CLUSTERIZACIÓN Y PREDICCIÓN DE SERIES TEMPORALES. APLICACIÓN EN EL MERCADO ELÉCTRICO EUROPEO

TRABAJO FIN DE MÁSTER

MÁSTER EN INGENIERÍA DE SISTEMAS DE DECISIÓN

ALICIA BURGOS CARRASCÓN

Junio de 2022

Supervisado por:

Ana Elizabeth García Sipols
Clara Simón de Blas
Mayte Santos Martín

Agradecimientos

Todo lo que aquí se expone no habría sido posible sin la dedicación de mis tutoras: Ana, Clara y Mayte, y de Óscar Generoso Gutiérrez, cuya ayuda ha sido fundamental en estas últimas semanas. Gracias por vuestra generosidad, por aceptar mi propuesta y por nutrir mi curiosidad por el ámbito de las series temporales.

También merece una mención especial mi Familia, en mayúsculas, tanto la de sangre como la de corazón, a quienes dedico estas líneas en señal de agradecimiento. Todas las buenas construcciones necesitan de pilares sólidos, y ellos son los míos. Con su buen hacer, su apoyo y sus consejos he conseguido sortear las piedras del camino y alcanzar, con el trabajo y esfuerzo del que son partícipes, los objetivos que me he ido proponiendo. Gracias por acompañarme en esta carrera de fondo que es el gusto por las Matemáticas o, en esencia, la locura de la razón.

Índice general

Abstract

The lack of regularization in the price of electricity has led to the rise of data mining techniques in its own market. Therefore, the extraction of useful information in time series of market prices can be of great interest to electricity producers. In this work, known clustering techniques will be used for the latter, in a sample taken from different countries of the European Union. Not only the behavior of the system will be represented, but also a novel contribution will be made by the prediction of time series.

Resumen

La falta de regularización en el precio de la electricidad ha propiciado el auge de las técnicas de Minería de Datos en dicho mercado. Por ello, la extracción de información útil en series temporales de los precios del mercado puede ser de gran interés para las productoras de energía eléctrica. En este trabajo, se utilizarán técnicas de clustering conocidas para esto último, en una muestra tomada de distintos países de la Unión Europea. No solamente se representará el comportamiento del sistema, sino que también se realizará una contribución novedosa mediante la predicción de series temporales.

Introducción

La electricidad y el mercado eléctrico

La electricidad es una de las fuentes principales de energía utilizadas en la actualidad. A pesar de que su uso abarca desde las actividades domésticas hasta las industriales, es innegable, a día de hoy, la estrecha relación entre una economía sólida y un sistema eléctrico robusto y de precios competitivos. Esta fuerte relación ha propiciado, de forma global, que la energía eléctrica sea un factor estratégico para los diferentes países, que lleva implícitas consecuencias económicas y sociales. Es por ello que se puede decir que el mundo actual está nutrido de la electricidad.

Académicamente, se ha catalogado a la energía eléctrica como una fuente de energía secundaria, ya que su obtención proviene de otro tipo de fuentes, llamadas primarias, como son los combustibles fósiles, o las energías hidráulica y nuclear. Este condicionamiento en la obtención de electricidad hace que el mercado eléctrico cuente con dos características fundamentales, descritas en [14]:

1. La primera, que la energía eléctrica no puede ser almacenada, independientemente de la fuente primaria utilizada. Es cierto que, gracias a los avances realizados, pueden almacenarse pequeñas cantidades de energía eléctrica mediante pilas químicas o mecanismos semejantes, pero aún no se conoce ninguna manera eficiente de almacenamiento de la generación de una central estándar.
2. En la mayoría de los casos, la energía eléctrica posee un matiz productivo irremplazable a corto plazo. El suministro de energía eléctrica es esencial para el funcionamiento de nuestra sociedad, resultando infactible la adaptación de los aparatos eléctricos a otras fuentes de energía, por lo menos a corto plazo.

Tal y como se indica en [9], estas características hacen que el mercado eléctrico cuente con matices relevantes. Su precio es un factor decisivo de la competitividad de buena parte de las economías. El desarrollo tecnológico de la industria eléctrica y su estructura de aprovisionamiento de materias primas determinan la evolución de otros sectores de la industria.

Por otra parte, la distribución de electricidad constituye un monopolio natural: se trata de una actividad intensiva en capital, que requiere conexiones directas con los consumidores, cuya demanda es no almacenable y varía en períodos relativamente cortos de tiempo. Además, la imposibilidad de almacenar electricidad requiere que la oferta sea igual a la demanda en cada instante de tiempo, lo que supone necesariamente una coordinación de la producción de energía eléctrica, así como la coordinación entre las decisiones de inversión en generación

y en transporte de energía eléctrica. Todas estas características técnicas y económicas hacen del sector eléctrico un sector necesariamente regulado.

En el caso de España, la Red Eléctrica se encarga de garantizar las condiciones técnicas para que la electricidad circule de forma ininterrumpida desde los núcleos de generación hasta los núcleos de consumo. También regula la oferta y la demanda de energía mediante el Operador del Mercado Eléctrico, cuya titularidad es privada, al igual que en los países miembros de la Unión Europea (UE). En este último caso, los mercados eléctricos de la UE caminan hacia la liberalización de los mismos para alcanzar una mayor competitividad. Así, se pretende realizar una transición de los monopolios estatales a un supuesto mercado eléctrico europeo.

En este contexto, se pretende analizar el precio de la energía de los últimos años en distintos países miembros. Si bien es cierto que la variabilidad de la demanda aporta cierta incertidumbre al sistema eléctrico, y a su vez, al precio de su suministro, se trata de conseguir agrupar distintas entradas por comportamientos similares y obtener una serie de conclusiones.

Técnicas multivariantes

El análisis cluster es un conjunto de técnicas multivariantes que tienen como objetivo agrupar a un conjunto de casos o individuos en conglomerados o clúster.

La clasificación forma parte del hilo conductor del desarrollo del ser humano: a la hora de encontrar un nuevo objeto, siempre se ha tratado de analizar sus características propias para poder establecer comparaciones con objetos conocidos, a través de reglas de similitud. Dentro de la clasificación no supervisada, se encuentra la técnica del análisis de conglomerados o clustering. Con ella, se consigue separar un grupo de datos sin etiquetar en subgrupos, más o menos homogéneos, de acuerdo con una medida de similitud, cumpliéndose que la similitud entre los objetos pertenecientes a un subgrupo es mayor a la similitud entre los objetos de otros subgrupos.

La predicción o la modelización de comportamientos son dos de los cometidos más frecuentes en el campo del análisis de datos, una herramienta fundamental para comprender el comportamiento de una gran cantidad de procesos actuales. A la hora de trabajar con un volúmen de datos extenso, una de las técnicas frecuentes es la de realizar una estimación de los datos futuros una vez son conocidas las agrupaciones de los mismos, es decir, partiendo de la base obtenida de la técnica de clusterización. De esta forma, se puede estimar el comportamiento futuro de cada grupo de observaciones, en lugar de hacerse para cada observación.

Objetivos

Este Trabajo de Fin de Máster tiene por objetivo realizar un análisis de la serie temporal del precio del megavatio en distintos países de la Unión Europea, haciendo uso de diferentes técnicas de Minería de Datos. En particular, se aplicarán técnicas de clustering para tratar de verificar la existencia de patrones similares dentro de un mercado que, a priori, parece presentar un comportamiento puramente aleatorio. Tras la clasificación, se intentará predecir su comportamiento a corto plazo, lo cual podría ser de gran interés dada la existente y evolucionada desregularización de los mercados de energía eléctrica europeos.

Para ello, se estructura el Trabajo de la siguiente manera:

En el capítulo 1, se describen cada uno de los requerimientos para la tarea de clustering de series temporales, y se aplican a las series del precio de la energía eléctrica en un grupo de países miembros de la Unión Europea. Principalmente, se describen una serie de pasos fundamentales, como son: la metodología usada para la obtención de los datos y de las series temporales asociada a ellos, la definición de las medidas de distancia, la formación de grupos o la extracción de representantes de cada grupo.

En el capítulo 2, se estudian un conjunto de técnicas de predicción de series temporales. Por un lado, se encuentran modelos más clásicos, entre los que se destacan SARIMA (Seasonal Autoregressive Integrated Moving Average) o TBATS (Trigonometric regressors for Box-Cox transformations with Arma errors and Trend and Seaonal components) y, por otro lado, se toman modelos más novedosos, basados en Aprendizaje Automático, como ARNN (Autoregressive Neural Networks), KNN (K-Nearest Neighbors) o SVM (Support Vector Machines).

En el capítulo 3, se interpretarán los principales resultados obtenidos en los dos capítulos previos, sintetizando así la información útil extraída a partir de los datos originales. También se comparará el comportamiento para cada una de las metodologías empleadas, en vista de sus medias de error cuadrático y absoluto.

Finalmente, en el capítulo 4, se expondrán una serie de conclusiones obtenidas en el trascurso de este Trabajo.

Capítulo 1

Análisis de técnicas de clustering aplicadas a las series temporales

> Time is not a line, but a series of now-points
>
> *Taisen Deshimaru*

La tarea de agrupar una serie de objetos en clases similares se denomina clustering. Así, un cluster estará formado por un conjunto de datos de comportamientos parecidos entre sí y distintos a los datos pertenecientes a otros clusters, pudiendo ser tratado, colectivamente, como un único grupo en un gran número de prácticas.

Las técnicas de clustering son métodos de clasificación no supervisada de patrones en conjuntos llamados clusters. A pesar de haber sido abordadas desde un gran número de disciplinas, radican en un problema complejo de análisis experimental de datos. En este primer capítulo, se introduce una visión global de las técnicas para el análisis de conglomerados tradicional o hard clustering. Con dicho esquema, se expone la elección particular de las técnicas usadas en esta Memoria: desde el cálculo de distancias hasta la elección de representantes de cada cluster.

Requerimientos de la tarea de clustering

Como se ha comentado, el análisis de clusterización se engloba dentro del Aprendizaje Automático no supervisado, pues estudia la estructura intrínseca de los datos, que no dependen de las clases previamente definidas, para abordar problemas como la agrupación o la reducción de la dimensionalidad. Para su correcta aplicación, es preciso que las tareas de clustering sigan una serie de pasos, que presentamos a continuación:

1. Metodología. Consiste en elegir el número, tipo y escala de las características de los datos a los que se va a aplicar el clustering. Su obtención se puede conseguir a través de dos técnicas: la elección de características, mediante la cual se identifica el subconjunto más efectivo de los datos para poder agruparlos, y la extracción de características, mediante la cual se modifican características existentes de los datos para obtener otras nuevas.

2. Definición de la medida de proximidad. Se basa en escoger una medida de proximidad entre los patrones que sea adecuada al dominio de los datos. Como medida de proximidad suele tomarse una función de distancia definida para pares de patrones.
3. Formación de clusters. La agrupación en clusters puede ser realizada de varias formas. El clustering de salida puede ser duro (*hard*) o difuso (*fuzzy*), en función de si se realiza una partición de los datos en grupos o si cada patrón tiene una probabilidad de pertenecer a cada uno de los clusters resultado. Por otro lado, el clustering también puede dividirse en jerárquico (si son una serie jerarquizada de particiones bajo un criterio de formación de clusters según su similitud) o particional (si identifican la partición que optimiza un criterio de agrupación, normalmente de forma local).
4. Extracción de representantes de cada cluster. Se trata de abstraer el conjunto de los datos de una forma sencilla y compacta. A menudo, se escoge un prototipo o representante de cada cluster.
5. Evaluación de los resultados. Realiza el análisis de la validez de la salida obtenida. La validación, que se obtiene de forma más objetiva mediante métodos estadísticos, puede ser de tres tipos: externa (compara la estructura obtenida con otra realizada a priori), interna (valida si la estructura resultante es adecuada para los datos) o relativa (compara dos estructuras, midiendo la calidad relativa de ambas).

En las secciones que se presentan a continuación, se describen de forma más minuciosa cada uno de los puntos anteriores.

1.1. Metodología

En cualquier proyecto que involucre la toma de datos es importante realizar una revisión de las bases de datos para verificar que es cumplen unas especificaciones prestablecidas de cara a su posterior estudio o análisis. En esta sección, se va a detallar la metodología que se ha seguido para la obtención de los datos que se analizan en este Trabajo: desde la recopilación de los datos hasta la depuración de los mismos. El resultado de este proceso de eliminación de información redundante son las series temporales a analizar en las secciones próximas.

1.1.1. Obtención de los datos

Si hay alguna ventaja por el hecho de vivir en la era del Data Science, ésta es la facilidad con la que se puede encontrar gran cantidad de la información hoy en día. De hecho, muchos gobiernos o instituciones cuentan con estándares y normativas vigentes en relación a la accesibilidad de sus sitios web, aunque pueden no estar adaptados en su totalidad. Este es el caso de Eurostat, la oficina de Estadística de la Unión Europea, encargada de publicar estadísticas e indicadores a escala europea que permiten hacer comparaciones entre países y regiones.

Para la obtención de los datos, se recurre a la base de datos Electricity Data dentro del explorador de datos de su propia página web. En ella se encuentran, entre otros, los registros de los precios de la electricidad y el gas natural en forma de gráficos interactivos y tablas, clasificando el precio del MWh por meses y por países. Se seleccionan el total de países disponibles, que son, en concreto, aquellos 27 que pertenen a la UE en 2022.

En cuanto al número de observaciones, se toman los registros mensuales de los últimos 7 años, es decir, desde el primer mes de 2015 hasta el último mes de 2021, aunque existen constantes actualizaciones de precios disponibles en la web de Ember [6]. Sin embargo, observamos que tenemos datos faltantes: los datos comienzan en octubre de 2016 para Bulgaria, octubre de 2017 para Croacia y enero de 2017 para Serbia.

El estudio comienza en enero de 2015, descartando la presencia de estos tres últimos países, puesto que se necesitan las series de datos completas.

1.1.2. Obtención de las series temporales

Después de haber obtenido los datos, lo próximo a realizar es su tratamiento o depuración. En el caso que nos ocupa, utilizaremos tanto Excel como el paquete estadístico R para la obtención de las series temporales asociadas a los datos. Se dispone de 24 series temporales asociadas a cada uno de los países en el estudio, en el periodo de tiempo enero 2017 - diciembre 2021.

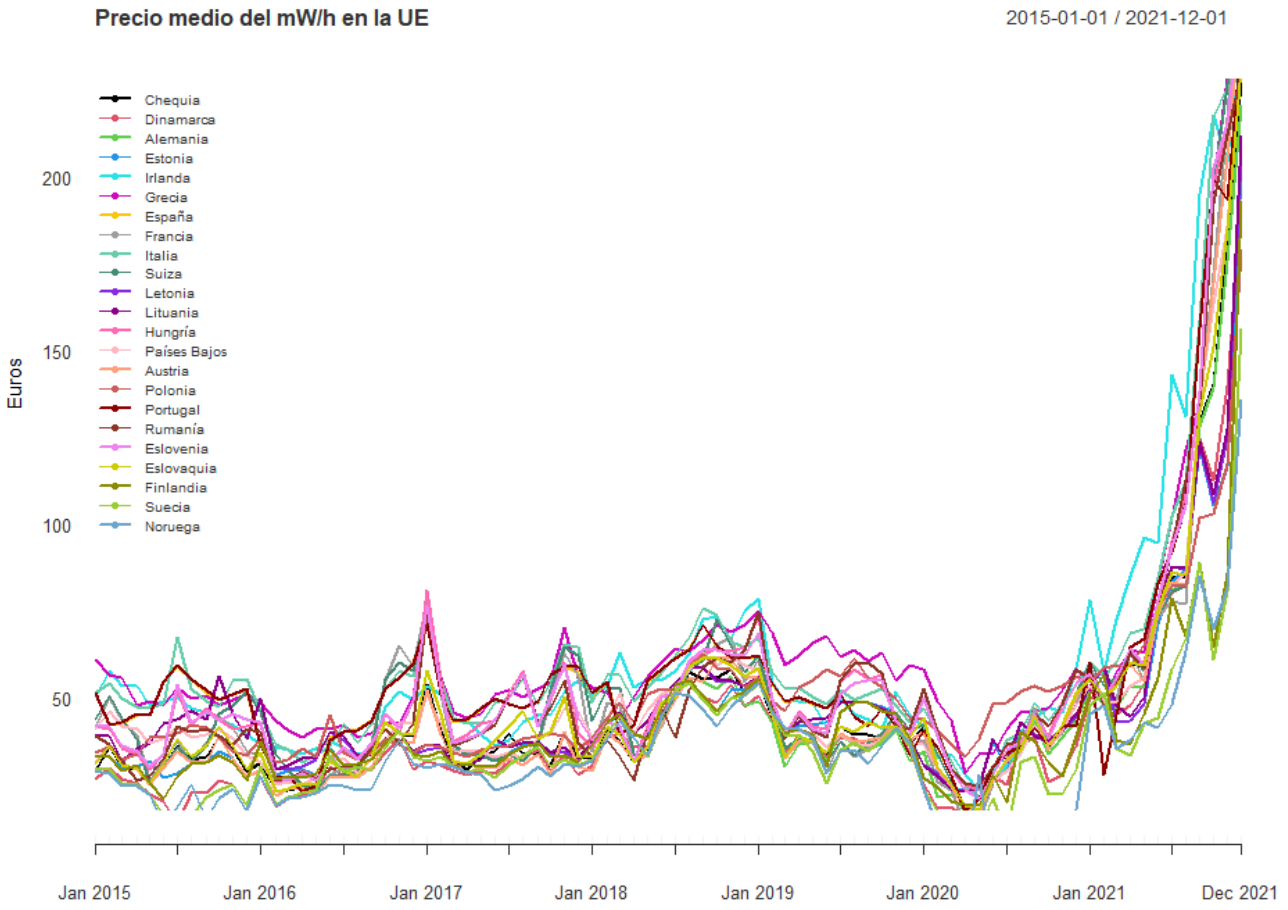


Figura 1.1: Representación de las 24 series temporales.

1.2. Definición de la medida de proximidad

Una de las claves en el Análisis de Cluster es la determinación de una medida de similaridad entre dos objetos, existiendo gran cantidad métodos para el caso particular en el que se dispone de datos medidos en el tiempo. Lo más recurrido es el cálculo del concepto contrario, es decir, la diferencia o disimititud entre dos elementos, utilizando la medida de la distancia en un espacio de características, entendiendo que, cuanto mayor sea la distancia que los separa, menor similitud tendrán entre sí. En esta sección, se recuerdan los fundamentos matemáticos de algunas de las distancias más utilizadas y se aplican al conjunto de los datos.

Además, a modo de recordatorio y si no se especifica de otra manera, $\mathbf{X}_T = (X_1, \dots X_T)$ y $\mathbf{Y}_T = (Y_1, \dots Y_T)$ van a denotar los valores de dos procesos reales $X = \{X_t, t \in \mathbb{Z}\}$ y $Y = \{Y_t, t \in \mathbb{Z}\}$ en el tiempo. Se asume también la misma longitud T para ambas.

1.2.1. Nociones básicas de distancia y similaridad

Tras considerar que el objetivo del clustering es el de conseguir agrupaciones naturales del conjunto de los elementos de una muestra, surge la necesidad de definir en base a qué criterio se considera que dos grupos son más o menos similares. Esta cuestión implica otras dos, que son:

- Cómo medir la similitud entre dos elementos de la muestra.
- Cómo evaluar cuándo dos clusters pueden ser agrupados o no.

A partir de ahora, se estudiarán con más detalle las posibles funciones que pueden tomarse para la medida de similitud entre los grupos que se van construyendo, diferenciando en primer lugar entre distancias (métricas) y similaridades.

Definición 1.1. *Sea U un conjunto finito o infinito de elementos. Una distancia es una función $d : U \times U \longrightarrow \mathbb{R}$ tal que $\forall x, y \in U$ se cumple:*

1. $d(x, y) \geq 0$
2. $d(x, x) = 0$
3. $d(x, y) = d(y, x)$

Si además, se cumple que

1. $d(x, y) = 0 \iff x = y$
2. $d(x, z) \leq d(x, y) + d(y, z)$, $\forall z \in U$,

hablaremos de distancia métrica.

De manera análoga, se tiene la siguiente definición de similaridad.

Definición 1.2. *Sea U un conjunto finito o infinito de elementos. Una similaridad es una función $s : U \times U \longrightarrow \mathbb{R}$ tal que $\forall x, y \in U$ se cumple:*

1. $s(x, y) \leq s_0$
2. $s(x, x) = s_0$
3. $s(x, y) = s(y, x)$

Si además, se cumple que

1. $s(x,y) = s_0 \implies x = y$
2. $|s(x,y) + s(y,z)| \cdot s(x,z) \geq s(x,y) \cdot s(y,z)$, $\forall z \in U$,

hablaremos de similaridad métrica.

Nótese que el segundo punto en la definición de similaridad métrica corresponde al hecho de que la máxima similaridad solo la poseen dos elementos idénticos.

Una vez hecha esta distinción, estamos en condiciones de centrarnos en su aplicación. En la práctica, el paquete *TSclust* de R implementa un conjunto de distancias, tal y como se puede consultar en [15]. Las distancias estudiadas -la distancia de Minkowski, la distancia DTW, la distancia CORT y la distancia COR- vienen implementadas en este paquete y son independientes de los modelos ajustados.

1.2.2. Distancia de Minkowski

La distancia de Minkowski de orden q, también llamada distancia de la norma L_q, con q un entero positivo, es una métrica en un espacio vectorial normalizado. Se define como

$$d_{L_q}(\mathbf{X}_T, \mathbf{Y}_T) = \left(\sum_{t=1}^{T} (X_t - Y_t)^q \right)^{1/q}.$$

La distancia de Minkowski es una generalización de la distancia de Manhattan (con $q = 1$) o de la distancia Euclídea (con $q = 2$), siendo esta última la más utilizada:

$$d_2(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{\sum_{t=1}^{T} (X_t - Y_t)^2}.$$

1.2.3. Distancia DTW

La distancia DTW (por sus siglas en inglés, *dynamic time warping distance*) o distancia de deformación dinámica de tiempo se suele utilizar para encontrar patrones en series temporales. Sea M el conjunto de todas las posibles secuencias de m pares preservando el orden de las observaciones en la forma

$$r = ((X_{a_1}, Y_{b_1}), \ldots, (X_{a_m}, Y_{b_m})),$$

con $a_i, b_j \in \{1, \ldots, T\}$ tal que $a_1 = b_1 = 1, a_m = b_m = T$ y $a_{i+1} = a_i$ o a_{i+1} y $b_{i+1} = b_i$ o b_{i+1} para $i \in \{1, \ldots, m-1\}$.

Entonces, la DTW trata de encontrar un mapeo r entre series de manera que se minimice una medida de distancia específica entre los pares de observaciones (X_{a_i}, Y_{b_i}). Su definición es:

$$d_{DTW}(\mathbf{X}_T, \mathbf{Y}_T) = \min_{r \in M} \left(\sum_{i=1}^{m} |X_{a_i} - Y_{b_i}| \right).$$

1.2.4. Distancia de correlación temporal

Recientemente se ha introducido esta medida de disimilitud, que engloba tanto la medida convencional de la proximidad en las observaciones, como la estimación de proximidad en el comportamiento. Esto se evalúa por medio del coeficiente de correlación temporal de primer orden, definido para el intervalo $[-1, 1]$ como

$$CORT(\mathbf{X}_T, \mathbf{Y}_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}},$$

de forma que la medida de disimilitud entre ambas series temporales viene dada por:

$$d_{CORT}(\mathbf{X}_T, \mathbf{Y}_T) = \phi[CORT(\mathbf{X}_T, \mathbf{Y}_T)]d(\mathbf{X}_T, \mathbf{Y}_T).$$

Notemos que $CORT(\mathbf{X}_T, \mathbf{Y}_T) = 1$ si las series tienen un comportamiento dinámico similar en la misma dirección (es decir, su tendencia es la misma en los mismos espacios temporales), $CORT(\mathbf{X}_T, \mathbf{Y}_T) = -1$ si las series tienen un comportamiento dinámico similar en direcciones opuestas y $CORT(\mathbf{X}_T, \mathbf{Y}_T) = 0$ si ambas series están incorreladas en el tiempo.

En este caso, $\phi(\cdot)$ es una función que modula de forma automática una distancia convencional $d(\mathbf{X}_T, \mathbf{Y}_T)$, pudiento tomarse, entre otras, cualquiera de las distancias vistas en las secciones 1.2.2 o en 1.2.3. Esta función, propuesta en [15], se define como

$$\phi_k(u) = \frac{2}{1 + \exp(ku)},\ k \geq 0.$$

1.2.5. Distancias basadas en correlación

El siguiente criterio de disimilitud considera el coeficiente de correlación de Pearson entre dos series temporales,

$$COR(\mathbf{X}_T, \mathbf{Y}_T) = \frac{\sum_{t=1}^{T}(X_t - \overline{\mathbf{X}_T})(Y_t - \overline{\mathbf{Y}_T})}{\sqrt{\sum_{t=1}^{T}(X_t - \overline{\mathbf{X}_T})^2}\sqrt{\sum_{t=1}^{T}(Y_t - \overline{\mathbf{Y}_T})^2}},$$

donde $\overline{\mathbf{X}_T}$ y $\overline{\mathbf{Y}_T}$ son los valores medios de las series $\mathbf{X}_T$ e $\mathbf{X}_T$ respectivamente.

En base a esto, es posible construir algoritmos de clasificación usando las dos distancias basadas en correlación que se presentan a continuación:

$$d_{COR.1}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{2(1 - COR(\mathbf{X}_T, \mathbf{Y}_T))},$$

$$d_{COR.2}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{\left(\frac{1 - COR(\mathbf{X}_T, \mathbf{Y}_T)}{1 + COR(\mathbf{X}_T, \mathbf{Y}_T)}\right)^{\beta}},\ \beta \geq 0.$$

Notemos que, por definición de $COR(\mathbf{X}_T, \mathbf{Y}_T)$, $d_{COR.1}(\mathbf{X}_T, \mathbf{Y}_T)$ tiene por intervalo de definición el $[0, 2]$. Si entre las series temporales no existe correlación lineal, el valor de dicha distancia será igual a $\sqrt{2}$. Intuitivamente, podemos pensar en este valor central como un indicador de cuán correlacionadas están las series: cuanto más se acerque $d_{COR.1}$ a los extremos del intervalo, mayor correlación lineal, y cuanto más se acerque $d_{COR.1}$ a $\sqrt{2}$, menor

correlación habrá entre las series.

Además, en $d_{COR.2}$ se introduce el parámetro β para regular el decremiciento rápido de la distancia. En cuanto a sus propiedades, se destaca que $d_{COR.2}(\mathbf{X}_T, \mathbf{Y}_T)$ diverge si $COR(\mathbf{X}_T, \mathbf{Y}_T) = -1$.

1.3. Formación de clusters

El análisis de cluster concierne a la identificación de grupos y su objetivo es el de realizar una partición de una serie de observaciones en un número distinto de grupos o clusters de tal forma que todas las observaciones dentro de un grupo sean similares, y observaciones de distintos grupos no lo sean.

Existen gran cantidad de algoritmos de formación de clusters, cuya técnica depende tanto del tipo de datos de los que se dispone, como del objetivo de su aplicación. Por lo general, los métodos de cluster se pueden clasificar en los siguientes:

1. Métodos particionales. Son aquellos que, dada una base de datos de n objetos, construyen k grupos con los datos, con $k \leq n$, en los que cada partición representa un cluster. En otras palabras, clasifica a los datos en k grupos, cumpliendo que:
   - Cada grupo contiene, al menos, un elemento.
   - Cada elemento pertenece únicamente a un grupo.

   Una vez se elige el número de particiones k, estos métodos relizan, tras una partición inicial, una técnica iterativa de recolocación con el fin de mejorar la partición mediante el movimiento sucesivo de los objetos. Para ello, se suele emplear una de las heurísticas más conocidas, como son:
   - Algoritmo de K-means, que forma agrupaciones de los datos de manera que la varianza de los distintos clusters sea similar, minimizando un concepto conocido como inercia, que es la suma de las distancias al cuadrado de cada objeto del cluster a un centroide (punto medio de todos los objetos del cluster).
   - Algoritmo de K-medians, en el que se calculan las medianas en vez de los centroides.

   Es preciso mencionar que, dentro de los métodos particionales, existe una clasificación de algoritmos en vista al tipo de partición que se realice y que merecerían un análisis más profundo. Estos pueden ser:
   - Algoritmos de hard clustering, que asignan cada elemento a un grupo o cluster definido.
   - Algoritmos de fuzzy clustering, que asignan, para cada elemento, un grado de pertenencia a cada uno de los clusters.

2. Métodos jerárquicos. Son aquellos en los que o bien se agrupan clusters para la formación de uno nuevo, o bien se separa alguno ya existente para dar lugar a otros dos. De esta manera, si se repite el proceso de forma sucesiva, se minimiza la distancia o se maximiza alguna medida de similitud. Pueden ser de dos tipos:

- Aglomerativos o ascendentes. Si comienzan el análisis con tantos grupos como elementos haya. A partir de estas unidades individuales se forman grupos, de manera ascendente, hasta que al final del proceso todos los elementos están englobados en un mismo conglomerado.
- Disociativos o descendientes. Constituyen el proceso inverso al anterior. Se comienza con un conglomerado que contiene a todos los elementos y, partiendo de ese grupo inicial, se van formado grupos cada vez más pequeños a través de sucesivas divisiones. Al final, se tienen tantas agrupaciones como elementos.

3. Métodos basados en densidad. Son técnicas de clustering que se suelen utilizar para eliminar ruido y formar clusters considerando la agrupación espacial de los datos. Su nombre se debe a que aumentan el tamaño del cluster hasta que la densidad (número de datos) en su vecindad supere una cantidad preestablecida, siendo el método DB-SCAN el más conocido.

Para la creación de los grupos de las series temporales, se decide utilizar el **clustering jerárquico aglomerativo**, basado en la técnica de escalado multidimensional de representación de las distancias en el plano.

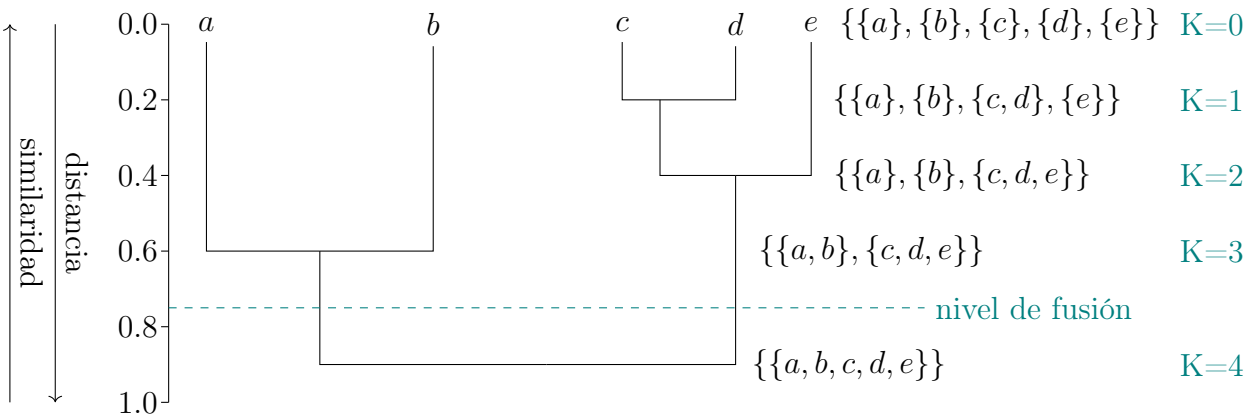
1.3.1. Métodos Jerárquicos de Análisis de Cluster aglomerativo

Para asentar conceptos, nos centramos en los métodos aglomerativos. Sea n el número de observaciones de la muestra, donde el nivel $K = 0$ posee n grupos. En el próximo nivel, se agruparían el par de observaciones que de mayor similitud (análogamente, de menor distancia), resultando en $n - 1$ grupos. Seguidamente, y continuando con dicha estrategia, se agrupan en el siguiente nivel aquel par de elementos os (o clusters recién formados) de mayor similitud, de forma que en el nivel L haya $n - L$ grupos. Si se continúa de dicha forma, en el nivel $L = n - 1$ solo hay un grupo, formado por todos los elementos de la muestra. A dicho cluster se le denomina cluster *trivial*.

Esta forma de agrupación tiene la peculiaridad de que una vez se agrupan dos clusters en un determinado nivel, quedan jerárquicamente agrupados para los siguiente niveles.

Además, los métodos jerárquicos permiten la construcción de un árbol de clasificación, un arbol binario típicamente llamado *dendograma*, que indica el proceso de unión gráficamente. En él se muestran los niveles en los que grupos que se van agrupando, así como el valor de la medida de asociación entre los grupos cuando se unen, que se conoce como *nivel de fusión*.

Las operaciones de los métodos jerárquicos aglomerativos son sencillas, pues se parte de tantos grupos como elementos haya, y se selecciona una medida de similitud de forma que se agrupen los grupos con mayor similitud. Se reitera el proceso hasta que, o bien se forme un solo grupo, o bien se obtenga un número de grupos prefijado. Además, se contrasta estadísticamente que no haya razones para continuar agrupando clusters.



A continuación, se estudian algunas de las estrategias empleadas al agrupar clusters en distintos niveles de un proceso jerárquico en R. Los agrupamientos estudiados son el agrupamiento simple, el completo, el de promedio ponderado, el de centroides y el de Ward. Sin embargo, existen muchos otros, los cuales se recomienda consultar en [21].

Agrupamiento simple

El agrupamiento simple o *simple linkage* considera que la similitud o distancia entre dos grupos vienen dada por la mínima distancia (o máxima similitud) entre sus componentes, respectivamente.

De esta forma, tras la etapa K-ésima se han formado $n-K$ grupos y la distancia entre los clusters C_i, de n_i elementos, y C_j, de n_j elementos viene dada por la expresión:

$$d(C_i, C_j) = \min_{\substack{x_l \in C_i \\ x_m \in C_j}} \{d(x_l, x_m)\}, \quad l = 1, \ldots, n_i, \quad m = 1, \ldots, n_j.$$

Para el nivel $K+1$-ésimo, se unirán los clusters C_i y C_j si

$$d(C_i, C_j) = \min_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \{d(C_{i1}, C_{j1})\}$$

$$= \min_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \left\{ \min_{\substack{x_l \in C_{i1} \\ x_m \in C_{j1}}} \{d(x_l, x_m)\} \right\}, \quad l = 1, \ldots, n_{i1}, \quad m = 1, \ldots, n_{j1}.$$

Observación 1.1. *Análogamente, si se quisiera definir el agrupamiento en términos de similitud, bastaría con hallar el máximo de la similitud entre sus componentes en cada uno de los casos.*

Agrupamiento completo

En esta estrategia, también conocida como *complete linkage*, la distancia o similitud entre dos clusters viene dada por la máxima distancia (o mínima similitud) entre sus componentes, respectivamente.

Siguiendo con el esquema del método previo, tras la etapa K-ésima se han formado $n-K$ grupos y la distancia entre los clusters C_i, de n_i elementos, y C_j, de n_j elementos cumple que:

$$d(C_i, C_j) = \max_{\substack{x_l \in C_i \\ x_m \in C_j}} \{d(x_l, x_m)\}, \quad l = 1, \ldots, n_i, \quad m = 1, \ldots, n_j.$$

La estrategia seguida en el nivel $K+1$ será la unión entre os clusters C_i y C_j si

$$d(C_i, C_j) = \min_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \{d(C_{i1}, C_{j1})\}$$

$$= \min_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \left\{ \max_{\substack{x_l \in C_{i1} \\ x_m \in C_{j1}}} \{d(x_l, x_m)\} \right\}, \quad l = 1, \ldots, n_{i1}, \quad m = 1, \ldots, n_{j1}.$$

Agrupamiento de promedio ponderado

El método de agrupamiento ponderado o *average linkage* toma como distancia entre clusters la media aritmética entre la distancia -o equivalentemente, similitud- de las componentes de dichos clusters. Por lo tanto, la distancia entre un cluster C_i, de n_i elementos (que a su vez está formado por dos clusters C_{i1} y C_{j1}, de órdenes n_{i1} y n_{j1}) y otro cluster C_j, de n_j elementos viene determinada por:

$$d(C_i, C_j) = \frac{d(C_{i1}, C_j) + d(C_{i2}, C_j)}{2}.$$

Como este método no considera el tamaño de los clusters presentes en el cálculo, ambas distancias, $d(C_{i1}, C_j)$ y $d(C_{i2}, C_j)$, son igual de relevantes. No obstante, existe una versión no ponderada del agrupamiento, como puede consultarse en [23].

Agrupamiento de centroides

En los métodos de agrupamiento basados en el centroide o *centroid linkage*, la distancia entre grupos viene dada por la distancia entre sus centroides, es decir, los vectores de medias de las variables medidas sobre sus elementos. Además, también poseen variantes ponderadas, que consideran el tamaño de los clusters al efectuar los cálculos, y no ponderadas, que los desprecian.

1. Método del centroide ponderado. Supongamos que se quiere medir la distancia entre los clusters C_i (formado por los clusters C_{i1} y C_{i2}, de n_{i1} y n_{i2} elementos, respectivamente) y C_j. Sean m^j, m^{i1} y m^{i2} los centroides de los clusters anteriores.
   Como los centroides son vectoes n-dimensionales, se puede expresar el centroide del cluster C_i de forma vectorial como:

   $$m^i = \frac{n_{i1} m^{i1} + n_{i2} m^{i2}}{n_{i1} + n_{i2}}.$$

   Sus componentes, entonces, se expresan como:

   $$m^i_l = \frac{n_{i1} m^{i1}_l + n_{i2} m^{i2}_l}{n_{i1} + n_{i2}}, \quad l = 1, \ldots, n.$$

Se puede comprobar, mediante una serie de operaciones básicas, que el cuadrado de la distancia euclídea entre los clusters C_i y C_j atiende a la expresión:

$$d_2^2(C_i, C_j) = \sum_{l=1}^{n} \left(m_l^j - m_l^i\right)^2$$

$$= \frac{n_{i1}}{n_{i1}+n_{i2}} d_2^2(C_{i1}, C_j) + \frac{n_{i2}}{n_{i1}+n_{i2}} d_2^2(C_{i2}, C_j) - \frac{n_{i1}n_{i2}}{(n_{i1}+n_{i2)^2}} d_2^2(C_{i1}, C_{i2}).$$

Observación 1.2. *Aunque se ha estudiado el caso particular de la distancia euclídea, esta relación se cumple para cualquier distancia cuya norma provenga de un producto escalar. Esta hipótesis puede relajarse todavía más, hasta considerear distancias cuyas normas cumplan la ley del paralelogramo*

$$||x+y||^2 + ||x-y||^2 = 2[||x||^2 + ||y||^2],$$

pues, en dichas circunstancias, puede definirse un producto escalar como:

$$< x, y >= \frac{1}{4}[||x+y||^2 - ||x-y||^2].$$

2. Método del centroide no ponderado. Para evitar que el centroide m^i esté excesivamente influenciado por la componente de mayor tamaño del cluster C_i (sobre todo, en el caso de que los tamaños n_{i1} y n_{i2} sean muy dispares), se sigue la estrategia de la distancia mediana. En ella, se considera que $n_{i1} = n_{i2}$, de forma que C_i se sitúa entre los clusters C_{i1} y C_{i2}. Así, el centroide del cluster (C_i, C_j) cae en el baricentro del triángulo formado por C_{i1}, C_{i2} y C_j.
   A excepción de esta distinción, la estrategia de la distancia mediana es similar a la anterior y sus características son parecidas. Así pues, la distancia entre los clusters C_i y C_j viene dada por:

$$d(C_i, C_j) = \frac{1}{2}[d(C_{i1}, C_j) + d(C_{i2}, C_j)] - \frac{1}{4} d(C_{i1}, C_{i2}).$$

Agrupamiento de Ward

El método de Ward es un procedimiento jerárquico en el que, para cada etapa, se unen los dos clusters para los que se tenga el menor incremento del valor total de la suma de los cuadrados de las diferencias -dentro de cada cluster- de cada elemento al centroide del cluster.

Se recomienda consultar [21], donde se realiza un desarrollo más extenso de este método. La idea general radica en que el menor incremento de los errores cuadráticos es proporcional al cuadrado de la distancia euclídea entre los centroides de los clusters que se unen.

Así, siguiendo con la notación empleada anteriormente, si C_t es el cluster resultante de la uniñon de otros dos, llamémosles C_p y C_q, y sea C_r un cluster distinto a los anteriores. Entonces, el incremento que se produciría con la unión de C_r y C_t es:

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t} \sum_{j=1}^{n} (m_j^r - m_j^t)^2.$$

1.3.2. Escalado multidimensional

Para descubrir las agrupaciones dentro de los datos, hemos estudiado medidas de cercanía, que indican el grado de asociación o similitud entre dos grupos. Ahora bien, el análisis de conglomerados es una herramienta de clasificación de objetos en grupos que no depende de la representación geométrica de los objetos en un espacio de dimensión baja.

Con el fin de explorar la dimensionalidad del espacio, se utiliza el escalado multidimensional, una técnica de reducción de los datos que toma una matriz de proximidad o distancias y trata de encontrar un conjunto de construcciones de menor dimensión, basada en las disimilitudes de los objetos estudiados.

Así, los pasos a seguir en un algortimo de escalado multidimensional o MDS clásico son:

1. Partiendo de la matriz de distancias, calcular $A = -\frac{d_{ij}^2}{2}$.
2. A partir de A, calcular $B = a_{ij} - a_{i.} - a_{.j} + a_{..}$, donde $a_{i.}$ es la media de las entradas a_{ij} recorriendo j.
3. Encontrar los p mayores valores propios $\lambda_1 > \lambda_2 > \ldots \lambda_p$ de B, con sus correspondientes vectores propios normalizados $L = (L_{(1)}, \ldots, L_{(p)})$.
4. Las coordenadas de los objetos a representar son las filas de L.

Aplicación al conjunto de datos

A continuación, se ilustra cómo han sido aplicadas estas técnicas de representación a las distancias ya calculadas en la sección 1.2 para el conjunto de datos del que se dispone.

Se puede notar que, para la distancia euclídea (1.2.2) y la distancia COR (1.2.5), se crea un grupo claro de países, mientras que hay tres de ellos -Noruega, Suecia y Finlandia- que se encuentran alejados del resto. Dentro del grupo grande de países, se puede notar que la representación de algunos de ellos es coincidente, como ocurre para: España y Portugal, Alemania, Eslovaquia y Austria, y Letonia y Lituania, de forma respectiva. Esto significa que dichos países están próximos en distancia, pero se debe analizar su comportamiento, ya que puede diferir.

Para la distancia DTW (1.2.3) se identifican grupos con facilidad, de hecho, se puede encontrar un conglomerado en la esquina inferior izquierda, en la que se ve una mayor concentración de países que en el resto de regiones del plano. Parece que por un lado, España y Portugal- o Letonia y Lituania- y por otro lado, Austria, Alemania y Eslovaquia tienen un comportamiento similar, al igual que ocurría en la gráfica anterior

En el caso de la distancia CORT (1.2.4), se aprecian dos regiones de agrupación aproximadas, pero no demasiado delimitadas. Además del solapamiento de España y Portugal y Letonia y Lituania en la representación, se observa una nueva similitud para el caso de la República Checa, Eslovaquia y Austria. Es preciso recordar que este método tiene en cuenta la similitud de comportamientos, además de la proximidad en distancia de las series temporales, lo cual puede resultar interesante.

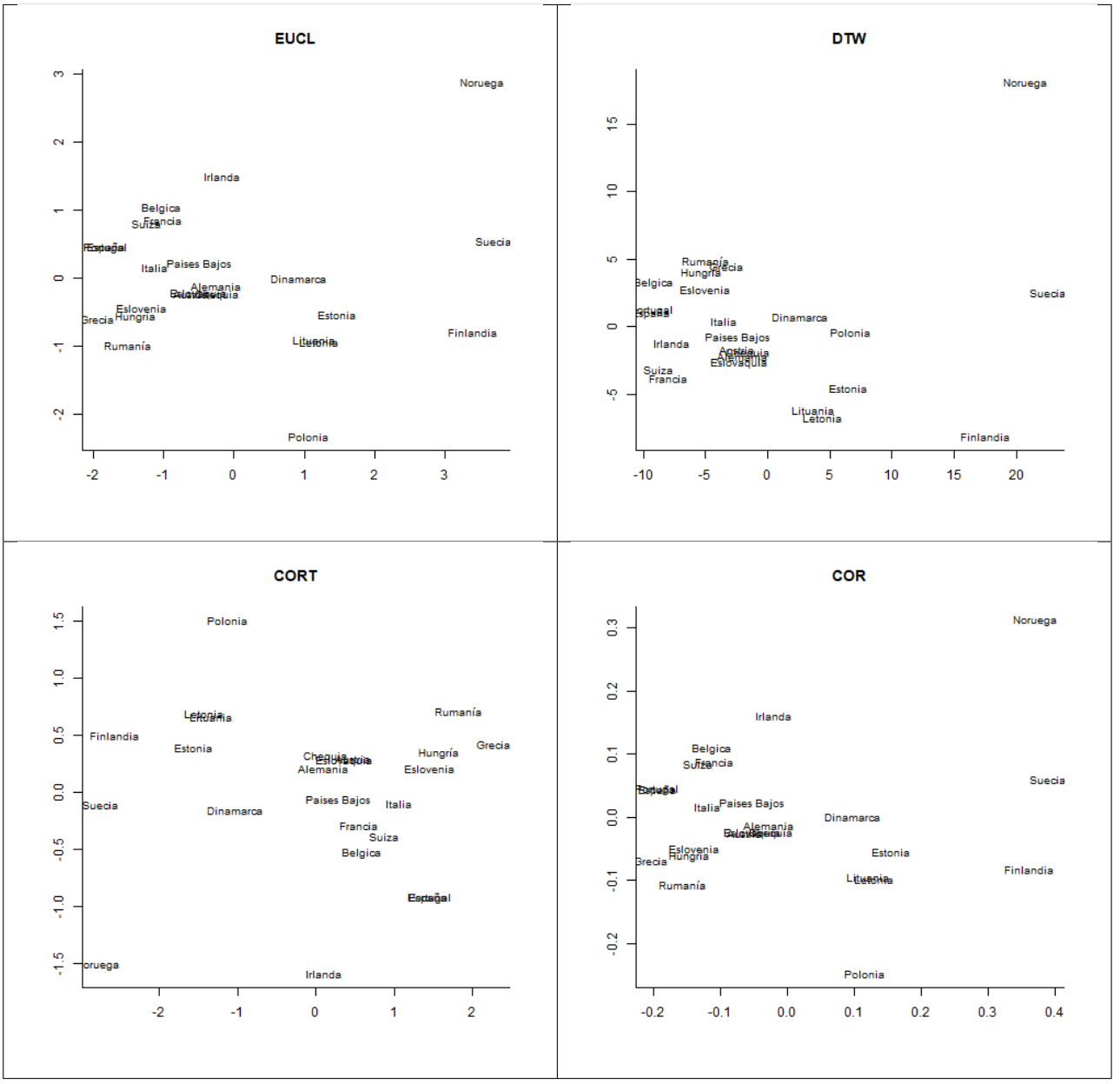


Figura 1.2: Representación mediante escalado multidimensional de los datos.

En R se representan los datos de acuerdo con los cuatro tipos de distancias elegidas y se aplica la metodología jerárquica aglomerativa para crear sus respectivos dendogramas. La correcta visualización de los mismos se hace posible tras la generación de documentos PDF en el directorio de trabajo. En ellos se pueden observar los resultados de cada algoritmo para cada tipo de distancia, facilitando su comparación.

Para una determinada distancia, es posible programar un algoritmo que genere de forma automática los posibles grupos en función de la cantidad de cortes, que se visualizan también mediante un archivo de salida PDF. Se decide, mediante una previa aplicación del Método del Codo o *Elbow Method* (ver figura 1.3 y [20]) a los datos escalados, que el número idóneo de clusters sea cuatro, por lo que los cortes estarán dentro del rango $(2, 4)$.

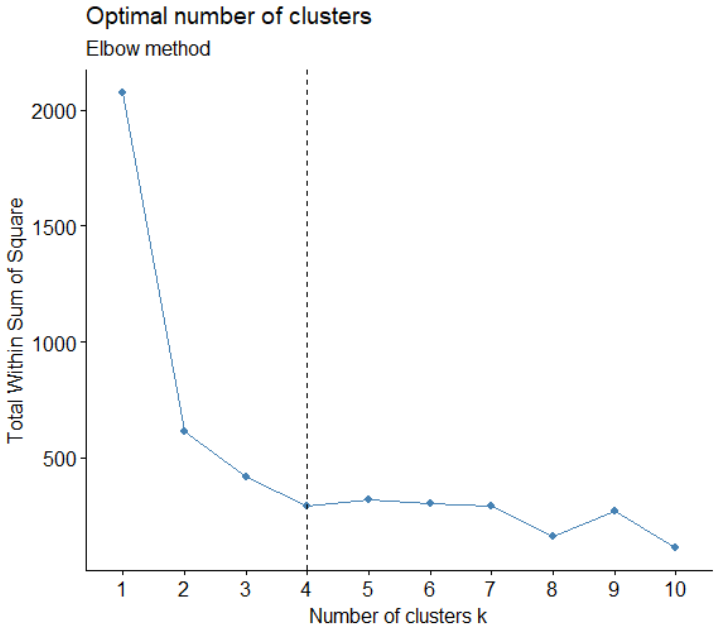


Figura 1.3: Selección del número de clusters mediante el Método del Codo.

En la figura 1.4, se muestran los cuatro grupos obtenidos con el encadenamiento completo, tras la aplicación de dicha técnica para la distancia CORT, escogida por la adecuación que presenta al problema.

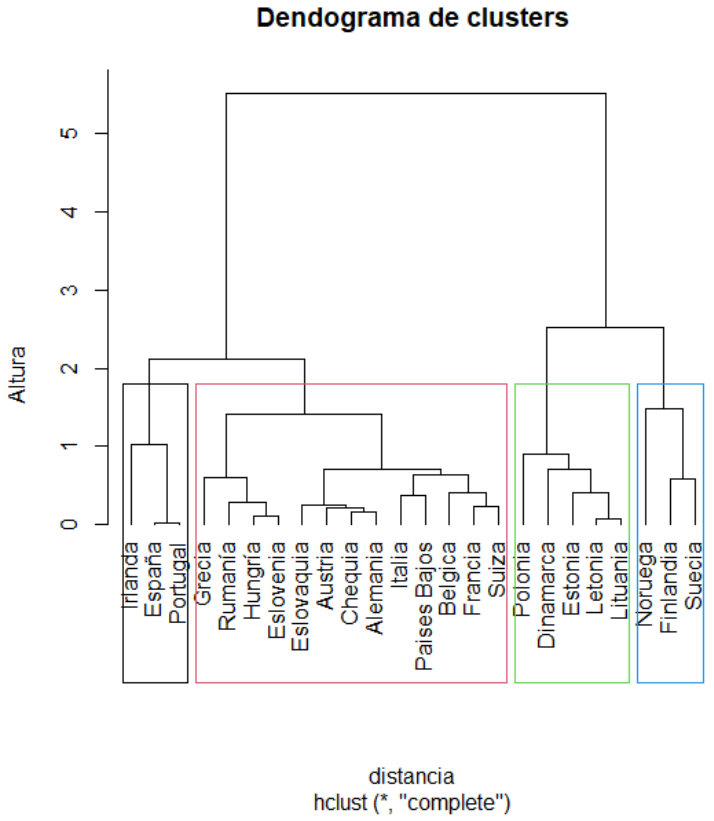


Figura 1.4: Dendogramas y grupos generados para d_{CORT} por el agrupamiento completo.

En vista del número de grupos generados, se tienen cuatro agrupaciones de los datos. En consecuencia, los grupos de países cuyos precios tienen comportamientos similares en el tiempo han quedado determinados de la siguiente manera:

- Grupo 1: España, Irlanda y Portugal.
- Grupo 2: Grecia, Rumanía, Hungría, Eslovenia, Eslovaquia, Austria, República Checa, Alemania, Italia, Países Bajos, Bélgica, Francia y Suiza.
- Grupo 3: Polonia, Dinamarca, Estonia, Letonia y Lituania.
- Grupo 4: Noruega, Suecia y Finlandia.

1.4. Extracción de representantes de cada cluster

Dado que ya se han conseguido los clusters de las series temporales mediante clustering jerárquico aglomerativo basado en la distancia de correlación temporal, es preciso tomar una serie que mejor represente a cada grupo, con el fin de completar la reducción de la dimensionalidad.

Para ello, se va a elegir como representante de cada cluster a aquel elemento que minimice la distancia CORT a la serie media del cluster, la cual se define como:

$$\left(\frac{\sum_{i=1}^{k} y_1^i}{k}, \frac{\sum_{i=1}^{k} y_2^i}{k}, \ldots, \frac{\sum_{i=1}^{k} y_T^i}{k} \right), \tag{1.1}$$

con k el número total de series temporales y T el número de registros u observaciones en la serie temporal. En este caso, $k = 24$ y $T = 84$, pues se han tomado las series temporales del valor del MWh de 24 países a lo largo de $12 * 7$ registros mensuales.

Los representantes obtenidos para cada grupo de series temporales han sido los siguientes:

- Representante Grupo 1: España.
- Representante Grupo 2: Eslovenia.
- Representante Grupo 3: Estonia.
- Representante Grupo 4: Suecia.

1.5. Resultados

Con el fin de obtener una validación interna de la solución, se puede notar que los resultados obtenidos para la agrupación en cuatro clusters mediante el encadenamiento completo de las distancias CORT de los datos son adecuados al problema que se estudia. De cada uno de los grupos, se ha obtenido un representante, tal y como se esperaba. Sería de gran interés completar esta validación con otra externa o relativa -como se hará en el capítulo 3 para el caso de las predicciones-, haciendo uso de test estadísticos que las soporten, o incluso cotejar los rendimientos de los distintos encadenamientos y distancias. Sin embargo, este objetivo supera al de este TFM, por lo que se toma el análisis realizado hasta ahora como correcto.

Capítulo 2

Modelos predictivos de series temporales

Tras completar el primer objetivo de este Trabajo, que consistía en la agrupación de las series temporales originales, en este capítulo se estudian distintas técnicas de predicción para las series temporales de los representantes de cada uno de los cuatro clusters obtenidos en la sección 1.4 del capítulo previo.

Este apartado está estucturado de la siguiente forma: en primer lugar, se nombran una serie de consideraciones generales sobre los modelos predictivos, para dar lugar a un estudio más profundo de cada una de las técnicas empleadas en el estudio, que se basan tanto en modelos estadísticos, como en modelos basados en Minería de Datos. Las fuentes consultadas para ello han sido extensas, pero destacan especialmente las publicaciones [13], [12] y [10].

En cuanto a la notación, se debe tener en cuenta que se usará la variable y en minúscula, provocando una diferencia con la notación empleada hasta ahora. Además, se empleará el subíndice y_t de forma indistinta, ya que la variable y siempre va a consistir en una serie dependiente del tiempo t.

2.1. Consideraciones generales

Como se ha comentado, las técnicas predictivas que se estudian en este capítulo abarcan desde modelos estadísticos hasta una serie de modelos más vanguardistas, inspirados en técnicas de un aspecto fundamental de la Minería de Datos, como es el Aprendizaje Automático. En primer lugar y antes de hacer una distinción entre ambos tipos de modelos, se repasan una serie de prácticas comunes a todos ellos: la segmentación de los datos en distintos conjuntos, y otras nociones básicas sobre modelos de regresión, como la estructura general de un modelo, las variables predictivas o las medidas de error cometido.

2.1.1. Regresión

En esta sección, se estudian los conceptos básicos sobre modelos de regresión, ya que la idea principal se basa en predecir una serie de nuestro interés $\hat{y}$ asumiendo que tiene una relación lineal con otra serie temporal y. La variable $\hat{y}$ suele conocerse como variable explicada y las variables y se conocen como variables explicativas.

Modelo lineal

En el caso más simple, el modelo de regresión permite una relación lineal entre la variable explicada $\hat{y}$ y la -única- variable explicativa y:

$$\hat{y}_t = \beta_0 + \beta_1 y_t + \epsilon_t.$$

Los coeficientes β_0 y β_1 indican, respectivamente, la ordenada en el origen y la pendiente de la recta. Es decir, β_0 representa el valor predicho $\hat{y}$ cuando $y = 0$, y β_1, el promedio del cambio predicho en $\hat{y}$ cuando se incrementa y en una unidad.

Sin embargo, cuando se usa un modelo de regresión lineal, se están asumiendo una serie de premisas sobre las variables de la ecuación: por ejemplo, que los errores tienen media distinta de cero, no están autocorrelados, y son independientes de las variables explicativas; o que la variable y no es aleatoria.

Para el caso multivariante, el modelo de regresión lineal múltiple puede escribirse como:

$$\hat{y}_t = \beta_0 + \beta_1 y_{1,t} + \beta_2 y_{2,t} + \cdots + \beta_k y_{k,t} + \varepsilon_t,$$

con ε_t de media cero y varianza σ^2.

Modelo no lineal

Aunque asumir linealidad entre las variables normalmente es adecuado, existen casos en los que una función no lineal produce un mejor ajuste. Normalmente, se puede aplicar una transformación logarítmica a la ecuación de regresión simple, aunque puede no adecuarse siempre, lo cual propicia la construcción de modelos no lineales. Suponiendo que solo se cuenta con una variable explicativa y, para la construcción de un modelo no lineal se requiere que:

$$\hat{y} = M(y) + \varepsilon,$$

con M una función no lineal. Para el caso de la regresión lineal estándar, $\beta_0 + \beta_1 y = M(y)$.

Estimación por mínimos cuadrados

En la práctica, ocurre que los valores de los coeficientes $\beta_0, \ldots, \beta_k$ son desconocidos y deben ser estimados a partir de los datos. El principio de mínimos cuadrados conforma una manera de elegir los coeficientes de forma efectiva minimizando la suma de los errores al cuadrado. Esto es, se eligen $\beta_0, \ldots, \beta_k$ que minimizan la expresión:

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 y_{1,t} - \beta_2 y_{2,t} - \cdots - \beta_k y_{k,t})^2.$$

Al hecho de encontrar los mejores estimadores de los coeficientes se le conoce como ajuste del modelo a los datos, aprendizaje o entrenamiento del modelo. Normalmente, a los coeficientes estimados se les asocia la notación $\hat{\beta}_0, \ldots, \hat{\beta}_k$.

Algunos predictores de interés

Existen multitud de predictores que surgen con frecuencia de la regresión de datos de series temporales. En particular, para la programación de este Trabajo se han usado una serie de variables predictivas, entre las que se destacan:

- Variables dummies. Son una serie de variables categóricas que toman dos valores (0/1), según correspondan a un "no" o a un "sí", que se pueden introducir a un modelo de regresión para estudiar el comportamiento en una determinada estación o el comportamiento de *outliers*. A modo de ejemplo, si se quisiera modelar un conjunto de datos anuales, que poseen un determinado patrón cuatrimestral, mediante un modelo de regresión lineal, se podrían definir variables dummies cuatrimestrales,

  $$\hat{y}_t = \beta_0 + \beta_1 y_t + \beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t} + \varepsilon_t,$$

  donde $d_{i,t} = 1$ si t está en el cuatrimestre i y 0 en caso contrario. En el caso del presente estudio, se han seleccionado variables dummies mensuales en los métodos basados en Máquinas de Vectores de Soporte.

- Variables de medias móviles. Ayudan a determinar la tendencia de la serie. Se obtienen dividiendo el total de los datos en bloques de una determinada longitud y calculando su promedio.

Medidas regresivas de error

A la hora de seleccionar una medida de error de un modelo, es importante tener en cuenta el ámbito del estudio, el conjunto de los datos, así como el papel que juegan los errores dentro del mismo, pues las distintas medidas de error describen características distintas de los modelos predictivos y de los datos. En particular, se consideran dos de los estadísticos más usados para problemas de regresión, que son el Error Absoluto Medio (MAE, por sus siglas en inglés) y la raíz del Error Cuadrático Medio (RMSE).

Definición 2.1. *Sean y_i los valores esperados de una serie temporal de tamaño T e $\hat{y}_i$ la predicción realizada por del modelo. Se define la raíz de su error cuadrático medio como:*

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2}.$$

Como se puede ver, el error cuadrático medio es el promedio de los errores de un conjunto de valores observados al cuadrado. Además, su valor es siempre positivo, valiendo cero en el caso contar de un modelo ajustado perfectamente a los datos. En especial, es de gran utilidad cuando hay valores atípicos en los datos, aunque puede llegar a sobreestimar la bondad del mismo, en el caso de que el total de los errores sean menores de uno, en módulo.

Definición 2.2. *Sean y_i los valores esperados de una serie temporal de tamaño T e $\hat{y}_i$ la predicción realizada por del modelo. Se define su error absoluto medio como:*

$$MAE = \frac{1}{T}\sum_{i=1}^{T}|y_i - \hat{y}_i|.$$

El Error Absoluto Medio es el promedio de las diferencias absolutas entre un conjunto de valores observados y sus predicciones. Es más robusto y menos sensible a los valores atípicos, ya que todas las diferencias individuales ponderan de la misma manera en el promedio.

2.1.2. Aprendizaje Automático

Como su nombre indica, el Aprendizaje Automático surge por la necesidad de conseguir que los ordenadores realicen determinadas tareas, de forma automática y a través de un método que les permita corregir sus propios errores de forma autónoma (aprender), hasta la obtención del resultado deseado. La principal diferencia entre estas técnicas y los algoritmos habitualmente usados es la de poder responder ante nuevas situaciones. Por ejemplo, ¿Cómo programamos un algoritmo capaz de estimar valores futuros de una serie de tiempo, basándose en los comportamientos pasados de sus indicadores? La cantidad de posibilidades que hay es amplia y ante una situación futura, el algoritmo no sabría qué responder, de ahí que surja la necesidad de encontrar un modelo que habiéndole mostrado una serie de observaciones, se comporte favorablemente ante situaciones nunca vistas.

Definición 2.3. *Sea $g^* : X \to Y$ una aplicación suprayectiva[1]. Un conjunto de entrenamiento de la función g^* es un conjunto finito $D \subset X \times Y$. Al conjunto X se le llama el conjunto de entradas de la función y a Y el conjunto de salidas.*

En esencia, a la hora de aproximar cierta función $g^* : X \to Y$ se distinguen dos caminos a seguir como los más importantes dentro del aprendizaje automático: el *aprendizaje supervisado* y el *aprendizaje no supervisado*.

- Cuando el conjunto de entrenamiento sea de la forma: $\{x_t, g^*(x_t)\}_{t \leq T}$, $T \in \mathbb{N}$, estaremos hablando de aprendizaje supervisado. Es decir, proporcionamos información completa de nuestra función g^* y a través de esos ejemplos debemos ser capaces de predecir de forma correcta $g^*(x)$ para cada $x \in X$. Como ejemplo podemos citar los modelos de regresión lineal y logística, los arboles de decisión, las redes neuronales y KNN (k -nearest neighbors).
- Cuando el conjunto de entrenamiento sea de la forma: $\{x_t\}_{t \leq T} \times \emptyset$, estaremos hablando de aprendizaje no supervisado. Aquí no conocemos cuáles son las salidas de g^*, por lo que las estrategias, en este caso, consisten en encontrar cierta similitud entre nuestras entradas. Ejemplos de este tipo de métodos serían el agrupamiento jerárquico y k-means.

Definición 2.4. *Sea $g^* : X \to Y$ una aplicación suprayectiva. Un conjunto de test de la función g^* es un conjunto finito $D' \subset X \times Y$. Con X las entradas e Y las salidas.*

Los modelos de aprendizaje supervisado toman un conjunto de entrenamiento D y otro de test D' disjuntos. Son entrenados con D para luego medir su precisión usando D' y así comprobar que ante ejemplos nunca vistos devuelven una salida correcta.

Definición 2.5. *Sean $g^* : X \to Y$ una aplicación y $\{x_t, g^*(x_t)\}_{t \leq T}$ un conjunto de entrenamiento. Diremos que $f : X \to Y$ es una aproximación exacta de g^* sobre ese conjunto de entrenamiento si se verifica:*

$$g^*(x_t) = f(x_t); \ \ \forall t \in \{1, ..., T\}$$

[1] $g^*(X) := \{g^*(x); x \in X\} = Y$.

2.2. Modelos predictivos estadísticos

A continuación, profundizaremos es este tipo de métodos, cuya idea radica en la creación de un modelo basado en ecuaciones matemáticas que representen la interacción ente las variables involucradas en el mismo. En particular, nos centraremos en los modelos SARIMA y TBATS, que pueden implementarse en R haciendo uso de la librería *forecast*. Con el fin de simplificar la notación, las variables explicadas para un instante t de tiempo, con $t \leq T$, serán nombradas como y_t.

2.2.1. ARIMA estacional

Antes de profundizar en el modelo SARIMA (Autorregresivo Estacional Integrado de Medias Móviles), es preciso aclarar algún concepto previo, lo cual pasa por la explicación de cada una de estas nociones. Se comenzará construyendo esta definición por su componente más simple, la autorregresión.

Estructura del modelo autorregresivo

En un modelo de regresión múltiple, se predicen las variables de interés usando una combinación lineal de los predictores. En un modelo autorregresivo, se predice la variable de interés usando una combinación lineal de los valores pasados de los predictores. El término autorregresión indica que se realiza una regresión de una variable con respecto a sí misma.

El modelo predictivo autorregresivo (AR) estudia las observaciones pasadas de un proceso con el fin de predecir observaciones futuras. En este proceso, el valor en un determinado instante puede ser calculado en función de los valores pasados de la serie temporal, siempre que ésta sea estacionaria (es decir, sus propiedades estadísticas no dependen del instante en el que se producen las observaciones, o sea, no están correlacionadas).

Un modelo AR que depende de los p valores más recientes cumple lo siguiente:

$$AR(p) \qquad \hat{y}_t = \alpha_1 + \beta_1 y_{t-1} + \alpha_2 + \beta_2 y_{t-2} + \cdots + \alpha_p + \beta_p y_{t-p} + \epsilon_t,$$

donde ϵ_t es el ruido blanco. Podemos pensar en el modelo AR como en una regresión múltiple que toma como predictores los valores retardados o *laggeados* de y_t.

Estacionariedad y diferenciación

Como es evidente, las propiedades de una serie temporal estacionaria no dependen del instante en el que ésta es observada. Así, las series temporales que siguen algún determinado patrón, o que tienen estacionalidades, no son estacionarias. En general, una serie estacionaria se caracteriza por mostrar una varianza constante en su representación, y no tiene patrones predecibles a largo plazo.

Sin embargo, existe un proceso por el cual transformar una serie no estacionaria en estacionaria, llamado diferenciación. La diferencia de una serie indica el cambio entre observaciones de la misma, y puede realizarse según distintos órdenes. Se exponen a continuación las diferencias más comunes:

- Primer orden. Busca transformar una serie no estacionaria, a través del cambio entre observaciones consecutivas de la misma. Como resultado, se obtiene una serie estacionaria

de $T-1$ valores.

$$y'_t = y_t - y_{t-1}.$$

- Segundo orden. En el caso de que una diferencia de primer orden no sea suficiente, se vuelve a aplicar una diferenciación, por la cual se obtiene una serie estacionaria de $T-2$ valores.

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}.$$

- Estacional. Una diferencia estacional es la diferencia entre una observación y la observación previa de la misma estación. Si m es el número de estaciones,

$$y'_t = y_t - y_{t-m}.$$

Estructura del modelo de medias móviles

En lugar de usar los valores pasados de una variable explicada en una regresión, el modelo de Medias Móviles (MA) usa los valores de los errores pasados cometidos.

$$MA(q) \qquad \hat{y}_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q},$$

donde ϵ_t es el ruido blanco. Al modelo ajustado por la ecuación anterior se le conoce como modelo de medias móviles de orden q.

Estructura del modelo ARIMA no estacional

Si se combinan los conceptos de autorregresión, diferenciación y medias móviles, surge el modelo ARIMA no estacional. ARIMA es un acrónimo para Autorregresivo Integrado de Medias Móviles (en este contexto, la integración es el proceso inverso a la diferenciación). El modelo completo se conoce como ARIMA(p,d,q), con:

$$ARIMA(p,d,q) \qquad \hat{y}'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} + \varepsilon_t, \tag{2.1}$$

donde p es el orden de la parte autorregresiva, d es el grado de la diferenciación , q es el orden de la parte de medias móviles e y'_t es la serie diferenciada. Además, comprueba que los predictores del lado derecho incluyen tanto los retardos de la variable y_t como los retardos de sus errores.

Una vez se ha comenzado a combinar componentes de esta manera, la utilización de la notación de retroceso es más sencilla. Por ejemplo, la ecuación 2.1 puede reescribirse mediante el operador diferencia B como:

$$\begin{array}{cccc} (1-\phi_1 B - \cdots - \phi_p B^p) & (1-B)^d y_t & = & c + (1+\theta_1 B + \cdots + \theta_q B^q)\varepsilon_t \\ \uparrow & \uparrow & & \uparrow \\ \text{AR}(p) & d \text{ diferencias} & & \text{MA}(q) \end{array} \tag{2.2}$$

No obstante, R usa una parametrización ligeramente distinta, que es preciso presentar:

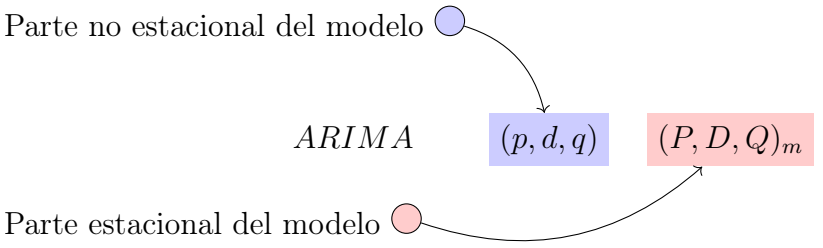
$$(1-\phi_1 B - \cdots - \phi_p B^p)(y'_t - \mu) = (1+\theta_1 B + \cdots + \theta_q B^q)\varepsilon_t, \tag{2.3}$$

donde $y'_t = (1-B)^d y_t$ y μ es la media de y'_t. Para convertir esta parametrización a la forma 2.2, basta con tomar $c = \mu(1-\phi_1 - \cdots - \phi_p)$.

SARIMA

Aunque en bibliografía los autores se centren en la explicación de en datos y modelos ARIMA no estacionales antes que en otros de distinto tipo, es conveniente saber que los modelos ARIMA también son capaces de modelar una gran variedad de datos estacionales.

Un modelo ARIMA estacional, o SARIMA, por sus siglas en inglés, se forma añadiendo términos estacionales a un modelo ARIMA. Se suelen representar de la siguiente forma:



(2.4)

donde

- p es el parámetro para la componente no estacional de $AR(p)$.
- q es el parámetro para la componente no estacional de $MA(q)$.
- d es el orden de la diferencia por la cual el proceso $ARIMA$ se convierte en un $ARMA(p, q)$.
- P es el parámetro para la componente estacional de $AR(P)$.
- Q es el parámetro para la componente estacional de $MA(Q)$.
- D es el orden de la diferencia para la componente estacional.
- m es la frecuencia del parámetro estacional.

Como se indica, se usa notación mayúscula para la parte estacional del modelo, y minúscula para la parte no estacional. Por ejemplo, un ARIMA$(p, d, q)(P, D, Q)_{12}$, para datos mensuales, como es el caso, puede expresarse mediante notación de retroceso por:

$$(1 - \phi_1 B)\ (1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)\ (1 + \Theta_1 B^{12})\varepsilon_t.$$

2.2.2. TBATS

El modelo TBATS, llamado así por sus siglas en inglés (Trigonometric regressors for Box-Cox transformations with Arma errors and Trend and Seasonal components), es introducido como alternativa a los modelos ARIMA por [11] para el pronóstico de series de tiempo estacionales complejas. Como novedad, se presenta notación trigonométrica con el fin de llevar a cabo una descomposición exitosa de las mismas, y una correcta identificación y extracción de las componentes estacionales.

Para ello, en primer lugar se extienden los modelos de suavizamiento exponencial (puede consultarse [16] en mayor detalle) para dar lugar al modelo BATS, como se indica:

$$y_t^{(\omega)} = \begin{cases} \dfrac{y_t^{\omega} - 1}{\omega} & \omega \neq 0 \\ \log y_t & \omega = 0 \end{cases}$$

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = b_{t-1} + \beta d_t$$

$$s_t^{(i)} = s_{t-m_i}^{(i)} + \gamma_i d_t$$

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t,$$

donde

- $y_t^{(\omega)}$ denota la transformación de Box-Cox, que corrije la asimetría de una variable y la transforma a una nueva que siga una distribución normal.
- $m_1, \ldots, m_T$ denotan los periodos estacionales.
- l_t representa la componente de nivel de la serie en el tiempo t.
- b_t representa la componente de tendencia de la serie en el tiempo t.
- $s_t^{(i)}$ representa la i-ésima componente estacional en el tiempo t.
- d_t denota un proceso ARMA(p, q).
- ε_t es el proceso de ruido blanco Gaussiano de media cero y varianza constante σ^2.
- $\alpha, \beta, \gamma_i, \quad i = 1, \ldots, N + L$, son los parámetros de suavizamiento.
- ϕ es el parámetro de amortiguación, que aporta un mayor control sobre la extrapolación de tendencia, especialmente cuando su componente está amortiguada.

La notación usada para denotar este tipo de modelo es BATS$(P, Q, m_1, \ldots, m_T)$. Como se ha comentado, B se refiere a la transformación de Box-Cox; A, a los residuos ARMA; T, a la componente de tendencia del modelo y S, a la componente estacional.

Finalmente, se introduce una nueva representación trigonométrica de las componentes estacionales, basada en las series de Fourier. Así, se puede reemplazar el valor de $s_t^{(i)}$ en el modelo BATS por las siguientes expresiones, para dar lugar al modelo TBATS:

$$s_t^{(i)} = \sum_{j=1}^{k_i} \alpha_{j,t}^{(i)} \cos(\lambda_j^i t) + \beta_{j,t}^{(i)} \sin(\lambda_j^i t)$$

$$\alpha_{j,t}^{(i)} = \alpha_{j,t-1}^{(i)} + \kappa_1^{(i)} d_t$$

$$\beta_{j,t}^{(i)} = \alpha_{j,t-1}^{(i)} + \kappa_1^{(i)} d_t,$$

donde

- $\kappa_1^{(i)}$ y $\kappa_2^{(i)}$ son los parámetros de suavizado.
- $\lambda_j^i = \dfrac{2\pi j}{m_i}$.

Mediante la sustitución de la componente estacional $s_t^{(i)}$, se obtiene una nueva clase de modelos, llamados modelos trigonométricos de suavizado exponencial, bajo la notación TBATS$(p, q, \{m_1, k_1\}, \ldots, \{m_T, k_T\})$. Para una mirada más profunda a la aplicación de dichos modelos, se sugiere una lectura de [26].

2.3. Modelos predictivos de aprendizaje automático

En esta sección se estudiarán los modelos predictivos de aprendizaje automático más utilizados a la hora de predecir series de tiempo. Los tres modelos estudiados se basan en algoritmos de aprendizaje supervisado.

2.3.1. Redes neuronales autorregresivas

En esta sección vamos a explicar este caso particular de red neuronal, cuyas técnicas están basadas en aprender ciertos ejemplos y responder de forma correcta ante nuevas situaciones y por tanto, formando parte del aprendizaje supervisado. Se dan por sentadas algunas nociones sobre redes neuronales elementales, que pueden cotejarse en [13] y [18], preferiblemente antes de la lectura del modelo ARNN.

Originalmente, una neurona biológica recibe y transmite señales eléctricas (impulsos nerviosos) hacia otras células. Matemáticamente, podemos pensar que una neurona recibe una serie de entradas $x \in \mathbb{R}^n$ para las cuales devuelve una salida $y = y(x) \in \mathbb{R}$.

Los perceptrones multicapa (MLP, por sus siglas en inglés) son la arquitectura de redes neuronales artificiales que se usa con mayor frecuencia a la hora de predecir series temporales no lineales. Recordemos que un perceptrón es una red neuronal cuya función de activación es una función característica. Entre sus aplicaciones destacan la predicción en el ámbito energético, como pueden ser la demanda de energía eléctrica o los precios de la electricidad.

Las redes neuronales autorregresivas o ARNN, por sus siglas en inglés, surgen tras combinar un MLP con un modelo lineal autorregresivo. Inicialmente basadas en un contraste estadístico de no linealidad para los modelos anteriores, las ARNN son una buena alternativa a los MLP en la predicción de series temporales debido a su componente lineal autorregresiva.

Tal y como se expone en [28], en un modelo ARNN, la variable dependiente y_t se obtiene como una función no lineal de sus p valores pasados, y_{t-p}, para $p = 1, \ldots, P$, y de sus h variables ocultas, para $h = 1, \ldots, H$ de acuerdo con la ecuación:

$$\hat{y}_t = \eta + \sum_{p=1}^{P} \varphi_p y_{t-p} + \sum_{h=1}^{H} \beta_h G(\omega_h + \sum_{p=1}^{P} \alpha_{p,h} y_{t-p}),$$

donde $G(\cdot)$ es la función sigmoidal adaptativa:

$$G(u) = \left[\frac{1}{1 + \exp(-u)} \right]^M.$$

Observación 2.1. *Notemos que la función de φ_p se representa en la figura 2.1 color aguamarina, la de $\alpha_{p,h}$ en negro, β_h en oliva, ω_h en naranja y η en morado.*

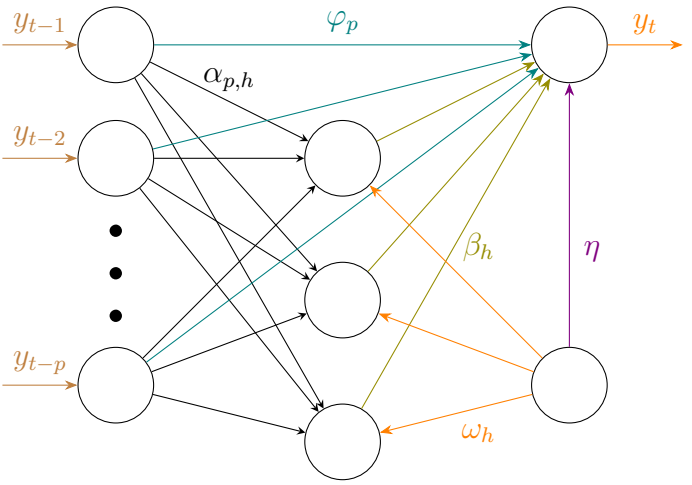


Figura 2.1: Arquitectura de una red neuronal autorregresiva.

Imponiendo la restricción $\varphi_1 = \varphi_2 = \cdots = \varphi_p$, esta ecuación se reduce a un perceptrón multicapa; o a un modelo autorregresivo mediante la imposición $H = 0$. Además, los parámetros $\eta, \varphi_p, \beta_h, \omega_h, \alpha_{p,h}$ y M se estiman mediante la minimización del error de regularización

$$\lambda \hat{E},$$

donde λ es un parámetro externo, definido a conveniencia del usuario. Así, se consigue que el error dependa únicamente de los parámetros del modelo:

$$\hat{E} = |\eta| + \sum_{h=1}^{H}(|\beta_h| + |\omega_h|) + \sum_{p=1}^{P}|\varphi| + \sum_{p=1}^{P}\sum_{h=1}^{H}|\alpha_{p,h}|.$$

En la práctica, el paquete *arnn* de R es capaz de implementar una red neuronal artificial como la que se viene describiendo. Consta de dos funciones principales: *arnn*, para la esimación del modelo, y *forecast*, para su predicción. En el capítulo posterior se ejemplificará su uso en la serie de los datos representantes.

2.3.2. K vecinos más cercanos

Esta estrategia de aprendizaje supervisado consiste en la búsqueda de elementos similares al que se desea predecir dentro de un conjunto de datos, utilizando una medida de distancia, como las mencionadas en el capítulo anterior. De entre todos ellos, se escogen aquellos k elementos que guarden una mayor similitud con el elemento a predecir. Finalmente, el valor siguiente de esos elementos determinará la predicción realizada.

La técnica del KNN es muy versátil y puede ser utilizada para problemas tanto de clasificación, como de regresión. Entre estos últimos se encuadra la predicción de series temporales, que es la principal tarea de este capítulo. Su particularización puede implementarse en R mediante el paquete *tsfknn*, y se describe de la siguiente manera ([7]):

1. Se transforma la serie temporal original de longitud T en elementos de longitud d. Así, se obtiene una serie definida como $y_t^d = \{y_{t-d+1}, \ldots, y_{t-1}, y_t\}$.

2. Se calculan las distancias de todos los elementos (anteriores) de la serie al elemento que se busca predecir: $y_t^d = \{y_{t-d+1}, \ldots, y_{t-1}, y_t\}$.
3. Se ordenan los elementos en función de la distancia escogida y se toman aquellos k más cercanos: $y_{t_1}^d, y_{t_2}^d \ldots, y_{t_k}^d$.
4. Se obtienen los valores siguientes a cada uno de los k elementos seleccionados y se calcula su media ponderada para realizar la predicción:

$$\hat{y}_{t+1} = \frac{\sum_{i=1}^{k} w_i y_{t_i}^d}{\sum_{i=1}^{k} w_i},$$

donde $\hat{y}_{t+1}$ es la predicción en el $t+1$-ésimo instante temporal y w_i el peso asociado al valor siguiente del i-ésimo vecino en la media ponderada.

Para su pronóstico, se decide contar con dos estrategias de predicción de series temporales de tipo *multi-step-ahead*, la estrategia recursiva y la estrategia MIMO (por sus siglas en inglés, Multiple Input Multiple Output). Su explicación puede encontrarse en [8].

Estrategia recursiva

Sigue la línea de los modelos SARIMA o TBATS en el sentido en el que, para el pronóstico de varios momentos temporales, se toma un modelo de una única variable respuesta, y se reitera el proceso tantas veces como variables se quieran predecir, sustituyendo el valor explicado por el explicativo en cada paso.

En esta estrategia, un único modelo M se entrena para arrojar una predicción de un solo paso, por ejemplo

$$y_{t+1} = M(y_t, \ldots, y_{t-d-1}) + \varepsilon,$$

con $t \in \{d, \ldots, T-1\}$. Para predecir l adelantos, se empieza prediciendo el primer adelanto aplicando el modelo. Seguidamente, se usa el valor que se acaba de predecir como una parte de las variables de entrada para la predicción del siguiente adelanto mediante el mismo modelo, y se repite el proceso sucesivamente.

Sea $\hat{M}$ el modelo entrenado. Entonces las predicciones vienen dadas según el esquema:

$$\hat{y}_{T+l} = \begin{cases} \hat{M}(y_T, \ldots, y_{T-d+1}), & \text{si} \quad l = 1 \\ \hat{M}(\hat{y}_{T+l-1}, \ldots, \hat{y}_{T+1}, y_T, \ldots, y_{T-d+l}), & \text{si} \quad l = \{2, \ldots d\} \\ \hat{M}(\hat{y}_{T+l-1}, \ldots, \hat{y}_{T-d+l}), & \text{si} \quad l = \{d+1, \ldots L\} \end{cases}$$

Dependiendo del ruido presente en la serie temporal y del horizonte de predicciones, la estrategia recursiva puede tener un bajo rendimiento al predecir más de un valor. Esto especialmente ocurre si el horizonte predictivo l supera la cantidad d, llamada dimensión de incrustación. También es sensible a la acumulación de errores, por lo que puede dar lugar a predicciones inexactas. No obstante y, a pesar de las limitaciones, esta estrategia proporciona muy buenos resultados cuando se usan modelos de Aprendizaje Automático, como son las redes neuronales o, en este caso, los vecinos más próximos.

Estrategia MIMO

Hasta ahora, se han estudiado estrategias de única salida, que modelan datos mediante una función de múltiple entrada pero única salida (o *Multiple Input, Single Output*). La introducción de la estrategia de MIMO (o *Multiple Input, Multiple Output*) tiene por objetivo mejorar la modelización de única salida, debido a las desventajas comentadas anteriormente.

La estrategia MIMO aprende un modelo de múltiple salida M a partir de una serie temporal $\{y_1, \ldots, y_T\}$, donde

$$\{y_{t+L}, \ldots, y_{t+1}\} = M(y_T, \ldots, y_{T-d+1}) + \varepsilon,$$

con $t \in \{d, \ldots, T-L\}$, $M : \mathbb{R}^d \to \mathbb{R}^L$ es una función evaluada en vectores y $\varepsilon \in \mathbb{R}^L$ es un vector de ruido.

Así, las predicciones tras un paso en un modelo MIMO M siguen el esquema

$$\{\hat{y}_{t+L}, \ldots, \hat{y}_{t+1}\} = \hat{M}(y_T, \ldots, y_{T-d+1}) + \varepsilon,$$

2.3.3. Máquinas de Vector Soporte

Como se ha explicado en la definición del error cuadrático medio 2.1, una de las limitaciones de la estimación por mínimos cuadrados es la suposición de que todas las observaciones tengan el mismo peso o importancia en el modelo. Es por ello que surgen técnicas de mejora, con el fin de aportar robustez al modelo regresivo lineal, entre las que se encuentra la regresión basada en las Máquinas de Vector Soporte o SVM.

Mientras que la regresión lineal busca minimizar una función de error (generalmente, el cuadrático en todos los puntos del conjunto de entrenamiento), en la regresión mediante SVM se define una región en torno al hiperplano de separación en la que se ignoran los errores, de anchura ε, que se busca maximizar.

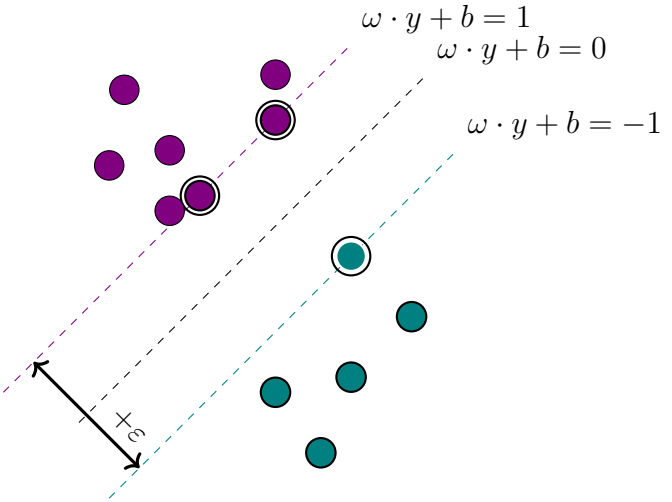


Figura 2.2: Estructura de una SVM, donde ω es el vector de pesos y b el sesgo.

Para una mejor comprensión, es precisa la explicación de una serie de conceptos, que se indican a continuación:

- Kernel. En este método, la regresión se hace para una dimensión superior a la dimensión del problema. La asignación entre los puntos de un espacio de menor dimensión a otro de mayor dimensión es posible gracias a la función de kernel, que devuelve un hiperplano en un espacio de mayor dimensión.
- Hiperplano. Los hiperplanos juegan un papel distinto para los problemas basados en SVM: en regresión, son las rectas que ayudan a predecir el valor objetivo.
- Líneas limítrofes. Son las rectas dibujadas a una distancia ε del hiperplano. Dan lugar a una región alrededor de la función de kernel llamado tubo, que es la región de aceptación de la predicción. Por ello, los errores cometidos entre ambas líneas son despreciados.
- Vector de soporte. Es la unidad sobre la que se define el hiperplano.

Dada una serie de tiempo con regresores y_t, para la que se poseen D ejemplos representativos, se puede aproximar mediante una SVM a través de la función:

$$\hat{y}_{t+1} = b + \sum_{d=1}^{D} \omega_d \cdot k(y_t, y_d),$$

donde ω_d es el vector de pesos y b el sesgo, y $k(y_t, y_d)$ la función de kernel. Así, una SVM es una combinación lineal del mapeo de y_t en un espacio definido por los puntos x_d y la función de transformación no lineal $k(\cdot, \cdot)$.

La estimación de esta función se basa en la minimización de la función de riesgo regularizado (para mayor detalle, se recomienda encarecidamente consultar [29]). Su solución puede ser obtenida mediante la teoría de los multiplicadores de Lagrange, y puede ser implementada en R mediante el paquete *e1071*.

En cuanto al kernel, existen varias funciones que son típicamente utilizadas. Se destacan las siguientes:

- Lineal, donde $k(x, y) = xy$.
- Polinomial, donde $k(x, y) = (xy + 1)^d$.
- Gaussiana, donde $k(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right)$.

2.4. Método predictivo combinado

Para terminar, se presenta una técnica que realiza una predicción mediante la combinación de modelos previos, implementada en R mediante el paquete *forecastcomb*.

Como se indica en [30], cuya lectura se recomienda, el método combinado proporciona una alternativa a escoger el mejor método de entre una serie de modelos, ya que incluye información de todos ellos con el fin de mejorar la exactitud de la aproximación.

A continuación, se exponen tres maneras sencillas de combinar distintas predicciones: la combinación mediante la Media Aritmética Simple (SA), la combinación de Bates-Granger (BG) y la combinación de Mínimos Cuadrados Restringidos (CLS). La notación está ajustada de la siguiente forma: se llama f^c a la predicción combinada y f_i a la predicción obtenida para cada modelo, con $i = 1, \ldots, P$, donde P el número de modelos considerados.

Combinación de Media Aritmética Simple (SA)

El enfoque más intuitivo para la combinación de predicciones es el uso de todas ellas, en la misma medida. La aproximación dada por la Media Aritmética Simple (o *Simple Average*) proporciona un buen punto de partida para la predicción combinada. Su expresión es la siguiente:

$$f^c = \frac{1}{P}\sum_{i=1}^{P} f_i.$$

Combinación de Bates-Granger (BG)

La aproximación de Bates-Granger usa los elementos diagonales de la matriz error cuadrático medio estimado para computar los pesos de la combinación:

$$f^c = \frac{\hat{\sigma}^{-2}(i)}{\sum_{j=1}^{P}\hat{\sigma}^{-2}(j)}\sum_{i=1}^{P} f_i,$$

donde $\hat{\sigma}^{-2}(i)$ es el error cuadrático medio de la predicción del modelo i.

Combinación de Mínimos Cuadrados Restringidos (CLS)

La idea de este método es que la predicción combinada sea una función lineal de las predicciones individuales, donde los pesos se determinan mediante una regresión de las mismas en la variable explicada:

$$y = \alpha + \sum_{i=i}^{P} \omega_i f_i + \varepsilon.$$

Usando una parte de las predicciones para entrenar el modelo de regresión, se pueden estimar los coeficientes de la ecuación anterior, pues minimizan la suma de sus errores al cuadrado. La predicción combinada ordinaria viene dada por:

$$f^c = \hat{\alpha} + \sum_{i=i}^{P} \hat{\omega}_i f_i.$$

Si restringimos las soluciones de esta ecuación a aquellas que además verifiquen que $\sum_{i=1}^{P}\omega_i = 1$ y $\omega_i \geq 0$, se obtiene la predicción combinada de mínimos cuadrados restringidos o *Constrained Least Squares*.

Capítulo 3

Resultados

A continuación, mostramos los resultados de aplicar los distintos modelos para el ajuste y predicción de los datos de los representantes obtenidos en el capítulo 1. Antes de profundizar en el rendimiento de los distintos modelos sobre nuestro problema, consideramos necesaria la elaboración de una serie de comentarios previos.

3.1. Comentarios previos

Como hemos explicado, la separación de los datos en un conjunto de entrenamiento y un conjunto de test es una parte importante en la evaluación de modelos de Minería de Datos. Tras usar el conjunto de entrenamiento para procesar un determinado modelo, es posible evaluarlo a través de la realización de predicciones frente al conjunto de test. Debido a que los datos en el conjunto de test contienen valores de los atributos que se busca predecir, no es complicado determinar si las predicciones del modelo con correctas.

Es por eso que se va a dividir el conjunto inicial de los T datos en dos subconjuntos, intentando que el conjunto de entrenamiento contenga cerca del 80 % del total de los datos, y el de test, alrededor del 20 %. Se decide la siguiente diferenciación, de forma que $T = N + L$:

- Conjunto de entrenamiento: Los $N = 82$ primeros datos de las series representantes son los datos que se usarán para crear los modelos y el estadístico.
- Conjunto de test: Los $L = 2$ últimos datos de cada serie se utilizarán para indicar el error real cometido con los modelos creados.

Es preciso aclarar que contar con tan solo dos datos para la validación de los modelos es ciertamente escaso. Sin embargo, se ha tomado esta decisión tras realizar una serie de pruebas, con el objetivo de intentar recoger de la mejor manera posible el efecto del pico de subida los precios, muy pronunciado en los ultimos meses de 2021. De otra forma, esta tendencia inusual no se contemplaría, dando lugar a unos resultados poco acordes a los datos observados, o aproximaciones poco exactas.

Una vez realizada esta observación, estamos en disposición de presentar los resultados obtenidos.

3.2. SARIMA

En primer lugar, presentamos el rendimiento de este modelo en nuestro conjunto de datos. En la figura 3.1, se puede notar el buen ajuste que proporciona el modelo SARIMA a los datos de las series temporales de los representantes de los clusters obtenidos. En particular, se ha podido modelizar con éxito las series de Eslovenia, Estonia, España y Suecia.

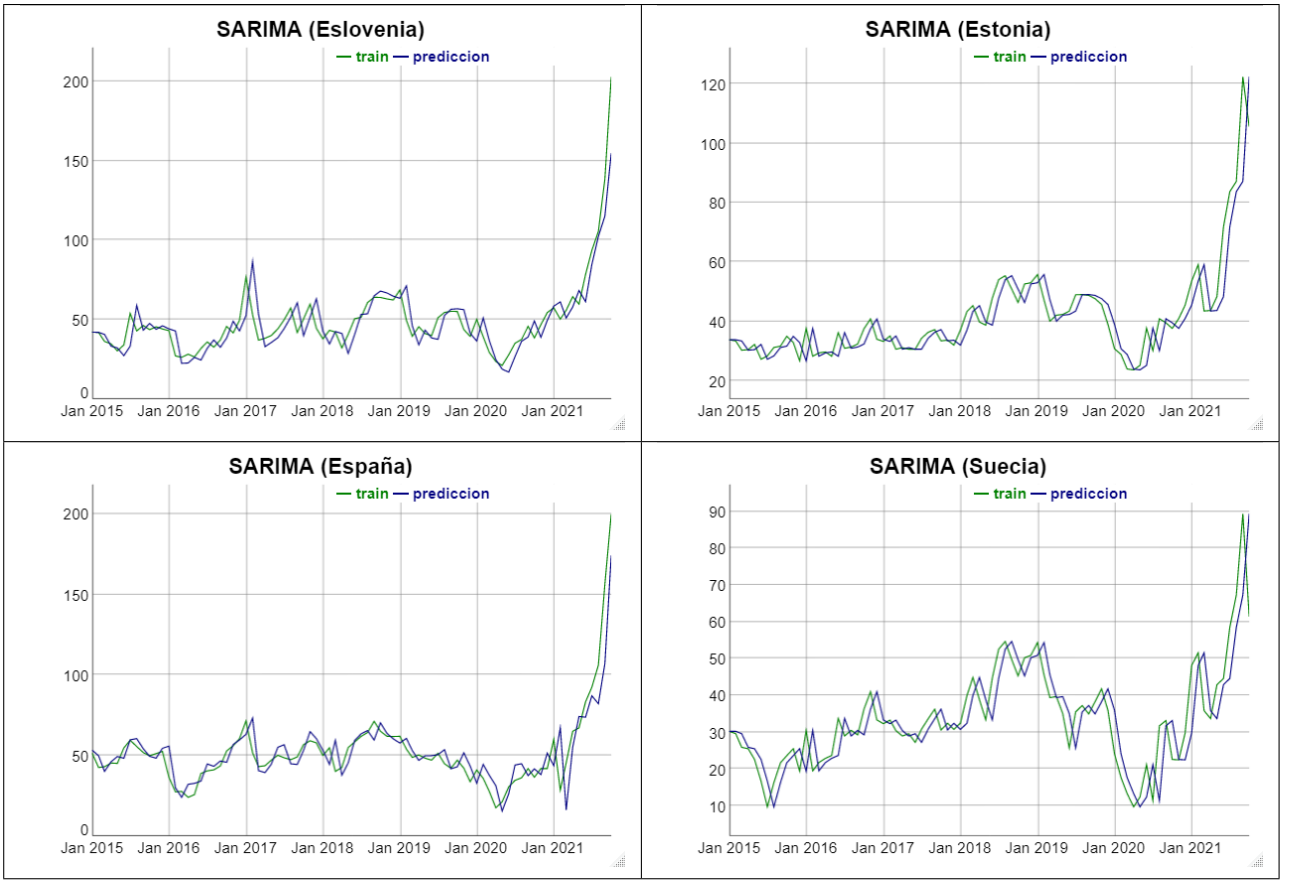


Figura 3.1: Ajuste de SARIMA a los datos de las series de Eslovenia, Estonia, España y Suecia.

En la figura 3.2, se muestra el conjunto de los datos de cada país segmentado en dos partes: por un lado, en verde se pueden ver los primeros 82 registros del precio a estudiar, que conforman el conjunto de entrenamiento; y por otro lado, en morado se representan los 2 datos del conjunto de test y sus respectivas prediciones mediante SARIMA en azul. Notemos que la representación a priori tiene sentido, pues el conjunto de entrenamiento y el de test no se solapan.

La predicción del modelo SARIMA no es demasiado buena para el caso de Estonia y Suecia, donde se obtiene una predicción lineal por el ARIMA$(0,0,1)(0,0,0)_{12}$, que proporciona los mayores errores. En el caso de Eslovenia, la predicción es muy buena, pues está cerca de los datos de test y sigue su misma tendencia. Para España, se produce una buena predicción para el primero de los elementos del test, pero no para el segundo, cuya tendencia ascendente parece no capturarse.

A simple vista, podemos afirmar que SARIMA proporciona un buen ajuste y una predicción, en general, mejorable.

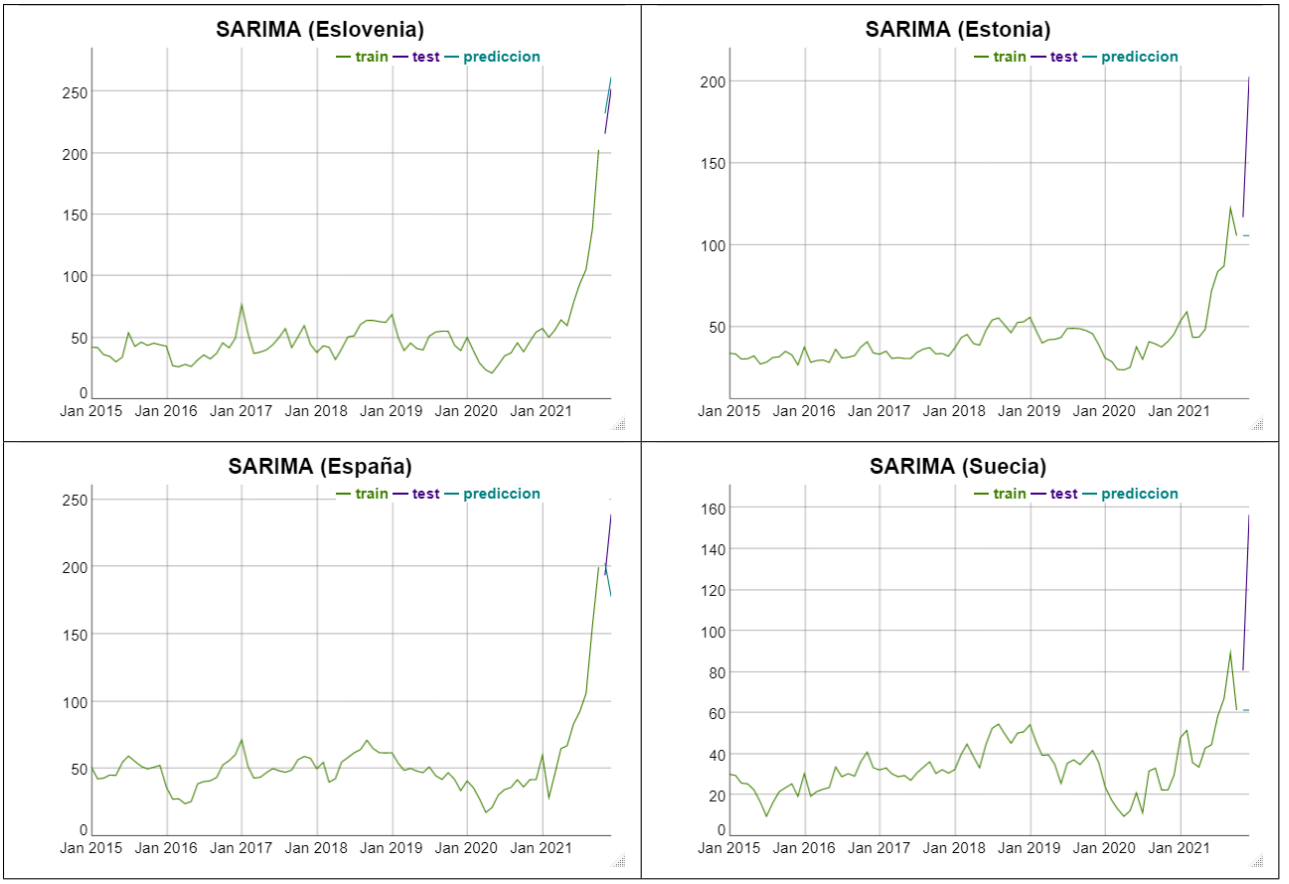


Figura 3.2: Predicción mediante SARIMA de los datos de las series de Eslovenia, Estonia, España y Suecia.

Para evaluar la bondad del modelo de forma rigurosa, se calculan las medidas de error definidas en 2.1 y 2.2.

País/Error	$RMSE$	MAE
Eslovenia	$13,44$	$12,98$
Estonia	$69,01$	$54,11$
España	$43,60$	$35,05$
Suecia	$68,74$	$57,32$

Cuadro 3.1: RMSE y MAE obtenidas de las predicciones del modelo SARIMA.

Estos estadísticos sustentan nuestras observaciones: todas las predicciones, salvo la de Eslovenia, se alejan de los valores reales en estos dos últimos meses, donde el escalado de los precios alcanza su pico máximo.

3.3. TBATS

En segundo lugar, presentamos el rendimiento de este modelo en nuestro conjunto de datos. En la figura 3.3, se puede notar el buen ajuste que proporciona el modelo TBATS a los datos de las series temporales de los representantes de los clusters obtenidos. En particular, se han podido modelizar con éxito las series de los cuatro países.

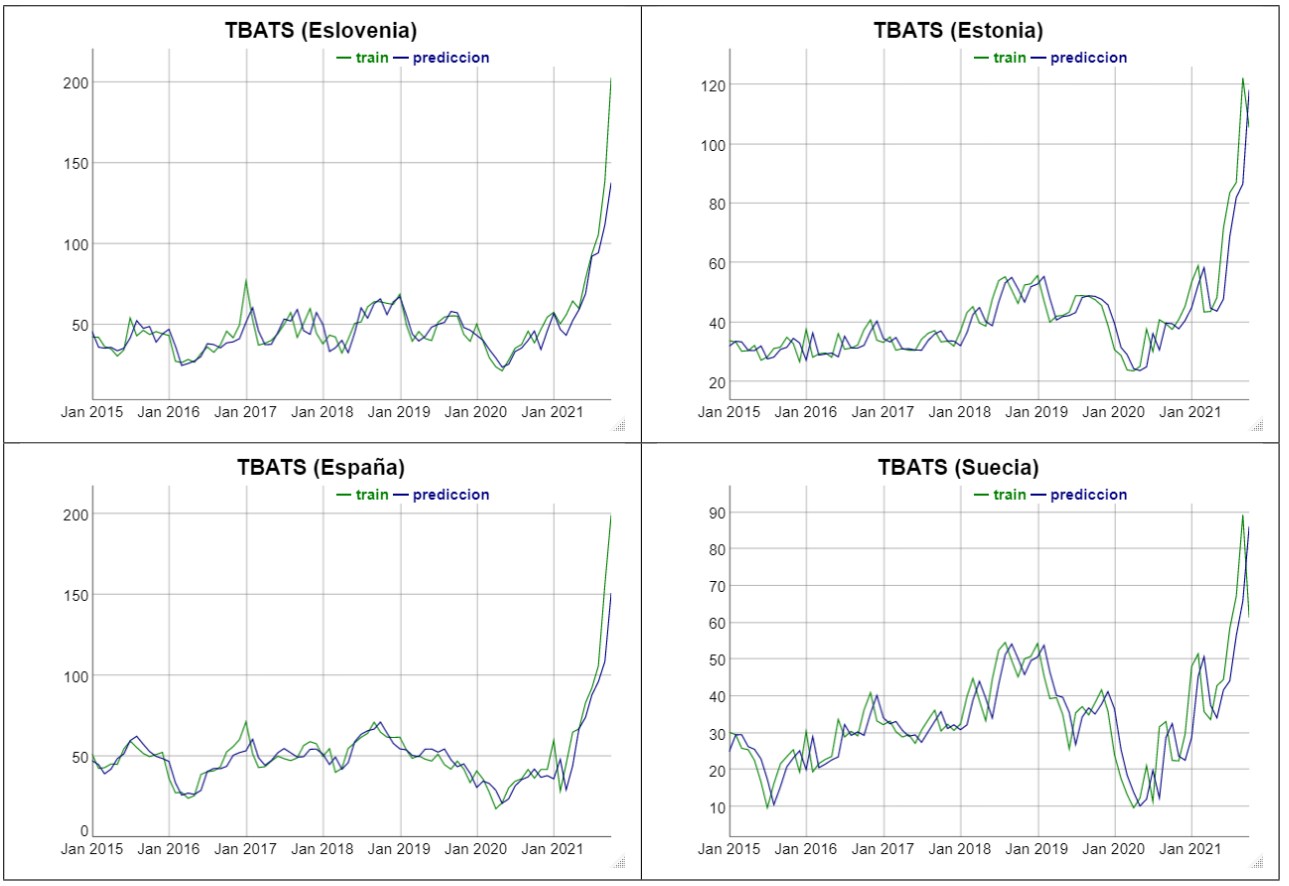


Figura 3.3: Ajuste de TBATS a los datos de las series de Eslovenia, Estonia, España y Suecia.

En la figura 3.4 se muestra el conjunto de los datos de cada país segmentado en dos partes: por un lado, en verde se pueden ver los 82 registros que conforman el conjunto de entrenamiento; y por otro lado, en morado se representan los 2 datos del conjunto de test y sus respectivas prediciones mediante TBATS en azul.

La predicción del modelo TBATS no es demasiado buena para ninguno de los cuatro países: en el caso de Estonia y Suecia, se obtiene una predicción lineal, al igual que con SARIMA. En el caso de Eslovenia, la predicción está lejos de los datos de test, pero sigue su misma línea. Para España, se produce una buena predicción para el primero de los elementos del test, pero no para el segundo, ya que no logra captar su tendencia ascendente.

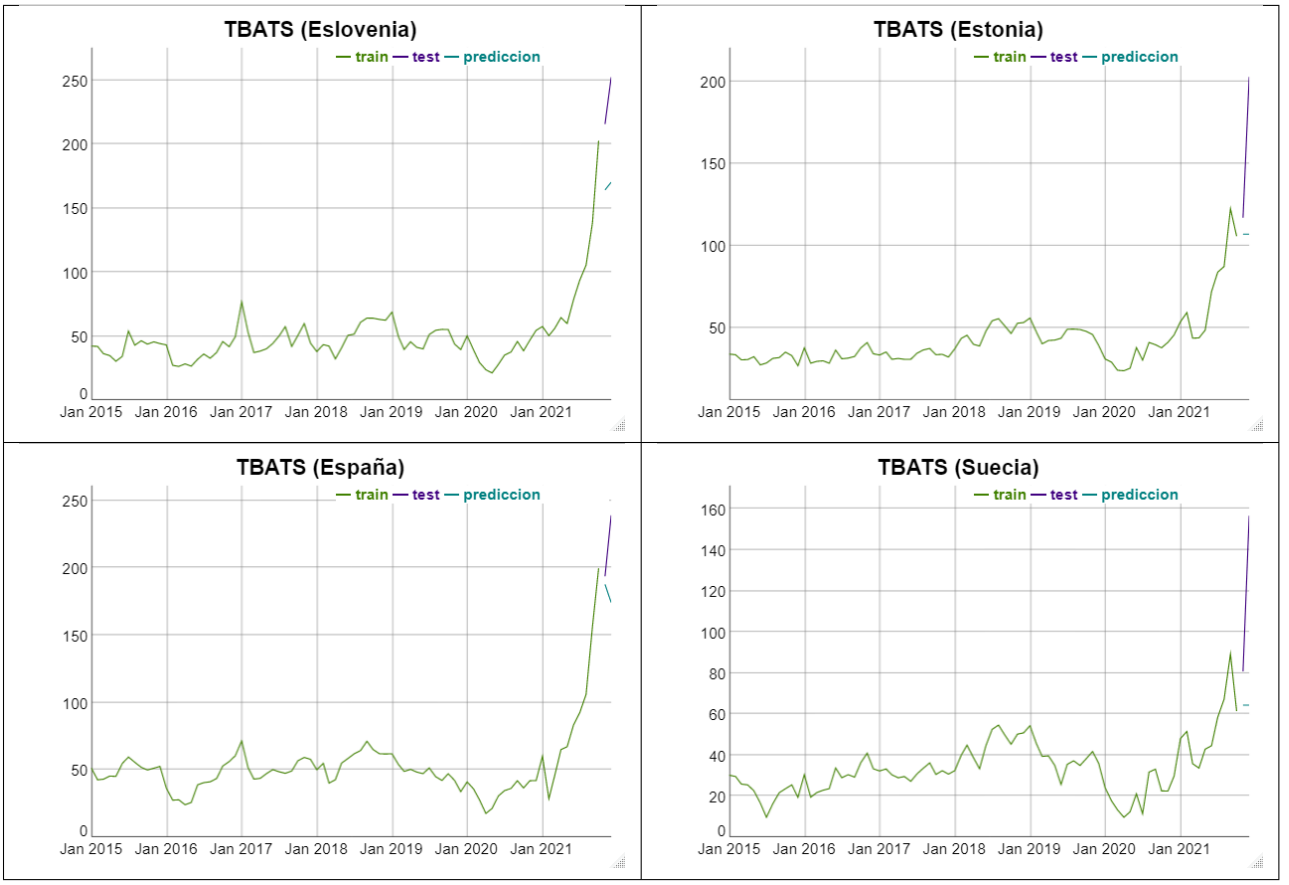


Figura 3.4: Predicción mediante TBATS de los datos de las series de Eslovenia, Estonia, España y Suecia.

Para evaluar la bondad del modelo de forma rigurosa, se calculan las medidas de error definidas en 2.1 y 2.2.

País/Error	$RMSE$	MAE
Eslovenia	$65,54$	$66,82$
Estonia	$68,08$	$52,91$
España	$46,01$	$35,40$
Suecia	$66,31$	$54,37$

Cuadro 3.2: RMSE y MAE obtenidas de las predicciones del modelo TBATS.

Estos estadísticos sugieren lo que previamente habíamos notado: el modelo TBATS no logra captar el escalado de los precios en los últimos meses, salvo en la primera predicción para España, efecto que se refleja en el cuadro 3.2.

3.4. ARNN

En tercer lugar, presentamos el rendimiento de este modelo en nuestro conjunto de datos. En la figura 3.5, se puede notar el buen ajuste que proporciona el modelo ARNN a los datos

de entrenamiento de los países representantes, correspondientes a los precios registrados desde enero de 2015 hasta octubre de 2021.

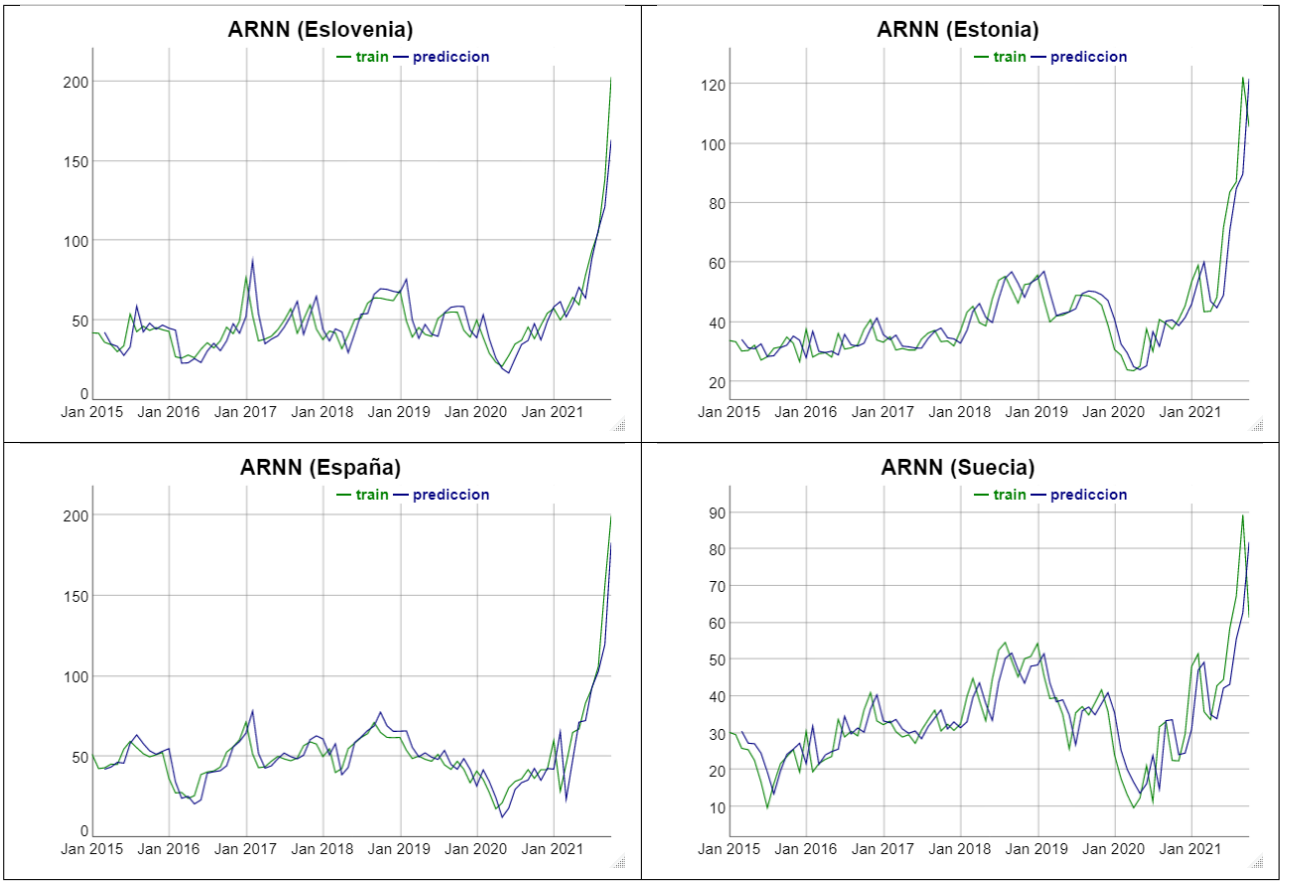


Figura 3.5: Ajuste de ARNN a los datos de las series de Eslovenia, Estonia, España y Suecia.

En la figura 3.6, se muestra la predicción de los precios para noviembre y diciembre de 2021 en cada país. Vemos que los valores arrojados se ajustan poco a los datos correspondientes, aunque el modelo ARNN proporciona una mejora visible con respecto a los estudiados anteriormente (SARIMA y TBATS). Esto es así porque ARNN logra adquirir la tendencia ascendente de los precios en el total de los casos, hecho que se refleja en las predicciones.

La raíz del Error Cuadrático Medio y el Error Absoluto Medio obtenidos en esta ocasión son los siguientes:

País/Error	$RMSE$	MAE
Eslovenia	$30,30$	$28,25$
Estonia	$63,99$	$50,13$
España	$35,09$	$31,19$
Suecia	$77,11$	$68,53$

Cuadro 3.3: RMSE y MAE obtenidas de las predicciones del modelo ARNN.

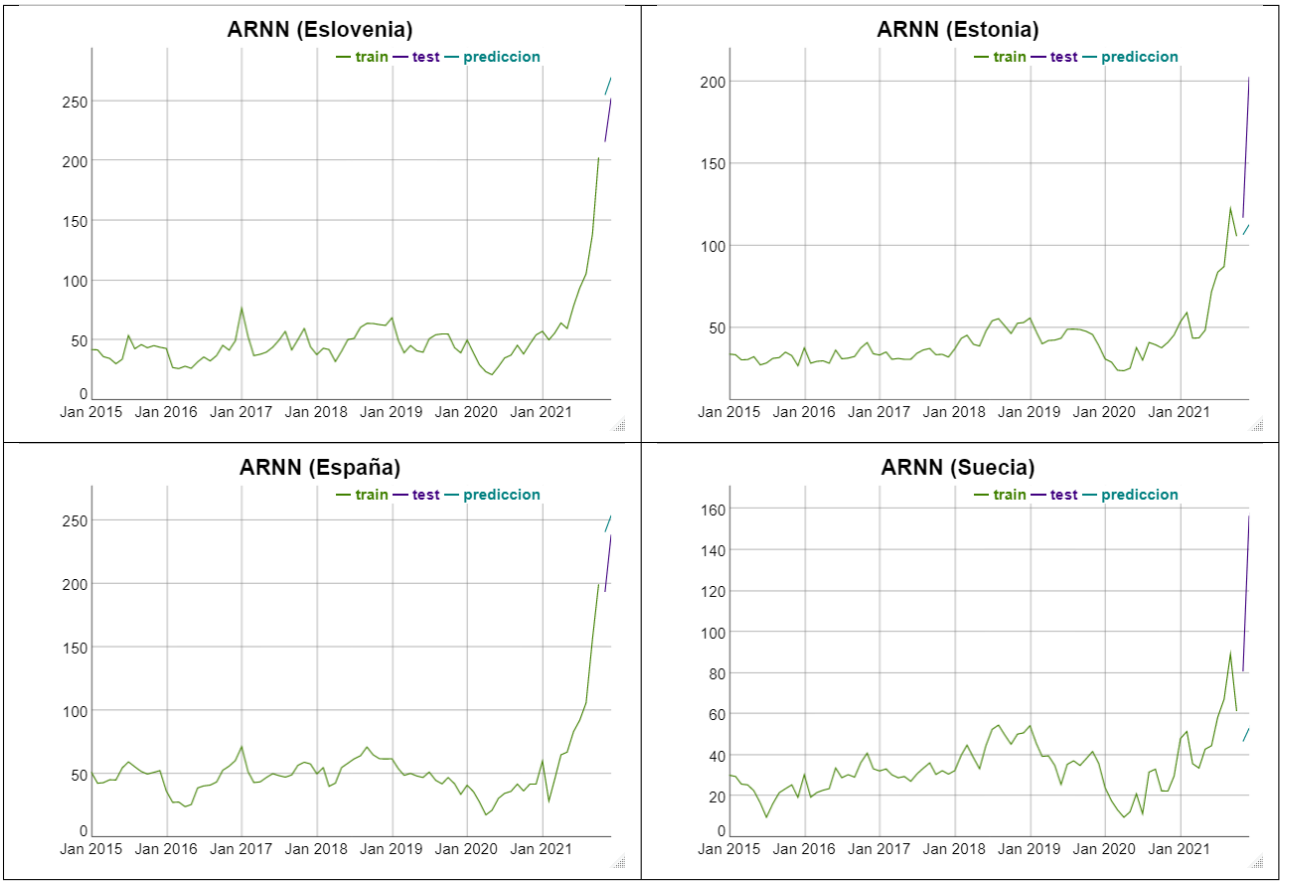


Figura 3.6: Predicción mediante ARNN de los datos de las series de Eslovenia, Estonia, España y Suecia.

3.5. K-NN

Para evitar complejidad al implementar este modelo, en vez de dividir el conjunto de los datos para el entrenamiento y la validación del modelo, se va a testar directamente la predicción del modelo a los datos que lo conforman. Esto se va a conseguir mediante la aplicación de las estrategias recursiva y MIMO, explicadas en la sección 2.3.2.

Para la predicción mediante el modelo de KNN, se fragmenta la serie temporal de entrenamiento en grupos de variables explicativas junto con una variable explicada, con los que se entrena el modelo M, atendiendo a la siguiente ecuación:

$$y_t = M(y_{t-1}, y_{t-2}, \ldots, y_{t-p}).$$

Elección de parámetros

La elección de los parámetros del modelo puede realizarse mediante el apoyo de técnicas de optimización, como RandomSearch o GridSearch. Sin embargo, en este caso se toma como número de elementos a considerar es $k = 9$, cuyo valor al cuadrado está próximo al valor total de los datos de entrenamiento. Por otro lado, se define d como la cantidad de observaciones consecutivas de la serie que conforman un elemento. Un valor adecuado para este parámetro es aquel que permita obtener vecinos significativos en la predicción. Se decide tomar $d = 2$, ya que cada bimestre, se realizan estudios exhaustivos de los registros del precio de la energía,

como se puede comprobar en el portal Eurostat de la Comisión Europea [25]. Además, la ponderación aplicada para calcular la predicción tomará pesos medios, es decir, $w_i = 1$.

Mostramos a continuación la predicción obtenida para esta elección de parámetros del modelo KNN, aplicando las estrategias recursiva y MIMO.

Estrategia recursiva

La predicción fuera de muestra para los países representantes (que se representa en la figura3.8) tan solo capta la tendencia en el caso de Estonia y Suecia. Para Eslovenia, se arroja una predicción lineal, pudiendo indicar que la estacionalidad de la serie no fue captada. Como viene sucediendo, las primera predicción de los precios en España es muy buena, aunque se desencamina después, pues no se consigue seguir la línea ascendente.

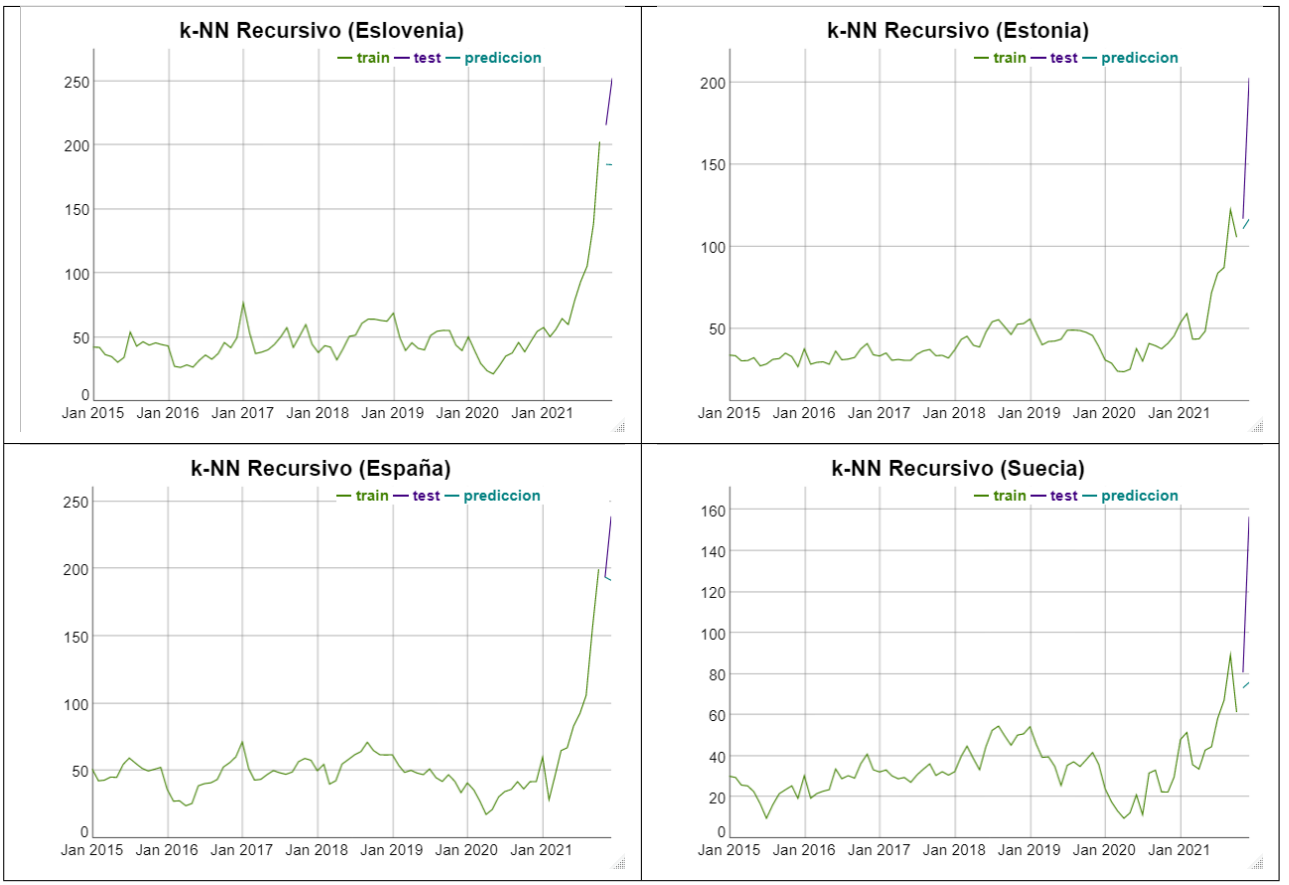


Figura 3.7: Predicción mediante KNN Recursivo de los datos de las series de los representantes.

Antes de incluir las medidas de error cometido, se muestra un ejemplo de la aplicación de la estrategia recursiva al entrenamiento del modelo. Se toman $p = 2$ retardos y, tras el entrenamiento del modelo M, se realiza la predicción obtenida en 3.7 sustituyendo los valores de los retardos. Para ello, segmentamos de la serie temporal en bloques de 2 variables explicativas y 1 variable explicada, para estimar las variables 83 y 84 a partir de sus 2 retardos correspondientes. La siguiente tabla muestra un ejemplo de ello:

Variables explicativas	Variable explicada
y_1, y_2	y_3
y_4, y_5	y_6
...	...
y_{80}, y_{81}	y_{82}
y_{81}, y_{82}	$\hat{y}_{83}$
$y_{82}, \hat{y}_{83}$	$\hat{y}_{84}$

Cuadro 3.4: Predicción de las dos últimas variables mediante estrategia recursiva.

Los errores cometidos en esta ocasión son los siguientes:

País/Error	$RMSE$	MAE
Eslovenia	48, 19	44, 84
Estonia	62, 83	47, 49
España	30, 76	22, 96
Suecia	58, 57	45, 20

Cuadro 3.5: RMSE y MAE obtenidas de las predicciones del modelo KNN recursivo.

Estrategia MIMO

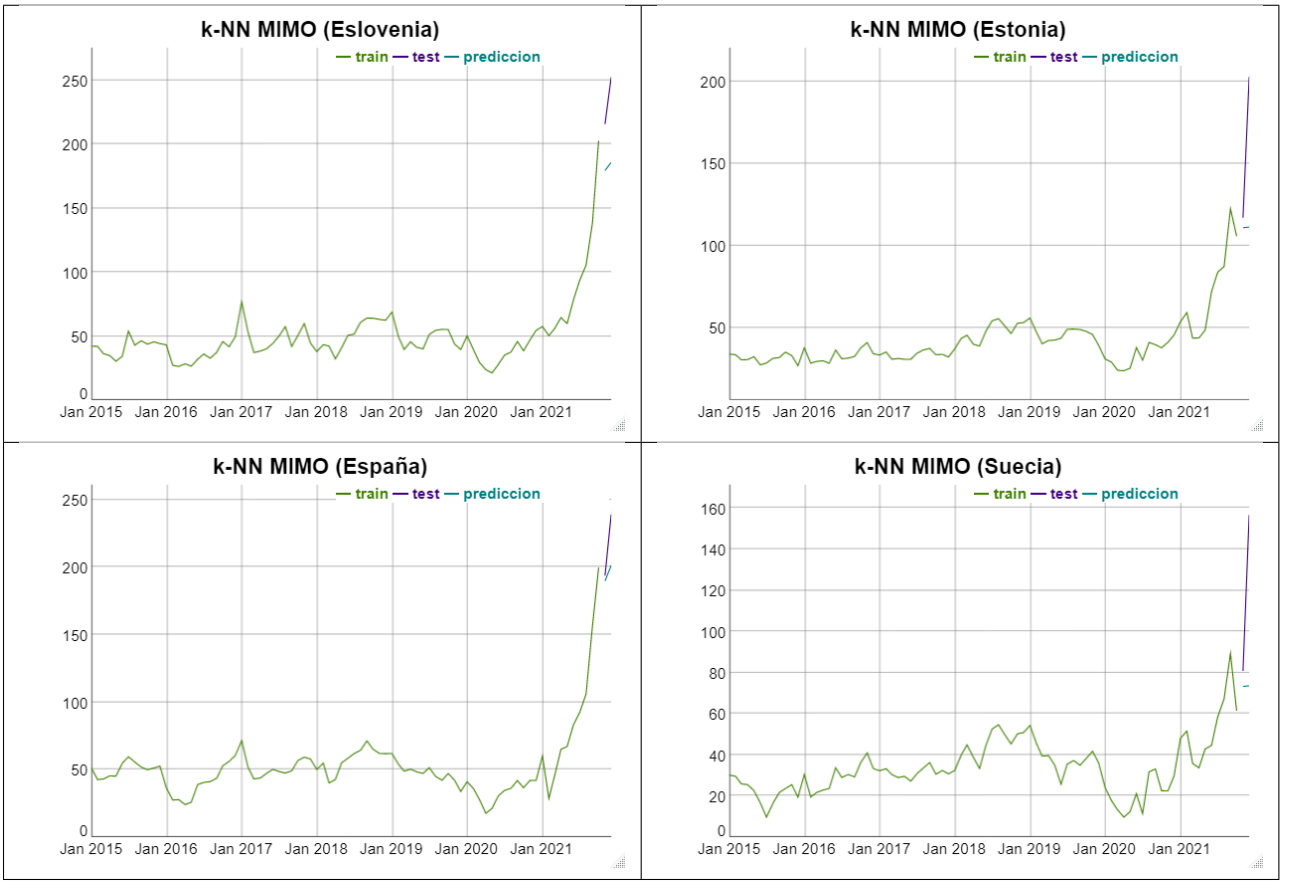


Figura 3.8: Predicción mediante KNN MIMO de los datos de las series de los representantes.

En la figura 3.8 se observa que la estrategia MIMO proporciona unos resultados ligeramente distintos a los de la estrategia recursiva. En concreto, mientras que en la estrategia recursiva se conseguía captar la tendencia de Estonia y Suecia, en esta ocasión se arroja una estimación lineal, que poco tiene que ver con los valores reales de los datos. En el caso de Eslovenia y España, ocurre el efecto contrario: mediante la recursividad, se obtenía cierta linealidad en las predicciones y ahora, se adquiere bien la tendencia, aunque con ciertos errores.

No obstante, en vista de las medidas de error, ambos modelos son equiparables, pues proporcionan estadísticos muy similares:

País/Error	$RMSE$	MAE
Eslovenia	$49,90$	$48,20$
Estonia	$65,88$	$49,65$
España	$28,42$	$22,01$
Suecia	$57,76$	$44,63$

Cuadro 3.6: RMSE y MAE obtenidas de las predicciones del modelo KNN recursivo.

3.6. SVM

Con el fin de proporcionar un mejor ajuste de este modelo, se han tomado como predictores un conjunto de variables dummies, autorregrresivas y de medias móviles, explicadas en 2.1.1. En concreto, hemos contado con:

- Una variable de retardo, correspondiente a la parte autorregresiva del modelo SARIMA, con $p = 1$.
- Una variable de medias móviles, con $l = 2$.
- Doce variables dummies mensuales (una para cada mes), para aportar la estacionalidad deseada. Estas variables se almacenan en memoria como vectores de tamaño 84 con entradas nulas, a excepción de las correspondientes al mes $i-$ésimo, que toman el valor 1.
- Tres variables dummies adicionales:
    - Una variable dummy para el cambio de tendencia en los precios, cuyas 5 últimas componentes tienen valor 1 (se considera que la escalada comienza a pronunciarse de forma excepcional a partir de agosto de 2021).
    - Dos variables dummies representando los meses de verano (con unos de junio a septiembre) y de invierno (con unos de octubre a enero), donde puede haber algún tipo de patrón.

Como se puede ver en 3.9, la predicción dentro de la muestra es perfecta, siendo el SVM el modelo que mejor ajusta los datos de entrenamiento, con gran diferencia.

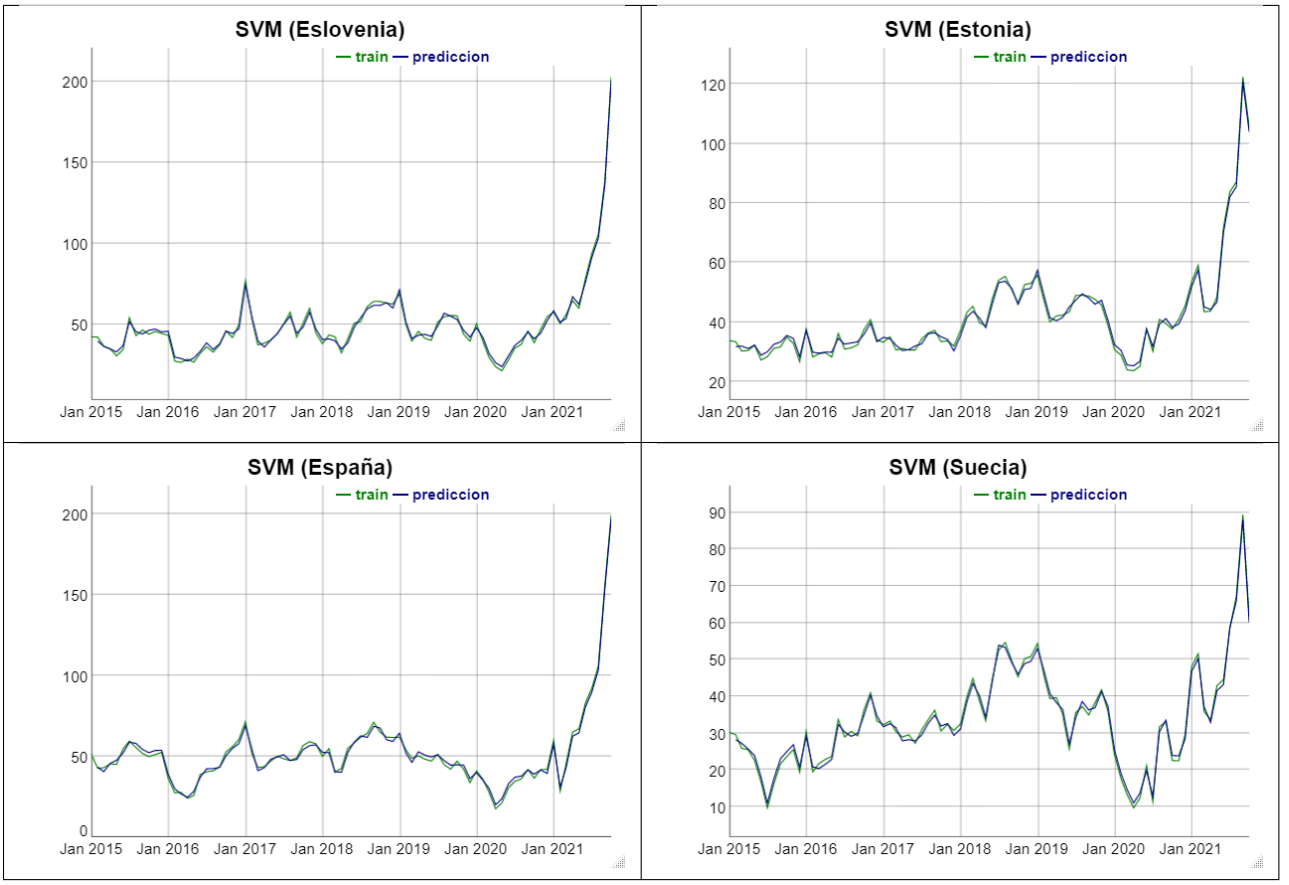


Figura 3.9: Ajuste de SVM a los datos de las series de Eslovenia, Estonia, España y Suecia.

Por último, en la figura 3.10 se muestra que no solo el ajuste del modelo en el conjunto de entrenamiento es muy exacto, sino que también lo son las predicciones para el conjunto de test de las series temporales de Eslovenia, Estonia, España y Suecia.

Vemos que los valores predichos y los reales son, a simple vista, los mismos, pues sus puntos son coincidentes. En vista de esto, no nos sorprende que los errores cometidos sean ínfimos.

País/Error	$RMSE$	MAE
Eslovenia	$3,23$	$2,61$
Estonia	$7,12$	$5,22$
España	$3,93$	$3,75$
Suecia	$5,78$	$4,96$

Cuadro 3.7: RMSE y MAE obtenidas de las predicciones del modelo SVM.

En definitiva, el modelo SVM proporciona un ajuste y una predicción buena de los datos que nos ocupan. Esto puede ser, en parte, por la introducción de las variables dummies en la regresión.

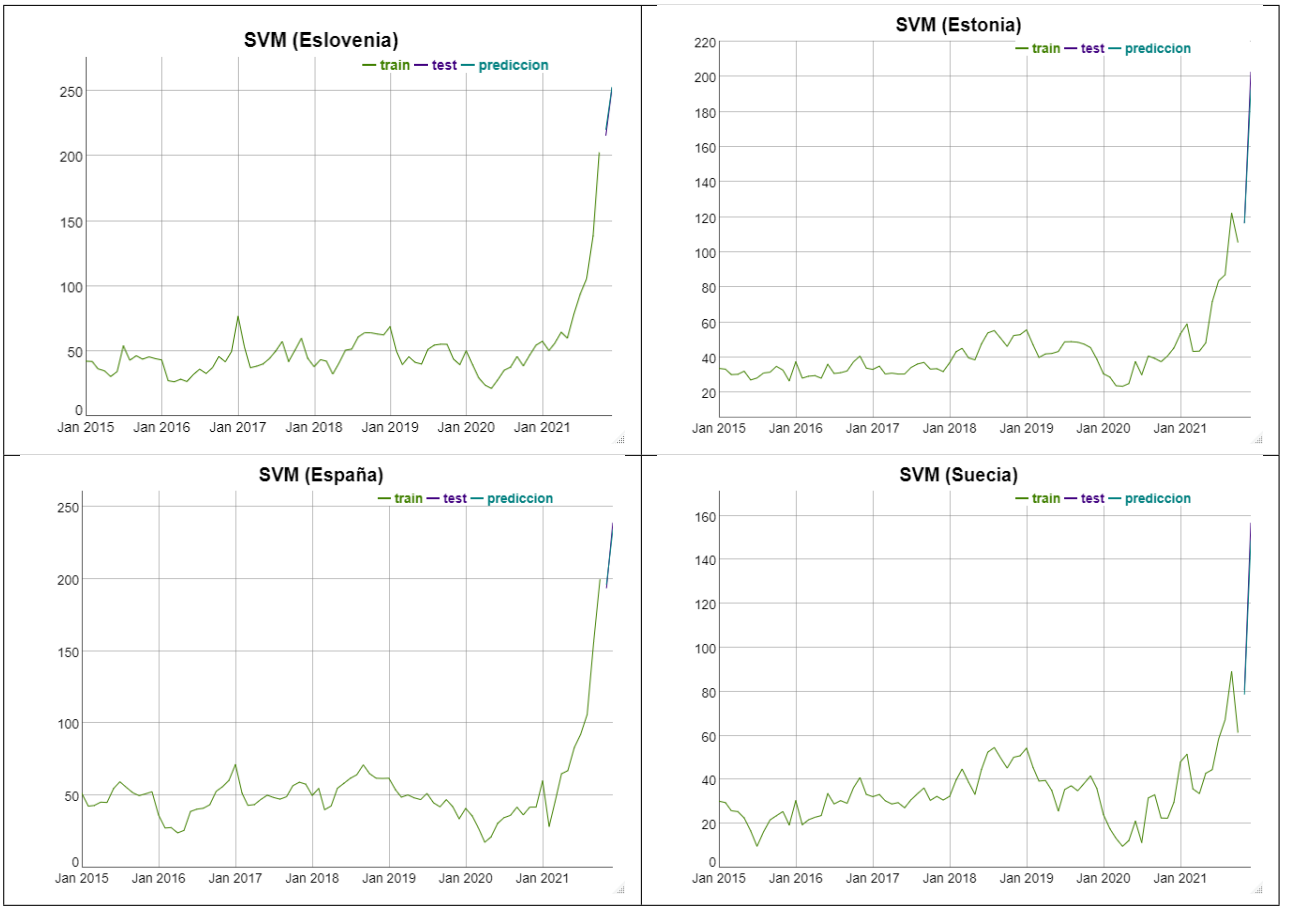


Figura 3.10: Predicción mediante SVM de los datos de las series de Eslovenia, Estonia, España y Suecia.

3.7. Método combinado

Para finalizar, y dado que ya se tienen ajustados los modelos anteriores, se presentan los resultados obtenidos mediante el modelo que los combina. Notemos que no se va a realizar un ajuste para los datos de entrenamiento, ya que en uno de los modelos que explican el modelo combinado (KNN) no lo hemos hecho.

Mostramos a continuación las predicciones obtenidas para las tres ponderaciones estudiadas: SA, BG y CLS.

Combinación SA

En la figura 3.11 podemos ver que las predicciones basadas en la Media Aritmética Simple van bien encaminadas, en comparación con los datos de test. Es notable el efecto del modelo SVM sobre el resto, pues el primer valor predicho es muy exacto. Sin embargo, quizás debido al efecto del resto de modelos, cuya ponderación -recordemos- es la misma, la segunda predicción se queda corta, en especial en el precio estimado del MWh en dicienbre de 2021 para Estonia y Suecia.

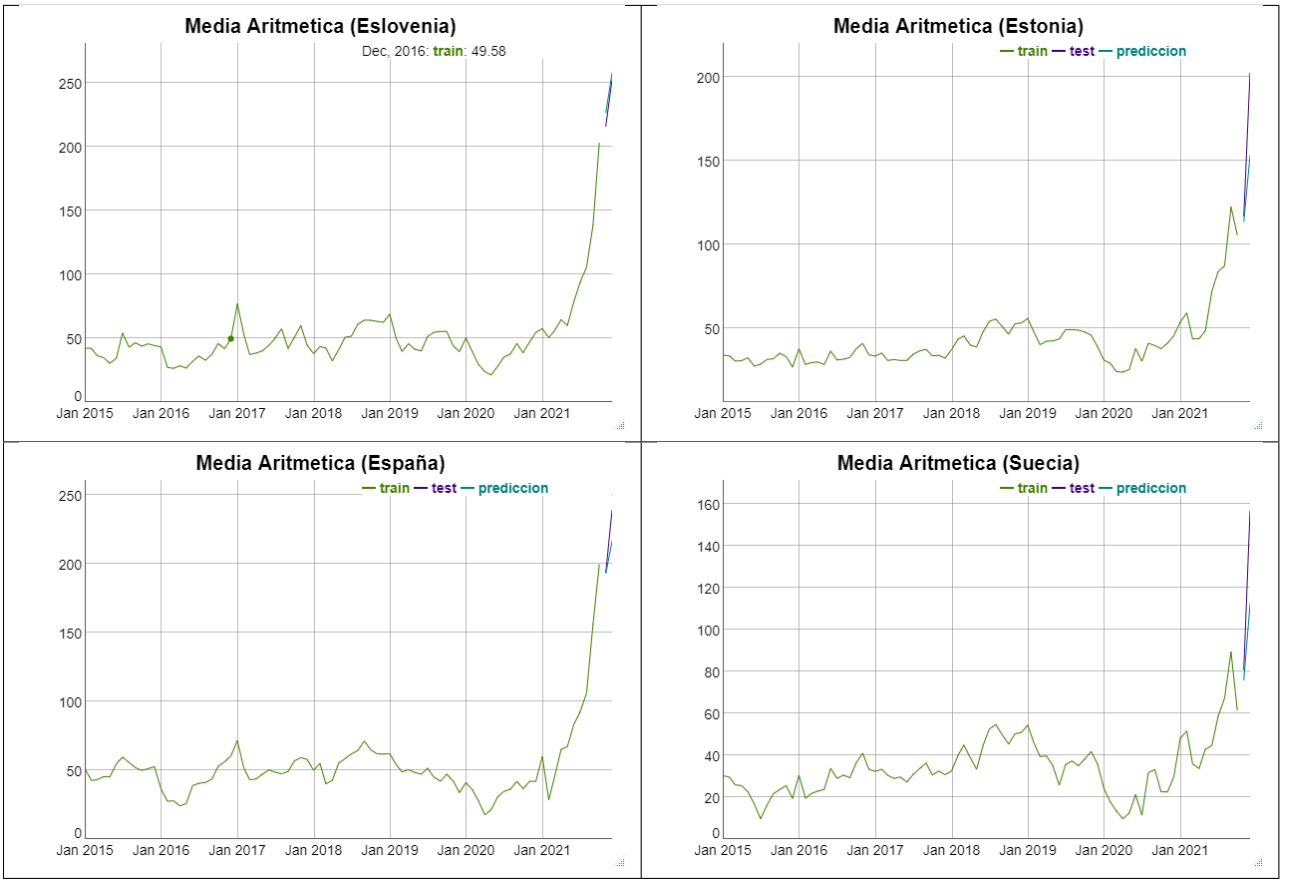


Figura 3.11: Predicción mediante Media Aritmética de los datos de las series de Eslovenia, Estonia, España y Suecia.

Este resultado queda recogido en las medidas de error, las cuales presentamos en su cuadro correspondiente.

País/Error	$RMSE$	MAE
Eslovenia	$8,25$	$7,79$
Estonia	$34,96$	$23,35$
España	$15,88$	$11,58$
Suecia	$31,74$	$24,79$

Cuadro 3.8: RMSE y MAE obtenidas de las predicciones del modelo SA.

Combinación BG

En cuanto a la predicción mediante el método de Bates-Granger, vemos una mejora sustancial con respecto al anterior, ya que parece que la elección de pesos que minimicen su Error Cuadrático Medio suaviza el efecto a evitar: que ciertos modelos que lo componen arrastren aproximaciones poco exactas, condicionando el resultado. Sus gráficas se muestran en 3.12.

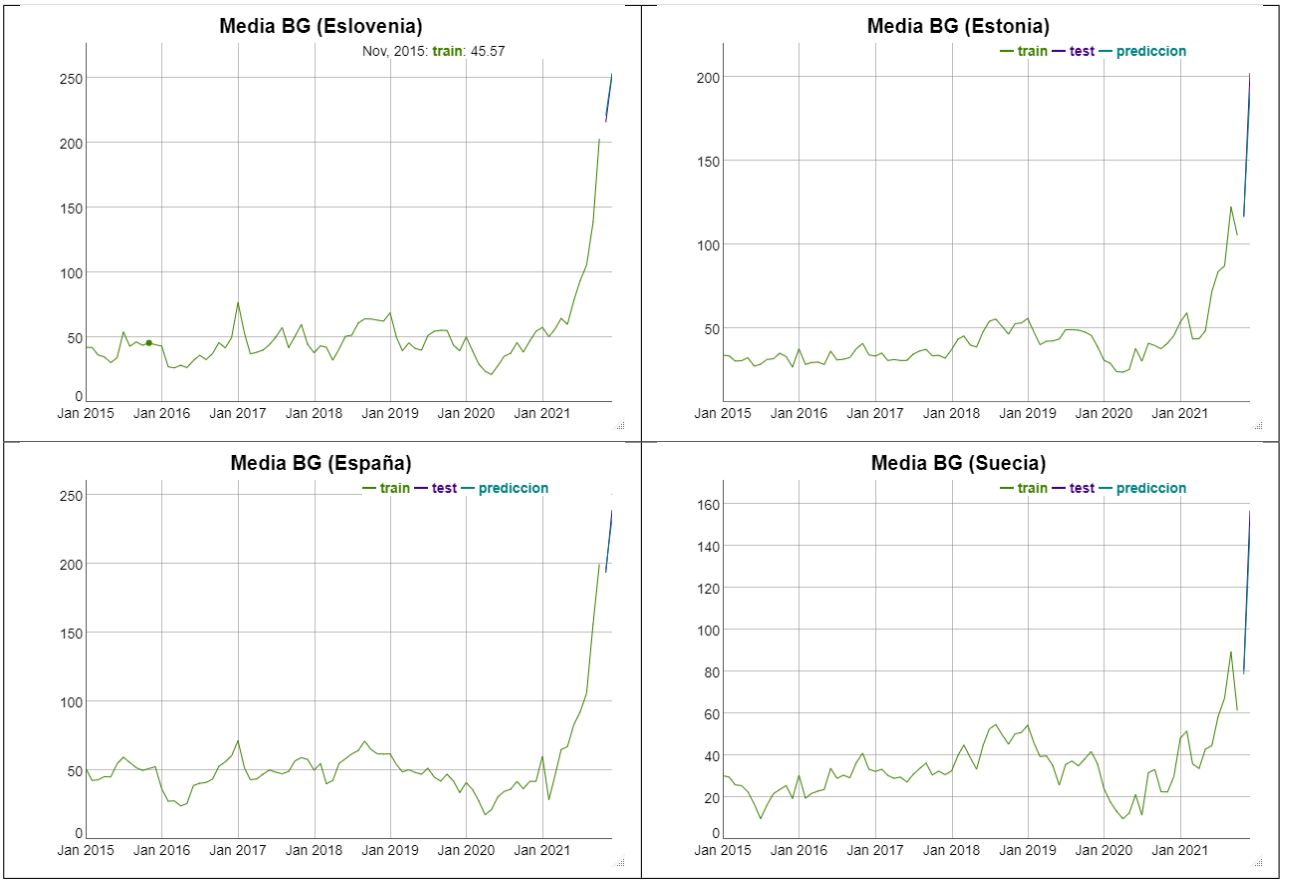


Figura 3.12: Predicción mediante Media BG de los datos de las series de Eslovenia, Estonia, España y Suecia.

Los errores cometidos en la aplicación de esta combinación de predicciones con distintos pesos son los siguientes:

País/Error	$RMSE$	MAE
Eslovenia	$3,75$	$3,17$
Estonia	$7,83$	$5,75$
España	$4,30$	$4,02$
Suecia	$6,29$	$5,36$

Cuadro 3.9: RMSE y MAE obtenidas de las predicciones del modelo BG.

Como vemos, esta técnica proporciona unos excelentes resultados, pues sus medidas de error cometidas son muy pequeñas, un poco mayores que las obtenidas con la regresión basada en SVM.

Combinación CLS

Por último, vemos el resultado de la aplicación de la combinación de modelos basada en la Restricción de Mínimos Cuadrados. La figura 3.13 muestra que los pesos escogidos proporcionan un resultado idóneo, pues las estimaciones están completamente ajustadas a los datos reales.

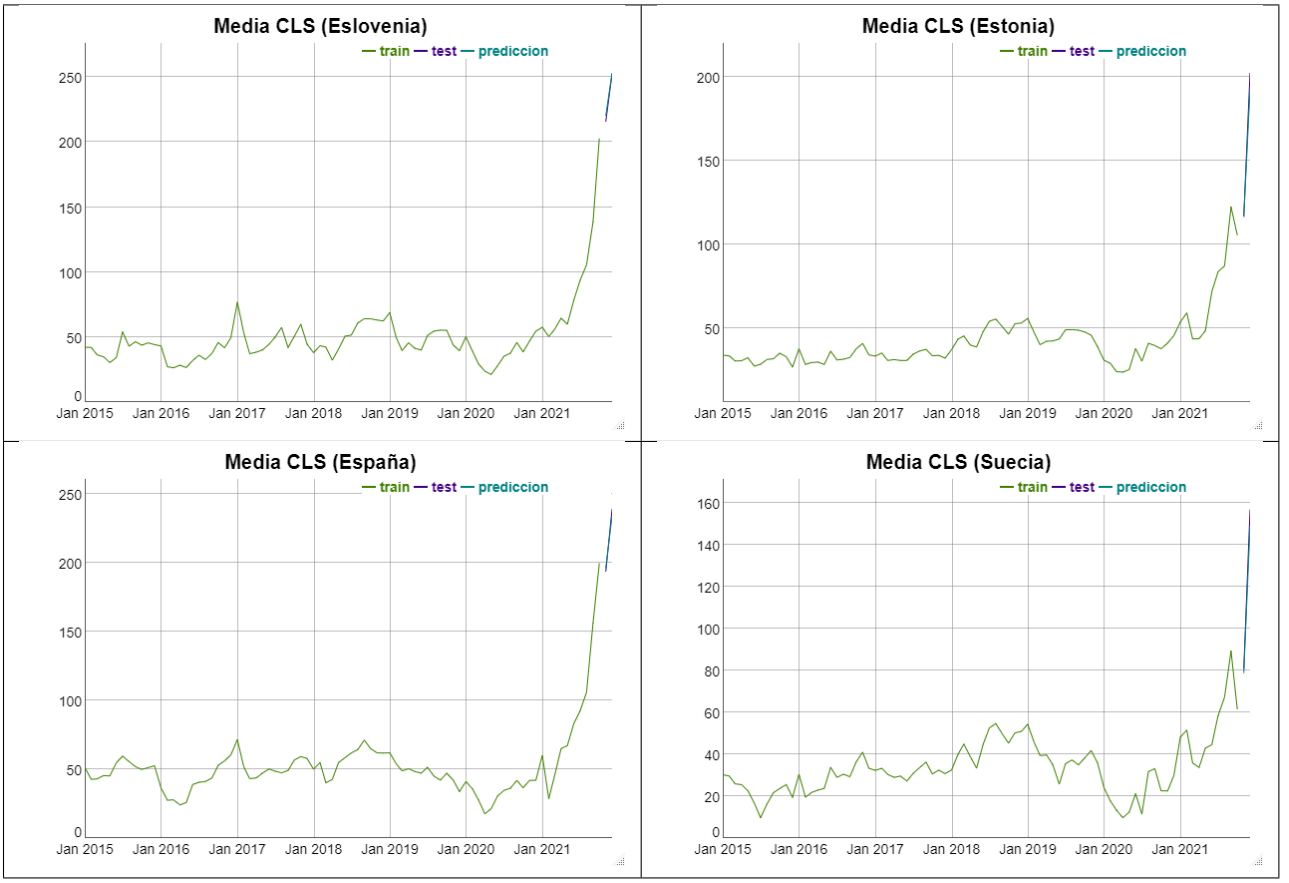


Figura 3.13: Predicción mediante Media CLS de los datos de las series de Eslovenia, Estonia, España y Suecia.

Los errores cometidos en la aplicación de esta combinación de predicciones con distintos pesos son los siguientes:

País/Error	$RMSE$	MAE
Eslovenia	$3,23$	$2,61$
Estonia	$7,12$	$5,22$
España	$3,93$	$7,54$
Suecia	$5,78$	$4,96$

Cuadro 3.10: RMSE y MAE obtenidas de las predicciones del modelo CLS.

Como vemos, esta técnica proporciona unos excelentes resultados, pues sus medidas de error cometidas son muy pequeñas, del orden de las obtenidas con la regresión basada en SVM.

3.8. Comparación de los modelos

Con el fin de realizar una comparación del total de los modelos, y así poder dar una propuesta del más indicado para la predicción del precio del MWh en los dos últimos meses de 2021, vamos a presentar una tabla con sus respectivos errores.

	Eslovenia		Estonia		España		Suecia	
Modelo	RSME	MAE	RSME	MAE	RSME	MAE	RSME	MAE
SARIMA	13,44	12,98	69,01	54,11	43,60	35,05	68,74	57,32
TBATS	65,54	66,82	68,08	52,91	46,01	35,40	66,31	54,37
ARNN	30,30	28,25	63,99	50,13	35,09	31,19	77,11	68,53
KNN recursivo	48,19	44,84	62,83	47,49	30,76	22,96	58,57	45,20
KNN MIMO	49,90	48,20	65,88	49,65	28,42	22,01	57,76	44,63
SVM	3,23	2,61	7,12	5,22	3,93	3,75	5,78	4,96
SA	8,25	7,79	34,96	23,35	15,88	11,58	31,74	24,79
BG	3,75	3,17	7,83	5,75	4,30	4,02	6,29	5,36
CLS	3,23	2,61	7,12	5,22	3,93	3,75	5,78	4,96

Cuadro 3.11: RMSE y MAE obtenidas de las predicciones de los distintos modelos

En primer lugar, nos percatamos de que el modelo CLS ha otorgado pesos nulos a todos los modelos excepto a SVM, por eso se obtienen los mismos errores en ambos métodos.

Como podemos ver, el modelo basado en SVM es el que proporciona un menor error en las predicciones de cada uno de los países, representantes de los cuatro grupos en los que hemos dividido las series temporales. Este hecho se puede comprobar ejecutando la línea *prediccion_estruct* del paquete *ForecastComb* en R, que devuelve la estructura de la combinación de predicciones que queda al final. Vemos que el modelo común a todas las combinaciones es SVM, con lo cual, es el que elegiremos como óptimo para el modelo de nuestros datos.

Capítulo 4

Conclusiones

En este apartado mostraremos, por una parte, las aportaciones realizadas en este Trabajo Fin de Máster y por otra, algunas cuestiones que quedan abiertas y ofrecen nuevas vías de estudio. Comenzaremos por una pequeña reflexión sobre la subida extraordinaria del precio de la luz en el último año, elemento condicionante en el desarrollo de este TFM.

La crisis energética detrás de la escalada de precios

En vista de los datos analizados está que, en 2021, las tarifas de electricidad de los países de la UE se dispararon. La causa principal radica en el encarecimiento del gas natural, materia prima que establece el precio marginal del mercado eléctrico. Al precio del gas se le añade la tasa de derechos de emisión de CO2, cuyo precio se multiplica a lo largo de los años. Además, en 2021 hubo, en general, escasez de viento, por lo que la subasta de energía eólica no fue suficiente para paliarlo.

El gas condiciona el precio de la electricidad: por cada euro que sube el gas, sube 2 euros la electricidad. Su valor se disparó a principios de verano, como hemos podido notar, debido a la gran demanda de China y al recorte de la oferta de los países productores, como Rusia, a consecuencia de razones geopolíticas. A todo esto, se tiene que sumar la suspensión del gasoducto Nord Stream 2 por parte de Alemania. Este problema no solo parece no terminar, sino que puede verse agravado en los últimos meses debido al conflicto Ruso-Ucraniano y al cierre del gasoducto entre Argelia y Marruecos, hecho que condicionaría el precio futuro de la luz en España.

La justificación de la situación anómala estudiada es clara: la escalada del precio se debe a una serie de hechos ocurridos de forma simultánea en los últimos meses. Sin embargo, el aumento de la demanda y del precio del gas crean un contexto grave de dependencia de la UE, que importa prácticamente la totalidad del gas natural que consume (la mitad, de Rusia). Más allá de los posibles consensos entre los distintos países que se puedan proponer por parte de los Organismos Europeos para paliar este efecto, es bien sabido que la solución pasa por potenciar la energía renovable.

En este trabajo, hemos analizado datos eléctricos y hemos propuesto una serie de modelos para explicar el precio de la luz de forma matemática. Sin embargo, hemos comprobado que, a pesar del buen ajuste de los modelos, en muchos casos los datos reales superaban a

las predicciones. Este hecho solo puede ser combatido mediante la reducción de fuentes de combustibles fósiles y, sobre todo, el auge de las renovables.

Sirva este TFM para poner sobre la mesa la grave volatilidad que tiene en el contexto actual el mercado eléctrico, que en esencia, viene determinado por el del gas. No olvidemos que esto repercute no solo a la modelización de su precio -como hemos intentado hacer, a pesar de contar con datos poco comunes-, o a los bolsillos de los ciudadanos, sino que es un problema para la sociedad en general.

Aportaciones

En cuanto a las aportaciones de este TFM, destacamos las siguientes:

- Hemos trabajado con referencias que presentan varias opciones para abordar el tema de la clusterización de series temporales, aunque hemos querido seguir un esquema claro: metodología- definición de distancia-clustering-elección de representantes. En este caso, nos hemos decantado por el estudio del clustering jerárquico aglomerativo, todo ello desde un enfoque analítico. Destacamos en este sentido las definiciones dadas para cada elemento que íbamos presentando, como es el caso de las medidas de proximidad, donde hemos querido hacer incapié en los conceptos de distancia y similaridad. En definitiva, hemos presentado matemáticamente la adecuación de este tipo de estructuras para los aspectos de la matemática más aplicada, como es el análisis de series temporales.
- Una vez obtenidos los representantes de las series mediante las técnicas previas, hemos pasado a estudiar distintos tipos de modelos para predecir series temporales. Hemos contado tanto con modelos mas tradicionales, basados en estadística, como son los SARIMA, como con modelos basados en técnicas de Aprendizaje Automático, como SVM o KNN.
- En el último capítulo, que es el más práctico, hemos presentado los resultados obtenidos al aplicar los distintos modelos al conjunto de nuestros datos. Hemos hecho uso de un lenguaje ampliamente usado, con el que también hemos trabajado durante el Máster, como es R. Además hemos explorado paquetes y funciones de R que nos han permitido crear un entorno favorable para conseguir los propósitos establecidos en este TFM.
- En general, todo lo estudiado es fundamental en problemas de Business Analytics. El análisis de los componentes de un mercado es una de las herramientas más potentes para preveer las tendencias en el mismo, lo cual supone una enorme optimización en la toma de decisiones empresariales.

Líneas futuras

Llegados a este punto, se abre un amplio abanico de posibilidades para seguir profundizando en estos temas:

- Una vía abierta es la del estudio de conceptos básicos de lógica difusa para analizar las redes neurodifusas, que han mostrado recientemente unos muy buenos resultados

para problemas de predicción. Sería interesante realizar por una comparativa entre el resultado de los métodos de tradicionales y los *fuzzy*, lo cual podría conformar un camino más vanguardista hacia la mejora de las técnicas de predicción.

- Otra posibilidad sería el estudio de los denominados *structural breaks*, pues se ha demostrado que su efecto limita la predicción de series temporales. Su presencia en nuestros datos podría justificar por qué algunas predicciones que hemos hecho se alejan de la realidad. `https://www.aptech.com/structural-breaks/`
- A modo de mejora: hemos estudiado el precio de la luz, que ha sufrido un incremento inusual desde el verano pasado. Hemos supuesto que esto guarda una relación directa con la crisis energética Europea, lo cual hemos justificado mostrando que casi todas las predicciones se alejan de los valores reales en esos últimos meses. Otra alternativa podía haber sido el análisis de los datos anteriores a la subida, para tener un mejor modelo, predecir valores post-subida con ese modelo, que se dan entre enero y abril de 2022 y comparar ambas series. Lamentablemente, cuando empezamos este TFM no contábamos con esos últimos datos, con lo cual no pudimos dar forma a esta alternativa.

Nuestro verdadero trabajo ha sido el de conseguir una pequeña transición entre el enfoque más clásico del problema de clasificación y regresión matemática, a uno más interdisciplinar, abierto y necesario en la era digital en la que nos encontramos, como es el análisis de series temporales en el mercado eléctrico.

Bibliografía

[1] A. Alonso, P. d'Urso, C. Gamboa, V. Guerrero. *Cophenetic-based fuzzy clustering of time series by linear dependency*. International Journal of Approximate Reasoning. Vol. 137, pp. 114-136, 2021.

[2] A. Alonso, P. Galeano, D. Peña. *A robust procedure to build dynamic factor models with cluster structure*. Journal of Econometrics, Vol. 216, no. 1, pp. 35-52, 2020.

[3] A. Alonso, E. Maharaj. *Comparison of time series using subsampling*. Computational Statistics and Data Analysis, Vol. 50, no. 10, pp. 2589-2599, 2006.

[4] A. Alonso, D. Peña. *Clustering time series by linear dependency*. Statistics and Computing, Vol. 29, no. 4, pp. 655–676, 2019.

[5] C. Arias. *Análisis avanzado y predicción de series temporales aplicados en un caso de Business Analytics*. Universidad Rey Juan Carlos, 2021.

[6] Base de datos energéticos de *Ember*. `https://ember-climate.org/data/`

[7] D. Bastarrica. *Predicción de series temporales mediante el método k-NN: explicabilidad y algoritmos de ensamblado* . Facultad de Informática. Universidad Complutense de Madrid, 2020.

[8] S. Ben Taieb, G. Bontempi, A. F. Atiya y A. Sorjamaa. *A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition*. Expert Systems with Applications, Vol. 39, nº 8, pp. 7067-7083, 2012.

[9] *BOE-A-1997-25340 Ley 54/1997, de 27 de noviembre, del Sector Eléctrico*.

[10] G. Camps i Valls. *Predicción de series temmporales*. Departamento de Ingeniería Electrónica. Universidad de Valencia, 2003. `https://isp.uv.es/courses/manuals/04_PST_secure.pdf`

[11] A. De Livera, R. Hyndman, R. Snyder. *Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing*. Journal of the American Statistical Association Vol. 106, pp. 1513-1527, 2010.

[12] A. Fierro. *Predicción de Series Temporales con Redes Neuronales*. Facultad de Informática. Universidad Nacional de la Plata, 2020. `http://sedici.unlp.edu.ar/bitstream/handle/10915/114857/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y`

[13] R. Hyndman, G. Athanasopoulos. *Forecasting principles and practice*. Monash, OTexts, 2018. https://otexts.com/fpp2/

[14] F. Martínez. *Análisis de las series temporales de los precios del mercado eléctrico mediante técnicas de clustering*. Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, 2007. http://www.lsi.us.es/docs/doctorado/memorias/Martinez,%20Francisco.pdf

[15] P. Montero, J. Villar. *TSclust: An R Package for Time Series Clustering*. Journal of Statistical Software, Vol. 61, no. 1, pp. 1-43, 2014.

[16] R. Nau. *Statistical forecasting: notes on regression and time series analysis*. Fuqua School of Business, Duke University. https://people.duke.edu/~rnau/411avg.htm

[17] D. Peña. *Análisis de las series temporales*. Madrid, Alianza Editorial, 2010.

[18] G. Santamaría. *Interpretación de teoremas que sustentan técnicas de aprendizaje automático*. Universidad de la Rioja, 2020. https://investigacion.unirioja.es/documentos/5fbf7e48299952682503c305

[19] Sitio web de *Search RProject*. Instituto de estadística y matemáticas de la Universidad de Viena. https://search.r-project.org/CRAN/refmans/TSclust/html/diss.COR.html

[20] Sitio web de *R-Bloggers*. Cluster Analysis in R. https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/

[21] Sitio web de *R-Bloggers*. Extended Agglomerative Hierarchical Clustering. https://search.r-project.org/CRAN/refmans/mdendro/html/linkage.html

[22] Sitio web de *Data Side of Life*. LSTM time series prediction in R. http://datasideoflife.com/?p=1171

[23] Sitio web de *A. Gallardo*. Universidad de Granada. www.ugr.es/~gallardo/pdf/cluster-g.pdf

[24] Sitio web de *Aptech*. Structural Breaks. https://www.aptech.com/structural-breaks/

[25] Sitio web de *Eurostat*. Energy. https://ec.europa.eu/eurostat/web/energy

[26] Sitio web de *MQL5*. Predicción de series temporales usando el ajuste exponencial. https://www.mql5.com/es/articles/318

[27] N. Timm. *Applied multivariate analysis*. New York, Springer, 2002.

[28] J. Velásquez, C. Zambrano, L. Velez. *ARNN: un paquete para la predicción de series de tiempo usando redes neuronales autorregresivas*. Avances en Sistemas e Informática, Vol. 8. pp. 177-182, 2011.

[29] J. Velásquez, Y. Olaya, C. Franco. *Predicción de series temporales usando máquinas de vectores de soporte*. Ingeniare. Revista chilena de ingeniería, Vol. 18, no. 1, pp. 64-75, 2010.

[30] C. Weiss, E. Raviv, G. Roetzer. *Forecast Combinations in R using the ForecastComb Package*. The R Journal, Vol. 10, no. 2, pp. 262-281, 2018. `https://journal.r-project.org/archive/2018/RJ-2018-052/RJ-2018-052.pdf`