales, comprobando en cada caso la proporción de curvas reales que se encuentran abarcadas por su región correspondiente. Es decir, fijados unos parámetros, se calcularán las PR de todos los días de 2012 y se comprobará cuántas de ellas contienen al verdadero dato que se observó ese día. Se tendrá en cuenta también la amplitud (el valor de $q(\alpha)$) de cada región. Fijada una norma, estos dos datos se recogerán tanto para la región calculada para la demanda como la calculada para el precio, comparando las diferencias que puedan existir en función del tipo de día (laborable, sábado, domingo o total), el nivel de confianza (0.95, 0.90, 0.80) y el modelo de regresión considerado (FNP o SFPL). También se recogerán sus homólogos, para cada combinación, calculados mediante el método propuesto como benchmark.

Estos datos se agruparán y mostrarán a continuación mediante tablas, elaborando un total de dos para cada norma (una para los datos de demanda y otra para los de precio).

3.2.1. Norma $\|\cdot\|_1$

	Método bootstrap			**Método benchmark**		
FNP	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.908(54707)	0.850(43267)	0.735(32424)	0.893(50454)	0.816(39964)	0.655(28750)
Sábados	0.846(72458)	0.769 (59939)	0.692 (43990)	0.808 (65020)	0.769(53323)	0.577(37932)
Domingos	0.906(72784)	0.755(57964)	0.547(43599)	0.717(61060)	0.623(48577)	0.434(36094)
Total	0.90(59851)	0.83(47764)	0.70(35686)	0.86(54060)	0.78(43110)	0.61(31119)
SFPL	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.935(52478)	0.877(41617)	0.732(31408)	0.912(50693)	0.843(40029)	0.674(28946)
Sábados	0.846(72361)	0.788(60587)	0.731(46729)	0.808(62344)	0.750(53245)	0.615(39368)
Domingos	0.868(68449)	0.736 (55979)	0.585(43545)	0.679(54530)	0.585(43509)	0.434(33487)
Total	0.92(57616)	0.85(46392)	0.71(35343)	0.87(52904)	0.79(42411)	0.63(31085)

Cuadro 3.1: Cobertura y (entre paréntesis) amplitud de las PR para la demanda eléctrica del año 2012 calculadas utilizando la norma $\|\cdot\|_1$.

	Método bootstrap			**Método benchmark**		
FNP	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.874(246.6)	0.774 (199.4)	0.663 (153.7)	0.651(227)	0.732(180.2)	0.582(136.6)
Sábados	0.788 (245.1)	0.731 (206.6)	0.519 (168.8)	0.712 (211.2)	0.577 (180.7)	0.481(148.0)
Domingos	0.868 (347.8)	0.698 (279.6)	0.472 (214.6)	0.755 (302.8)	0.585(239.6)	0.396(179.5)
Total	0.85(260.15)	0.76(211.90)	0.62(164.58)	0.82(235.69)	0.69(188.94)	0.54(144.43)
SFPL	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.897(243.5)	0.793(196.9)	0.667(148.8)	0.889(237.2)	0.785(191.6)	0.636(139.7)
Sábados	0.788(197.2)	0.769(175.2)	0.635(150.5)	0.769(172.1)	0.615(152.7)	0.558(135.3)
Domingos	0.774(243)	0.698(208.7)	0.604(175.2)	0.698(203.1)	0.623(178.1)	0.509 (154.6)
Total	0.87(236.84)	0.78(195.49)	0.65(152.89)	0.85(223.04)	0.74(184.10)	0.61(141.25)

Cuadro 3.2: Cobertura y (entre paréntesis) amplitud de las PR para el precio eléctrico del año 2012 calculadas utilizando la norma $\|\cdot\|_1$.

De 3.1 y en especial de 3.2 se concluye que las coberturas del modelo SFPL mejoran a las del FNP. Además, la amplitud media de las regiones es sensiblemente más pequeña. Llama la atención que en los datos del método benchmark ya se pueda apreciar esta mejora significativa. El tiempo de computación, eso sí, es mucho más lento en el modelo SFPL: mientras que calcular la región para un día laborable siguiendo el modelo FNP apenas tarda un par de minutos, en el modelo SFLP puede llegar a tardar hasta 90.

3.2.2. Norma $\|\cdot\|_2$

	Método bootstrap			**Método benchmark**		
FNP	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.923(12809)	0.851(10400)	0.720(7802)	0.908(12021)	0.808(9603)	0.655(6920)
Sábados	0.827 (15856)	0.788(13172)	0.712(9922)	0.808(14259)	0.769(11777)	0.558 (8567)
Domingos	0.868 (15778)	0.755(12850)	0.547 (9875)	0.736(13363)	0.585(10942)	0.434(8213)
Total	0.90(13672)	0.83(11148)	0.70(8404)	0.87(12534)	0.77(10106)	0.61(7341)
SFPL	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.935 (12232)	0.866(9902)	0.747(7542)	0.919(11952)	0.828(9470)	0.697(7011)
Sábados	0.846(15695)	0.788(13291)	0.750(10476)	0.788(13531)	0.750(11645)	0.654(8804)
Domingos	0.849 (14968)	0.717 (12496)	0.585(9950)	0.679(12148)	0.604(9766)	0.472(7698)
Total	0.91(13120)	0.84(10760)	0.73(8308)	0.87(12205)	0.79(9823)	0.66(7365)

Cuadro 3.3: Cobertura y (entre paréntesis) amplitud de las PR para la demanda eléctrica del año 2012 calculadas utilizando la norma $\|\cdot\|_2$.

	Método bootstrap			**Método benchmark**		
FNP	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.854(62.85)	0.801(50.42)	0.661(38.53)	0.847(58.51)	0.751(46.05)	0.567(34.43)
Sábados	0.827(58.86)	0.731(50.23)	0.519(40.89)	0.731(52.29)	0.596(44.45)	0.481(36.13)
Domingos	0.849(78.79)	0.660(64)	0.491 (50.49)	0.698(68.39)	0.547(55.85)	0.434(42.14)
Total	0.85(64.44)	0.77(52.33)	0.62(40.55)	0.81(59.06)	0.70(47.24)	0.54(37.80)
SFPL	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.904(61)	0.805(48.77)	0.682(37.42)	0.900(59.90)	0.782(46.44)	0.644(35.42)
Sábados	0.827(49.31)	0.769(43.44)	0.615(36.83)	0.788(43.81)	0.673(38.78)	0.500(32.72)
Domingos	0.736(57.45)	0.698(49.97)	0.604(42.03)	0.679(49.59)	0.604(43.74)	0.5089(36.97)
Total	0.87(58.83)	0.79(48.19)	0.66(38.00)	0.85(56.12)	0.74(44.96)	0.61(35.26)

Cuadro 3.4: Cobertura y (entre paréntesis) amplitud de las PR para el precio eléctrico del año 2012 calculadas utilizando la norma $\|\cdot\|_2$.

En esta ocasión, si bien tanto en 3.3 como en 3.4 se observa mejoría tanto en las coberturas como en las amplitudes, no se ve tanta diferencia como utilizando la norma anterior. En este caso, y a falta de realizar un estudio sobre la elección de la ventana h, no parece muy recomendable utilizar el modelo SFPL, pues los tiempos de computación son mucho más lentos.

3.2.3. Norma $\|\cdot\|_\infty$

	Método bootstrap			Método benchmark		
FNP	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.916(4638)	0.835(3834)	0.701(2983)	0.889(4329)	0.789(3537)	0.636(2669)
Sábados	0.865(5249)	0.827(4450)	0.692(3515)	0.827(4654)	0.769(3925)	0.615(3087)
Domingos	0.792 (5085)	0.755(4286)	0.585(3486)	0.736(4357)	0.641(3690)	0.416(2959)
Total	0.89(4790)	0.82(3987)	0.68(3132)	0.86(4379)	0.77(3614)	0.60(2771)
SFPL	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.916(4334)	0.847(3614)	0.736(2897)	0.908(4190)	0.831(3469)	0.693(2716)
Sábados	0.885(5245)	0.846(4528)	0.731(3663)	0.827(4597)	0.808(3999)	0.596(3043)
Domingos	0.792(4943)	0.755(4221)	0.585(3434)	0.736(4053)	0.566(3433)	0.358(2681)
Total	0.90(4552)	0.84(3832)	0.72(3084)	0.87(4229)	0.79(3539)	0.63(2757)

Cuadro 3.5: Cobertura y (entre paréntesis) amplitud de las PR para la demanda eléctrica del año 2012 calculadas utilizando la norma $\|\cdot\|_\infty$.

En este caso sorprende la igualdad entre los métodos FNP y SFPL. Si bien en el método benchmark sí se aprecian diferencias significativas entre las coberturas de uno y otro modelo, en el método bootstrap estas son muy similares entre sí. Las amplitudes medias tampoco son mucho menores utilizando el modelo SFPL. En 3.1, 3.2 y 3.3 se observa que, tomando un día al azar, las regiones calculadas por uno y otro método son muy similares, independientemente del nivel de confianza.

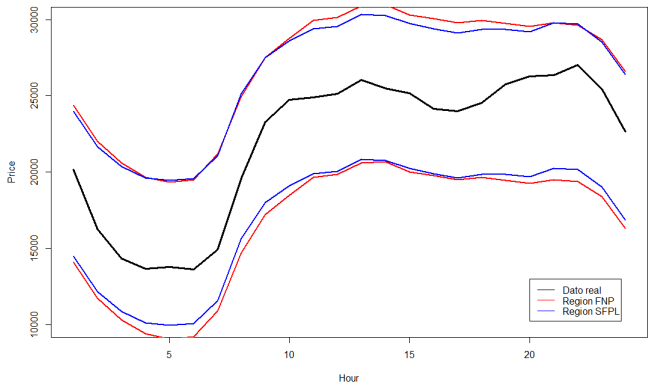

Figura 3.1: Regiones de predicción al 95 % siguiendo ambos modelos.

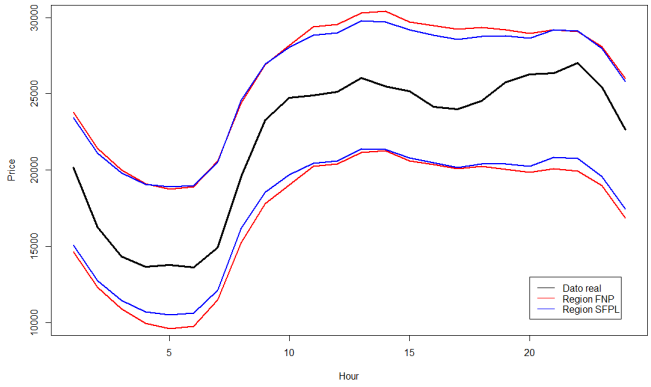

Figura 3.2: Regiones de predicción al 90 % siguiendo ambos modelos.

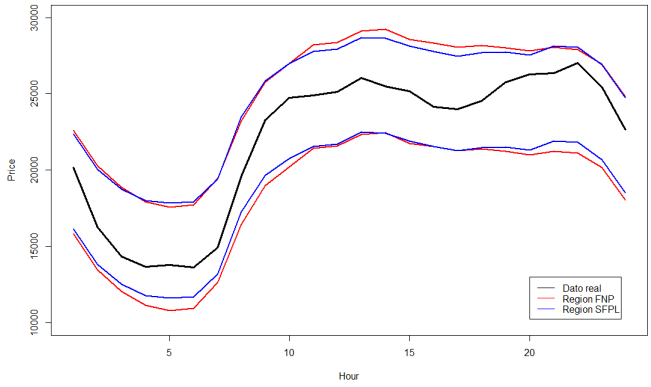

Figura 3.3: Regiones de predicción al 80 % siguiendo ambos modelos.

El día elegido al azar fue laborable. En 3.10 y 3.4 se toma respectivamente un sábado y un domingo. Las diferencias continúan siendo mínimas.

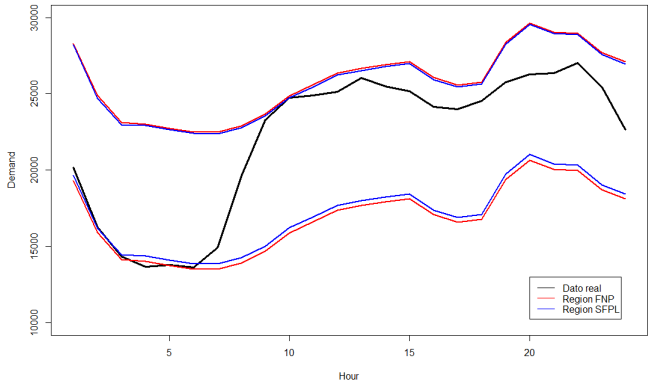

Figura 3.4: Regiones de predicción al 95 % siguiendo ambos modelos para un sábado.

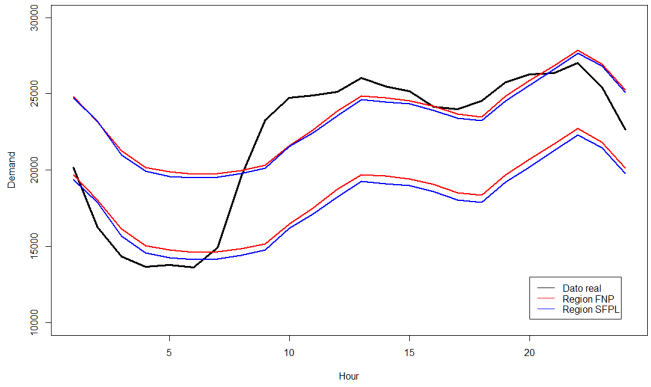

Figura 3.5: Regiones de predicción al 80 % siguiendo ambos modelos para un domingo.

En lo respectivo al precio, como se puede apreciar en 3.6, las coberturas de los días laborables no son significativamente mejores en el modelo SFPL que en el FNP. En sábados y domingos se aprecian mayores diferencias, aunque probablemente se deba a los reducidos tamaños muestrales. Una vez, en la interpretación gráfica las regiones de predicción son muy similares.

Para sábados y domingos se representan los tres niveles de confianza en un mismo gráfico, apreciándose bien que las líneas rojas prácticamente se solapan con las azules.

	Método bootstrap			Método benchmark		
FNP	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.916 (27.48)	0.831 (22.1)	0.690(16.34)	0.900(25.68)	0.774 (20.27)	0.613(14.58)
Sábados	0.846(24.59)	0.827(20.82)	0.500(16.22)	0.808(22.52)	0.692(18.84)	0.481(14.19)
Domingos	0.811 (28.67)	0.642(23.55)	0.472(19.16)	0.660(25.67)	0.528(20.70)	0.358(16.05)
Total	0.89(27.23)	0.81(22.11)	0.63(16.71)	0.85(25.23)	0.73(20.13)	0.59(14.54)
SFPL	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.2$
Laborables	0.920(26.37)	0.835(20.95)	0.701(15.44)	0.920(26.18)	0.801(19.97)	0.655(14.68)
Sábados	0.904(22.03)	0.788(18.52)	0.635(15.2)	0.865(20.21)	0.673(16.49)	0.404(13.43)
Domingos	0.792(23.10)	0.698(19.80)	0.547(16.78)	0.679(19.93)	0.604(17.41)	0.396(15.03)
Total	0.90(27.28)	0.81(20.44)	0.67(15.60)	0.88(24.42)	0.76(19.11)	(0.58(14.55)

Cuadro 3.6: Cobertura y (entre paréntesis) amplitud de las PR para el precio eléctrico del año 2012 calculadas utilizando la norma $\|\cdot\|_\infty$.

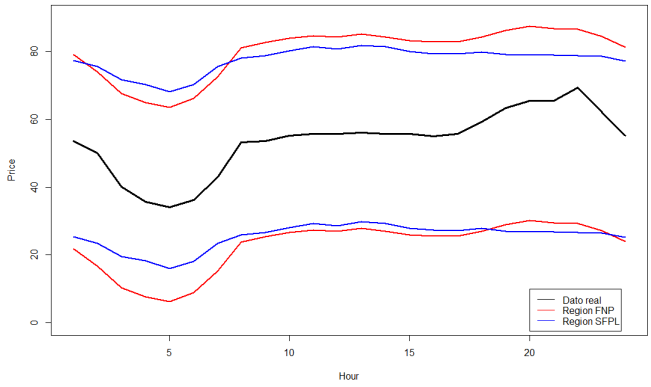

Figura 3.6: Regiones de predicción al 95 % siguiendo ambos modelos para un laborable.

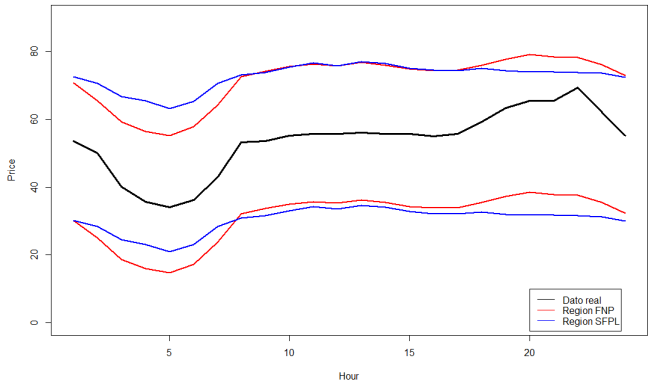

Figura 3.7: Regiones de predicción al 90 % siguiendo ambos modelos para un laborable.

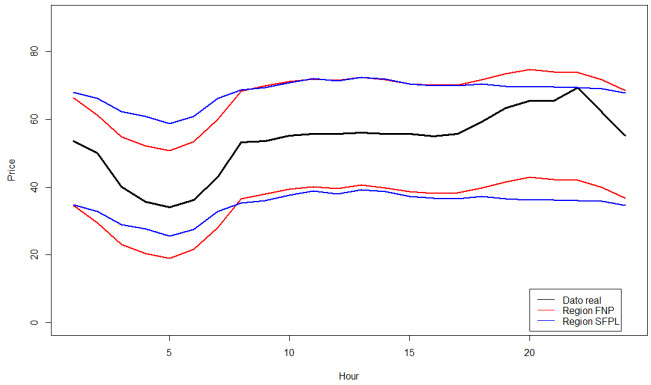

Figura 3.8: Regiones de predicción al 80 % siguiendo ambos modelos para un laborable.

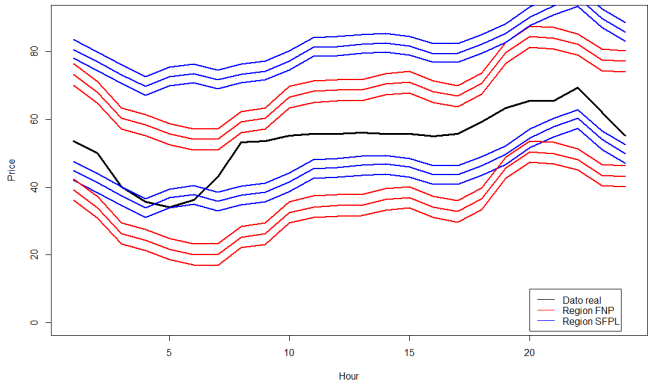

Figura 3.9: Regiones de predicción al 80 %, 90 % y 95 % siguiendo ambos modelos para un sábado.

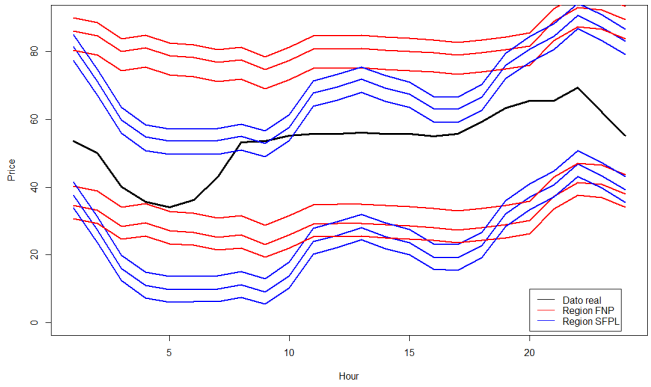

Figura 3.10: Regiones de predicción al 80 %, 90 % y 95 % siguiendo ambos modelos para un domingo.

Capítulo 4

Predicción funcional vs puntual

Se cuenta con los datos obetenidos en Vilar y cols. (2018) donde, para estos mismos datos, se calcularon intervalos de predicción puntuales para cada hora del día (también tanto para demanda como para precio). Resulta interesante comparar estos intervalos con las regiones que se han obtenido utilizando la norma $\|\cdot\|_\infty$. Se usarán los datos obtenidos para $1-\alpha = 0.95$, puesto que es el único nivel de confianza para el que se dispone de los intervalos puntuales. Aun así, hay que tener en cuenta una cosa: en el caso de los intervalos, el nivel de confianza hace referencia a la probabilidad de que un intervalo contenga al verdadero dato, es decir, si en una hora determinada el verdadero valor está fuera del intervalo, el modelo estaría fallando en 1/24 datos de ese día. Por otro lado, al trabajar con regiones de predicción los datos son las curvas diarias completas, por lo que si un dato horario se sale del recinto ya se considera que se ha fallado en la predicción de todo el dato. Por este motivo es de esperar que las regiones de predicción, a pesar de estar construidas con el mismo nivel de confianza, sean más anchas.

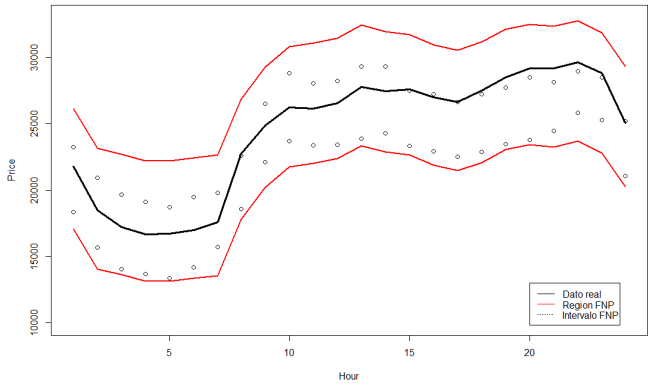

Figura 4.1: Región e intervalos de predicción para la demanda al 95 % siguiendo ambos métodos el modelo FNP.

En las figuras 4.1, 4.2, 4.3 y 4.4 se observa como, en efecto, las regiones de predicción restringidas a una hora generarían intervalos más amplios que los calculados. Sin embargo, sobre todo en el caso de la demanda, sí se observa una mejor cobertura.

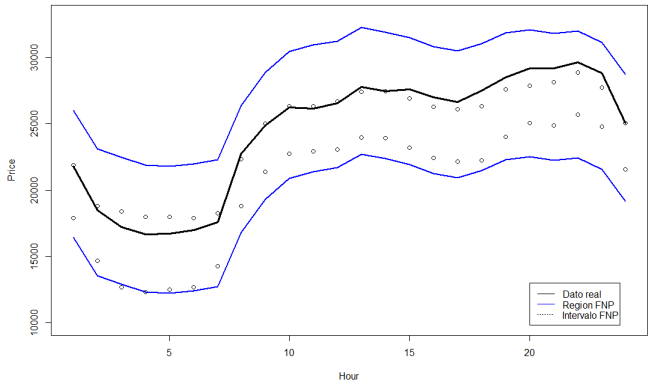

Figura 4.2: Región e intervalos de predicción para la demanda al 95 % siguiendo ambos métodos el modelo SFPL.

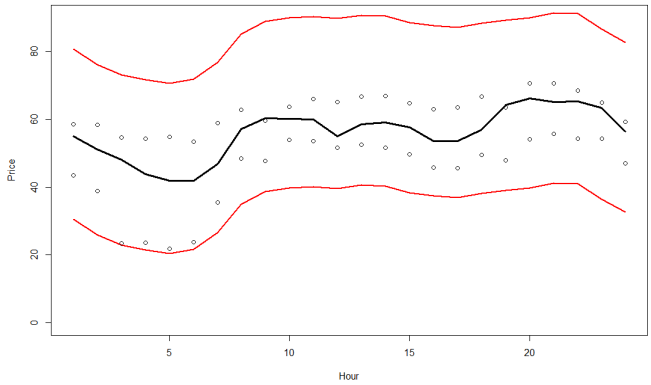

Figura 4.3: Región e intervalos de predicción para el precio al 95 % siguiendo ambos métodos el modelo FNP.

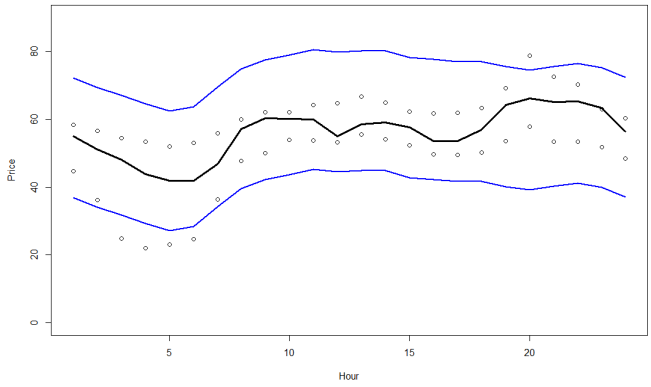

Figura 4.4: Región e intervalos de predicción para el precio al 95 % siguiendo ambos métodos el modelo SFPL.

Referencias

Aneiros, G., Vilar, J., Cao, R., y Muñoz San-Roque, A. (2013). Functional prediction for the residual demand in electricity spot markets. *IEEE Trans Power Syst*, *28*, 4201-4208.

Aneiros, G., Vilar, J., y Raña, P. (2016). Short-term forecast of daily curves of electricity demand and price. *Electr Power Energy Syst*, *80*, 96-102.

Febrero, G.-P., M, y González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, *19*, 331-345.

Ferraty, F., Laksaci, A., Tadj, A., y Vieu, P. (2011). Kernel regression with functional response. *Electr J of Statist*, *5*, 159-171.

Hyde, O., y Hodnett, P. (2015). Modeling the effect of weather in short-term load forecasting. *Math Engrg Indust*, *6*, 155-169.

Raña, P., Aneiros, G., y Vilar, J. (2015). Detection of outliers in functional time series. *Environmetrics*, *26*, 178-191.

Raña, P. (2016). *Pointwise forecast, confidence and prediction intervals in electricity demand and price* (Tesis). Universidade da Coruña.

Taylor, J., y Buizza, R. (2003). Using weather ensamble predictions in electricity demand forecasting. *Int J Forecast*, *19*, 57-70.

Taylor, J., de Menezes, L., y McSharry, P. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *Int J Forecast*, *22(1)*, 1-16.

Vilar, J., Raña, P., y Aneiros, G. (2018). Prediction intervals for electricity demand and price using functional data. *Electr Power Energy Syst*, *96*, 457-492.

Zhu, T., y Politis, D. N. (2017). Kernel estimates of nonparametric functional autoregression models and their bootstrap approximation. *Electrn J Statist*, *11(2)*, 2876-2906.