BIG DATA

PREDICCIÓN Y DECISIONES ECONÓMICAS CON *BIG DATA*

Daniel Peña
Pilar Poncela
Eva Senra
(editores)

funcas

BIG DATA

PREDICCIÓN Y DECISIONES ECONÓMICAS CON *BIG DATA*

Daniel Peña
Pilar Poncela
Eva Senra
(editores)

funcas

Impreso en España
Edita: Funcas
Caballero de Gracia, 28, 28013 - Madrid

ISBN impreso: 978-84-17609-79-5

ISBN digital: 978-84-17609-80-1
Depósito legal: M-13795-2024
Maquetación: Funcas
Imprime: Cecabank

Contenido

Presentación

Los nuevos datos masivos han abierto la posibilidad de incrementar nuestro conocimiento en todos los campos y, en particular, mejorar las decisiones económicas y empresariales. Para ello, un componente imprescindible es ampliar nuestra capacidad de predicción. Con este objetivo, este libro analiza varias contribuciones que se presentaron en las jornadas del mismo nombre celebradas en Funcas el 7 de noviembre de 2023. Las jornadas fueron organizadas por los tres editores de esta monografía y tuvieron una amplia participación presencial y *online*. El lector interesado puede visualizar las presentaciones orales de los trabajos incluidos en esta monografía en el canal de Funcas en YouTube: https://www.youtube.com/watch?v=5EW1vVhdga0youtube

Los rápidos cambios económicos y sociales de las últimas décadas han generado una creciente incertidumbre respecto al futuro. Tenemos ejemplos muy recientes, como la epidemia del COVID-19, que hundió las economías de todo el mundo, o las guerras en Ucrania y Gaza, que están teniendo importantes repercusiones económicas. Estos cambios han generado la necesidad de incluir más información en los modelos de predicción, así como de desarrollar procedimientos resistentes a posibles no linealidades, como cambios estructurales. Las predicciones racionales se basan siempre en extrapolar de manera inteligente el pasado hacia el futuro, pero la evolución de la economía depende mucho de factores sociales y políticos que resultan muy difíciles de prever y, además, pueden tener un gran impacto en los resultados económicos. Resulta importante, por tanto, incorporar estas nuevas variables a los modelos de predicción.

El primer bloque de trabajos de esta monografía incluye los capítulos I, II y III, que presentan contribuciones donde se analizan nuevos tipos de datos que pueden mejorar la predicción. Entre ellas están las obtenidas con el análisis del lenguaje natural o de sentimientos reflejando cuestiones como la percepción subjetiva de la conyuntura económica o la importancia del cambio climático y que pueden tener impacto en las decisiones económicas. También se analiza el importante problema de cómo el análisis con variables desagregadas complementa el análisis global sobre el conjunto de variables originales.

El capítulo I está escrito por **Álvaro Ortiz** y **Tomasa Rodrigo** y estudia la incorporación de información obtenida con el lenguaje escrito para mejorar las decisiones empresariales. Los autores muestran cómo gracias al desarrollo de los modelos de lenguaje natural, los

textos digitales se están convirtiendo en una fuente muy importante de información para guiar estrategias económicas. Entre los modelos de generación de lenguaje, ChatGPT, que se ha convertido en un referente en los métodos de inteligencia artificial aplicados al análisis automático de datos no estructurados, es un ejemplo destacado.

En el capítulo II, **María Alló** y **Maria Loureiro** analizan la evolución del sentimiento social hacia el cambio climático en España y su posible efecto sobre las decisiones económicas. Aunque es generalmente aceptado que el cambio climático representa un desafío crucial para la sociedad actual, la percepción de su importancia ha sido poco estudiada. En este trabajo se utiliza información obtenida de redes sociales para medir la percepción sobre su evolución temporal, incluyendo técnicas de procesamiento de lenguaje natural. Los resultados analizan también el efecto de cambios estructurales, como la pandemia debida al COVID-19.

Otra de las formas de incorporar información adicional es utilizar datos más desagregados de los habituales en los modelos económicos. **Antonio Montañés** estudia, en el capítulo III, las diferencias en la evolución regional de los precios en España utilizando datos desagregados tanto en el espacio, por provincias, como en el tiempo, por periodos. Se realiza un estudio para contrastar si existe un patrón de comportamiento único, o varios diferentes en el espacio y en el tiempo, detectando la presencia de grupos o clústeres de series temporales. Tener en cuenta estas variables es importante para obtener predicciones más detalladas y precisas.

El segundo bloque se centra en el análisis de datos masivos. Por un lado, en el capítulo IV, **Josep Mestres Domènech** presenta el portal *Economía en Tiempo Real*, de CaixaBank Research. A partir de información de CaixaBank, como los datos de tarjetas de gasto y reintegro de tarjetas de crédito, ingresos (nóminas y subsidios de paro) o los pagos efectuados en España en sus TPV con las tarjetas emitidas por entidades extranjeras, se describen nuevos indicadores de consumo, acceso a la vivienda, salarios, turismo o desigualdad.

En los capítulos V y VI, el enfoque se centra en la utilización de técnicas capaces de aprovechar las oportunidades de los datos masivos. Por un lado, en el capítulo V, **Isabel Casas, Juan Miguel Marín** y **Helena Veiga** exploran la eficacia de distintos modelos en la predicción de la varianza diaria realizada a partir de datos intradiarios de Bitcoin, NASDAQ y S&P500. Este capítulo muestra cómo la disponibilidad de datos, a frecuencias muy altas, constituye una gran ventaja, posibilitando la estimación de la volatilidad de forma consistente.

Por otro lado, en el capítulo VI, **María Ángeles Carnero, Ángel León** y **Trino Ñíguez** abordan la capacidad del acceso y análisis de datos en tiempo real para predecir la rentabilidad de una cartera de inversión y cómo usar dichas predicciones para seleccionar los activos y su ponderación dentro de la cartera de inversión. El análisis se realiza sobre el índice Russell 1000, que refleja de forma continua las 1.000 empresas más grandes en el mercado de valores estadounidense.

El último bloque está comprendido por los capítulos VII y VIII y está dedicado al mercado eléctrico, de gran relevancia económica y social y cuyo análisis se está sofisticando debido a la actual abundancia de datos. En ambos capítulos se analizan, desde distintos puntos de vista, las curvas de oferta de los productores eléctricos.

En el capítulo VII, **Jorge Arias Martí** y **Andrés M. Alonso Fernández** estudian las ofertas realizadas por los productores de electricidad del mercado español. El objetivo es agrupar los conjuntos de oferta mediante técnicas de clasificación no supervisada utilizando la distancia de Hausdorff para realizar los grupos. Una vez obtenidos los distintos grupos, se relacionan con variables de producción energética, además de las habituales variables temporales, como hora, día de la semana y mes.

En el último capítulo de la monografía, **Antonio Muñoz, José Portela, Eugenio Sánchez-Úbeda** y **Guillermo Mestre** introducen modelos de series temporales funcionales inspirados en la metodología Box-Jenkins para modelar y predecir las curvas de oferta del mercado eléctrico. Esta propuesta permite introducir un enfoque probabilístico en las estrategias de oferta de los agentes del mercado. La caracterización obtenida con estas técnicas de las curvas de oferta permite captar su dinámica, que incluye varias estacionalidades en el mismo modelo. La propuesta se aplica al mercado eléctrico diario italiano.

Queremos agradecer a todos los autores su contribución a este trabajo y al director general de Funcas, Carlos Ocaña, todo su apoyo a las actividades de *big data* y, en particular, a la realización de las jornadas que han dado lugar a esta monografía. Por último, agradecemos a Myriam González su eficiencia en su publicación.

Daniel Peña, Pilar Poncela y **Eva Senra**

Marzo 2024

CAPÍTULO I

Economía, mercados y geopolítica: el papel de los modelos de lenguaje natural en las ciencias sociales

Alvaro Ortiz
Tomasa Rodrigo

La digitalización de la información y el desarrollo de la inteligencia artificial están propiciando un cambio sin precedentes en la disponibilidad de nuevos datos. Gran parte de la información a la que hoy en día podemos acceder se produce de manera no estructurada. Gracias al desarrollo de los modelos de lenguaje natural, el texto se ha convertido en una de las principales fuentes de información y la posibilidad de trasladar "texto a números" se está convirtiendo en una poderosa herramienta de análisis en las ciencias sociales. El último exponente son los modelos generativos del lenguaje (*LLM,* por sus siglas en inglés). En este capítulo exploramos a través de varios ejemplos el papel que estos desarrollos pueden jugar dentro del análisis de las ciencias sociales con un especial foco en el ámbito económico.

Palabras clave: procesamiento de lenguaje natural, análisis de sentimento, *big data,* economía, geopolítica, mercados.

1. INTRODUCCIÓN

En los últimos años hemos sido testigos del rápido crecimiento y desarrollo de la inteligencia artificial (IA), que ha pasado a formar parte de nuestro día a día como, por ejemplo, a través de asistentes de voz, recomendadores de compras virtuales o plataformas digitales de entretenimiento. Este crecimiento no es un fenómeno aislado, sino que es el resultado del desarrollo y avances de distintos factores coincidentes en el tiempo.

En primer lugar, el principal factor es el aumento exponencial de la información con la digitalización y la capacidad de almacenar todos estos datos en la nube. Otro factor relevante es el crecimiento sustancial en el poder computacional y las capacidades de procesamiento de la información, que durante las últimas décadas ha avanzado de forma exponencial, permitiendo procesar cada vez una mayor cantidad de información en menos tiempo. Esto, además, ha venido acompañado de una reducción de los costes computacionales y de procesamiento de los datos, que hacen posible, y escalable, el uso de los mismos. Estos avances han impulsado a su vez el rápido desarrollo de algoritmos, metodologías y nuevos modelos adaptados a la naturaleza de los datos en la nube, que permiten sacar el máximo provecho a la alta frecuencia y granularidad de la información.

En la actualidad, la mayor parte de nuestra actividad diaria está digitalizada a partir de nuestra interacción con nuestro móvil, el ordenador, la tarjeta bancaria o nuestras redes sociales, entre otros muchos ejemplos. Mucha de esta información está no estructurada en forma de texto, vídeos, imágenes o voz (como, por ejemplo, el contenido de redes sociales, blogs, documentos personales, correos electrónicos, mensajes de texto, búsquedas en internet, fotografías, audio, vídeo...). Gracias al desarrollo de los algoritmos de procesamiento de lenguaje natural (PLN) y la inteligencia artificial, como los algoritmos de redes neuronales para el procesamiento de imágenes y vídeos, estos datos pueden ahora convertirse en datos estructurados para ser procesados y analizados.

Además de los individuos, las entidades público/privadas también generan ingentes cantidades de datos. El sector público, por ejemplo, genera datos sustanciales en forma de registros públicos, mientras que el sector privado produce información muy detallada fruto de su actividad, de alto valor para el análisis, como las señales que emite un terminal móvil, la huella que dejan las transacciones financieras o las interacciones en aplicaciones y plataformas digitales.

En los últimos años, el uso de las técnicas de procesamiento de lenguaje natural se ha visto revolucionado por la nueva generación de los modelos de redes neuronales profundas, conocidos como transformadores, que pueden detectar patrones sutiles y significados semánticos en el lenguaje. Esta reciente explosión de los grandes modelos de lenguaje (*Large Language Models* en inglés, *LLM*) ha generalizado el análisis de estos datos no estructurados en todas las disciplinas.

En este capítulo exploramos el uso de estas técnicas de PLN, que convierten el texto en números, detallando numerosas aplicaciones desarrolladas en BBVA Research para enriquecer y complementar el análisis económico, financiero, social y geopolítico.

En la primera sección, hacemos un repaso a la trayectoria del desarrollo de las técnicas de análisis de texto y procesamiento de lenguaje natural, desde la minería de texto a los grandes modelos de lenguaje, mostrando ejemplos en la literatura de cómo se han utilizado cada una de estas técnicas en economía. En la segunda sección, enumeramos las distintas aplicaciones desarrolladas por BBVA Research en el ámbito económico, empresarial, geopolítico y político. Finalmente, en la tercera sección, concluimos, enumerando los retos a futuro.

2. EL USO DEL LENGUAJE EN LAS CIENCIAS SOCIALES: SU EVOLUCIÓN DE LA MINERÍA DE TEXTO A LOS GRANDES MODELOS DEL LENGUAJE

El análisis de texto no es un área nueva de estudio, de hecho su uso se remonta a hace más de un siglo. Sin embargo, ha sido en la última década cuando ha experimentado una evolución transformadora, convirtiéndose en una poderosa fuente de valor para investigadores y analistas de datos lingüísticos.

Hasta las últimas dos décadas, el análisis de texto se centraba básicamente en la interpretación de una lectura humana profunda y detallada, la cual no se podía escalar a los grandes volúmenes de texto que tenemos disponibles hoy día. La creciente disponibilidad de información digitalizada en forma de texto ha propiciado el desarrollo y sofisticación de las técnicas de procesamiento del lenguaje para el análisis del mismo. Ash y Hansen (2023) ofrecen una visión detallada de los métodos utilizados en economía para el análisis de texto a partir de distintas metodologías, así como sus limitaciones, especialmente en el campo de la validación de resultados generados por los algoritmos de PLN.

En esta sección, hacemos un repaso a la evolución de las herramientas de PLN, que han evolucionado desde aplicaciones frecuentistas de conteo de palabras que se analizan de forma independiente a tener en cuenta el contexto y contenido semántico de las mismas, mejorando la capacidad de los modelos para representar la estructura temática subyacente en los datos y ayudando a una comprensión más profunda del lenguaje.

2.1. Bolsa de palabras (*Bag of Words*)

El modelo bolsa de palabras (*Bag of Words, BoW,* por sus siglas en inglés) es la forma más sencilla de representar documentos, que consiste en convertir el texto en un formato numérico donde cada documento se representa como un vector y cada dimensión del vector es una palabra que toma como valor el conteo de la ocurrencia de la misma en el documento.

Esta metodología cuantifica la presencia y frecuencia de palabras dentro de los textos, pero sin identificar ninguna relación entre ellas, es decir, no reconoce palabras derivadas, sinónimos, etcétera.

Para aplicar esta metodología, el primer paso se basa en la limpieza y preparación del texto, que se organiza en unidades básicas llamadas tokens[1] en la literatura de PLN, que normalmente son palabras, pero también pueden ser caracteres o subpalabras. De este conjunto de tokens se eliminan los caracteres no alfabéticos como los signos de puntuación y palabras de uso común (como "el/la", "en", "y"), con poco valor analítico y que no agregan significado al texto, conocidas en inglés como *stopwords*. Este paso ayuda a reducir el tamaño del conjunto de datos y a mejorar el tiempo de procesamiento de los mismos.

A continuación, se convierten todas las letras en minúsculas y se aplica una técnica de normalización del texto conocida como *stemming* que consiste en reducir las palabras a su forma raíz, cortando prefijos y sufijos según reglas del lenguaje (por ejemplo, de "niña" y "niñez" a "niñ").

Finalmente, se calcula la conocida matriz TF-IDF (Frecuencia de Término – Frecuencia Inversa de Documento, comúnmente conocida como matriz documento-término), que es una medida estadística para clasificar lo relevante que es una palabra para un documento en una colección de documentos. En dicha matriz, cada fila representa un documento y cada columna representa una palabra de toda la colección de documentos. Así, las celdas de la matriz contienen la frecuencia de cada palabra en cada documento. Esta métrica aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero se compensa con la frecuencia de la palabra en el corpus, lo que ayuda a ajustar el hecho de que algunas palabras aparecen más frecuentemente en general.

Tras las modificaciones enumeradas previamente, se crea la bolsa de palabras, que contiene todos los tokens que aparecen en el texto, ignorando el orden y el contexto en el que aparecen. Cada token representa una característica en el modelo.

El siguiente paso es la vectorización. Cada documento se transforma en un vector. Esto implica contar la frecuencia con la que cada token de la bolsa de palabras aparece en el documento. El resultado es un vector donde cada índice representa un token de la bolsa y cada valor representa la frecuencia de ese token en el documento analizado.

Estos vectores se utilizan para entrenar modelos de aprendizaje automático como la clasificación de textos o análisis de sentimiento a partir de diccionarios asistidos. Son numerosas las aplicaciones de esta técnica en economía como primera aproximación para convertir texto en números y analizar estas nuevas fuentes de información. Permite, por ejemplo, cuantificar la frecuencia y presencia de palabras específicas relevantes en textos de diversa índole para analizar políticas económicas, identificando prioridades, cambios de política o impacto de

[1] Esta técnica también implica eliminar las características lingüísticas únicas del idioma, como los acentos y las eñes en español.

las mismas; monitorizar indicadores económicos donde la frecuencia y cambio en términos económicos específicos pueden señalar cambios en dichos indicadores antes de que se publiquen las cifras oficiales; o analizar comunicados de prensa o informes corporativos. Baker *et al.* (2016) se basan en esta metodología para construir un índice que mide la incertidumbre de la política económica a partir de la búsqueda de palabras clave en artículos de prensa de los principales periódicos estadounidenses y europeos. Este trabajo ha sido de las primeras iniciativas en el uso de las técnicas de análisis de texto en economía, convirtiéndose en un referente es este campo. Loughran y McDonald (2011 y 2016) realizan un estudio exhaustivo de cómo se utiliza el análisis de texto y los modelos de bolsa de palabras en finanzas, analizando su aplicación en el análisis de los comunicados financieros, informes de resultados y otros documentos corporativos.

De esta forma, el modelo de bolsa de palabras se puede usar como herramienta para transformar datos textuales no estructurados en datos estructurados y analizables, facilitando el análisis de texto. No obstante, aunque el modelo de bolsa de palabras es intuitivo y muy sencillo de implementar, es bastante limitado dado que ignora el orden de las palabras y el contexto, dificultando la comprensión más profunda del lenguaje.

2.2. Modelos de temas (*Topic Models*)

Los modelos de temas (*topic models,* en inglés) han supuesto un avance significativo en el análisis de texto y el procesamiento de lenguaje natural frente a los modelos de bolsa de palabras. Esta técnica se basa en la reducción de la dimensionalidad agrupando los tokens en temas subyacentes a partir de una colección de documentos, ayudando a identificar de qué se habla en cada documento.

Una de las primeras técnicas de reducción de la dimensionalidad es el análisis semántico latente (*LSA,* por sus siglas en inglés) (Deerwester *et al.,* 1990). Reconociendo las limitaciones de los modelos de bolsa de palabras para captar el significado y los temas de los documentos, esta metodología *LSA* fue el primer intento de incorporar información semántica. *LSA* utiliza técnicas de componentes principales en la matriz documento-término (que muestra la frecuencia relativa de cada palabra en un conjunto de documentos, como se define en la sección anterior) para reducir su dimensionalidad, identificando así las relaciones latentes entre palabras y documentos. Estas relaciones nos dan una pista de los temas subyacentes, pero no tienen un marco probabilístico asociado, lo que dificulta la interpretación de sus resultados (para más detalle consultar Hash y Hansen (2023)).

Para cubrir esta problemática, Hofmann (1999) desarrolló el análisis semántico latente probabilístico (*pLSA*), que introduce un modelo probabilístico que asume que un documento es una mezcla de temas y un tema es una mezcla de palabras. Esto permite una comprensión más intuitiva de que los documentos y las palabras se generan a partir de temas latentes. Cada documento se modela como una distribución probabilística sobre temas, y cada tema es una

distribución sobre palabras. No obstante, el modelo *pLSA* también tiene limitaciones, dado que no proporciona un modelo generativo completo y tiende a sobreajustarse con un gran número de parámetros como citan Ash y Hansen (2023).

Blei *et al.* (2003) resuelven esta problemática con la asignación de Dirichlet Latente (*LDA,* por sus siglas en inglés), que se trata de un modelo probabilístico basado en *pLSA,* pero incorpora la distribución de Dirichlet para las distribuciones de temas dentro de documentos y distribuciones de palabras dentro de temas, lo que ayuda a manejar el sobreajuste y proporciona un modelo más robusto. De esta forma, los componentes identificados representan el contenido temático latente que se encuentran en los patrones de co-ocurrencia de términos en cada documento que hacen referencia a un mismo tema.

Este modelo *LDA* asigna una distribución de temas a cada documento, lo que significa que un documento puede estar compuesto por varios temas en proporciones diferentes. Por ejemplo, un artículo de noticias podría contener un 70 % del tema "economía" y un 30 % del tema "finanzas".

El paso final es interpretar los temas identificados y utilizarlos para el análisis posterior. La interpretación de los conjuntos de tokens relacionados que conforman un tema es labor del analista y requieren una interpretación cuidadosa, dado que suelen depender del contexto y del conocimiento del dominio.

Una vez identificados estos componentes en temas, se puede estudiar la prevalencia de los diferentes temas a lo largo del tiempo, entre diferentes fuentes de información o como parte de un análisis más amplio como el análisis de sentimiento, identificando cómo cambia el sentimiento e importancia de cada uno de estos temas a lo largo del tiempo.

Esta técnica se ha convertido en la más común para la reducción de la dimensionalidad en procesamiento de lenguaje natural por su interpretabilidad y robustez. Son numerosos los trabajos que emplean *LDA* en economía, como Hansen *et al.* (2018), que usan un modelo *LDA* para entender las transcripciones del Comité de Mercado Abierto de la Reserva Federal, destacando cómo puede utilizarse esta metodología para entender las temáticas más relevantes en el lenguaje de bancos centrales y sus estrategias de comunicación o Bybee *et al.* (2023), que utilizan *LDA* para identificar de qué se habla en noticias financieras y empresariales del *Wall Street Journal* y su importancia para medir el ciclo económico.

Como variantes basadas en *LDA,* Blei y Lafferty (2006) desarrollaron los modelos de temas dinámicos (*DTM,* por sus siglas en inglés) para capturar la evolución de temas a lo largo del tiempo, introduciendo en el modelo *LDA* unos parámetros que evolucionan suavemente para estimar la prevalencia a lo largo del tiempo como parte del propio modelo de temas. Posteriormente, Roberts *et al.* (2013) presentaron el modelo temático estructural (*STM,* por sus siglas en inglés) que incorpora la estructura del corpus al modelo temático estándar del *LDA*. De esta forma, se tiene en cuenta no solo la prevalencia de los temas sino el contenido

temático. En la siguiente sección comentaremos algunas de las aplicaciones desarrolladas en BBVA Research basadas en el uso de *LDA* y *STM*.

Pese a su amplia aplicación y utilidad, estos modelos también cuentan con limitaciones dado que ignoran el orden de las palabras y no tienen en cuenta el contexto de forma que no captan frases con significado similar, pero expresadas de forma distinta, el sarcasmo o la ironía.

2.3. Representaciones vectoriales de palabras (*Word Embeddings*)

Una técnica más avanzada en el campo del procesamiento de lenguaje natural utilizada para representar palabras como vectores en un espacio multidimensional son los *word embeddings,* representaciones vectoriales de palabras o incrustaciones de palabras en español. Estos modelos permiten que las relaciones entre palabras informen del significado de las mismas y sus vectores capten la información semántica y sintáctica de las palabras.

Para ello, cada palabra se representa como un vector en un espacio de alta dimensionalidad. Estos vectores no son arbitrarios, sino que están diseñados de tal manera que palabras con significados o contextos similares tienen representaciones vectoriales similares. De esta forma, en un modelo de *word embeddings,* las palabras "gato" y "perro" estarán ubicadas cercanamente en el espacio vectorial, reflejando su relación, mientras que ambas tendrán una representación distinta a "sombrero", dado que no guardan ninguna relación semántica con esta última palabra. Estos modelos permiten, además, que el significado venga determinado por las palabras vecinas y funcionan mejor cuando se entrenan con grandes volúmenes de texto. Su buen funcionamiento depende en gran medida de la calidad y diversidad del corpus de entrenamiento utilizado, dado que es durante el entrenamiento cuando el modelo aprende a asociar palabras con su contexto.

Entre los modelos más populares para generar *word embeddings* está el *Word2Vec*, desarrollado por Mikolov *et al.* (2013a y 2013b), que se basa en la idea de que el significado de una palabra se puede inferir del contexto en el que aparece. Para ello, se usan redes neuronales poco profundas para incrustar palabras en un espacio vectorial continuo. Las redes neuronales se aplican para aprender de las representaciones vectoriales de palabras basadas en su contexto. La arquitectura de la red neuronal utilizada en los *embeddings* de palabras generalmente implica entrenar la red para predecir tanto las palabras de contexto circundantes dada una palabra objetivo, como al contrario. De esta forma, las incrustaciones de palabras generadas por *Word2Vec* capturan relaciones sintácticas y semánticas. Este modelo se basa en la arquitectura Skip-Gram que predice palabras de contexto a partir de las palabras objetivo. Así, para cada palabra objetivo, mira una ventana de palabras de contexto circundantes e intenta predecirlas. Como comentamos anteriormente, esta metodología funciona mejor cuanto mayor es el volumen de información.

Otro conocido modelo basado en *word embeddings* son los vectores globales para la representación de palabras (*Global Vectors for Word Representation, GloVe,* por sus siglas en inglés) (Pennington *et al.,* 2014). Este modelo *GloVe* se basa en otra aproximación para la incrustación de palabras, centrándose en las matrices de co-ocurrencia palabra-palabra, diseñado así para construir vectores de palabras que codifican la co-ocurrencia local. De esta forma, se construye una matriz de co-ocurrencia global en la que se recoge la frecuencia de palabras que aparecen juntas en todo el corpus analizado. No obstante, *GloVe* también incorpora información del contexto local de cada palabra, considerando la probabilidad de co-ocurrencia de palabras. Esto le ayuda a capturar tanto las relaciones semánticas entre palabras en todo el texto, matizando estas relaciones a partir del contexto local. Este modelo proporcionó un enfoque complementario a *Word2Vec* y mostró cómo se pueden combinar diferentes tipos de información (como la factoriación matricial global con métodos de contexto local) para obtener representaciones de palabras robustas.

Los modelos basados en *word embeddings* son utilizados en una variedad de aplicaciones de PLN, incluyendo la traducción automática, la extracción de información, el análisis de sentimiento y los chatbot, dado que entienden mejor el significado de las palabras en diferentes contextos. En economía, Bloom *et al.* (2021) utilizan los *word embeddings* para identificar frases asociadas a las nuevas tecnologías en patentes, ofertas de empleo e informes de resultados para estudiar la difusión de empleos relacionados con las nuevas tecnologías. Por otro lado, Ash y Gennaro (2023) utilizan esta metodología de *word embeddings* para construir una escala de emocionalidad en los discursos políticos pronunciados en el Congreso de los Estados Unidos durante 1858-2014.

No obstante, los *word embeddings* también tienen ámbitos de mejora como la dificultad para capturar el significado de palabras polisémicas (palabras con múltiples significados) o manejar palabras nuevas (fuera del vocabulario incluido en el contexto analizado).

2.4. Grandes modelos de lenguaje (*Large Language Models*)

Finalmente, los grandes modelos de lenguaje (*Large Language Models, LLM,* por sus siglas en inglés) representan la evolución más disruptiva y reciente en el procesamiento del lenguaje natural e inteligencia artificial. Estos modelos son sistemas avanzados de aprendizaje automático diseñados para entender, interpretar y generar texto, ofreciendo una comprensión profunda del mismo y una capacidad de respuesta lingüística sin precedentes.

Sus miles de millones de parámetros les permiten procesar y generar lenguaje con gran precisión y están entrenados con ingentes cantidades de texto para identificar patrones lingüísticos, gramaticales y semánticos del idioma, adaptándose además a diferentes estilos y dialectos y resolviendo el problema de la polisemia o palabras nuevas que presentan los *embeddings*. Los *LLM* no solo son capaces de comprender el texto de entrada, sino también de generar texto de salida coherente y contextualmente relevante, resolviendo preguntas de distinta índole, resumiendo documentos o traduciendo los mismos, generando ideas, etcétera.

Muchos de estos modelos están basados en arquitecturas de redes neuronales avanzadas, como los modelos *Transformers*[2], que han revolucionado el procesamiento del lenguaje entrenando a los algoritmos para que también "presten atención" a las características relevantes del contexto específico. Estos modelos *Transformers* fueron desarrollados por Vaswani *et al.* (2017) en el documento de trabajo seminal en la literatura de PLN *Attention is All You Need.* La innovación clave de los modelos *Transformers* es el uso de mecanismos de autoatención, que permiten al modelo ponderar la importancia de diferentes palabras en una frase o secuencia, enfocándose dinámicamente en diferentes partes de los datos de entrada al hacer predicciones o generar datos de salida en forma de texto.

En esta tecnología *Transformers* se basan los conocidos modelos de lenguaje como es el caso de BERT (Devlin *et al.,* 2019), desarrollado por Google. Su lanzamiento supuso un cambio significativo en el análisis con PLN frente a modelos previos al incorporar una profunda comprensión del contexto y sutileza del lenguaje. Este modelo aprende a partir de grandes corpus a predecir palabras ocultas en función de su contexto, así como predecir si una frase es continuación lógica de otra para entender las relaciones entre ellas. Posteriormente, Facebook AI desarrolló RoBERTa (Liu *et al.,* 2019), una versión optimizada de BERT que mejora significativamente el rendimiento del modelo y se entrena con volúmenes de corpus significativamente mayores. Sin embargo, han sido en los últimos dos años cuando estos modelos *Transformers* han evolucionado a velocidad de vértigo, cambiando la forma en la que podemos usar la inteligencia artificial (IA) generativa en nuestro día a día. En 2022, Google presentó PALM (Chowdhery *et al.,* 2022) como modelo de IA generativa que mejora los modelos anteriores en términos de escalabilidad, comprensión y flexibilidad. Recientemente, Google junto con DeepMind, han presentado PALM 2, que soporta más de cien idiomas y también es capaz de escribir código, y finalmente han lanzado a final de 2023 Gemini (Gemini y Google, 2024), su modelo más avanzado de inteligencia artificial generativa a partir de tareas multimodales que introduce notables capacidades de comprensión de imágenes, audio, vídeo y texto. No ha sido menos significativa la evolución de la familia de modelos *GPT* (*Generative Pre-trained Transformer,* por sus siglas en inglés) (Radford *et al.,* 2018; Brown *et al.,* 2020; OpenAI, 2023), culminado con su última versión ChatGPT 4, diseñado para entender el lenguaje humano y generar respuestas como si de una persona se tratara. ChatGPT 4 se ha convertido en los últimos meses en la solución de IA generativa más utilizada del mercado basado, al igual que Gemini, en modelos multimodales para trabajar también con imágenes y código. No obstante, la evolución exponencial de estos modelos y la competencia de las grandes tecnológicas como Open AI, Google y Facebook AI por conseguir la mejor solución en AI evidencian el enorme potencial actual y futuro de estas herramientas para el análisis.

Estos modelos ya han sido aplicados a distintos campos dentro de la economía. Korinek (2023) describe distintos casos de uso en los que los *LLM* son de utilidad para la investigación económica en la generación de ideas, escritura, investigación, análisis de datos y codificación. Shapiro *et al.* (2022) muestra el uso de Bert entre otras metodologías de PLN para el análisis

[2] Véase Phuong y Hutter (2022) para una explicación más detallada y precisa de los modelos *Transformers*.

de sentimiento y desarrolla una nueva medida del sentimiento económico a partir de artículos de prensa económica y financiera de enero de 1980 a abril de 2015. Hansen *et al.* (2023a) usan modelos *LLM,* ajustando un modelo estándar basado en la metodología *transformers* para que tenga en cuenta la estructura lingüística específica de las ofertas de empleo con el objetivo de medir y caracterizar el cambio al trabajo a distancia producido tras la pandemia. Por otro lado, Hansen *et al.* (2023b) evalúan la capacidad de los modelos *GPT* para clasificar la orientación política de los anuncios del Comité Federal de Mercado Abierto en Estados Unidos en relación con la valoración humana.

Mas allá del ámbito académico, los grandes modelos del lenguaje también han sido aplicados en el ámbito corporativo para aprovechar el potencial de estas tecnologías. El trabajo de Dabravolski *et al.* (2023) describe el modelo *GPT* personalizado por Bloomberg con 50.000 millones de parámetros, entrenado con datos financieros recopilados a lo largo de cuarenta años con el objetivo de ayudar a la compañía en sus aplicaciones con análisis de texto.

A pesar de su avanzada capacidad, los modelos *LLM* también tienen algunos desafíos aún pendientes de resolver. Estos modelos pueden replicar o amplificar sesgos presentes en los datos de entrenamiento. Además, la interpretación de matices lingüísticos y culturales, así como la gestión de ambigüedades y polisemias del idioma, sigue siendo un área en desarrollo.

3. APLICACIONES DE LAS TÉCNICAS DE LENGUAJE NATURAL EN LAS CIENCIAS SOCIALES EN BBVA RESEARCH

El procesamiento del lenguaje natural se ha convertido en una herramienta muy valiosa en las ciencias sociales, ofreciendo nuevas perspectivas que complementan y enriquecen el conocimiento de la sociedad. Estas técnicas, descritas en la sección anterior, ayudan al investigador en la comprensión del comportamiento humano y los patrones sociales.

Desde BBVA Research son numerosas las aplicaciones desarrolladas a partir del uso de distintas metodologías de procesamiento de lenguaje natural descritas anteriormente. A continuación, enumeramos algunas de ellas en el campo de la economía, mercados y ámbito empresarial, así como en el área geopolítica y política.

3.1. El uso del análisis de texto en economía

En BBVA Research hemos desarrollado una amplia gama de indicadores en tiempo real con técnicas de procesamiento de lenguaje natural y análisis de sentimiento para monitorizar la evolución de temas candentes con impacto en la economía y construir indicadores difícilmente medibles con datos numéricos.

3.1.1. De noticias globales a indicadores en tiempo real

La información disponible en medios de comunicación en forma de noticias *online*, radio y televisión se han convertido en los últimos años en una rica base de datos, altavoz de temas de índole muy diversa de interés social. Gracias a las técnicas de PLN podemos analizar de qué se habla, cómo, dónde y cuándo se habla de cualquier tema, persona u organización.

En BBVA Research llevamos más de una década trabajando con una base de datos en la nube llamada GDELT (Base de Datos Global de Eventos, Lenguaje y Tono)[3], que extrae, procesa y analiza noticias en medios de comunicación a nivel mundial en más de cien idiomas diariamente, desde fuentes de medios globales, nacionales, regionales hasta locales, todos ellos traducidos al inglés automáticamente.

Esta fuente de datos *big data* utiliza diferentes diccionarios para identificar miles de temas, entre los que se incluye todo el glosario de temas del Banco Mundial[4], para clasificar y categorizar la información. Los algoritmos utilizados por GDELT también identifican emociones, organizaciones, ubicaciones, fuentes de noticias y eventos en todo el mundo. Además, generan un sentimiento promedio de cada pieza de información.

Centrándonos en el sentimiento, GDELT aplica más de 40 diccionarios diferentes que clasifican palabras asociadas con tonos positivos y negativos para calcular el tono o sentimiento promedio de todos los documentos que contienen una o más menciones de los eventos o temas que queremos monitorizar de acuerdo a la siguiente fórmula:

$$\overline{S_h} = \frac{\Sigma w_h^+ - \Sigma w_h^-}{\Sigma w_h} * 100 \qquad [1]$$

Donde S es el sentimiento por pieza de información h, y W representa las palabras en cada artículo. La puntuación varía de -100 (extremadamente negativo) a +100 (extremadamente positivo), aunque los valores comunes se sitúan entre -10 y +10, siendo 0 indicativo de neutral.

Usando esta base de datos de noticias globales, estudiamos la evolución de temas económicos difícilmente medibles como la incertidumbre de política económica (*EPU*). Para ello, monitorizamos el sentimiento medio y la cobertura mediática de este tema incluido en la taxonomía de GDELT[5]. La construcción del índice es un producto ponderado formado por la

[3] Véase www.gdelt.org para más información.

[4] Véase el siguiente enlace para acceder a la lista detallada de temas definidos por el Banco Mundial https://vocabulary.worldbank.org/taxonomy/1737.html

[5] Dentro de la taxonomía de GDELT monitorizamos el tema "epu_economy" que incluye noticias relacionadas principalmente con política fiscal y gasto público, política monetaria y tipos de interés, política comercial y aranceles, política regulatoria y asuntos legales, pero también indicadores económicos y tendencias del mercado, acontecimientos internacionales y tensiones geopolíticas, crisis de salud pública y pandemias, desastres naturales y fenómenos meteorológicos, política energética y medioambiental, política laboral y de empleo, crecimiento económico y recesión, inflación y deflación, mercados bursátiles y financieros, banca e instituciones financieras, beneficios empresariales e informes financieros, confianza y gasto de los consumidores, fabricación y producción, vivienda y sector inmobiliario, política de infraestructuras y transporte.

cobertura relativa del tema, medida como el número de noticias relacionadas con el término de búsqueda de *EPU* ese día sobre el número total de noticias publicadas ese mismo día, multiplicado por el sentimiento medio de estos artículos. Multiplicamos el indicador resultante por -1 para facilitar la interpretación, de forma que valores positivos indican mayor incertidumbre por un aumento de la cobertura y/o un empeoramiento del sentimiento medio. La construcción del índice se puede resumir con la siguiente fórmula:

$$EPU_{t,i} = \frac{\Sigma_{s \in G} COV_{t,i}}{\Sigma_G COV_{t,i}} * avg\left(-\overline{S_{t,i}}\right) \quad [2]$$

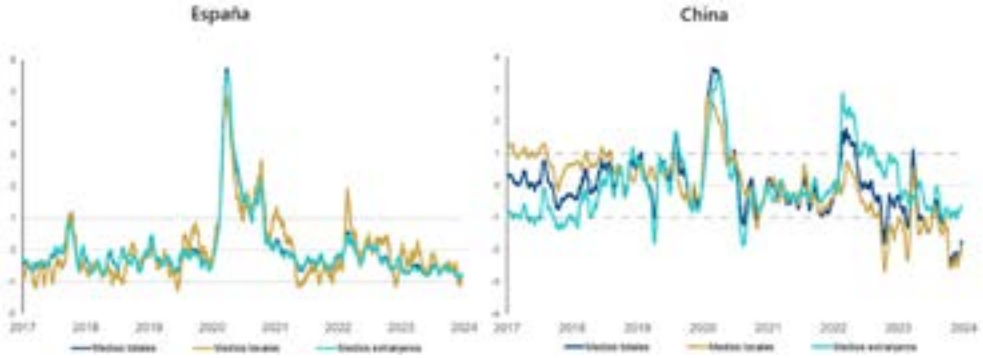
Donde *t* es el tiempo (día) e *i* es el país. *s* se refiere a los artículos de noticias sobre *EPU* y *G* a toda la base de datos. *COV* es a la cobertura mediática y *S* es el sentimiento como vimos en la fórmula [1].

La información se recoge a nivel país. Para presentar indicadores homogéneos, todos los índices se normalizan y se transforman aplicando un promedio móvil ponderado de 28 días[6] para reducir el ruido de la información diaria e identificar señales más claramente.

Figura 1.

Índices *big data* BBVA Research de incertidumbre de política económica

Media móvil 28 días. Indicador normalizado desde 2017



Nota: Valores positivos (negativos) indican mayor (menor) incertidumbre relativo a la media del periodo de 2017 hasta la actualidad.

Fuentes: BBVA Research y www.gdelt.org

La figura 1 muestra el indicador *Big Data* BBVA Research de Incertidumbre de Política Económica para el caso de España y China. Estos indicadores por país nos muestran que a mayor valor (bien por una mayor cobertura al tema, por un empeoramiento del sentimiento

[6] Se aplica media móvil 28 días para incluir cuatro semanas completas con el mismo número de días de la semana y evitar el ruido que pueda generar que no todos los días de la semana hay el mismo volumen de información.

o por la combinación de ambas), mayor es la preocupación por temas de incertidumbre de política económica. Además, podemos distinguir la evolución del indicador por la fuente de información, clasificándolos en medios locales del país, fuentes extranjeras y medios totales, que nos ayudan a identificar las diferentes percepciones o sesgos de los medios de comunicación en determinadas regiones. En el caso de España existe una mayor similitud en las percepciones de incertidumbre de política económica en prensa local y extranjera que en el caso de China. Sobre estas diferencias por origen de los medios de comunicación, en el caso de España, destaca que en medios locales españoles el índice empeora frente a la percepción en prensa extranjera en enero de 2021 debido a las implicaciones económicas de la última oleada monitorizada del COVID-19 con la variante ómicron, acompañada del primer mes de datos de inflación positivos tras el estadillo del COVID-19 y los cuellos de botella en las cadenas globales de valor. El indicador local también sobresale a principios de 2022 con el estadillo de la guerra de Ucrania y más recientemente en 2023 con el periodo electoral, primero por las elecciones municipales en mayo, seguidas de las elecciones generales en julio. En el caso de China, destaca la brecha significativa entre las señales de mayor incertidumbre que arroja el indicador en prensa extranjera frente al de prensa local desde inicios de 2022, al contrario que sucedía en 2017 y 2018.

Utilizando la información de los medios de comunicación también analizamos un tema candente relacionado con las cadenas globales de valor como es la crisis de los semiconductores, impulsada por la guerra comercial entre EE. UU. y China, los problemas relacionados con la capacidad de la industria después del COVID-19, el mal tiempo durante 2020-2022 y eventos geopolíticos clave como las tensiones entre China y Taiwán y la guerra entre Rusia y Ucrania.

Figura 2.

Indicador Global *Big Data* BBVA Research de Semiconductores (2018-2023)
Media móvil 28 días. Indicador normalizado desde 2018

Nota: Valores positivos (negativos) indican mayores (menores) tensiones relativo a la media del periodo de 2018 hasta la actualidad.

Fuentes: BBVA Research y www.gdelt.org

Los semiconductores son esenciales para las tecnologías modernas y desempeñan un papel clave en algunas industrias estratégicas. Con el uso del *big data* de noticias globales y técnicas de PLN e inteligencia artificial, desarrollamos indicadores de sentimiento mediático en tiempo real para analizar su evolución en el tiempo (figura 2) siguiendo la metodología descrita en el caso anterior de los indicadores *EPU* (Fórmula [2]). El indicador es, por tanto, un índice ponderado de la cobertura relativa del tema de los semiconductores por el sentimiento medio y multiplicado por -1 de tal forma que valores mayores indican mayor riesgo, tensiones o peor sentimiento.

El Indicador Global Big Data BBVA Research de Semiconductores muestra que la crisis de los semiconductores y su posterior normalización se ha visto impulsada por varios factores, como la guerra comercial entre Estados Unidos y China, la escasez relacionada con el COVID-19 y los problemas de capacidad, el mal tiempo, las tensiones geopolíticas y los avances en la diplomacia internacional. La guerra comercial entre EE. UU. y China en 2018-2019 generó un aumento del indicador a la zona de riesgo a mediados de 2019. Posteriormente, el COVID-19 provocó alteraciones sin precedentes en las cadenas de suministro mundiales que, junto con un aumento de la demanda de productos electrónicos debido al trabajo a distancia, generaron un aumento del indicador debido a la escasez de oferta. Pero fue la guerra entre Rusia y Ucrania en 2022 el mayor amplificador de los problemas en las cadenas de suministro mundiales, lo que llevó al mayor aumento del indicador en el horizonte temporal estudiado. Desde mediados de 2023, el indicador muestra una relajación en el mercado gracias al alivio de las tensiones entre EE. UU. y China y una mejora de los cuellos de botella en la producción de los mismos.

El uso del PLN y el análisis de texto en medios de comunicación también nos permite entender relaciones e interconexiones entre temas, geografías, personas, organizaciones… teniendo en cuenta la co-ocurrencia de los mismos en prensa a través del análisis de redes. Para entender las posibles implicaciones económicas de un hipotético conflicto entre China y Taiwán tiene especial relevancia el peso de Taiwán en la Industria Mundial de Semiconductores a través de la empresa Taiwan Semiconductor Manufacturing Company (TSMC), que es particularmente relevante para la producción de chips para los dispositivos más avanzados, utilizados especialmente por los países desarrollados. Para analizar el papel sistémico de TSMC en la industria de semiconductores, construimos una red de noticias globales con GDELT basada en la co-ocurrencia de empresas y sectores con la empresa TSMC en la misma noticia (figura 3).

Los nodos o vértices de la red en la figura 3 representan empresas y las aristas indican la relación entre ellas medida como el conteo de noticias donde aparecen ambas empresas mencionadas. Existen varias medidas de centralidad, como el grado, la proximidad y la interrelación o intermediación. Con el objeto de analizar el papel sistémico de TSMC, implementamos un algoritmo de centralidad de autovectores en la red para ponderar el tamaño de cada nodo en función del número de aristas o relaciones, así como la relevancia de estas conexiones de aristas, dadas sus propias relaciones con otros nodos en la red. Un mayor tamaño del nodo significa que un nodo está conectado a muchos otros nodos que, a su vez, tienen también

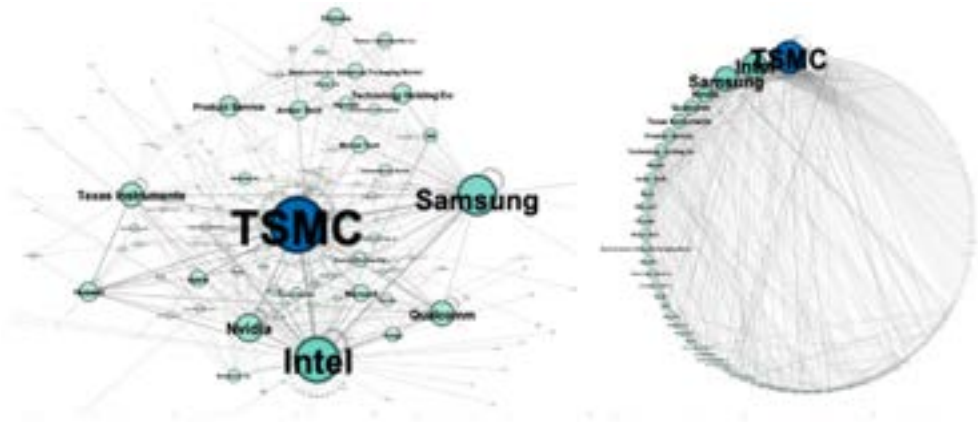
muchas relaciones con otros nodos, lo que refleja mejor la naturaleza sistémica de un nodo en la red.

Esta red de interconexiones entre compañías nos permite describir las posibles relaciones en las noticias de la crisis de semiconductores de Taiwán con el resto de la industria. En ella vemos la importancia y centralidad de TSMC en la industria. La red también muestra la relación de TSMC con el resto de las empresas de la industria donde grandes fabricantes de dispositivos integrados (conocidos como *IDMs,* por sus siglas en inglés, *Integrated device manufacturers*) como Intel, Samsung, Texas Instruments y Micron Tech tienen un peso relevante por su relación con TSMC, dado que subcontratan parte de su producción para tecnologías avanzadas. Se identifican también en la red clientes relevantes de TSMC como Nvidia y Qualcomm, seguidas por Broadcom y AMD. Adicionalmente, encontramos grandes empresas tecnológicas como Apple, Huawei y Microsoft, que dependen de TSMC para la producción de chips personalizados.

Figura 3.

Análisis de redes de empresas de semiconductores y TSMC (2022)

Algoritmos Multi-gravedad Force Atlas (izquierda) y esquema circular (derecha)



Nota: El tamaño de los nodos está ponderado por la centralidad de los autovectores.

Fuentes: BBVA Research y www.gdelt.org

En definitiva, el papel de TSMC como proveedor clave de chips avanzados y su posición en la cadena de valor de semiconductores hacen que TSMC se convierta en un elemento de riesgo sistémico importante en la red para la industria de semiconductores y las economías avanzadas. Su interconexión con importantes actores globales refleja la complejidad y la interdependencia del sector dada la importancia de TSMC en la cadena de suministros globales y su papel crítico en la producción de tecnologías avanzadas. Estos vínculos son fundamentales para entender la dinámica del mercado y las posibles repercusiones de cualquier perturbación en la industria.

3.1.2. Análisis del lenguaje de los bancos centrales a partir de los comunicados de política monetaria

Además de las noticias globales, las técnicas de procesamiento de lenguaje natural permiten analizar cualquier texto, sea cual sea su naturaleza. Los comunicados de política monetaria de los bancos centrales se han convertido en una herramienta clave para controlar las expectativas de inflación y analizar dichos comunicados nos ayuda a entender mejor la estrategia de los mismos en sus decisiones de política monetaria y, por tanto, su impacto en la economía real.

Siguiendo la metodología aplicada por Hansen y McMahon (2016), analizamos el lenguaje de los bancos centrales sobre política monetaria a partir de sus comunicados de prensa o declaraciones, actas y discursos publicados en la web. Tras limpiar, transformar y procesar el texto convertido en vectores y resumido en la matriz documento-término, utilizamos los modelos de temas dinámicos basados en la asignación de Dirichlet Latente (*LDA*) que, como explicamos en la sección anterior, tienden a distribuir palabras en un conjunto reducido de temas para maximizar las probabilidades de las palabras de aparecer juntas para cada tema dado. Es un algoritmo no supervisado donde el investigador tiene que interpretar cada tema examinando la colección de palabras clave en cada componente. Una vez que tenemos identificados los temas, realizamos análisis de sentimiento basado en la aproximación del léxico, que consiste en analizar dentro de estos temas cuantas palabras positivas y negativas le acompañan de acuerdo a diccionarios asistidos para obtener un sentimiento medio del tema a lo largo del tiempo. Para ello, usamos el diccionario Loughran-McDonald (2011), que fue creado específicamente para analizar textos de índole financiera y el diccionario de la FED para la estabilidad financiera (Correa *et al.,* 2017). Estos diccionarios identifican palabras negativas y positivas, incluyendo palabras específicas de la jerga financiera y nos dan una mejor aproximación en el sentimiento que diccionarios creados a partir de lenguajes más generales.

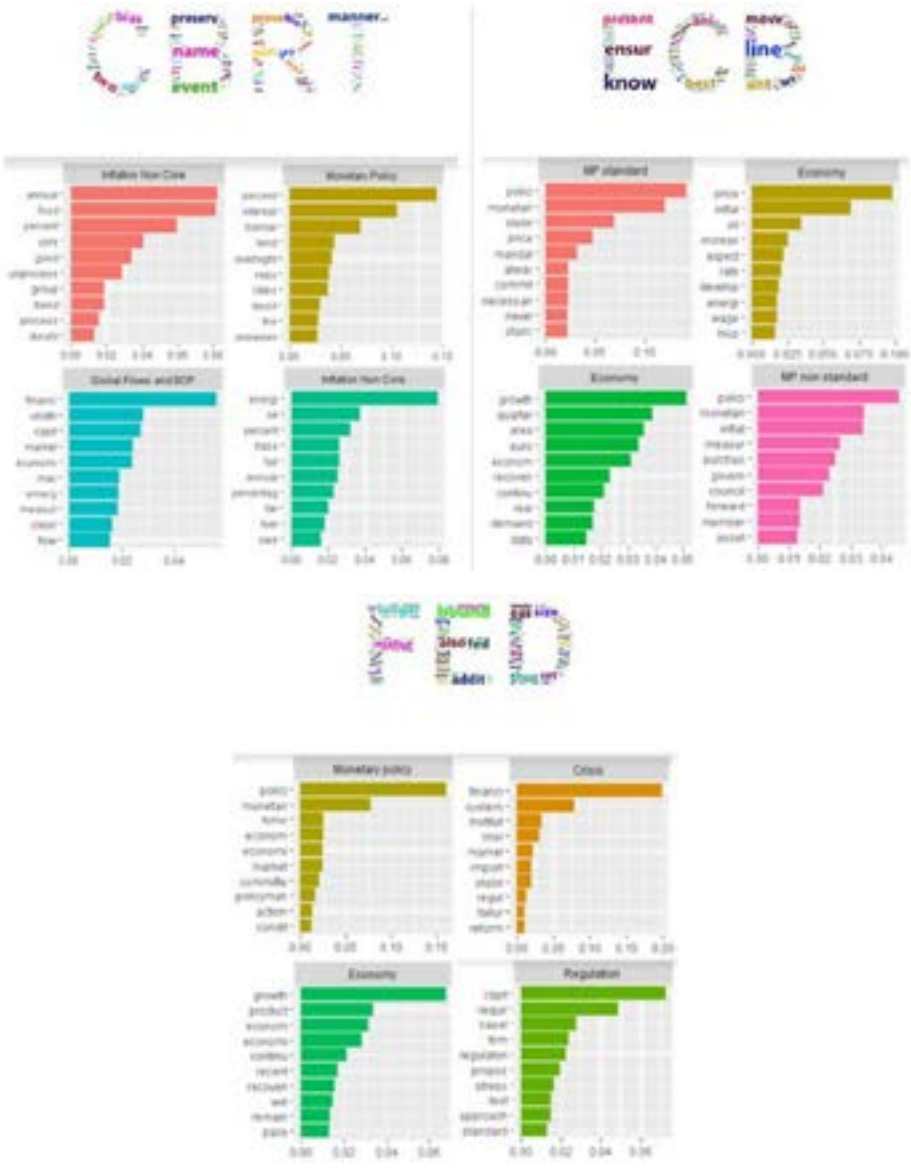
Con esta aproximación, analizamos los comunicados de varios bancos centrales, entre ellos el Banco Central Europeo (ECB, por sus siglas en inglés) o la Reserva Federal (FED) dentro de los países desarrollados y el Banco Central de Turquía (CBRT, por sus siglas en inglés) como país emergente durante los años 2016-2018. Para todos ellos obtenemos los principales temas identificados en sus comunicados como una composición de los términos relacionados más importantes y frecuentemente utilizados.

En la figura 4 podemos ver para el corpus analizado de cada banco central, qué temas emergen para cada uno de ellos. Términos de política monetaria como tasas de interés son un denominador común en todos los comunicados de bancos centrales, aunque existen matices relevantes dado que usan distintas herramientas de política monetaria. En el caso del BCE, destaca el uso de la política monetaria no estándar durante el periodo estudiado. En el caso del Banco Central de Turquía, destaca el tema de los flujos de capital por su importancia para las economías emergentes, especialmente tras la crisis financiera global.

Figura 4.

Principales temas identificados para el Banco Central de Turquía, Banco Central Europeo y la Reserva Federal (2016-2018)

Modelo estructural de temas basado en *LDA* para identificar los componentes



Fuentes: BBVA Research a partir de los comunicados e informes de política monetaria del CBRT, ECB y la FED.

Figura 5.

Evolución de los temas de política monetaria y sentimiento de los comunicados del Banco Central de Turquía a partir del análisis de texto

Modelo estructural de temas basado en *LDA* para identificar los componentes, la evolución de su peso en el tiempo y análisis de sentimiento

Fuentes: BBVA Research a partir de los comunicados e informes de política monetaria del CBRT.

Una vez identificados los temas, estudiamos su evolución en el tiempo, analizando cómo varía el peso o importancia de cada tema en cada comunicado, así como el sentimiento medio asociado. La figura 5 resume este análisis para el Banco Central de Turquía. El primer gráfico muestra cómo identificamos a partir de las técnicas de *LDA* y *STM* de qué se habla en las minutas del CBRT, mostrando tres ejemplos de temas identificados de política monetaria, inflación y actividad. De todos estos temas, gracias al modelo *STM,* podemos estudiar su evolución a lo largo del tiempo en cuanto al peso relativo de cada tema en los comunicados. Así por ejemplo, el segundo gráfico muestra la evolución de los temas de política monetaria en detalle, donde vemos cómo a lo largo del tiempo con la crisis financiera global, la política monetaria tradicional fue disminuyendo su importancia frente a las herramientas de política monetaria macroprudencial, que habían estado prácticamente inexistentes antes de la crisis. De todas estas series, tenemos, además, la evolución del sentimiento medio asociado. El tercer gráfico muestra el tono medio en todos los comunicados del CBRT e identificamos una estrecha relación del sentimiento con la postura del banco central, donde valores negativos van asociados a una posición de endurecimiento de la política monetaria y positivos con una flexibilización de la misma.

3.2. El uso de los informes corporativos como indicador de la evolución económica por sector de actividad

En el ámbito empresarial, la aplicación de las técnicas de PLN es de gran utilidad para analizar la información publicada por las empresas en sus informes trimestrales y anuales,

donde tienen que reportar su actividad y desempeño financiero, incluyendo datos financieros y resultados, información sobre segmentos de mercado, nuevos planes de productos e investigación y actividades de desarrollo en programas futuros. Es, por ello, que son una fuente de información de gran utilidad para entender el desempeño, perspectivas de futuro e interpretaciones del pasado de las empresas, así como los principales riesgos y oportunidades que se perciben en cada sector de actividad y, por ello, en la economía.

Con este fin, en BBVA Research analizamos los informes trimestrales y anuales de las empresas estadounidenses, las cuales están obligadas a presentar información veraz y detallada sobre su salud financiera a la SEC (Comisión de Bolsa y Valores de EE. UU.). La SEC pone a disposición pública todos los archivos electrónicos de estos informes corporativos a través del sistema de Recopilación, Análisis y Recuperación de Datos Electrónicos (EDGAR). Nos centramos en la información de las empresas del SP500, dado que son las 500 empresas de gran capitalización y cotizadas en la bolsa de valores de EE. UU. que cubren aproximadamente el 80 % del mercado de acciones estadounidense por capitalización.

Analizamos de ellas más de 21 millones de documentos desde 1994 a 2019 a los que les aplicamos dos enfoques diferentes para obtener información sobre los principales problemas abordados en el texto. En primer lugar, aplicamos análisis de temas utilizando la asignación de Dirichlet Latente como algoritmo de aprendizaje automático no supervisado para identificar temáticas en función de la relación de palabras en el texto. En segundo lugar, usamos la metodología *Word2Vec*, que es una red neuronal que obtiene las relaciones no lineales entre palabras para agruparlas de acuerdo a su contexto, por lo que nos dirá, dada una palabra o conjunto de palabras, el resto de las palabras similares que aparecen en el mismo contexto, generando un diccionario adaptado a las particularidades del texto analizado.

Finalmente, después de obtener los temas identificados y los diccionarios a partir de palabras clave, aplicamos el sentimiento utilizando el diccionario de Loughran y McDonald (2011) para obtener el tono o sentimiento promedio de los documentos.

Así, la combinación de ambas técnicas complementa la comprensión más profunda y detallada del texto. Por tanto, dejamos que los datos hablen para identificar de qué están hablando las empresas por sector de actividad y cómo están hablando a partir del sentimiento medio, y además identificamos, para términos estratégicos como el riesgo o la incertidumbre, cómo se relacionan en estos informes y su contexto.

Los principales temas identificados en los informes hacen referencia a sectores de actividad, dado que empresas del mismo sector tienden a comentar contextos similares. La figura 6 resume los resultados de los temas identificados con *STM* donde se destacan algunos de ellos con nubes de palabras en las que el tamaño y el color hacen referencia a la mayor probabilidad de la palabra de pertenecer a ese grupo. De esta forma, en la primera nube de palabras, los tokens de energía, electricidad o transmisor nos ayudan a identificar el sector energético. Del mismo modo, a partir de estos tokens, identificamos temas financieros, comercio al por menor, sector inmobiliario o automotriz entre otros.

Figura 6.

Principales temas identificados en los informes corporativos de las empresas SP500 (1994-2019)

Modelo estructural de temas basado en *LDA* para identificar temas

Fuentes: BBVA Research a partir de los informes corporativos de las empresas del SP500.

Para complementar este análisis, aplicamos la metodología *Word2Vec* para generar un conjunto de diccionarios con el fin de entender el contexto de términos económicos y financieros clave como la inflación, el crecimiento económico, la depreciación, el riesgo o la incertidumbre. La figura 7 muestra los resultados de este análisis. La distribución de palabras asociadas con cada término de interés se ilustra en el gráfico de dispersión (segundo gráfico). Los resultados muestran que hay una red significativa de interrelacio-

Figura 7.

Distribución de palabras en la generación de diccionarios y su evolución en el tiempo

Modelo *Word2Vec* para la generación de diccionarios de acuerdo a la proximidad entre palabras

Fuentes: BBVA Research a partir de los informes corporativos de las empresas del SP500.

nes entre palabras que hacen referencia al mismo contexto y, en general, las palabras clave escogidas están bien diferenciadas entre ellas, puesto que no hay muchas palabras mezcladas entre contextos. Inflación y depreciación guardan una mayor similitud, al igual que riesgo e incertidumbre. Sobre estos dos últimos términos, mostramos en la figura 7 la nube de palabras del diccionario creado para riesgo (primer gráfico) e incertidumbre (tercer gráfico), donde el tamaño (mayor) y el color (más oscuro) de las palabras hacen referencia a la proximidad (mayor) entre ellas. Vemos así qué riesgo está más asociado con términos financieros y de mercado, mientras que incertidumbre se relaciona más con preocupaciones y fragilidades económicas. Con estos diccionarios, podemos monitorizar con mayor robustez (dado que se han creado específicamente para el corpus analizado), la evolución de estas temáticas a partir de palabras clave a lo largo del tiempo, sectorialmente y por tipo de empresa, entendiendo mejor la heterogeneidad entre temas y sectores de actividad económica.

3.3. Monitorización de indicadores geopolíticos y políticos a partir de los medios de comunicación

En el campo geopolítico, el uso de estas técnicas de PLN se hace especialmente relevante dado que la información cuantitativa existente es escasa. El análisis de texto en las noticias nos permite rastrear múltiples eventos, proporcionando respuestas rápidas a preguntas cada vez más complejas en un mundo cada vez más fragmentado.

En BBVA Research llevamos casi una década utilizando estos métodos en el campo geopolítico para obtener respuestas en tiempo real y de forma muy granular y detallada. Así, hemos desarrollado desde indicadores de intensidad de conflictos, protestas, riesgo geopolítico y tensiones políticas, hasta mapas dinámicos de flujos migratorios con mucho detalle para ver el origen y destino de los migrantes durante la crisis humanitaria con la guerra de Siria.

Empezando por los indicadores de intensidad de protesta y conflicto, utilizamos la información de GDELT, donde cada evento se codifica según el sistema de codificación de eventos CAMEO (*Conflict and Mediation Event Observations*) desarrollado por Gerner *et al.* (2002). CAMEO es un esquema de codificación ampliamente utilizado para sistematizar el análisis de eventos políticos y sociales y dividirlos en una escala que va desde la cooperación material y verbal hasta el conflicto verbal y material. A partir de este sistema de CAMEO, identificamos todos los eventos relacionados con protesta y conflicto y monitorizamos la cobertura mediática de los mismos a través de GDELT.

De esta forma, construimos unos indicadores de intensidad de protesta y de conflicto, que capturan el volumen total de artículos de noticias por día que incluyen cualquier men-

ción de estos eventos identificados en CAMEO sobre protestas y conflictos. El número total de eventos de protestas y conflictos cada día y en cada país se divide por el número total de todos los eventos registrados por GDELT para ese día en ese país con objeto de elaborar un índice de intensidad por país. De esta manera, el indicador rastrea el nivel de prevalencia de la actividad de protesta y conflicto a lo largo del tiempo, corrigiendo por el aumento exponencial en la cobertura mediática en los últimos años y las diferencias de la cobertura mediática entre geografías.

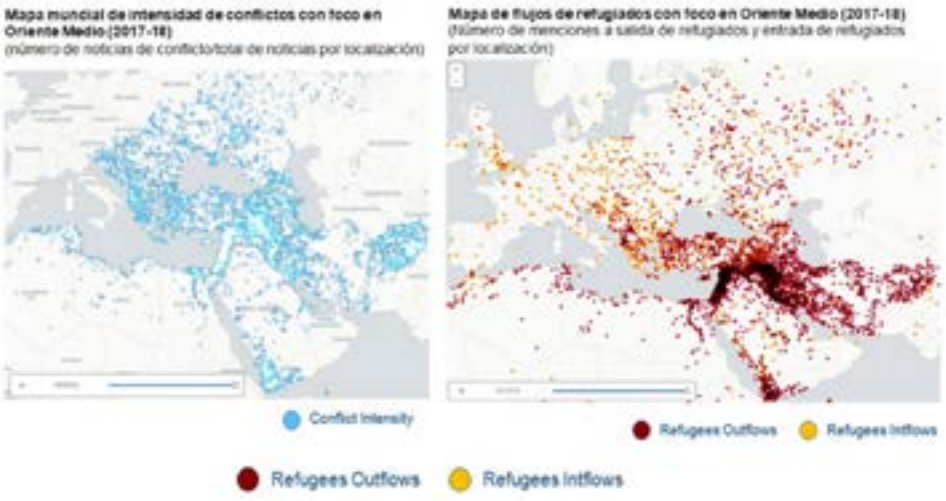
En los últimos años no hemos estado exentos de conflictos locales con efectos globales como la guerra de Siria, la guerra de Ucrania o la más reciente invasión de Hamas. Medir su evolución nos permite cuantificar su impacto social y económico.

En las figuras 8 y 9 podemos ver cómo hemos medido estos conflictos con distintos focos. En el caso de la guerra de Siria, construimos un mapa dinámico del indicador de intensidad de conflicto durante 2017 y 2018 con foco en Europa y Oriente Medio para identificar donde se localizan estos conflictos y cómo evolucionan en el tiempo (figura 8). El principal foco de conflicto se situaba en Siria e Irak, pero también destacan conflictos congelados como en Yemen y en Afganistán (primer mapa). Como consecuencia de la guerra de Siria, se produjo un éxodo masivo de refugiados con importantes implicaciones sociales y económicas para la región y para Europa. En el segundo mapa de la figura 8 mostramos

Figura 8.

Evolución de los conflictos en Oriente Medio y flujos migratorios a raíz de la Guerra de Siria (2017-2018)

Indicadores de intensidad de conflictos y de salida/entrada de refugiados



Fuentes: BBVA Research y www.gdelt.org

estos flujos migratorios donde en color rojo recogemos el volumen de noticias que hablan de salidas de refugiados en esa geografía y en amarillo la llegada de refugiados. Turquía fue el principal receptor de estos flujos migratorios con más de 3,5 millones de migrantes, como muestra el mapa, que refleja muy bien la trazabilidad de estos movimientos. La dinámica en el tiempo nos muestra el paso de los migrantes por Turquía, reflejando cómo una gran parte de ellos se quedaban en el país y otros continuaban su camino por los Balcanes hasta llegar a Europa Central.

La evolución temporal de los indicadores de conflicto por país y en frecuencia diaria permite identificar rápidamente cambios de tendencia como, por ejemplo, con la invasión rusa en Ucrania en febrero de 2022 o la reciente invasión de Hamas a inicios de octubre de 2023 (figura 9). En el caso de la guerra de Ucrania, el indicador tanto para Rusia como para Ucrania, muestra que el conflicto continúa vivo con valores de riesgo alto (por encima de una desviación típica con respecto a su media histórica desde 2019), sin signos de desescalada (primer gráfico). El aumento de las tensiones en los indicadores de conflicto ha sido especialmente significativo en el caso de Hamas (segundo gráfico), que además tiene implicaciones importantes para la región con riesgo de escalada e importantes efectos globales y económicos.

Aparte de conflictos y protestas, utilizando la taxonomía de GDELT, podemos rastrear otros temas geopolíticos, sociales y económicos como el riesgo geopolítico o las tensiones políticas. De esta forma, en BBVA Research desarrollamos un indicador de riesgo geopolítico basado en la metodología de Caldara e Iacoviello (2022) para iden-

Figura 9.

Indicador Big Data BBVA Research de intensidad de conflictos: Rusia, Ucrania y Oriente Medio (2017-2024)

Número de noticias de conflicto/total de noticias por geografía

Fuentes: BBVA Research y www.gdelt.org

tificar palabras clave relacionadas con el riesgo geopolítico como violencia, disturbios, relaciones exteriores[7],...

A partir de las noticias identificadas, generamos un índice basado en el tono y la cobertura relativa de estas palabras clave en la base de datos de GDELT de acuerdo a la fórmula [2] detallada en esta sección. De la misma manera, para las tensiones políticas, recopilamos todos los artículos de noticias que mencionan temas relacionados con política siguiendo la taxonomía de GDELT, como elecciones y campañas, partidos políticos y políticos, instituciones gubernamentales, políticas y escándalos políticos entre otros y construimos un indicador por país en frecuencia diaria de acuerdo a la misma metodología descrita en la fórmula [2].

Con todo ello, generamos un conjunto de indicadores que nos permiten monitorizar el panorama geopolítico en alta frecuencia y con gran precisión (figura 10), no solo identifi-

Figura 10.

Cuadro resumen de los principales indicadores *big data* geopolíticos de BBVA Research antes del conflicto Hamas – Israel y en la actualidad

Media móvil 28 días. Indicadores normalizados desde 2017

Fuentes: BBVA Research y www.gdelt.org

[7] El Índice de Riesgo Geopolítico de BBVA Research recoge la intersección de dos grupos de palabras clave. Las búsquedas deben contener al menos un tema incluido en la taxonomía GDELT de cada grupo (grupo 1 y grupo 2). El grupo 1 incluye los siguientes temas: guerra, conflicto, violencia revolucionaria, rebelión, disturbios violentos, mantenimiento de la paz, acuerdos de reconocimiento mutuo, alto el fuego, tratados, parlamento y legislaturas, militares, tropas, energía nuclear, energía hidroeléctrica, terror, rebeldes guerrilleros e insurgentes, secuestro, alianza, comité de resistencia popular de grupo, insurgencia, resistencia social de grupo, cooperación militar, armada y rebeldes entre los términos más relevantes. El grupo 2 incluye los siguientes temas: acto perjudicial, anunciante, riesgo, preocupación en todo el mundo, especie en peligro, crisis, problema, disputas, procedimientos de despido, boicot, perturbación, refuerzo militar, sanciones, bloqueo, vulnerabilidad y riesgos financieros, cuarentena, ultimátum, declaración, brote, anunciador, armero, persecución, choque, redada, conflicto armado, acto de fuerza, amenaza de bomba, asesinato, huelga. Véase Caldara e Iacoviello (2022) para mayor detalle.

cando puntos de tensión, sino posibles efectos contagio entre geografías o sectores. Su seguimiento en frecuencia diaria nos permite contar con una herramienta de alerta temprana para entender que está pasando en el mundo. En el primer gráfico de la figura 10 mostramos cómo estaban estos indicadores antes de la invasión de Hamas como referencia al conflicto armado más reciente, y en el gráfico de la derecha una imagen más actualizada. La figura muestra que se ha producido un aumento generalizado de los focos de conflicto y el riesgo geopolítico, acompañado en el caso de Oriente Medio y en algunos países latinoamericanos de un aumento de las tensiones políticas. El descontento social y el aumento de incertidumbre de política económica se mantiene contenido hasta el momento.

4. CONCLUSIONES

La digitalización y la interacción social en tiempo real con dispositivos móviles, ordenadores, redes sociales y plataformas digitales ha generado ingentes volúmenes de datos no estructura- dos (textos, videos, imágenes, voz...). Gracias al avance de los algoritmos de procesamiento de lenguaje natural y las técnicas de inteligencia artificial, estos datos pueden convertirse en información estructurada para el análisis. Además, la disminución significativa de los costes de cómputo y procesamiento de la información y el desarrollo de modelos y algoritmos para tratar estos datos, han revolucionado la forma de hacer analítica en las ciencias sociales.

En este capítulo hacemos un seguimiento a la evolución de los algoritmos de análisis de texto, desde la bolsa de palabras a los grandes modelos de lenguaje, destacando sus ventajas y desafíos y enumerando ejemplos de cómo se han utilizado en economía. Finalmente, mostramos cómo en BBVA Research utilizamos estas técnicas de procesamiento de lenguaje natural para el análisis económico, financiero, social y geopolítico.

En economía, analizamos el lenguaje de bancos centrales, temas clave difícilmente cuantificables con datos tradicionales como la incertidumbre de política económica a partir de los medios de comunicación o la crisis de los semiconductores. En el ámbito financiero, mostramos el análisis de las empresas estadounidenses del SP500 para estudiar la evolución sectorial y términos económicos y financieros clave. Finalmente en el campo geopolítico, mostramos indicadores en tiempo real para monitorizar conflictos, protestas, riesgo geopolítico y tensiones políticas.

Todo ello muestra el potencial del análisis de texto en las ciencias sociales, que alcanzan su máximo exponente en la actualidad con los grandes modelos de lenguaje (*LLM*), que comprenden y analizan en profundidad el lenguaje humano. No obstante, existen desafíos relevantes que condicionarán su desarrollo y viabilidad futura como la importancia de asegurar un uso justo de los mismos, sin sesgos y éticamente correctos, que además tienen que ir acompañados del desarrollo de marcos regulatorios que garanticen la privacidad de los usuarios, la seguridad de los datos y la penalización de un mal uso del contenido generado

por los mismos. Si conseguimos hacer frente a estos desafíos, las oportunidades de futuro y el potencial de la inteligencia artificial es incalculable para el análisis, donde la imaginación es el límite.

Referencias

Ash, E. y Gennaro, G. (2023). Emotion and Reason in Political Language. *The Economic Journal, Volume 133,* Issue 650.

Ash, E. y Hansen, S. (2023). Text Algorithms in Economics. *Working Paper.*

Baker, S. R., Bloom, N. y Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics, 131*(4), pp.1593–1636. Oxford University Press.

Blei, D. y *Lafferty, J.* (2006). *Dynamic Topic Models. ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning, 2006.* 113-120. 10.1145/1143844.1143859.

Blei, D. M., Ng, A. Y. y Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*(null), pp. 993–1022.

Bloom, N., Hassan, T., Kalyani A., Lerner, J. y Tahoun, A. (2021). The diffusion of disruptive technologies. *CEP Discussion Papers,* dp1798. Centre for Economic Performance, LSE.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan,A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G.,Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen,M., Sigler, E., Lit- win, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. y Amodei, D. (2020). Language Models are Few-Shot Learners. *In Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc.

Bybee, L., Kelly, B., Manela, A. y Xiu, D. (2023). Business News and Business Cycles. *Journal of Finance.* Forthcoming.

Caldara, D. e Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review, 112*(4), pp. 1194-1225.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P. *et al.* (2022). PaLM: Scaling language modeling with Pathways. arXiv preprint arXiv:2204.02311.

Correa, R., Keshav, G., Londono-Yarce, J. M. y Nathan M. (2017). Constructing a Dictionary for Financial Stability. IFDP Notes. Washington: Board of Governors of the Federal Reserve System, June 2017. https://doi.org/10.17016/2573-2129.33

Dabravolski, V., Dredze, M., Gehrmann, S., Irsoy, O., Kambadur, P., Lu, S., Mann, G., Rosenberg, D. y Wu, S. (2023). *BloombergGPT: A Large Language Model for Finance.*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. y Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), pp. 391–407.

Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team Google (2023). Gemini: A Family of Highly Capable Multimodal Models Gemini Team. *Working Paper.*

Gerner, D., Jabr, R. y Schrodt, P. (2002). *Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions.*

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R. y Taska, B. (2023a). Remote Work across Jobs, Companies, and Space. *Working Paper.*

Hansen, S., Lundgaard, A. y Kazinnik, S. (2023b). Can ChatGPT Decipher Fedspeak? *Working Paper.*

Hansen, S. y McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99, S114–S133.

Hansen, S., McMahon, M. y Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics,133*(2), pp. 801–870.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99* (pp. 289–296). San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature, 61*(4), pp. 1281-1317.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. y Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.arXiv preprint arXiv:1907.11692.

Loughran, T. y McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance,* 66, pp. 35–65.

Loughran, T. y McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research, 54,* issue 4, p. 1187-1230.

Mikolov, T., Chen, K., Corrado, G. y Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. y Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.

OpenAI. (2023). GPT-4 Technical Report.

Pennington, J., Socher, R. y Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in naturallanguage processing* (*EMNLP*) (pp.532–1543).

Phuong, M. y Hutter, M. (2022). Formal Algorithms for Transformers. *Working paper.*

Radford, A., Narasimhan, K., Salimans, T. y Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *Working Paper, OpenAI Blog.*

Roberts, M. E., Stewart, B. M. y Airoldi, E. M. (2013). Structural Topic Models. *Working paper.*

Shapiro, A. H., Sudhof, M. y Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics, 228*(2), pp. 221–243.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. y Polosukhin, I. (2017). Attention is All you Need. In *Advances in NeuralInformation Processing Systems,* volume 30. Curran Associates, Inc.

CAPÍTULO II

Análisis de la evolución del sentimiento hacia el cambio climático en España

Maria Alló*
Maria Loureiro

El cambio climático representa un desafío crucial para la sociedad actual, y es imperativo implementar políticas efectivas para combatirlo. Sin embargo, hay una carencia de datos actualizados sobre la percepción pública al respecto. Este estudio subraya la importancia de analizar la información de las redes sociales para entender la percepción social en tiempo real (*nowcasting*) sobre el cambio climático, empleando técnicas de procesamiento de lenguaje natural. De acuerdo al análisis de tuits recogidos entre 2019-2022 en España, nuestros resultados muestran un pico de preocupación climática en España en los meses posteriores al levantamiento del primer confinamiento en 2020, seguido de una leve mejoría, aunque aún por debajo de los niveles pre-COVID-19. Sin embargo, durante el período más extremo de la crisis del COVID-19, los datos demuestran que la preocupación por el cambio climático se relaja, mostrando evidencia en favor de la "hipótesis de la existencia de una reserva de preocupaciones finita" (Weber, 2006). En resumen, estos resultados ponen de manifiesto el aumento de la preocupación por la cuestión climática en España en el período evaluado, a pesar de la gran incidencia de la crisis sanitaria del COVID-19.

Palabras clave: cambio climático, sentimiento, redes sociales, PLN, España.

* Las autoras agradecen la financiación recibida por el proyecto *Economía del Cambio Climático: Vulnerabilidad y Políticas de Adaptación y Mitigación en España,* convocatoria RETOS 2019, referencia PID2019-111255RB-I00, que ha permitido articular la recogida de datos.

1. INTRODUCCIÓN

El cambio climático es uno de los grandes retos a los que se enfrenta la sociedad actual y sus consecuencias son cada día más evidentes. Según el Servicio de Cambio Climático Copernicus (C3S, 2023), el año 2023 se caracterizó por ser un año récord en cuanto a temperaturas registradas, siendo la media global de 14,98 °C, lo que lo convierte en el año más caluroso registrado desde que hay datos. Esta cifra es 0,17 °C superior al anterior récord, establecido en 2016. Además, la concentración de gases de efecto invernadero en la atmósfera alcanzó un nuevo máximo histórico. Según la Organización Meteorológica Mundial (OMM, 2023), la concentración de dióxido de carbono en la atmósfera alcanzó los 421,3 partes por millón (ppm), lo que supone un aumento del 1,8 % respecto a 2022. Dichos datos confirman que el cambio climático es una realidad que está afectando al planeta de forma cada vez más severa.

Los eventos climáticos extremos, tales como las olas de calor, las sequías, las inundaciones y los incendios forestales también están aumentando en intensidad y frecuencia en países vulnerables, como por ejemplo, en España. En concreto, la ola de calor que afectó a Europa en el verano de 2023 fue especialmente grave en este país. Las temperaturas alcanzaron los 45 ºC en algunas zonas, lo que provocó la muerte de cientos de personas. La sequía también fue un problema importante en España en 2023, en especial en la zona sur y mediterránea. Y los incendios forestales también fueron más frecuentes e intensos en este último año. En Galicia, por ejemplo, se produjeron más de 200 incendios forestales, que arrasaron miles de hectáreas de terreno. Estos eventos causan anualmente grandes pérdidas humanas y económicas, y evidencian la vulnerabilidad de las sociedades humanas al cambio climático.

La comunidad científica ha advertido sobre la necesidad de tomar medidas urgentes para frenar el cambio climático. A este respecto, la reciente cumbre del clima COP28 ha propuesto como objetivo inminente el reducir el consumo de combustibles fósiles, especialmente el carbón, que se eliminará de la generación eléctrica en 2030. Además, se destaca también la necesidad de reducir las emisiones en un 43 % hasta 2030, y un 50 % hasta 2035. Entre otras medidas, se promocionan también las energías renovables, que ya representan el 46,4 % de la generación eléctrica en España; además de mejorar la eficiencia energética, que reducirá el consumo de energía en un 30 % para 2030.

Ante la urgencia y rapidez necesaria para la exitosa adaptación y mitigación del cambio climático, las distintas administraciones e instituciones están elaborando medidas urgentes de mitigación y adaptación al mismo. Sin embargo, debido a su contexto y problemática de bien global, la adaptación climática solo será exitosa si involucra, cuando menos, a una gran parte de la sociedad. Por ello, se hace relevante entender las percepciones sociales hacia la cuestión climática.

En este sentido, los datos más recientes publicados por el Pew Research Center (2022) destacan cómo el cambio climático es considerado la principal amenaza a la que se enfrentan los territorios en un estudio realizado en 19 países. Por tanto, estas estadísticas muestran una preocupación pública hacia el cambio climático en Europa y en todo el mundo. Los datos

de estas estadísticas internacionales para España, reflejan tendencias similares. Por ejemplo, según datos del Eurobarómetro (2023) un 86 % de la población considera el cambio climático un problema muy serio (frente a un 81 % en 2021). Además, esta encuesta también muestra un amplio apoyo a medidas de acción climática, como la promoción de energías renovables, la reducción de emisiones y la adopción de políticas para combatirlo. Por otra parte, los datos del Pew Research Center (2022) muestran cómo alrededor del 78 % de los españoles consideran el cambio climático como una de las principales amenazas. Los resultados de la *European Social Survey* (ESS, 2020) también reflejan que el 55,2 % de la población estaba muy o extremadamente preocupada por el cambio climático en el año 2020. En resumen, tanto el Eurobarómetro, como los datos de la *European Social Survey,* así como, las encuestas del Pew Research Center sugieren que la preocupación hacia el cambio climático ha ido en aumento en España y en todo el mundo, con una creciente conciencia sobre la importancia de tomar medidas para enfrentar este desafío ambiental.

Las encuestas y estudios realizados en España también ofrecen resultados que están en línea con los obtenidos por las encuestas internacionales. Un estudio del Observatorio de la Sostenibilidad en España encontró que, aproximadamente, el 80 % de los encuestados consideraba que el cambio climático era un problema muy serio. En esta misma dirección apuntan los resultados de las encuestas llevadas a cabo por el Centro de Investigaciones Sociológicas (CIS, 2023) y el Real Instituto Elcano (2019), que han reflejado niveles significativos de preocupación sobre este tema entre la población española. En particular, el Real Instituto Elcano (2019) ha indicado que los españoles perciben el cambio climático como la mayor amenaza a nivel mundial que sufre la población. Además, la última encuesta del CIS de diciembre 2023, pone de relieve que para el 7 % de la población española, el cambio climático es el mayor problema actualmente.

Estos estudios también han señalado un aumento en la sensibilización y la voluntad de apoyar medidas concretas para abordar el cambio climático, como la transición hacia energías renovables y la adopción de políticas ambientales más sólidas. Además, en España se han observado movimientos sociales y protestas que reflejan la preocupación y el deseo de acción con respecto al cambio climático, con una participación activa en marchas y campañas dirigidas a la concienciación y la exigencia de políticas más ambiciosas para combatir este problema.

Dada la magnitud y relevancia del proceso del cambio climático, es sorprendente que a día de hoy solo contemos con información esporádica recopilada por distintas instituciones, algunas ya comentadas, y que las encuestas nacionales no recojan de forma sistemática la importancia del clima en las preocupaciones ciudadanas y cómo estas están condicionando sus elecciones.

A este respecto, este trabajo presenta una fuente de datos alternativa, que consideramos, que tratada con el cuidado requerido y entendiendo sus limitaciones, puede ser muy útil, en tanto en cuanto nos facilita una comparativa de la evolución de la percepción ciudadana hacia la problemática climática. Esta información obtenida en redes sociales es analizada con

herramientas de procesado de lenguaje natural (PLN) basadas en *machine learning,* que de forma automática pueden tratar cantidades ingentes de texto y generar indicadores, tanto de polaridad de opiniones como de gradualidad de las mismas. Para tal fin, recogemos datos de la plataforma social Twitter (ahora conocida como X) entre 2019-2022, y vemos cómo han evolucionado las percepciones hacia el cambio climático en España durante este período. Los resultados más destacados hacen referencia a que la preocupación está aumentando y, por tanto, están en línea con las estadísticas anteriormente comentadas. Por otro lado, la crisis sanitaria del COVID-19 y, en concreto, el primer confinamiento que tuvo lugar en España supuso una caída en el nivel de preocupación. Sin embargo, desde el fin de ese primer confinamiento se recuperaron los niveles de preocupación e incluso el año 2020 fue, en media, el año con un mayor nivel de preocupación en el período 2019- 2022. Finalmente, este análisis del sentir de la población hacia el cambio climático empleando conjuntamente variables climáticas como son las temperaturas (en concreto, las temperaturas máximas) o la ocurrencia de eventos extremos parecen ser también importantes impulsores de este incremento de la preocupación de la población.

La estructura de este trabajo es la siguiente: a continuación, en la sección segunda, se ofrece una revisión bibliográfica, que aunque somera, centra el tema de los métodos aplicados; en la sección tercera, se comenta el aspecto de análisis de datos que describe las técnicas fundamentales basadas en PLN. A continuación, en la sección cuarta, se describe la base de datos recogida y objeto de análisis; mientras que los resultados obtenidos se presentan en la sección quinta. Finalmente, concluye resumiendo los hitos alcanzados además de indicar futuras nuevas áreas a tratar para superar las limitaciones encontradas (sección 6).

2. REVISIÓN DE LA LITERATURA

La importancia del sentimiento se ha estudiado en la economía neoclásica y neokeynesiana de forma extensa, en variedad de aplicaciones. Algunos ejemplos relevantes son los trabajos del laureado Shiller (2017), que ha examinado cómo el relato y los sentimientos pueden influir en los mercados financieros, generando burbujas o situaciones de pánico, y cómo la irracionalidad puede apoderarse de los mercados (Baker y Wurgler, 2007). Algunos de sus trabajos muestran cómo los precios de los activos pueden divergir de los que resultarían tomando como base los fundamentos económicos de cada empresa, destacando, por lo tanto, el papel que juegan las emociones y las expectativas irracionales de los inversores. Hirshleifer *et al.* (2020) han investigado la influencia del sentimiento en la toma de decisiones financieras, contribuyendo al área de las finanzas conductuales, mostrando cómo las expectativas y el sentimiento condicionan la rentabilidad de distintos activos financieros. Desde la perspectiva del análisis macroeconómico, la literatura que relaciona el índice del sentimiento del consumidor y el gasto en consumo es muy extensa. Carroll *et al.* (1994) concluyen que dicha relación es muy significativa, encontrando que valores pasados del índice de confianza del consumidor por si solos, pueden explicar aproximadamente el 14 % de la variación del crecimiento de los gastos personales de consumo en el período posterior a 1954 en EE. UU.

El campo del análisis de sentimiento ha experimentado recientemente un gran auge, debido fundamentalmente a la posibilidad de analizar grandes bases de datos de forma rápida y automática, gracias a distintas herramientas computaciones basadas en *machine learning.* A este respecto, la mayoría de los análisis de sentimiento se basan en el estudio del discurso o contenido vertido en memorándums, informes o dictámenes, en el caso del sentimiento experto (Born *et al.,* 2014) o en el estudio de periódicos y redes sociales, si consideramos el análisis del sentimiento más popular (Rosenberg *et al.,* 2023). En el ámbito de la economía ambiental, el análisis del sentimiento es un campo con un gran potencial, pues permite el análisis de gran cantidad de texto de forma masiva que puede ser usado como indicador del pulso u opinión y ser comparado entre países de forma simultánea (Hase *et al.,* 2021), o incluso un indicador del impacto de determinados eventos extremos recientes, tales como la crisis del COVID-19 (Evensen *et al.,* 2021), de los cuales no necesariamente existen datos previos u otras transacciones de mercado. Es por ello, que el análisis del sentimiento de las conversaciones sobre el cambio climático en las redes sociales de España nos parece especialmente relevante.

La literatura previa más reciente en lo que análisis de sentimiento se refiere aplicado a la problemática del cambio climático todavía no es muy extensa. Loureiro y Alló (2020) llevaron a cabo una evaluación sobre cómo se expresaron sentimientos y emociones hacia el cambio climático en redes sociales en el Reino Unido y España, encontrando que los mensajes en el Reino Unido relacionados con el cambio climático son menos negativos que en España. Baylis (2020) aplicó el análisis de sentimiento en el contexto de las conversaciones sobre el cambio climático en EE. UU. Los resultados mostraron un efecto negativo entre las temperaturas más altas y el sentimiento general expresado. De manera similar, Loureiro *et al.* (2022) utilizaron el análisis de sentimiento para analizar el impacto de los incendios forestales en la ciudadanía, encontrando una relación negativa entre distancias próximas, la calidad del aire y el humo en el sentimiento expresado.

En general, estos estudios sugieren que las preocupaciones hacia el cambio climático y las preferencias públicas por instrumentos políticos están conformadas por una amplia gama de factores, incluyendo percepciones sobre efectividad, equidad y costes relacionados. Al comprender estas preferencias, se pueden diseñar políticas climáticas que sean más propensas a ser aceptadas y apoyadas, y que puedan conducir a acciones relevantes para abordar este problema apremiante.

3. LAS TÉCNICAS DE PROCESADO DE LENGUAJE NATURAL (PLN)

El análisis de sentimiento es una técnica de procesamiento del lenguaje natural (PLN) que se utiliza para identificar y extraer información de textos. En términos generales, el análisis de sentimiento intenta determinar la actitud abiertamente declarada o subyacente con respecto a algún tema concreto. En este sentido, puede ofrecer una clasificación sobre la polaridad de un texto concreto, oración, o párrafo, donde se expresa un mensaje negativo, positivo o neutral. Este análisis de sentimiento se puede utilizar para una variedad de propósitos,

como por ejemplo, analizar las opiniones o percepciones sociales sobre temas candentes de interés, monitorear las tendencias sociales y de opinión, o ayudar a sintetizar opiniones y esto como ya se ha mencionado anteriormente, puede ayudar a las organizaciones a comprender el clima social y a identificar oportunidades o riesgos potenciales.

Existen dos enfoques principales para el análisis de sentimiento:

- Basado en reglas: este enfoque utiliza un conjunto de reglas predefinidas para identificar palabras y frases que se asocian con emociones positivas, negativas o neutras.

- Aprendizaje automático: este enfoque utiliza un algoritmo de aprendizaje automático para aprender a identificar las emociones en el texto.

El enfoque basado en reglas es más simple y eficiente, pero puede ser menos preciso que el aprendizaje automático. El aprendizaje automático es más complejo y requiere más datos de entrenamiento, pero puede ser más preciso en la identificación de emociones complejas.

En este capítulo, realizamos análisis de sentimiento utilizando las librerías de VADER (*Valence Aware Dictionary and Sentiment Reasoner*) (Hutto y Gilbert, 2014) y LabMT (*Language Assessment by Mechanical Turk*) (Dodds *et al.,* 2011). En ambos casos estamos usando diccionarios de análisis de sentimiento basados en reglas. Estos son una buena opción para el análisis de sentimiento de texto corto o conciso, como el de las redes sociales. Sin embargo, pueden ser menos precisos que los métodos basados en aprendizaje automático para el análisis de sentimiento de texto largo o complejo.

VADER es un modelo que analiza sentimientos basándose en reglas y léxico (tiene un conjunto de palabras con puntuaciones de sentimiento asignadas) y es sensible tanto a la polaridad como a la intensidad de la emoción. Es decir, nos permite identificar la polaridad de un texto, esto es, si el texto es positivo, negativo o neutral; pero también permite identificar la intensidad de la emoción expresada en el texto. Es especialmente útil para el análisis de datos que provienen de redes sociales dado que es capaz de entender el lenguaje coloquial, los emoticonos, y otras expresiones informales. Es decir, combina el uso de un diccionario de palabras junto con cinco heurísticas que tienen en cuenta la puntuación, el uso de mayúsculas, los modificadores de grado, el "pero", la doble negación. La puntuación del sentimiento que nos proporciona oscila entre -1 y 1, siendo -1 una puntuación muy negativa, y 1 una puntuación muy positiva.

En el caso de LabMT, el análisis se basa en un diccionario que contiene 10.000 palabras con puntuaciones que indican el nivel de felicidad expresada y, que previamente han sido evaluadas por individuos a través de la plataforma Amazon Mechanical Turk. Por tanto, el procedimiento consiste en asignar a cada palabra una puntuación y, posteriormente, calcular un promedio para indicar el nivel de felicidad del texto. La puntuación que proporciona LabMT para cada texto oscila entre 1 y 9, siendo 1 un sentimiento muy negativo o triste, y 9 un sentimiento muy positivo o feliz.

En ambos casos, el interés es obtener un indicador de sentimiento general de cada uno de los mensajes (o valencia general). Este se obtiene agregando individualmente las puntuaciones individuales de cada palabra contenida en cada mensaje. A mayores, en el caso de VADER también se consideran las reglas heurísticas pertinentes, que enfatizan tanto el sentimiento positivo como negativo de las palabras individuales.

En este capítulo, utilizamos análisis de sentimiento para entender la evolución de las percepciones hacia el cambio climático en la sociedad española durante 2019-2022. Por tanto, recopilamos conversaciones de las redes sociales (en concreto Twitter) y, posteriormente, les aplicamos análisis de sentimiento. Consideramos que este tipo de análisis es interesante, dado que no existen estudios previos que analicen las percepciones hacia el cambio climático durante un período anterior y posterior a la crisis sanitaria del COVID-19 que, por un lado, se focalicen en el efecto del confinamiento y, por otro lado, utilicen los datos de las redes sociales para tal fin.

4. DATOS

Con el fin de llevar a cabo el análisis de sentimiento aplicado al estudio de la percepción del cambio climático, se recopilan las conversaciones geoetiquetadas de la red social Twitter (ahora conocida como X) en España durante el período 2019-2022. En concreto, se analizan los mensajes geoetiquetados recuperados a partir de una cadena de búsqueda booleana que contiene las siguientes palabras clave "cambio climático", "incendios", "inundaciones", "sequía", "calentamiento global", "clima", "huracán", "tsunami", "tornado". La figura 1 muestra la geolocalización de los tuits en el territorio y, como se puede observar se dispone de infor-

Figura 1.

Geolocalización de los tuits analizados

Fuente: Elaboración propia.

mación para todo el país. Los indicadores de sentimiento pueden generarse a nivel usuario, o bien a un nivel de agregación superior, considerando para ello las coordenadas de las distintas conversaciones.

Una vez recopilados los tuits, se procede a limpiar y procesar los datos para poder analizar las conversaciones. El primer paso requiere limpiar el texto, y eliminar la información no relacionada con el cambio climático. Por ejemplo, se eliminan los tuits relacionados con canciones, refranes, que aunque incluyen palabras como "temperatura", "clima", "calor", es información que no está relacionada con el objeto de estudio. Tampoco se tienen en cuenta los "retuits" dado que son una repetición de otro tuit, es más, en el caso de que un tuit fuese una respuesta a otro tuit, se incluyeron ambos. Así, tras una primera limpieza, se procede a poner todo el texto en minúsculas, limpiar los acentos, eliminar las palabras vacías, obteniendo una base de datos final que contiene 2.932.421 tuits escritos durante el período 2019-2022.

Este conjunto de datos final contiene información sobre la fecha y hora en la que se escribe cada tuit, la ubicación desde la que se ha escrito (a través de las coordenadas geográficas), además del texto escrito. Posteriormente y, con objeto de analizar el sentimiento hacia el cambio climático, se procede con el cálculo de la puntuación de sentimiento para cada uno de los tuits. Es importante mencionar que los tuits se recogieron y analizaron en español, su idioma original, dado que los dos diccionarios empleados (VADER y LabMT) ya cuentan con versiones que permiten analizar directamente texto escrito en habla hispana. Así, a la base de datos inicial se le añade el sentimiento como una medida numérica.

Finalmente, y con el fin de explorar la posible relación entre el sentimiento expresado con referencia al cambio climático y variables climáticas importantes, se recogen datos de la temperatura máxima que estaba experimentado el usuario en el día y a la hora que estaba escribiendo el tuit (WeatherAPI, 2022). También se añade un indicador que refleja la intensidad de los eventos extremos relacionados con el cambio climático que tuvieron lugar en la zona desde la que se escribe el tuit. Estos eventos hacen referencia a los incendios, las olas de calor, las inundaciones o las sequías recopilados de la base de datos International Disaster Database (Guha-Sapir *et al.,* 2023). Sin embargo, y con el fin de incluir un número más amplio de eventos relacionados con el cambio en el clima, se recogieron otros fenómenos tales como borrascas (Celia, Armand, entre otras) o la intrusión de polvo sahariano en el mes de marzo de 2022 (estos son solo algunos ejemplos) a través de la herramienta de búsqueda de Google. Por consiguiente, hemos establecido un indicador que nos ayuda a discernir las áreas en España más impactadas por desastres derivados del cambio climático. Este indicador se relaciona potencialmente con las emociones y opiniones expresadas en las redes sociales sobre el mismo.

5. RESULTADOS

5.1. La temática de las conversaciones

Inicialmente, es fundamental plantearnos: ¿a qué nos referimos exactamente al hablar de cuestiones relacionadas con el cambio climático? ¿Cómo ha evolucionado este discurso

a lo largo del tiempo? El análisis de datos de las redes sociales nos permite identificar qué dinámicas hay detrás de estas conversaciones públicas. La figura 2 nos muestra las nubes de palabras obtenidas para cada uno de los años de estudio. Como se puede observar, los tópicos más importantes tras estas conversaciones están relacionados con los desastres naturales: incendios, huracanes, tsunamis, sequías, inundaciones, y otros eventos extremos. Por tanto, de forma general, en redes sociales la población expresa fundamentalmente su preocupación por desastres/eventos extremos que afectan gravemente a las personas y a los territorios (Cody *et al.,* 2015). Además, tal y como se observa en la figura, este contenido no ha variado de forma significativa a lo largo de los años, puesto que las palabras más repetidas permanecen invariables.

Figura 2.

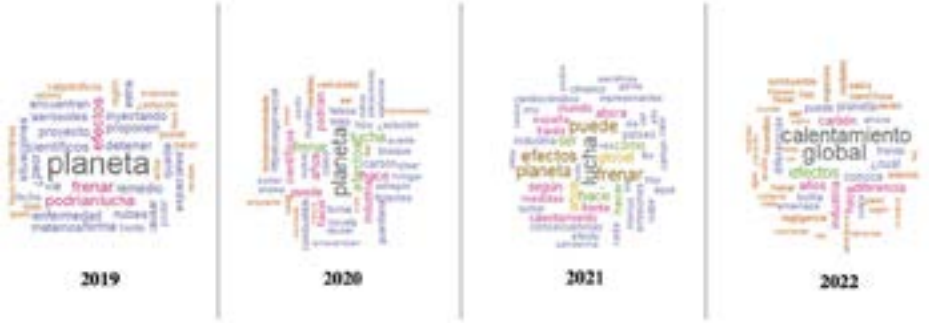
Nube de palabras 2019-2022

Fuente: Elaboración propia.

Si realizamos un análisis más detallado y nos centramos únicamente en las conversaciones que incluyen las palabras "cambio climático", entonces sí se observa un cambio en el discurso de la población española a lo largo de estos cuatro años (figura 3). Mientras que en el año 2019 se hablaba de "planeta", "efectos", "frenar", "lucha"; en el año 2020 aparecen palabras como "científicos" o "industria", es decir, las nubes de palabras sugieren que la población empieza a considerar en sus conversaciones parte de las causas, así como también soluciones potenciales. El rasgo común es que en ambos años la palabra más frecuente es "planeta". A partir del año 2021 domina la palabra "lucha", por tanto, la sociedad parece estar más activa y concienciada acerca de la necesidad y la importancia de luchar contra el cambio climático. Esto sugiere que quizás la pandemia de la COVID-19 pudo ser un impulsor, marcando un punto de inflexión, en la preocupación acerca de este problema. Finalmente, analizando las conversaciones del año 2022, la palabra más frecuente es "calentamiento global", por tanto, una vez más los datos sugieren un mayor nivel de concienciación de la sociedad y una mayor proactividad para, quizás adoptar medidas de lucha más contundentes. Drews *et al.* (2002) llevaron a cabo una encuesta en España antes y después de la pandemia del COVID-19, encontrando que el nivel de concienciación decrecía, pero el nivel de aceptabilidad de políticas aumentaba (los resultados después de la COVID-19 se referían a respuestas obtenidas en el mes de junio de 2020).

Figura 3.

Nube de palabras 2019-2022: cambio climático



Fuente: Elaboración propia.

5.2. Evolución del sentimiento hacia el cambio climático: el efecto de la crisis sanitaria COVID-19

Las estadísticas disponibles actualmente sugieren un creciente nivel de concienciación y preocupación en la población respecto al cambio climático. No obstante, ¿qué revela el análisis del sentimiento expresado en las redes sociales sobre esta temática? Este enfoque nos permite explorar más allá de las cifras generales, adentrándonos en la comprensión profunda de las actitudes y emociones que la sociedad manifiesta digitalmente acerca del cambio climático.

El cuadro 1 nos muestra la evolución del sentimiento en España, medido a través de dos diccionarios VADER y LabMT, hacia el cambio climático en los cuatro años de estudio. En ambos casos, cuanto mayor es la puntuación, mayor es el sentimiento, y nos indica que más ”feliz” está el individuo que está escribiendo el texto. En primer lugar, cabe comentar que los datos se concentran en el centro de la escala para ambas medidas (un valor de 5 en LabMT representaría un sentimiento neutral, y niveles superiores a 5 indican un sentimiento más positivo), estos resultan así para España, pero para el resto de los países para los cuales tenemos datos se obtienen cifras similares. En el caso de LabMT disponemos de más fuentes para comparar los datos, por ejemplo, Dodds *et al.* (2015) publican datos del sentimiento de las conversaciones en general (sin centrarse en el tema del cambio climático) en diferentes idiomas (inglés, español, francés, entre otros) y las puntuaciones oscilan, en general, entre un 5.6 y un 6,3. En segundo lugar y a la vista de los datos, el año 2020 es el año con menor sentimiento, por tanto, mayor preocupación hacia el cambio climático y todos sus efectos. Los años 2021 y 2022 muestran un nivel de sentimiento mayor al del año 2020, sin embargo, en ningún caso superior al nivel del año 2019. Los resultados están en línea con las estadísticas anteriormente mencionadas y confirman que el nivel de preocupación habría aumentado con el paso de los años. También en cierta medida, los resultados obtenidos producen evidencia acerca de la existencia del *finite pool of worry hypothesis,* esto es, la reserva limitada del número de preocupaciones, por las que el ser humano tiende a priorizarlas para sobrevivir y ganar calidad de vida (Weber, 2006; Evensen *et al.,* 2021), igual que ha sucedido en otros países europeos.

Cuadro 1.

Evolución del estimador de sentimiento hacia el cambio climático empleando LabMT y VADER

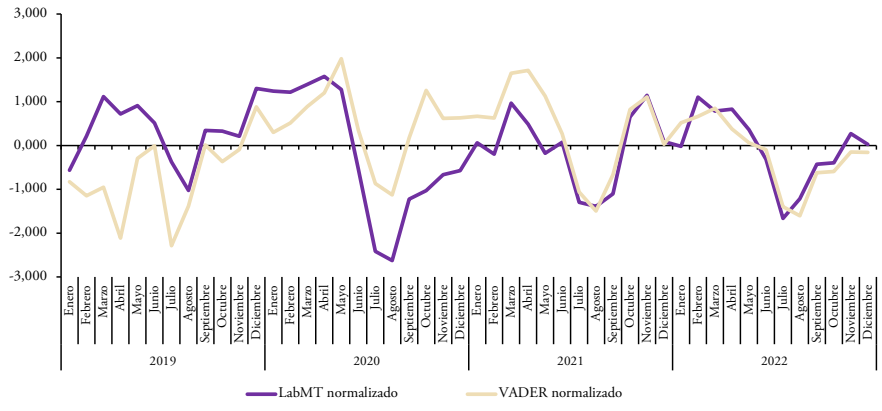
Año	*LabMT*	*VADER*
2019	5.768	-0.035
2020	5.731	-0.004
2021	5.741	-0.006
2022	5.742	-0.021

Fuente: Elaboración propia.

Con el fin de analizar el efecto de la pandemia, estudiamos más en detalle la evolución mensual del sentimiento. La figura 4 nos muestra los resultados detallados y nos permite concluir que durante el confinamiento que comenzó el 14 de marzo de 2020 y terminó el 21 de junio de 2020 en España, el nivel de sentimiento de la sociedad española en conversaciones sobre cambio climático aumentó, es decir, las conversaciones sobre cambio climático eran "más felices". Sin embargo, ya en el mes de junio de ese año el sentimiento cae indicando un incremento en el nivel de preocupación, una vez se termina el confinamiento (esto se confirma para las dos medidas analizadas). Estos resultados se alinean razonablemente con el contexto social durante el período de confinamiento. Durante esta fase, las discusiones sobre el cambio climático en la sociedad eran menos preocupantes, probablemente debido a la priorización de asuntos más inmediatos como la salud pública y las preocupaciones económicas. Sin embargo, es importante destacar que, paralelamente, la sociedad tomó conciencia de los

Figura 4.

Evolución del estimador de sentimiento



Fuente: Elaboración propia.

efectos positivos en el medio ambiente derivados de la reducción de la actividad económica, dado que al finalizar el período de cierre la preocupación se incrementó. Esta observación sugiere que, una vez levantadas las restricciones, pudo haber un creciente reconocimiento y sensibilización respecto a la importancia del cambio climático y la necesidad de adoptar medidas pertinentes.

Con el fin de profundizar en estas diferencias antes, durante y después del confinamiento, vamos a analizar más en detalle si existen diferencias estadísticamente significativas entre los índices de sentimiento estimados. El cuadro 2 nos muestra las medias para las dos medidas consideradas (VADER y LabMT), observando cómo durante el confinamiento la puntuación de sentimiento es mayor, es decir, en media, la preocupación durante el período de principal restricción fue menor que durante el período anterior y posterior. Esto se ha puesto de manifiesto en varios estudios internacionales, mostrando evidencia a favor de la reserva limitada de preocupaciones.

Cuadro 2.

El sentimiento hacia el cambio climático antes, durante y después del COVID-19
Durante: período de confinamiento 14/03-21/06/20

	LabMT	*VADER*
Antes	5.778	-0.019
Durante	5.831	0.013
Después	5.712	-0.016

Fuente: Elaboración propia.

Con el fin de testar si hay diferencias significativas entre los tres grupos considerados, realizamos la prueba de Kruskal-Wallis (1952)[1]. Empleamos esta estadística no paramétrica dado que no se cumplen los supuestos de normalidad y homogeneidad de varianza, además de que estamos analizando dos variables de sentimiento ordinales. La hipótesis nula es que no existen diferencias significativas entre las medianas de los tres grupos y la hipótesis alternativa es que al menos en un grupo difiere significativamente de los demás. Por tanto:

$$H_0: \mu_{antes} = \mu_{durante} = \mu_{después} \quad [1]$$

$$H_1: \textit{No todas las medianas } \mu_i \textit{ son iguales} \quad [2]$$

donde μ_{antes}, $\mu_{durante}$ y $\mu_{después}$ representan las medianas del sentimiento antes, durante y después del confinamiento, respectivamente.

A la vista de los resultados del cuadro 3, sí existen diferencias estadísticamente significativas entre al menos uno de los grupos.

[1] Se empleó la prueba de Kruskal-Wallis para analizar las diferencias entre grupos dadas las características de los datos, a pesar de que en el capítulo se comentan las medias, debido a la presencia de valores extremadamente bajos que resultaron en una mediana igual a cero en algunos grupos.

Cuadro 3.

Resultados de las pruebas de Kruskal-Wallis para LabMT y VADER

Variable	*Estadístico Chi-Cuadrado*	*Valor p*
LabMT	2533.6	< 0.001
VADER	666.22	< 0.001

Fuente: Elaboración propia.

Seguidamente y con el fin de identificar en cuáles de los grupos existen dichas diferencias, realizamos las pruebas post-hoc. En concreto, se utiliza la prueba Dunn-Bonferroni (Dunn, 1961). Los resultados muestran que las diferencias son estadísticamente significativas entre los tres grupos analizados. Por tanto, confirmamos que el nivel de preocupación por el cambio climático disminuyó durante el período de confinamiento; sin embargo, una vez finalizada la restricción, el nivel de preocupación aumenta y sigue en el año 2022 en niveles superiores a los del año 2019.

Cuadro 4.

Resultados de las pruebas post-hoc para LabMT y VADER

	LabMT	*VADER*
Antes vs. después	< 0,001	< 0,001
Durante vs. después	0,000	< 0.001

Fuente: Elaboración propia.

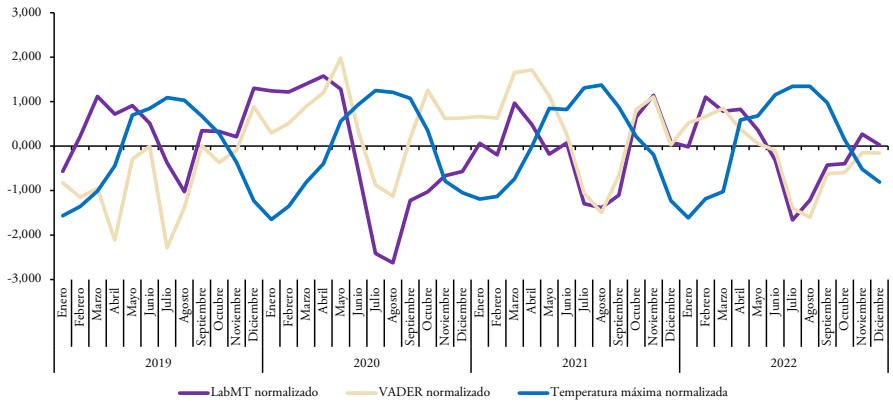
5.3. La evolución del sentimiento: las temperaturas máximas y los eventos extremos

Para profundizar en el estudio sobre la percepción del cambio climático, ahora nos enfocamos hacia el impacto potencial de las temperaturas máximas y los eventos climáticos extremos. Esta línea nos permitirá evaluar cómo estos factores ambientales específicos pueden influir en la opinión pública y en la sensibilización hacia el cambio climático. En la figura 5 se presenta la evolución del sentimiento empleando los dos diccionarios ya comentados, así como también la evolución de las temperatura máximas (medias mensuales) (las variables están normalizadas para que la media sea 0 y la desviación estándar 1, favoreciendo así la comparación). Dado que tenemos datos desagregados del sentimiento por individuo, lo que aquí se presenta es el sentimiento agregado a nivel de país y lo mismo sucede con las medias de la temperatura. Tres cuestiones son destacables: 1) las dos medidas empleadas siguen una tendencia similar; 2) en general, se detecta un incremento del nivel de preocupación (caída del sentimiento), es decir, las puntuaciones o los indicadores de sentimiento de LabMT y VADER parecen seguir una tendencia decreciente; 3) durante los meses de verano el nivel de preocupación aumenta de forma considerable. En concreto, los meses de julio y agosto marcan los mínimos de sentimiento, es decir, serían los niveles más altos de preocupación. Este último resultado coincide justamente con los picos de más calor, es decir, los

máximos de las temperaturas se producen en los meses de verano. Por tanto, es probable que la percepción de la sociedad sobre el aumento de las temperaturas esté contribuyendo a una mayor conciencia del peligro que representa el cambio climático. Este resultado está en línea con el obtenido por Baylis (2020) para EE. UU. En el caso de VADER se observa un máximo del indicador de sentimiento muy claro en los meses de inicio del COVID-19 y confinamiento, lo cual, esta medida parece haber detectado, en mayor medida, que las conversaciones sobre cambio climático en este período eran mucho menos preocupantes; sin embargo, a partir del mes de mayo el sentimiento comienza a caer, lo que implica que la preocupación vuelve a aumentar.

Figura 5.

Evolución del sentimiento y de la temperatura máxima promedio



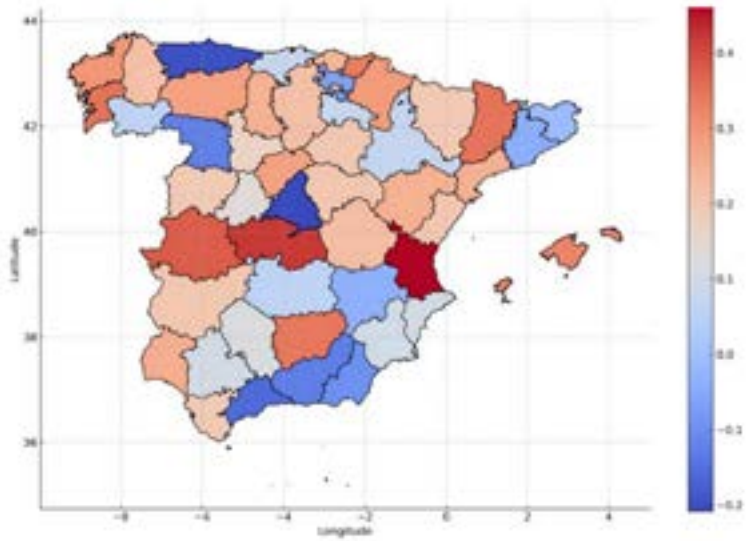
Fuente: Elaboración propia.

Para ahondar en la relación entre las temperaturas máximas y el nivel de sentimiento, la figura 6 muestra la correlación entre ambas variables a nivel de provincia y una media agregada para los cuatro años de estudio (se representa únicamente la medida obtenida con VADER por razones de simplicidad). Los colores más oscuros indican un mayor nivel de correlación, siendo lo colores azulados los que indican una correlación negativa y los colores más rojizos una correlación positiva. A la vista de la figura, se evidencia cómo Madrid o las provincias situadas más al sur, incluso zonas más al norte como Asturias (con gran número de incendios forestales), Ourense (zona que sufre altas temperaturas) muestran una mayor correlación negativa, es decir, a temperaturas más altas, mayor nivel de preocupación, por tanto, menor sentimiento.

La figura 7 ilustra la evolución de la percepción pública sobre el cambio climático en distintas regiones climáticas, definidas estas desde una perspectiva tradicional. Esto es, el clima oceánico comprende las CC. AA. de Galicia, Asturias, Cantabria y País Vasco. En el caso del clima mediterráneo estamos teniendo en cuenta a Extremadura, Andalucía,

Figura 6.

Correlación entre el sentimiento medido con VADER y la temperatura máxima promedio



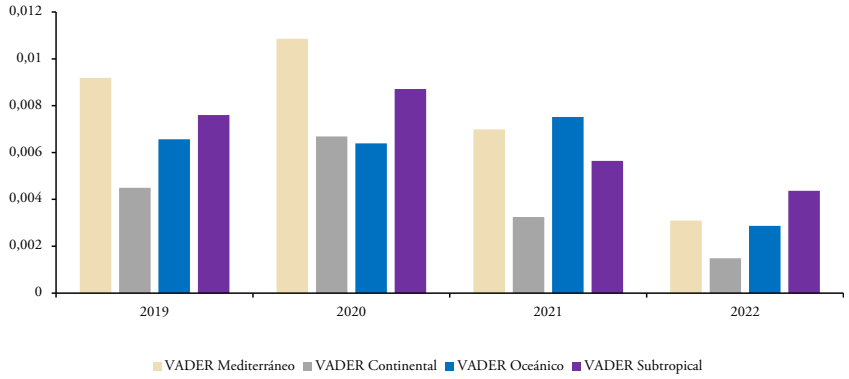
Fuente: Elaboración propia.

Comunidad Valenciana, Cataluña e Islas Baleares. Las zonas de Castilla-La Mancha, Castilla y León, Aragón, La Rioja y Navarra se encuadrarían en el clima continental, mientras que las Islas Canarias conformarían las áreas de un clima subtropical. En los últimos años y a consecuencia de los cambios en los patrones climáticos, las zonas climáticas tradicionales son cada más difusas. Por ejemplo, Del Río *et al.* (2005) señalaron cómo en el caso de Castilla y León, los cambios detectados permitían establecer un clima más oceánico que continental en esta zona. En este sentido y a la vista de la figura 7, se detecta que las zonas con un clima continental son las que representan un menor nivel de sentimiento con respecto al resto de áreas (es decir, muestran un mayor nivel de preocupación). Y una vez más, se observa cómo, si comparamos el nivel de sentimiento a lo largo de estos cuatro años de estudio, en general, ha decrecido, lo que nos indica que la población española está cada vez más preocupada por el cambio climático. En concreto, se observa cómo la zona con un clima mediterráneo es la zona donde el nivel de preocupación se ha incrementado de forma más considerable (mayor caída del nivel de sentimiento).

El segundo factor que nos parece interesante estudiar es el papel que juegan los eventos extremos como consecuencia del cambio climático en la opinión pública. A este respecto, la figura 8 muestra la correlación existente entre los eventos extremos y el nivel de preocupación (el sentimiento) a nivel provincial durante los años analizados en este trabajo. Es importante destacar que, en el caso de los desastres puede haber retardos, dado que

Figura 7.

Evolución del sentimiento según las regiones climáticas tradicionales (VADER)

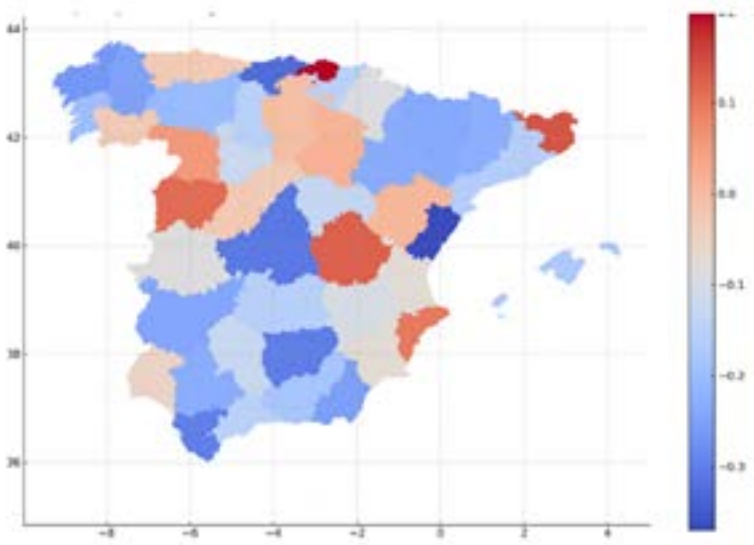


Fuentes Elaboración propia.

la sociedad puede estar escribiendo acerca de un episodio días o incluso meses posteriores al acontecimiento del mismo. En este caso, con el fin de reflejar los eventos extremos se recopila información de las zonas que sufrieron desastres relacionados con el clima (olas

Figura 8.

Correlación entre el sentimiento medido con VADER y los desastres climáticos



Fuente: Elaboración propia.

de calor, incendios, borrascas, etc.) en los años de estudio, tal y como ya se comentado previamente. Por tanto, lo que se tiene en cuenta es un indicador que nos identifica el grado de afectación de las provincias españolas por desastres climáticos junto con el sentimiento medio agregado para cada provincia. Analizando esta figura, hay que tener en cuenta que los colores azulados, representan una relación negativa entre ambos indicadores, siendo esta relación más fuerte cuanto más oscuro es el color. Los colores más rojizos nos mostrarían una relación positiva (que en ningún caso es estadísticamente significativa) y los colores grisáceos nos mostrarían ausencia de correlación. Así, en general, encontramos una relación negativa entre ambas variables, destacando una mayor correlación negativa en provincias como Madrid (zona que en los últimos años ha sufrido un incremento considerable de las temperaturas, temporales como Filomena, etc.), Asturias o Galicia (zonas que en los últimos años han sufrido un incremento en el número de incendios, así como de la severidad de los mismos, u olas de calor, entre otros).

Para finalizar, hay que destacar cómo este tipo de análisis nos proporciona información valiosa acerca de cómo las experiencias personales y colectivas relacionadas con el clima pueden influir en la percepción y actitudes ambientales.

6. CONCLUSIONES

Es evidente que el cambio climático se ha convertido en un tema de interés público y que las redes sociales desempeñan un papel crucial en la difusión de información y en la formación de opiniones. Considerando el análisis de la percepción de la ciudadanía española sobre el cambio climático a través de las redes sociales, se pueden extraer conclusiones significativas. En especial, la falta de datos actualizados requiere que busquemos fuentes alternativas, especialmente para entender circunstancias actuales (*nowcasting*). Esto hace que el estudio de sentimiento basado en las interacciones en plataformas digitales emerja como una herramienta valiosa y ágil para comprender las dinámicas de preocupación y evolución de actitudes hacia distintas problemáticas, incluyendo el cambio climático. Además, el análisis de tópicos nos permite entender la evolución de las distintas temáticas que motivan la percepción global.

En línea con otras fuentes de datos estadísticas más tradicionales, los resultados de este trabajo muestran una creciente conciencia acerca de los desafíos ambientales, y las redes sociales se revelan como un reflejo directo de dicha conciencia. La capacidad de obtener información prácticamente en tiempo real (*nowcasting*) nos permite comprender no solo el nivel de preocupación de la población española, sino también cómo este sentimiento ha variado a lo largo del tiempo. En esta aplicación concreta observamos cambios notorios en el período anterior y posterior a la emergencia sanitaria del COVID-19, mostrando evidencia a favor de la conocida "hipótesis de reserva finita de preocupaciones" (Weber, 2006).

En esta aplicación utilizamos dos diccionarios muy comunes en el análisis de sentimiento (VADER y LabMT). Sin embargo, la robustez de los resultados presentes también

puede ser evaluada con nuevos diccionarios que están ganando popularidad. A su vez, el desarrollo de nuevos métodos basados en inteligencia artificial y *machine learning* aplicados al análisis de sentimiento permitirán la mejora de las herramientas y técnicas actuales, de forma que el procesado de datos y su comprensión pueda ser aún más rápido y más útil para la investigación. En resumen, esperamos que estudios futuros complementen estos resultados aquí presentados.

Referencias

Baker, M. y Wurgler, J. (2007). Investor sentiment in the stock market. *The Journal of Economic Perspectives, 21*(2), pp. 129–151.

Baylis, P. (2020). Temperature and temperament: evidence from twitter. *Journal of Public Economics,* 184, 104161.

Born, B., Ehrmann, M. y Fratzscher, M. (2014). Central bank communication on financial stability. *Economic Journal,* 124, pp. 701–734.

Carroll, C. D., Fuhrer, J. C. y Wilcox, D. W. (1994). Does consumer sentiment forecast household Spending? If So, Why? *The American Economic Review, 84*(5), pp. 1397–1408.

Centro de Investigaciones Sociológicas (CIS). (2023). Barómetro de diciembre 2023. Estudio nº 3431.

Cody, R., Reagan, A. J., Mitchell, L., Dodds, P. S. y Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLOS ONE, 10*(8), pp. e0136092- e0136092.

Comisión Europea. (2023). Eurobarómetro Especial 538: Cambio Climático. Bruselas: Comisión Europea.

Dodds, P. S., Clark, E. M. y Desu, S. (2015). Human language reveals a universal positivity bias. *PNAS, 112*(8) pp. 2389-2394.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. y Danforth, C. M. (2011). Temporal Patterns of Happiness and Information in a Global-Scale Social Network: Hedonometrics and Twitter. *PLoS ONE, 6*(12), pp. e26752.

Drews, S., Savin, I., van den Bergh, C. J. M. y Villamayor-Tomás, S. (2022). Climate concern and policy acceptance before and after COVID-19. *Ecological Economics,* 199. pp. 107507.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association,* 56. pp. 52–64.

European Social Survey. (ESS). (2020). ESS round 10 - 2020. Democracy, Digital Social Contacts.

Evensen, D., Whitmarsh, L., Bartie, P., Devine-Wright, P., Dickie, J., Varley, A., ... y Mayer, A. (2021). Effect of "finite pool of worry" and COVID-19 on UK climate change perceptions. *Proceedings of the National Academy of Sciences, 118*(3), e2018936118.

Guha-Sapir, D., Hoyois, P., Wallemacq, P. y Below, R. (2023). EM-DAT: International Disaster Database. Université Catholique de Louvain (UCL)-CRED, D. Guha-Sapir, Brussels, Belgium. www.emdat.be

Hase *et al.* (2021). Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). *Global Environmental Change,* 70, 102353.

Hirshleifer, D., Jiang, D. y DiGiovanni, Y. M. (2020). Mood beta and seasonalities in stock returns. *Journal of Financial Economics, 137*(1), pp.272-295.

Hutto, C. J. y Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14).* Ann Arbor, MI, June 2014.

Kruskal, W. H. y Wallis, A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association , 47*(260), pp. 583–621.

Loureiro, M. L. y Alló. M. (2020). Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain. *Energy Policy,* 143, pp. 111490.

Loureiro, M. L., Alló, M. y Coello, P. (2022). Hot in Twitter: assessing the emotional impacts of wildfires with sentiment analysis. *Ecological Economics,* 200, pp. 107502.

Organización Meteorológica Mundial (OMM). (2023). https://wmo.int/es

Pew Research Center. (2022). Spring 2022 Global Attitudes Survey. Q10a-e. https://www.pewresearch.org/global/2022/08/31/climate-change-remains-top-global-threat-across-19-country-survey/

Real Instituto Elcano. (2019). *Los españoles ante el cambio climático. Apoyo ciudadano a los elementos, instrumentos y procesos de una Ley de Cambio Climático y Transición Energética.*

Rio, S. del, Penas, Á. y Fraile, R. (2005). Analysis of recent climatic variations in Castile and Leon (Spain). *Atmospheric Research, 73*(1-2), pp. 69-85.

Rosenberg *et al.* (2023). Sentiment analysis on Twitter data towards climate action. *Results in Engineering,* 19, 101287.

Servicio de Cambio Climático Copernicus (C3S). (2023). *European State of the Climate 2022 Report.* https://climate.copernicus.eu/esotc/2022

Shiller, R. (2017) Narrative Economics. *American Economic Review, 107*(4), pp. 967-1004.

Weather API. (2022). https://www.meteomatics.com/en/api/overview/

Weber, E. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change,* 77, pp. 103–120.

CAPÍTULO III

Diferencias provinciales en la evolución del índice de precios al consumo

Antonio Montañés*

Este documento aborda el análisis de la evolución de los precios en España utilizando datos desagregados tanto geográfica (52 provincias) como temporalmente (período 2002:1-2023:9). Para ello, se ha utilizado una metodología que permite contrastar tanto la presencia de un único patrón de comportamiento común (convergencia) como la presencia de varios clubes de convergencia. Los resultados muestran que, para el total del índice de precios al consumo, existen hasta cinco patrones de comportamiento significativamente diferentes. Este resultado se mantiene si se descompone este índice en sus componentes principales, a excepción de los de bebidas alcohólicas y tabaco (G02) y de transporte (G07).

Palabras clave: precios, disparidades provinciales, convergencia, factores comunes, Phillips-Sul.

* Agradezco la invitación cursada por D. Peña y M. P. Poncela para participar en este evento. La actual versión se ha visto beneficiada por los comentarios de un evaluador anónimo.

1. INTRODUCCIÓN

Uno de los avances más destacados en el campo del análisis cuantitativo es la amplia disponibilidad de datos a la que los investigadores tienen acceso en la actualidad. El aumento en la capacidad de almacenamiento, combinado con el mayor rendimiento de las computadoras, ha llevado a los investigadores a adaptarse a estas nuevas circunstancias. En muchos casos, esto ha dado lugar al desarrollo de nuevas técnicas que permiten el manejo de datos masivos. Es crucial destacar las mejoras significativas en los métodos de *machine learning,* las capacidades mejoradas de las técnicas de regresión en paralelo y los algoritmos de optimización más rápidos y eficientes en comparación con los enfoques tradicionales, entre otras ventajas.

Sin embargo, también resulta necesario reconocer que esta disponibilidad masiva de datos también ha planteado ciertos desafíos. Por ejemplo, los investigadores han tenido que desarrollar métodos para seleccionar de manera adecuada las variables a utilizar. Entre estos métodos, se deben mencionar en particular el LASSO o *Adaptative Lasso* (Zhou, 2006; Chatterjee *et al.,* 2015), así como el uso de factores comunes para resumir la información contenida en un conjunto amplio de variables, como lo hacen Stock y Watson (2002) en el ámbito de la predicción.

Esta abundancia de datos y las nuevas herramientas econométricas han beneficiado significativamente las investigaciones en áreas como el análisis financiero o la medicina, especialmente en epidemiología. No obstante, posiblemente el campo que ha aprovechado mejor estas bases de datos masivas sea la economía. En particular, los estudios de convergencia han experimentado un impulso significativo. Los trabajos pioneros de Baumol (1986) o Barro y Sala-i-Marti (1992) buscaban analizar la posible existencia de procesos de *catching-up,* mediante los cuales el producto interno bruto per cápita de una región tiende a acercarse con el tiempo al de otra región más próspera, reduciendo las distancias inicialmente existentes y, en algunos casos, llegando a converger con este último.

Recientemente, Phillips y Sul (2007) han desarrollado métodos que permiten estudiar la hipótesis nula de convergencia. Este método es muy flexible en cuanto a las características temporales de las variables y, lo que es crucial en nuestro caso, se adapta de manera eficiente a la existencia de volúmenes de información elevados, facilitando el estudio de un patrón de comportamiento común para datos muy desagregados tanto geográfica como temporalmente.

El objetivo de este trabajo es utilizar precisamente esta herramienta para analizar la evolución de los precios en España, abarcando un intervalo temporal extenso y, sobre todo, una desagregación territorial tan amplia como lo permitan los datos. En nuestro caso, se utilizan datos provinciales, pero la metodología es robusta y podría aplicarse a desagregaciones más detalladas, como comarcas o áreas metropolitanas, aprovechando las ventajas de disponer de datos de alta frecuencia.

Para lograrlo, el resto del trabajo se organiza de la siguiente manera. En la próxima sección se presentan los datos que se emplearán. En la sección tercera se expone la metodología de Phillips y Sul (2007), tanto para contrastar la hipótesis nula de convergencia (equivalente

a la existencia de un patrón de comportamiento común en los datos) como para analizar la posible existencia de clubes de convergencia. La sección cuarta presenta los resultados más destacados obtenidos al aplicar esta metodología. El trabajo termina con una revisión de las conclusiones más destacadas alcanzadas.

2. DATOS

Los datos utilizados han sido obtenidos del Instituto Nacional de Estadística, específicamente los índices de precios al consumo (IPC) mensuales para las 52 provincias españolas. La muestra abarca el período 2002:1-2023:9 y considera tanto el valor total del IPC como el de sus 12 grupos ECOICOP (*European Classification of Individual Consumption by Purpose*). De este modo, se descompuso el comportamiento total de los precios en componentes como (G01) Alimentos y bebidas no alcohólicas; (G02) Bebidas alcohólicas y tabaco; (G03) Vestido y calzado; (G04) Vivienda; (G05) Menaje; (G06) Medicina; (G07) Transporte; (G08) Comunicaciones; (G09) Ocio y cultura; (G10) Enseñanza; (G11) Hoteles, cafés y restaurantes y (G12) Otros bienes y servicios.

Dado que los datos están expresados en el año base 2021 y el objetivo es analizar posibles disparidades, se cambió la base a 2002, el inicio de la muestra, para evitar conclusiones espurias. Este ajuste aborda un problema detectado en el artículo de Phillips y Sul (2007), cuya metodología se sigue en este trabajo. Hay que tener en cuenta que, de otra manera, los índices de precios toman valores similares al final de la muestra, generando una falsa idea de convergencia. Después de este cambio de base, se presenta una breve descripción de los datos para una comprensión más clara. Los cuadros 1-3 muestran las tasas de crecimiento para varios períodos de interés, las 52 provincias españolas y la economía española en su totalidad.

Cuadro 1.

Tasas de crecimiento IPC total

Provincia	*g0203*	*g0219*	*g2023*	*g2223*
Álava	0,19	0,17	0,34	0,37
Albacete	0,19	0,16	0,38	0,47
Alicante	0,18	0,15	0,37	0,43
Almería	0,19	0,16	0,35	0,42
Ávila	0,19	0,16	0,40	0,48
Badajoz	0,17	0,14	0,38	0,43
Islas Baleares	0,19	0,16	0,37	0,44
Barcelona	0,21	0,19	0,32	0,38
Burgos	0,19	0,16	0,35	0,41
Cáceres	0,17	0,15	0,33	0,37

Cuadro 1. *(continuación)*

Tasas de crecimiento IPC total

Provincia	g_{0203}	g_{0219}	g_{2023}	g_{2223}
Cádiz	0,18	0,14	0,37	0,46
Castellón	0,19	0,15	0,39	0,45
Ciudad Real	0,19	0,16	0,39	0,45
Córdoba	0,19	0,16	0,38	0,41
La Coruña	0,19	0,17	0,37	0,41
Cuenca	0,19	0,15	0,38	0,45
Gerona	0,20	0,18	0,35	0,41
Granada	0,18	0,15	0,37	0,45
Guadalajara	0,19	0,15	0,39	0,45
Guipúzcoa	0,19	0,16	0,35	0,42
Huelva	0,19	0,15	0,38	0,45
Huesca	0,19	0,16	0,37	0,41
Jaén	0,18	0,15	0,37	0,43
León	0,20	0,16	0,41	0,47
Lérida	0,21	0,18	0,38	0,45
La Rioja	0,19	0,16	0,36	0,42
Lugo	0,19	0,16	0,37	0,43
Madrid	0,18	0,16	0,31	0,36
Málaga	0,20	0,16	0,38	0,46
Murcia	0,19	0,16	0,36	0,43
Navarra	0,19	0,16	0,37	0,44
Ourense	0,18	0,15	0,38	0,44
Asturias	0,18	0,15	0,34	0,41
Palencia	0,18	0,15	0,36	0,42
Las Palmas	0,16	0,12	0,33	0,42
Pontevedra	0,19	0,16	0,39	0,46
Salamanca	0,18	0,15	0,35	0,42
Santa Cruz de Tenerife	0,16	0,12	0,36	0,45
Cantabria	0,19	0,16	0,36	0,42
Segovia	0,19	0,16	0,38	0,42
Sevilla	0,18	0,15	0,36	0,43
Soria	0,19	0,16	0,38	0,43

Cuadro 1. *(continuación)*

Tasas de crecimiento IPC total

Provincia	*g0203*	*g0219*	*g2023*	*g2223*
Tarragona	0,18	0,15	0,35	0,39
Teruel	0,19	0,16	0,36	0,44
Toledo	0,19	0,15	0,41	0,46
Valencia	0,18	0,16	0,34	0,40
Valladolid	0,19	0,16	0,35	0,40
Vizcaya	0,19	0,17	0,36	0,42
Zamora	0,19	0,16	0,40	0,47
Zaragoza	0,18	0,16	0,35	0,38
Ceuta	0,17	0,14	0,36	0,45
Melilla	0,19	0,15	0,43	0,48
g *xxyy* es la tasa promedio de crecimiento entre los periodos 20xx y 20yy				

Fuentes: Elaboración propia.

En el cuadro 1, que se centra en el comportamiento del IPC total, se observa cierta heterogeneidad entre las provincias en cuanto al crecimiento promedio. El crecimiento mensual promedio del IPC de la economía española fue del 0,19 %. Las provincias de Lérida y Barcelona destacan con un crecimiento del 0,21 %, mientras que Las Palmas y Santa Cruz de Tenerife tienen un crecimiento inferior. Si se considera el período anterior al COVID-19 (2002:1-2019:12), los resultados son ligeramente diferentes. En primer lugar se observa que el crecimiento promedio de los precios para el total de la economía española fue del 0,16 %, si bien para las provincias de Lérida, Gerona y Barcelona fue ligeramente superior, mientras que para las ya mencionadas provincias de Las Palmas y de Santa Cruz de Tenerife fue netamente inferior (0,12 %).

El comportamiento pos-COVID-19 es realmente diferente. Podemos observar cómo el crecimiento promedio del período 2020:1-2023:9 fue de un 0.35 % para el total de la economía española, casi el doble de lo observado en el período precedente. Destacan los crecimientos de Melilla (0,43 %), León y Toledo, ambos con un 0,41 %, así como los de Madrid (0,31 %) y Barcelona (0,32 %). Incluso, si consideramos el período 2022:1-2023:9 las diferencias son más notables, por cuanto el crecimiento promedio de los precios de la economía española fue de un 0,41 %. Las mayores alzas en los precios ocurrieron en las provincias de Melilla y Ávila (0,48 %), mientras que las provincias donde los precios presentaron menor crecimiento promedio fueron las de Madrid (0,36 %), Álava y Cáceres (ambas con un 0,37 %).

Estos resultados se pueden complementar con el análisis de la figura 1 en el que se presenta la evolución temporal del coeficiente de variación del IPC total. En él se puede apreciar

cómo la dispersión muestral crece a lo largo de la muestra, si bien desde finales de 2019 hasta mediados de 2020 existe un claro retroceso en la dispersión. Posteriormente, vuelve a retomar su senda alcista hasta mediados de 2022, disminuyendo significativamente a partir de este período.

Figura 1.

σ-convergencia. Coeficiente de variación IPC total

Fuente: Elaboración propia.

Si en lugar de considerar el IPC total analizamos el comportamiento de los 12 grupos ECOCIP podemos encontrar nuevos resultados. En el cuadro 2 se presentan los crecimientos promedios desde el año 2002 hasta el final de la muestra, respectivamente, para cada una de las provincias y de los grupos que hemos considerados.

Cuadro 2.

Tasa de crecimiento promedio de la muestra de los grupos ECOICOP y del IPC total

Provincia	*G01*	*G02*	*G03*	*G04*	*G05*	*G06*	*G07*	*G08*	*G09*	*G10*	*G11*	*G12*	*Total*
Álava	0,28	0,33	0,06	0,21	0,13	0,09	0,21	-0,08	0,03	0,19	0,22	0,20	0,19
Albacete	0,24	0,36	0,08	0,24	0,13	0,04	0,24	-0,07	0,00	0,27	0,25	0,19	0,19
Alicante	0,24	0,35	0,09	0,20	0,12	0,03	0,24	-0,10	0,01	0,21	0,23	0,19	0,18
Almería	0,26	0,36	0,07	0,21	0,08	0,08	0,23	-0,06	0,00	0,21	0,20	0,17	0,19
Ávila	0,25	0,34	0,02	0,26	0,11	0,09	0,23	-0,08	0,01	0,21	0,27	0,19	0,19
Badajoz	0,25	0,36	0,03	0,22	0,08	0,02	0,21	-0,10	-0,05	0,23	0,21	0,18	0,17
Islas Baleares	0,24	0,32	0,07	0,21	0,11	0,07	0,23	-0,07	0,03	0,21	0,25	0,20	0,19

Cuadro 2. *(continuación)*

Tasa de crecimiento promedio de la muestra de los grupos ECOICOP y del IPC total

Provincia	*G01*	*G02*	*G03*	*G04*	*G05*	*G06*	*G07*	*G08*	*G09*	*G10*	*G11*	*G12*	*Total*
Barcelona	0,26	0,34	0,11	0,22	0,17	0,10	0,23	-0,10	0,08	0,25	0,24	0,23	0,21
Burgos	0,26	0,33	0,06	0,21	0,16	0,10	0,22	-0,07	0,01	0,21	0,22	0,18	0,19
Cáceres	0,22	0,35	0,02	0,20	0,10	0,06	0,22	-0,09	-0,03	0,18	0,21	0,19	0,17
Cádiz	0,25	0,35	0,04	0,20	0,09	0,07	0,23	-0,12	-0,02	0,26	0,23	0,17	0,18
Castellón	0,26	0,34	0,11	0,22	0,09	0,01	0,24	-0,11	-0,02	0,22	0,22	0,19	0,19
Ciudad Real	0,25	0,36	0,10	0,24	0,07	0,04	0,24	-0,08	-0,07	0,19	0,23	0,18	0,19
Córdoba	0,24	0,33	0,08	0,21	0,10	0,10	0,22	-0,09	0,00	0,23	0,24	0,18	0,19
La Coruña	0,25	0,31	0,09	0,24	0,15	0,10	0,23	-0,09	0,01	0,19	0,25	0,17	0,19
Cuenca	0,24	0,36	0,06	0,27	0,12	0,06	0,22	-0,08	-0,04	0,21	0,23	0,18	0,19
Gerona	0,26	0,32	0,11	0,24	0,13	0,11	0,24	-0,11	0,04	0,25	0,24	0,23	0,20
Granada	0,26	0,35	0,06	0,23	0,12	0,03	0,22	-0,07	-0,02	0,24	0,24	0,18	0,18
Guadalajara	0,26	0,36	0,04	0,22	0,11	0,10	0,22	-0,09	0,05	0,23	0,21	0,16	0,19
Guipúzcoa	0,25	0,31	0,03	0,22	0,12	0,10	0,24	-0,08	0,06	0,20	0,24	0,19	0,19
Huelva	0,24	0,35	0,06	0,22	0,12	0,09	0,23	-0,08	-0,01	0,23	0,21	0,16	0,19
Huesca	0,25	0,31	0,07	0,25	0,13	0,10	0,23	-0,06	-0,02	0,22	0,22	0,19	0,19
Jaén	0,25	0,34	0,07	0,22	0,08	0,08	0,24	-0,07	-0,04	0,23	0,21	0,15	0,18
León	0,26	0,33	0,11	0,25	0,15	0,10	0,22	-0,09	-0,03	0,21	0,26	0,19	0,20
Lérida	0,26	0,33	0,11	0,24	0,15	0,11	0,23	-0,07	0,04	0,24	0,24	0,22	0,21
La Rioja	0,24	0,33	0,09	0,23	0,16	0,09	0,23	-0,10	0,05	0,21	0,22	0,21	0,19
Lugo	0,24	0,30	0,07	0,21	0,14	0,08	0,23	-0,12	0,07	0,22	0,26	0,16	0,19
Madrid	0,25	0,35	0,08	0,19	0,12	0,04	0,22	-0,08	0,05	0,21	0,22	0,21	0,18
Málaga	0,25	0,36	0,05	0,24	0,13	0,06	0,23	-0,07	-0,02	0,24	0,26	0,19	0,20
Murcia	0,26	0,35	0,09	0,21	0,11	0,04	0,23	-0,09	0,02	0,22	0,23	0,19	0,19
Navarra	0,24	0,33	0,10	0,22	0,13	0,14	0,22	-0,10	0,05	0,22	0,22	0,20	0,19
Ourense	0,24	0,31	0,05	0,24	0,12	0,09	0,22	-0,10	0,02	0,23	0,23	0,15	0,18
Asturias	0,23	0,33	0,10	0,23	0,11	0,04	0,22	-0,10	0,00	0,20	0,23	0,19	0,18
Palencia	0,24	0,35	0,02	0,24	0,11	0,07	0,22	-0,10	-0,01	0,22	0,21	0,18	0,18
Las Palmas	0,22	0,36	-0,01	0,16	0,04	0,05	0,23	-0,11	-0,03	0,25	0,20	0,13	0,16
Pontevedra	0,24	0,33	0,07	0,24	0,15	0,10	0,23	-0,09	0,01	0,19	0,27	0,18	0,19
Salamanca	0,26	0,34	0,03	0,21	0,08	0,05	0,23	-0,10	-0,03	0,20	0,22	0,17	0,18
Santa Cruz de Tenerife	0,25	0,35	-0,06	0,16	0,04	0,02	0,27	-0,06	-0,01	0,18	0,19	0,15	0,16

Cuadro 2. *(continuación)*

Tasa de crecimiento promedio de la muestra de los grupos ECOICOP y del IPC total

Provincia	*G01*	*G02*	*G03*	*G04*	*G05*	*G06*	*G07*	*G08*	*G09*	*G10*	*G11*	*G12*	*Total*
Cantabria	0,25	0,34	0,06	0,23	0,10	0,08	0,23	-0,06	0,03	0,17	0,23	0,20	0,19
Segovia	0,26	0,33	0,11	0,23	0,08	0,10	0,22	-0,10	0,06	0,26	0,23	0,19	0,19
Sevilla	0,25	0,35	0,07	0,21	0,12	0,04	0,21	-0,11	-0,01	0,23	0,22	0,19	0,18
Soria	0,25	0,33	0,09	0,25	0,10	0,09	0,23	-0,06	0,00	0,20	0,23	0,16	0,19
Tarragona	0,24	0,34	0,03	0,21	0,07	0,09	0,24	-0,08	-0,01	0,19	0,22	0,20	0,18
Teruel	0,25	0,34	0,07	0,27	0,11	0,13	0,24	-0,09	-0,01	0,17	0,23	0,18	0,19
Toledo	0,25	0,35	0,07	0,26	0,08	0,09	0,21	-0,06	-0,01	0,24	0,23	0,18	0,19
Valencia	0,25	0,35	0,07	0,24	0,11	0,06	0,22	-0,12	0,03	0,18	0,25	0,18	0,18
Valladolid	0,24	0,35	0,08	0,22	0,14	0,09	0,22	-0,07	0,01	0,22	0,22	0,17	0,19
Vizcaya	0,26	0,33	0,11	0,23	0,17	0,09	0,22	-0,09	0,03	0,23	0,23	0,20	0,19
Zamora	0,27	0,35	0,08	0,23	0,11	0,09	0,23	-0,06	0,00	0,18	0,22	0,16	0,19
Zaragoza	0,24	0,34	0,07	0,23	0,12	0,10	0,22	-0,08	0,02	0,24	0,23	0,20	0,18
Ceuta	0,23	0,28	0,08	0,19	0,06	0,02	0,28	-0,05	0,01	0,14	0,19	0,15	0,17
Melilla	0,28	0,27	0,04	0,21	0,10	0,04	0,30	-0,09	-0,01	0,21	0,22	0,15	0,19
España	0,25	0,34	0,08	0,22	0,12	0,07	0,23	-0,09	0,03	0,22	0,23	0,20	0,19

Fuentes: Elaboración propia.

Si comenzamos analizando el cuadro 2, vemos que las tasas de crecimiento promedio son bastante heterogéneas. Si consideramos la media nacional, se observa que el menor crecimiento durante el período fue el del grupo 08 (Comunicaciones), con un crecimiento negativo promedio del -0,09 %. Frente a ello, el grupo 02 (Bebidas alcohólicas y tabaco) tiene un crecimiento promedio del 0,34 %. También hay que señalar que el grupo 01 (Alimentos y bebidas no alcohólicas), presenta el segundo crecimiento promedio más elevado (0,25 %).

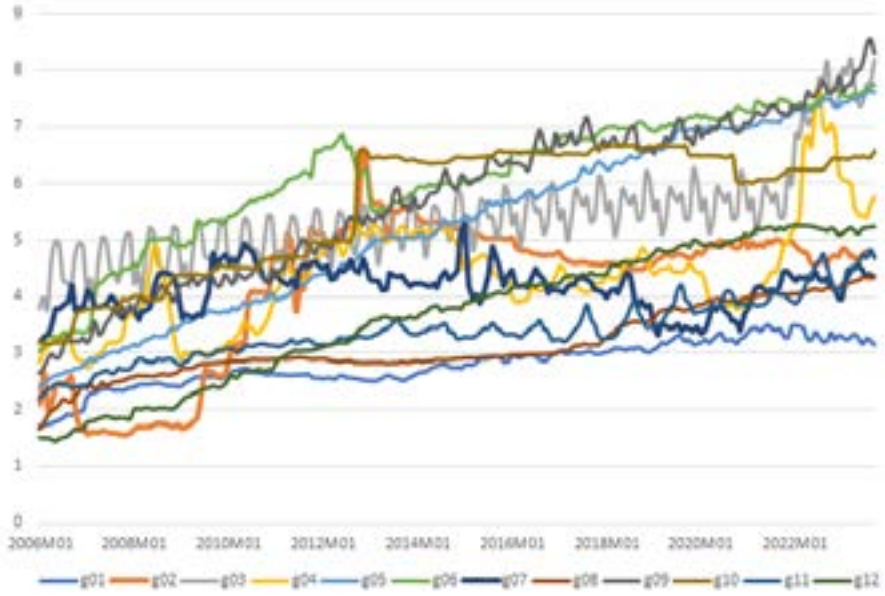
Si desagregamos por provincias, de nuevo se observa gran disparidad en los datos. En particular, el coeficiente de variación de las tasas de crecimiento promedio varía entre el 4,9 % del grupo 01 (Alimentos y bebidas no alcohólicas) y el 49,5 % del grupo 04 (Vestido y calzado). El máximo crecimiento promedio se encuentra en el grupo 02 (Bebidas alcohólicas y tabaco) en la provincia de Albacete, mientras que la menor tasa de crecimiento promedio es la del grupo 08 (Comunicaciones) en la provincia de Lugo (-0,12 %).

En la figura 2 se presenta la evolución de la dispersión para los diversos componentes del IPC. Como se puede observar, el comportamiento es claramente creciente para todos los grupos considerados, si bien el crecimiento promedio no es homogéneo. Así, los grupos 07 (Transporte) y 01 (Alimentos y bebidas no alcohólicas), presentan los menores crecimientos

promedios desde el año 2006 (1,4 % y 1,8 %), respectivamente. Por el contrario, los grupos 09 (Ocio y cultura), 05 (Muebles, artículos del hogar y artículos para el mantenimiento corriente del hogar) y 12 (Otros bienes y servicios) tienen crecimientos promedios por encima del 3 % (3.2 %, 3,2 % y 3,5 %, respectivamente). Además, se puede comprobar que los máximos valores de la dispersión se producen mayoritariamente alrededor del final de la muestra. Son claras excepciones los casos de los grupos 02 (Bebidas alcohólicas y tabaco), 04 (Vivienda, agua, electricidad, gas y otros combustibles) y 07 (Transporte), este último en menor medida. En todo caso, el hecho de presentar tendencias crecientes en la evolución del coeficiente de dispersión supone que existe cierta evidencia en contra de la existencia de σ-convergencia.

Figura 2.

σ-convergencia. Coeficiente de variación de los grupos del IPC



Fuente: Elaboración propia.

A partir de esta información inicial es sencillo concluir que la evolución de los precios en España no ha sido homogénea ni desde un punto de vista geográfico, ni desde una perspectiva temporal. Por lo que no resulta sencillo pensar que la evolución provincial de los precios en España exhiba un único patrón de comportamiento. Por el contrario, es más que probable que podemos concluir que existen disparidades significativas.

No obstante, los métodos utilizados en esta sección son meramente descriptivos, por lo que parece adecuado utilizar herramientas que sean capaces de ofrecer conclusiones más

robustas y, sobre todo, que permitan el uso de un número importante de datos, como los que aquí se usan. La metodología a emplear se presenta en la siguiente sección.

3. METODOLOGÍA

Antes de presentar la metodología que vamos a emplear, creemos necesario recordar la relación entre los análisis de convergencia y con los conceptos de β-convergencia y σ-convergencia, inicialmente propuestos por Barro y Sala-i-Martin (1990) al estudiar la dispersión del PIB per cápita. La β-convergencia sugiere que las economías menos prósperas tienden a crecer más rápidamente que las más ricas, reduciendo la brecha entre ellas, mientras que la σ-convergencia implica una disminución en la dispersión transversal de la variable a lo largo del tiempo.

Ambos métodos han recibido severas críticas; por ejemplo, la β-convergencia ha sido señalada por indicar convergencia incluso en casos donde podría no existir genuinamente, según De Long (1988) y Quah (1993). Además, ninguno de estos métodos prueba explícitamente la hipótesis nula de convergencia, lo que ha llevado a la propuesta de enfoques alternativos en la literatura.

Para solucionar este problema, Phillips y Sul (2007, 2009) desarrollaron una metodología asociada al concepto de σ-convergencia, que es la que se sigue en este trabajo. Estos autores diseñan un estadístico para contrastar la hipótesis nula de convergencia y determinar la existencia de clubes de convergencia si esta hipótesis es rechazada.

La metodología de Phillips-Sul resulta muy adecuada, en especial en un entorno de uso masivo de datos, por cuanto es muy flexible con respecto a las características temporales de las variables. Además, permite aprovechar las ventajas que ofrece el uso combinado de datos de series temporales para un número amplio de corte transversales. En consecuencia, supera claramente al análisis clásico de convergencia. Por lo tanto, no es sorprendente que la metodología esté ganando popularidad y se aplique ampliamente para analizar disparidades en la evolución de diversas variables. En concreto, en el caso de los precios hay que reseñar que el propio artículo original de Phillips y Sul (2007) estudia la evolución de los precios en los Estados Unidos. Otros ejemplos son los trabajos de Liu y Yeh (2012) quienes estudian los precios en las ciudades USA, Lin y Robberts (2023) que estudian el caso de los precios de la vivienda en el Reino Unido o Gil *et al.* (2023) quienes estudian los precios del mercado internacional de la madera aserrada, entre otros casos.

Siguiendo a Phillips-Sul, denominemos nuestra variable de interés (el logaritmo del índice de precios al consumo total y desglosado en sus doce grupos) como X_{it}, donde i representa las 52 provincias españolas, y t abarca el período 2006:1-2023:9. Como se puede apreciar, la muestra utilizada no se corresponde con el total de la muestra disponible. Phillips-Sul aconsejan eliminar algunas observaciones iniciales en el caso de los precios, para evitar el efecto base, ya que todas las provincias toman el valor 100 al comienzo de la muestra.

Entonces, podemos desagregar la variable de interés de la siguiente manera $X_{it} = \delta_{it}\mu_t$ donde δ_{it} es el componente idiosincrático y μ_t es el componente de tendencia común.

Dado que el número de incógnitas en el modelo es mayor que el número de observaciones en el panel, Phillips y Sul (2007) eliminan el factor común, obteniendo la variable h_{it}.

$$h_{it} = \frac{X_{it}}{N^{-1}\sum_{i=1}^{N} X_{it}} = \frac{\delta_{it}\mu_t}{N^{-1}\sum_{i=1}^{N} \delta_{it}\mu_t} = \frac{\delta_{it}}{N^{-1}\sum_{i=1}^{N} \delta_{it}} \qquad [1]$$

Esta variable h_{it} representa el parámetro relativo de transición (transition path) y mide la posición relativa de cada corte transversal con el promedio del panel en el período t. Por definición, la media transversal de h_{it} es la unidad. Por tanto, cuando existe convergencia, h_{it} debería aproximarse a la unidad para todos los i cuando t tiende a infinito.

La varianza transversal de h_{it} se define de la siguiente manera:

$$H_t = \frac{1}{N}\sum_{i=1}^{N}\left(h_{it} - 1\right)^2 \qquad [2]$$

Esta varianza, bajo el supuesto de convergencia, tiende hacia 0 conforme la muestra crece. Por tanto, para analizar la presencia de convergencia basta con estudiar si esta varianza decrece hacia 0. Para ello, Phillips y Sul (2007) proponen estimar el siguiente modelo:

$$\ln\frac{H_1}{H_t} - \ln L(t) = a + \beta \ln L(t) + u_t,\ t = [rT], [rT]+1, \ldots, T \ \ con\ r > 0 \qquad [3]$$

donde la variable dependiente del modelo está claramente relacionada con la varianza de sección-cruzada, de manera que se puede estudiar si la varianza del panel (H_t) tiene tendencia decreciente a partir del valor del parámetro β. Para mejorar las propiedades del método propuesto introducen $L(t)$ que es una función de t. A través de diversos ejercicios de simulación, los autores eligen usar $L(t)$= ln t. De igual manera, r es el primer porcentaje de los datos que se descarta. Según los experimentos de Monte Carlo realizados por Phillips y Sul (2007), r debería tomar valores en el intervalo (0,2, 0,3) dependiendo del tamaño de la muestra. En nuestro caso se ha usado el valor 1/3.

De acuerdo a este modelo, existe convergencia siempre que $\beta \geq 0$, mientras que si $\beta < 0$, entonces tenemos un comportamiento divergente. Para distinguir entre estas dos alternativas, PS proponen usar el denominado log t-ratio que no es otra cosa que el t-ratio para contrastar la hipótesis nula $\beta \geq 0$ en el modelo anterior. Estos autores recomiendan emplear métodos de estimación robustos a la presencia de autocorrelación y heterocedasticidad. Entonces, si el log t-ratio es inferior a −1,65, se rechaza la hipótesis nula de convergencia. No obstante, hay que tener en cuenta que podrían existir los denominados clubes de convergencia. Para tener en cuenta esta posibilidad, PS desarrollan un algoritmo para estimar dichos clubes de convergencia. El algoritmo consta de los siguientes pasos:

- Ordenar los miembros del panel según su último valor de la variable de interés. En este caso, se ordenan de mayor a menor las provincias en función del índice de precios del último período de la muestra.
- Efectuada esta ordenación, se debe generar un grupo de provincias inicial. Para ello, se van agregando provincias en función de la ordenación anterior mientras que el log t-ratio sea inferior a -1,65.
- Una vez formado este grupo inicial, se prueba con el resto de las provincias no incluidas y se reestima el modelo [3] para cada provincia añadida. Si el log t-ratio es inferior a -1,65, dicha provincia se descarta. En el caso contrario, la provincia se añade al grupo inicial. Una vez consideradas todas las provincias, se tiene formado el primer club de convergencia.
- Ahora, debemos analizar la posible existencia de múltiples clubes de convergencia. Por lo que se repiten los pasos anteriores para las provincias descartadas. Si alguna provincia no puede incluirse en ningún club, se considera que diverge y no se incluye en ningún club de convergencia.

Phillips y Sul (2007) señalan la posibilidad de encontrar más clubes de convergencia de los que realmente existen debido a la naturaleza conservadora del algoritmo. Para eliminar la sobreestimación de clubes convergentes, recomiendan realizar pruebas de regresión log-t entre clubes adyacentes para determinar si deben fusionarse en clubes más grandes. Este procedimiento de fusión se aplica hasta que no se pueden fusionar más clubes.

Finalmente, Phillips y Sul (2007) aconsejan aplicar el filtro de Hodrick y Prescott (1997) para eliminar cierto ruido de la variable y mejorar la potencia del estadístico. En nuestro caso, no se ha utilizado este filtro dado que la evidencia en contra de la hipótesis nula era notable y, además, debería haberse usado un valor elevado del parámetro de alisado dado que nuestros datos son mensuales. Este filtro podría distorsionar los resultados y, por tanto, hemos preferido usar los datos originales.

4. RESULTADOS

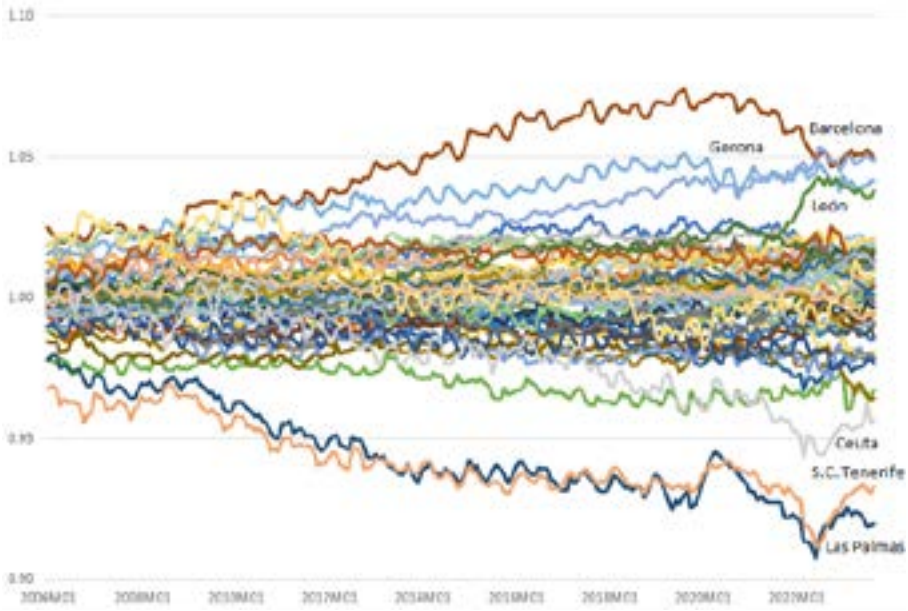
En primer lugar, hay que tener en cuenta el estadístico log-t para contrastar la hipótesis nula de convergencia. El valor del este estadístico es -28,81, inferior al valor crítico (-1,65), por lo que la evidencia en contra de la hipótesis nula es clara. En consecuencia, se puede afirmar que la evolución de los precios provinciales en España presenta diferencias significativas y no existe un único patrón de comportamiento.

Este hecho se puede confirmar analizando la figura 3, donde se presentan los valores de la variable *h*, definida en [1] y que habitualmente es denominada *transition path,* para las diversas provincias. Esta variable proporciona una medida de cuál es la distancia de cada una de las variables con respecto al promedio del panel. En la mencionada figura se observa

que existen provincias cuyos valores *h* del final de la muestra se agrupan en torno a valores claramente alejados de la unidad. Es el caso de Barcelona, Lérida, Gerona y León, con valores superiores a 1,038. Por el contrario, las provincias de Las Palmas y de Santa Cruz de Tenerife muestran valores ligeramente superiores a 0,92, mientras que el de Ceuta se sitúa en 0,96. Por lo tanto, parece que existen diversos patrones de comportamiento y que las provincias se pueden agrupar en torno a ellos. Para ello, parece conveniente utilizar el algoritmo propuesto por Phillips y Sul (2007) y analizar si pueden existir distintos clubes de convergencia.

Figura 3.

Transition paths



Fuente: Elaboración propia.

Los resultados de aplicar el citado algoritmo nos llevan a estimar la existencia de cinco clubes de convergencia. La composición provincial de cada uno de estos clubes de convergencia es la siguiente:

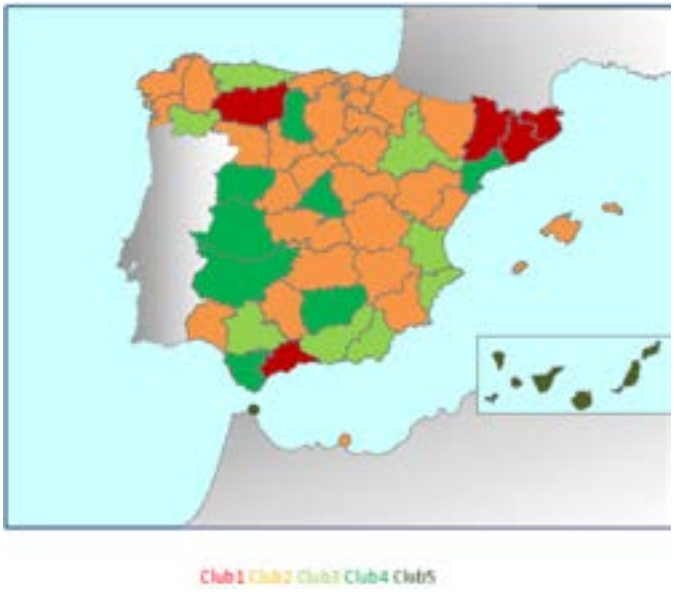
- Club 1: Barcelona, Gerona, León, Lérida y Málaga.
- Club 2: Álava, Albacete, Ávila, Islas Baleares, Burgos, Castellón, Ciudad Real, Córdoba, La Coruña, Cuenca, Guadalajara, Guipúzcoa, Huelva, Huesca, La Rioja, Lugo, Murcia, Navarra, Pontevedra, Cantabria, Segovia, Soria, Teruel, Toledo, Valladolid, Vizcaya, Zamora y Melilla.
- Club 3: Alicante, Almería, Granada, Orense, Asturias, Sevilla, Valencia y Zaragoza.

- Club 4: Badajoz, Cáceres, Cádiz, Jaén, Madrid, Palencia, Salamanca y Tarragona.
- Club 5: Las Palmas, Santa Cruz de Tenerife y Ceuta.

Si comparamos los resultados con el análisis de la figura 3, resulta directo asociar los clubes 1 y 5 con la evolución de los valores de la variable *h*. Para entender mejor la composición de los clubes, la figura 4 presenta la composición de los diferentes clubes de convergencia estimados en forma de mapa. La pertenencia de cada una de las provincias a los diferentes clubes estimados se caracteriza de un color, siendo los colores más cálidos (rojo y naranja) los primeros clubes, asociados a los valores más altos de los precios al final de la muestra, mientras que las provincias incluidas en los últimos clubes se colorean en tonos verdosos.

Figura 4.

Clubes de convergencia estimados



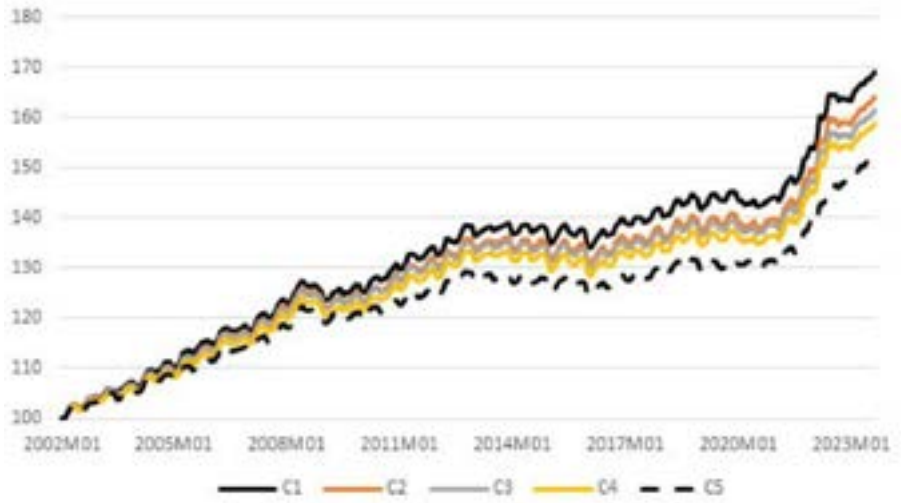
Fuente: Elaboración propia.

Como se puede apreciar en dicha figura, existe cierto grado de correlación espacial. Por ejemplo, las provincias incluidas en el club 1 están situadas en Cataluña, si bien hay que añadir a éstas los casos de León y Málaga. Frente a ellas, tenemos el caso de las provincias insulares canarias, que pertenece al club 5, conjuntamente con Ceuta. También es interesante resaltar que el club 2 es el que contiene el mayor número de provincias, mientras que cuatro de las cinco capitales de provincias más pobladas (Madrid, Sevilla, Valencia y Zaragoza) se encuentran en los clubes 3 y 4. Por el contrario, provincias con capitales escasamente pobladas, como Huesca y Teruel, están en el club 2, al igual que las provincias que rodean a Madrid.

La figura 5 presenta la media aritmética del índice de precios para las provincias incluidas en cada club estimado. Como se puede apreciar, las provincias del club 1 muestran las medias más altas, mientras que las incluidas en el club 5 son las más bajas. En todo caso, los perfiles de la evolución del IPC son muy similares para cada uno de los clubes estimados, mostrando un claro intervalo de escaso crecimiento durante el período 2012-2021. A partir de dicho año, los precios experimentan un claro período alcista, cuya fuerza llega hasta el final de la muestra utilizada.

Figura 5.

Evolución de los comportamientos medios de los clubes de emergencia. IPC total



Fuente: Elaboración propia.

El análisis descriptivo realizado con anterioridad mostró que la evolución de los diferentes componentes del índice de precios no era homogénea. Por ello, cabe plantearse si los clubes estimados son también válidos para cada uno de los 12 grupos en los que se puede desagregar el índice de precios al consumo. En el cuadro 3 se presentan los valores del log t-ratio para contrastar la hipótesis nula de convergencia en cada uno de estos grupos. Para ello, hemos estimado la ecuación [3] para cada uno de estos grupos y hemos obtenido el correspondiente log t-ratio para contrastar la hipótesis nula de convergencia. Se observa que dicha hipótesis se rechaza mayoritariamente, por lo que existe evidencia de disparidad en la evolución de los precios para las diferentes provincias. No obstante, también se ve que esta hipótesis nula no puede rechazarse para los grupos G02 (Bebidas alcohólicas y tabaco) y G07 (Transporte), por lo que para estos dos grupos, la evolución de los precios provinciales ha sido estadísticamente idéntica.

Para el resto de los grupos, tal y como hemos comentado, se rechaza la hipótesis nula de convergencia y, por tanto, no existe un único comportamiento. Pero, al igual que para el total

Cuadro 3.

Contraste de convergencia para los grupos provinciales del IPC

Grupos	*log t-ratio*	
G01. Alimentos y bebidas no alcohólicas	-0,94 (-16,84)	*
G02. Bebidas alcohólicas y tabaco	-0,03 (-0,29)	
G03. Vestido y calzado	-1,01 (-6,87)	
G04. Vivienda, agua, electricidad, gas y otros combustibles	-0,45 (-3,05)	
G05. Muebles, artículos del hogar y artículos para el mantenimiento corriente del hogar	-1,43 (-46,53)	
G06. Sanidad	-0,90 (-12,01)	*
G07. Transporte	-0,17 (-1,31)	
G08. Comunicaciones	-1,36 (-14,97)	
G09. Ocio y cultura	-1,20 (-22,57)	
G10. Enseñanza	-0,54 (-5,07)	
G11. Restaurantes y hoteles	-0,91 (-7,79)	
G12. Otros bienes y servicios	-1,34 (-39,51)	

Notas: La segunda columna presenta el log t-ratio para contrastar la hipótesis de convergencia entre paréntesis. *significa que no se puede rechazar la hipótesis nula de convergencia.

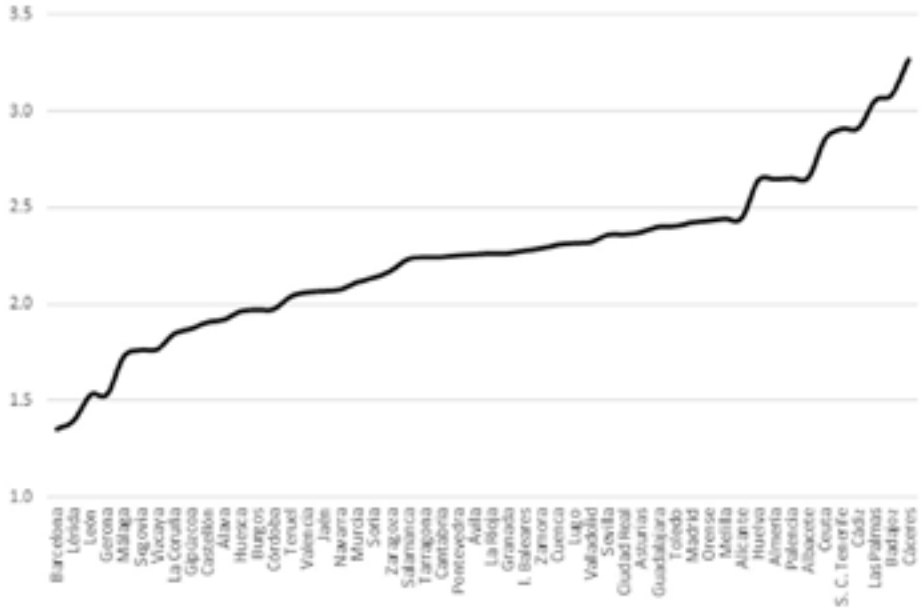
Fuentes: Elaboración propia.

del índice de precios al consumo, parece adecuado analizar la posible presencia de clubes de convergencia Para ello, hemos aplicado el algoritmo de Phillips-Sul, encontrando diversos resultados. Todos ellos quedan resumidos en la figura 6 en la que se presenta el valor promedio del ordinal del club de convergencia en el que cada provincia se ha incluido, multiplicado por la ponderación de cada uno de los grupos en el índice de precios de dicha provincia. Como se puede apreciar, los valores de las provincias de Barcelona, Lérida, Gerona y León, todas ellas incluidas en el club 1 para el total del índice de precios al consumo, son los más bajos. Por tanto, dichas provincias muestran valores del índice de precios al consumo elevados para cada uno de los componentes, no solamente para algunos de sus componentes. Del mismo modo, las provincias de Ceuta, Santa Cruz de Tenerife, Cádiz, Las Palmas, Badajoz y

Cáceres son las que presentan un valor medio ponderado más elevado. Hay que señalar que dichas provincias se incluyeron en los clubes de convergencia 4 y 5 para el total del índice de precios al consumo.

Figura 6.

Club de convergencia de las provincias (ponderado por peso grupos del IPC)



Fuente: Elaboración propia.

Como caso particular, se analiza de forma separada la evolución de los precios del grupo 01 (Alimentación y Bebidas). Este grupo presenta el peso más elevado dentro del total del índice de precios al consumo. Esta ponderación va desde el valor 0,17 (Madrid) hasta el valor más alto, 0,25, para Ceuta.

El algoritmo de clusterización propuesto por Phillips-Sul estima la existencia de cuatro clubs de convergencia diferentes, cuya composición es la siguiente:

- Club 1: Álava, Almería, Barcelona, Castellón, Gerona, Guadalajara, León, Lérida, Salamanca, Segovia, Zamora y Melilla.

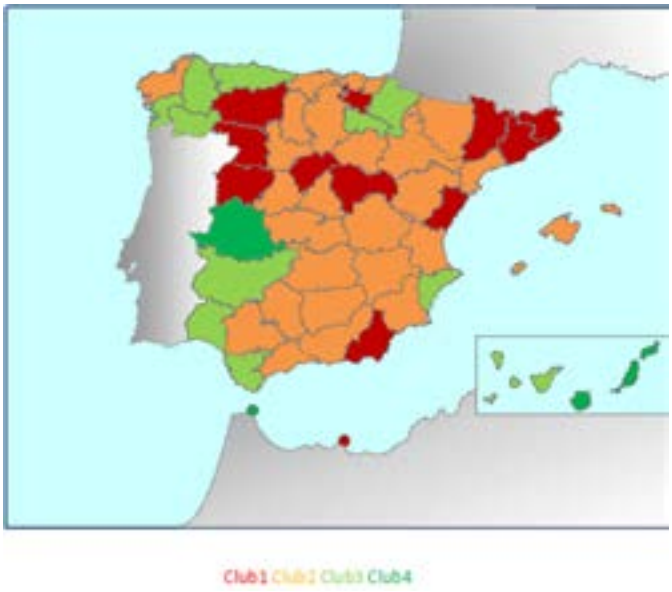
- Club 2: Albacete, Ávila, I. Baleares, Burgos, Ciudad Real, Córdoba, La Coruña, Cuenca, Granada, Guipúzcoa, Huesca, Jaén, Madrid, Málaga, Murcia, Palencia, Cantabria, Sevilla, Soria, Tarragona, Teruel, Toledo, Valencia, Valladolid, Vizcaya y Zaragoza.

- Club 3: Alicante, Badajoz, Cádiz, Huelva, La Rioja, Lugo, Navarra, Orense, Asturias, Pontevedra y Santa Cruz de Tenerife.
- Club 4: Cáceres, Las Palmas y Ceuta.

La figura 7 presenta la composición geográfica de cada uno de los clubes. Hay que destacar que en el club 1 se mantienen las provincias de Barcelona, Gerona, Lérida y León, a las que se añaden Álava, Almería, Castellón, Guadalajara, Salamanca, Segovia, Zamora y Melilla. Por el contrario, el club 4 contiene las provincias con los precios más bajos: Cáceres, Las Palmas y Ceuta. Hay que resaltar que el club 2 agrupa el mayor número de provincias, incluyendo las que tienen capitales de provincias más pobladas, a excepción de Barcelona.

Figura 7.

Clubes de convergencia estimados. Grupo 01

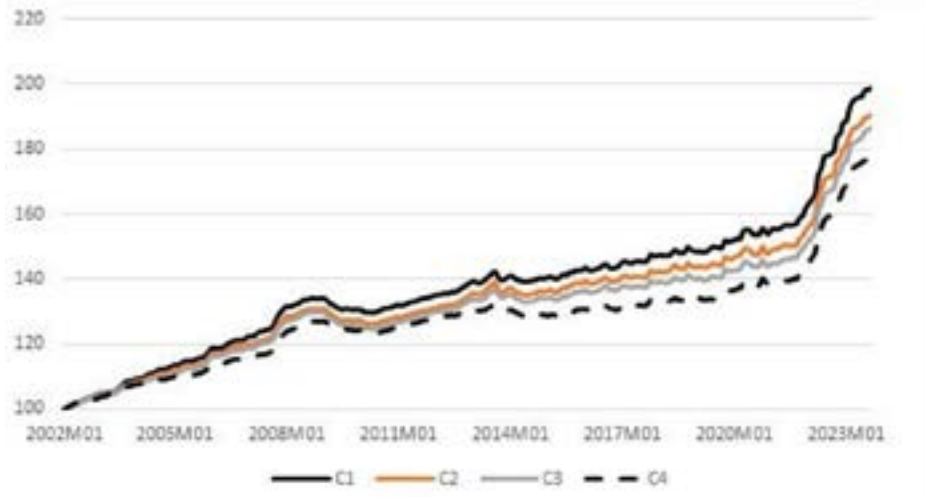


Fuente: Elaboración propia.

La evolución promedio de los índices de precios del Grupo 01 se presenta en la figura 8. Se puede observar que el perfil es similar para los cuatro clubes de convergencia, aunque la tasa de crecimiento promedio del período es del 0,6 %, 0,25 %, 0,24 % y 0,22 % para, respectivamente, los clubes de convergencia 1, 2, 3 y 4. Si nos centramos en la parte final de la muestra, se observa un crecimiento netamente superior al visto con anterioridad. En concreto, las tasas de crecimiento promedio desde 2022:1 son del 0,99 %, 1,00 %, 1,04 % y 1,04 % para los clubes de convergencia 1, 2, 3 y 4, respectivamente. Por tanto, el período inflacionista ocurrido en este grupo de productos al final de la muestra ha sido muy similar para todas las provincias.

Figura 8.

Evolución promedio de los clubes de convergencia estimados. Grupo 01



Fuente: Elaboración propia.

5. CONCLUSIONES

La presente investigación ha analizado la evolución de los precios provinciales en España mediante la aplicación de técnicas que posibilitan manejar volúmenes significativos de información. Específicamente, la metodología propuesta por Phillips y Sul (2007) es la que hemos usado para estudiar la presencia de patrones de comportamiento compartidos en los índices de precios al consumo de las diversas provincias españolas durante el período 2002:1-2023:9.

Los resultados obtenidos revelan que la trayectoria de los precios a nivel provincial en España no sigue una pauta uniforme. Por el contrario, se observan disparidades significativas en su evolución a lo largo del tiempo para las distintas provincias. En particular, la aplicación de la mencionada metodología ha conducido al rechazo de la hipótesis nula de convergencia, evidenciando la existencia de cinco clubes de convergencia para el índice de precios al consumo provincial. La composición de estos clubes indica que provincias como Barcelona, Gerona, Lérida, León y Málaga exhiben precios más elevados, mientras que el grupo conformado por las provincias de Las Palmas, Santa Cruz de Tenerife y Ceuta presenta precios más bajos.

Estas disparidades persisten al desagregar el índice total en sus 12 componentes ECOICOP, donde se rechaza la hipótesis nula de convergencia en la mayoría de estos grupos, con excepción de los Grupos 02 (Bebidas alcohólicas y tabaco) y 07 (Transporte).

En resumen, este estudio resalta la complejidad y diversidad inherente a la evolución de los precios provinciales en España. La implementación de técnicas estadísticas que facilitan el procesamiento masivo de datos ha permitido identificar patrones de comportamiento comunes, proporcionando una comprensión más profunda de las dinámicas subyacentes en los precios españoles a nivel provincial. Estas conclusiones poseen relevancia en el ámbito de la toma de decisiones económicas, en especial si consideramos las formulación de políticas diferenciadas espacialmente.

Referencias

Barro, R. J. y Sala-i-Martin, X. (1990). Economic growth and convergence across the United States. *NBER Working Paper,* (w3419).

Barro, R. J. y Sala-i-Martin, X. (1992). Convergence. *Journal of political Economy, 100*(2), pp. 223-251.

Baumol, W. J. (1986). Productivity growth, convergence, and welfare: what the long-run data show. *The American Economic Review, 76*(5), pp. 1072-1085.

Chatterjee, A., Gupta, S. y Lahiri, S. N. (2015). On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property. *Journal of Econometrics, 186*(2), pp. 317-324.

Gil, J. M., Montañés, A. y Vásquez-González, B. (2023). Are prices converging in the global sawnwood market? *Forest Policy and Economics,* 102998.

Hodrick, R. J. y Prescott, E. C. (1997). Postwar US business cycles: an empirical investigation. *Journal of Money, Credit, and Banking, 29*(1), pp. 1-16.

De Long, J. B. (1988). Productivity growth, convergence, and welfare: comment. *The American Economic Review, 78*(5), pp. 1138-1154.

Lin, P. T. y Robberts, A. (2023). Regional house price convergence: implications of monetary policy. *Regional Studies, 58*(5), pp. 1-13.

Liu, W. H. y Yeh, C. C. (2012). Convergence in price levels across US cities. *Economics Letters, 114*(3), pp. 245-248.

Phillips, P. C. y Sul, D. (2007). Transition modeling and econometric convergence tests. Econometrica, 75(6), pp. 1771-1855.

Phillips, P. C. y Sul, D. (2009). Economic transition and growth. *Journal of Applied Econometrics, 24*(7), pp. 1153-1185.

Quah, D. (1993). Galton's fallacy and tests of the convergence hypothesis. *The Scandinavian Journal of Economics, 95*(4), pp. 427-443.

Stock, J. H. y Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association, 97*(460), pp. 1167-1179.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association, 101*(476), pp.1418-1429.

CAPÍTULO IV

El análisis de la economía en tiempo real a partir de datos masivos de transacciones en cuentas bancarias

Josep Mestres Domènech

En este trabajo se describe la experiencia de CaixaBank Research en el análisis de datos masivos de transacciones financieras. Los movimientos de cuentas bancarias, de alta frecuencia y con un nivel de granularidad elevado, permiten obtener información económica de gran calidad y precisión de manera casi instantánea y con un enorme potencial para la investigación económica. Sin embargo, su análisis es complejo, puesto que se trata de un gran volumen de datos creado para finalidades distintas del análisis económico. El trabajo describe el proceso de aprendizaje a partir de los proyectos llevados a cabo, y en particular, la creación de un portal de Economía en Tiempo Real (https://realtimeeconomics.CaixaBankresearch.com/#/home). Este portal monitoriza la evolución de la economía española mediante 12 indicadores construidos con datos internos de CaixaBank, agregados mediante técnicas de *big data*.

Palabras clave: análisis de datos masivos en tiempo real, *big data,* transacciones financieras.

1. INTRODUCCIÓN

Si bien hace ya unos años que los economistas han ido incorporando los datos masivos (*big data*) a sus análisis, la pandemia del COVID-19 supuso una verdadera revolución de la economía en tiempo real. Disponer de información al momento era imprescindible para gestionar la crisis y poder tomar decisiones rápidamente. Por ejemplo, gracias a los datos de Google Mobility Report se pudo medir el impacto del COVID-19 en la movilidad de las personas tras las restricciones sanitarias y gracias a los datos de Opportunity Insights (https://www.tracktherecovery.org/), el impacto económico de la pandemia en EE. UU. (Chetty *et al.,* 2020). En CaixaBank Research, junto con investigadores de la Universitat Pompeu Fabra y el Institute of Political Economy and Governance (IPEG), desarrollamos un monitor pionero en medir en tiempo real la desigualdad y el papel del estado de bienestar en España (Aspachs *et al.,* 2022).

Las fuentes de información en tiempo real se generan tanto en el sector privado (compañías de telefonía, internet, financieras, etc.) como en el seno de la Administración pública, la cual atesora datos de registro que son una fuente inigualable de información económica. La Agencia Tributaria, por poner un ejemplo, posee los datos más detallados y actualizados sobre la evolución de las empresas españolas gracias a las declaraciones de impuestos.

La información que se obtiene a partir de los datos masivos es complementaria a la que ofrecen datos más tradicionales, como las encuestas o los experimentos económicos, y no la sustituye. Sin embargo, los datos masivos ofrecen una mayor granularidad e inmediatez que las fuentes más tradicionales no pueden ofrecer. Al estar creados para otras finalidades, los datos masivos ya están disponibles (no hace falta elaborar una encuesta, por ejemplo, "solo" se tienen que procesar y analizar), con lo que la información económica resultante se cosecha a un menor coste, con mayor celeridad y puede llegar a proporcionar muestras más representativas y de mayor tamaño.

La economía en tiempo real puede jugar también un importante papel en el diseño e implementación de las políticas públicas. Los datos masivos favorecen un diseño más ajustado de las políticas económicas y que se focalicen en aquellos colectivos que más lo necesitan (lo que se conoce como *smart policies*). También permiten un seguimiento *a posteriori* de las medidas, una vez implementadas, y ajustarlas a tiempo real cuando sea necesario. Y todo ello a un menor coste y una mayor disponibilidad que otro tipo de datos.

Por todo ello, CaixaBank Research lleva desde 2018 desarrollando investigación económica con los datos masivos de transacciones financieras, debidamente anonimizados. CaixaBank es un banco idóneo para llevar a cabo estos análisis puesto que es un banco generalista y cuenta con la mayor cuota de mercado en España. Por ejemplo, la cuota de CaixaBank en 2023 en depósitos corresponde al 24,9 % del total nacional, y las cuotas de mercado en ingresos de clientes son superiores, como en nóminas (36,7 % del total) o

en pensiones domiciliadas (34,3 %)[1]. Esto permite obtener una muestra representativa de la población española, cuestión de suma relevancia para la elaboración de indicadores económicos.

En este capítulo analizamos la experiencia de CaixaBank Research con el análisis de datos masivos a partir de la elaboración del portal de economía en tiempo real, explicando los obstáculos y limitaciones encontradas por el camino. Al ser parte de los estudios con los datos internos experimentales, no todos han conseguido su objetivo inicial, pero en todos los casos nos han ayudado a entender mejor los datos de transacciones y su potencial para analizar la economía española. El portal de economía en tiempo real es un proyecto pionero del equipo de CaixaBank Research[2] cuyo objetivo es monitorizar la evolución de la economía española a través de 12 indicadores construidos con datos internos de CaixaBank, agregados mediante técnicas de *big data*. Los indicadores se agrupan en cinco ámbitos: el consumo, la vivienda, el turismo, los salarios y la desigualdad. Cada uno de estos indicadores se puede consultar por distintas dimensiones, lo cual permite identificar diferencias entre colectivos (edad, sexo, ingresos, sector de actividad), regiones (CC. AA., municipios), etc. En total, publicamos más de 800 series de datos que actualizamos mensualmente.

2. EL ANÁLISIS DE LA ECONOMÍA EN TIEMPO REAL A PARTIR DE DATOS MASIVOS: LA EXPERIENCIA DE CAIXABANK RESEARCH

Los movimientos de cuentas bancarias, de alta frecuencia y con un nivel de granularidad elevado, permiten obtener información económica de gran calidad y precisión de manera casi instantánea y con un enorme potencial para la investigación económica. Sin embargo, su análisis es complejo, puesto que se trata de un gran volumen de datos creado para finalidades distintas del análisis económico. Por ello, la utilización de estas fuentes de datos requiere múltiples condiciones previas.

En primer lugar, los datos masivos creados por parte de empresas privadas deben estar disponibles para su análisis estadístico, idealmente en un repositorio único que sea accesible para todos. Esta cuestión no es obvia para muchas empresas que no se dedican a la ciencia de los datos, y se deben invertir recursos para que esto sea posible.

En segundo lugar, este repositorio de datos debe ser accesible para poder analizarlo. En CaixaBank Research iniciamos en 2018 el análisis de datos masivos de transacciones financieras, momento en que se pusieron a disposición del Servicio de Estudios. Sin embargo, un repositorio único de datos existía desde muchos años atrás, aunque no estaba destinado a

[1] https://www.caixabank.com/deployedfiles/caixabank_com/Estaticos/PDFs/Accionistasinversores/Informacion_economico_financiera/231027_Webcast_3T23_es.pdf

[2] El portal ha sido un esfuerzo colectivo en el que han participado Eduard Alcobe, Oriol Aspachs, Patricia Esteban, Alberto Graziano, Javier Ibáñez de Aldecoa, Eduard Llorens, Zoel Martín, Josep Mestres y Judit Montoriol. El presente artículo describe este trabajo colectivo, aunque cualquier error u omisión es responsabilidad mía y no del resto del equipo.

fines de investigación económica. Pero a partir de 2018 se puede acceder a esta capacidad tecnológica. En tercer lugar, se necesita un equipo humano capaz de analizar estos datos. Para ello, a partir de ese año se complementó la formación de los economistas del equipo con cursos en *data science,* a la vez que se incorporan *data scientists* al equipo dedicados plenamente a proyectos de *big data*.

El proceso de aprendizaje en análisis de datos masivos ha sido un largo camino. El conocimiento sobre los datos mismos, a menudo en tablas independientes creadas para fines distintos, y la información contenida en ellas, se fue obteniendo a medida que llevábamos a cabo distintos proyectos de *analytics* en CaixaBank Research. Los primeros análisis con datos internos de CaixaBank permitieron observar, de forma siempre anonimizada, los patrones de gasto de los turistas extranjeros (Campos y Montoriol, 2019) o, en el ámbito inmobiliario, entender las decisiones de compra o alquiler de vivienda (Montoriol, 2020), entre otros análisis. Pero fue la pandemia del COVID-19 la que supuso un revulsivo para su uso. En 2020, en plena pandemia, urgía entender cómo estaba afectando el COVID-19 a los colectivos más vulnerables y cómo estaba respondiendo a la crisis el estado de bienestar. De esa urgencia surgió el Monitor de Desigualdad, germen del portal actual de Economía en tiempo real, impulsado por investigadores de la Universitat Pompeu Fabra, el Institute of Political Economy and Governance y CaixaBank Research. El monitor nos permitió analizar la evolución de millones de nóminas cada mes para obtener una estimación precisa y representativa de la distribución de los salarios tanto para el conjunto de la sociedad, como para los colectivos más vulnerables. Y también constatamos el papel que han jugado las transferencias del sector público para amortiguar el duro golpe que supuso la pandemia (Aspachs *et al.,* 2021). Por primera vez, se pudo monitorear la desigualdad en España con una frecuencia mensual y con un extraordinario nivel de detalle.

Conscientes de las dificultades que entraña la predicción económica y del papel de los datos de alta frecuencia para ayudar a despejarlas, especialmente en un contexto tan complejo como el de la crisis del coronavirus, desarrollamos, también en 2020, el Monitor de consumo[3]. Mediante este indicador logramos tomar el pulso a la actividad española, semana a semana, a partir de los pagos con tarjetas. De nuevo, millones de transacciones al mes.

A partir de entonces, el análisis de la situación económica en nuestras publicaciones se nutrió cada vez más de los datos internos, siempre debidamente anonimizados. Algunos ejemplos de análisis exitosos que hemos llevado a cabo se han beneficiado de un tipo de información particular, como la domiciliación de recibos o de la pensión[4]. Tras el estallido de la guerra en Ucrania, analizamos los recibos de electricidad de clientes particulares domiciliados en CaixaBank para determinar cómo afectó el aumento de los precios en la economía de las

[3] https://www.caixabankresearch.com/es/publications/monitor-consumo

[4] Para más detalle y una lista completa de publicaciones, ver https://www.caixabankresearch.com/es/tendencias-fondo/economia-tiempo-real?_gl=1*1flsxkb*_ga*NjUyOTk5NDc1LjE3MDU1ODc5MjQ.*_ga_0S8KEKC-3M2*MTcwNTY4ODM5MS4zNy4xLjE3MDU2ODkxMzAuMC4wLjA

familias españolas. Hemos utilizado también los datos internos para analizar los patrones de consumo y ahorro tras la jubilación para el caso español. La utilización de datos internos permitía conocer los ingresos, el consumo y el ahorro de los individuos con precisión, así como identificar el momento de la jubilación mediante el primer cobro de la pensión. Otros proyectos no han conseguido el objetivo inicial de análisis, pero han servido como aprendizaje y muchos de los elementos se han usado posteriormente, para mejorar geolocalización de las transacciones o la identificación de tipos específicos de transferencias, para dar algunos ejemplos.

Con todo este aprendizaje se ha conseguido llevar a cabo el portal de Economía en Tiempo Real. El portal define unos indicadores económicos de calidad, publicados casi a tiempo real y con granularidad. Los indicadores publicados están construidos a partir de datos internos anonimizados de CaixaBank, sobre una base de más de 18 millones de clientes, 11.500 cajeros automáticos y 700.000 terminales de punto de venta. Se utilizan técnicas de *big data* para procesar millones de operaciones cada mes, lo que aporta una gran cantidad de información que permite observar la marcha de la economía española de manera prácticamente instantánea y con una alta precisión. Gracias a la digitalización y a la elevada penetración de CaixaBank en el territorio, los datos obtenidos son representativos del conjunto de la sociedad, lo cual permite detectar anticipadamente las nuevas tendencias e identificar diferencias entre colectivos y comunidades autónomas. El portal, pionero en España, fue lanzado en noviembre de 2022 y su acceso está abierto al público, al igual que el resto de las publicaciones de CaixaBank Research, para dar un servicio a la sociedad al crear y divulgar análisis económico y al democratizar el acceso a la información económica de calidad.

3. EL PORTAL DE ECONOMÍA EN TIEMPO REAL: CONSTRUCCIÓN Y REPRESENTATIVIDAD

Esta sección describe la metodología aplicada en la construcción de los indicadores publicados en el portal de Economía en Tiempo Real de CaixaBank Research. Los indicadores están construidos a partir de los datos internos de CaixaBank y se agregan mediante técnicas de *big data.* El objetivo principal fue elaborar indicadores económicos de calidad, representativos del conjunto de la economía española, y con la granularidad e inmediatez temporal que los datos masivos permiten. Su frecuencia de publicación es mensual, con un decalaje de pocos días después del cierre del mes, con la excepción del indicador de la accesibilidad a la vivienda. En este caso, el indicador se publica en bloques de tres meses tras la publicación del precio de la vivienda del MITMA.

El portal de Economía en Tiempo Real de CaixaBank Research ofrece información sobre 12 indicadores, agrupados en cinco ámbitos: consumo, vivienda, salarios, turismo y desigualdad. Cada indicador se muestra para distintas categorías (comunidad autónoma, franjas de edad, franjas de ingresos, sector de actividad, etc.) y se publican unas 850 series en total (ver cuadro 1 para más detalle).

Cuadro 1.

Indicadores disponibles en el portal de Economía en tiempo real

	Indicadores	*Definición*	*Dimensiones*
Consumo	Consumo total, consumo presencial y comercio electrónico	Consumo registrado a partir de pagos y reintegros con tarjetas de débito y de crédito tanto españolas como extranjeras	Nacional, CC. AA., edad, nivel de ingresos, tipo de gasto
Vivienda	Accesibilidad a la vivienda	Número de años de ingresos laborales netos que el hogar mediano debe destinar a la compra de una vivienda en una zona geográfica específica	Nacional, CC. AA. y capital de provincia
Salarios	Salarios	Ingresos salariales mensuales calculados a partir de las nóminas domiciliadas en las cuentas de CaixaBank	Nacional, CC. AA., edad, género, sector de actividad, sector público/privado
Turismo	Turismo doméstico, internacional y gasto exterior	Gasto registrado a partir de pagos y reintegros con tarjetas españolas o extranjeras	Nacional, edad, país, tipo de gasto
Desigualdad	Índice de Gini, percentiles de ingresos, distribución de ingresos, curva de Lorenz	Desigualdad salarial definida a partir de las nóminas domiciliadas así como de los cobros de la prestación y del subsidio de paro	Nacional, CC. AA., edad, género, país de nacimiento

Nota: Para mayor detalle sobre la construcción de las variables y su representatividad, véase la Nota Metodológica (https://www.caixabankresearch.com/es/nota-metodologica_rte).

Fuentes: CaixaBank Research, portal de Economía en tiempo real (https://realtimeeconomics.caixabankresearch.com).

3.1. Consumo

Los indicadores de consumo se refieren al consumo nominal (no deflactado) y contiene tres conjuntos de indicadores: consumo total, consumo presencial y comercio electrónico. Para el consumo doméstico, se considera una muestra cerrada de clientes de CaixaBank, mientras que para los extranjeros no se utiliza ningún filtro en la muestra. Se presenta la variación interanual del indicador, que recoge la mayor parte de la estacionalidad de la serie (no se realiza ningún ajuste adicional por los efectos de calendario).

El consumo total incluye el gasto y los reintegros con tarjetas de crédito y débito emitidas por CaixaBank, y gasto con tarjetas extranjeras en TPV CaixaBank y reintegros en cajeros de CaixaBank. El consumo total engloba el consumo presencial con tarjeta, las retiradas de efectivo y el comercio electrónico (*e-commerce*).

El consumo presencial doméstico incluye el gasto presencial y reintegros con tarjetas de crédito y débito emitidas por CaixaBank y solo se consideran personas físicas. Entre las categorías disponibles, el gasto se puede visualizar para el total nacional y para la C. A. del TPV, por franja de edad (16-29 años, 30-49 años, 50-64 años, 65 años y más), franja de ingre-

sos (ingresos bajos, medios y altos) y por sector de actividad. Los sectores de actividad están clasificados a partir del sector del comercio propietario del TPV de la siguiente manera:

a. Primera necesidad (alimentación, farmacias).

b. Bienes duraderos (tecnología, menaje, textil, deportes, jugueterías y joyerías).

c. Ocio y restauración (restaurantes, casinos, clubs deportivos, espectáculos, museos, parques temáticos).

d. Transporte (transporte de viajeros, autopistas, parkings y gasolineras).

e. Turismo (hoteles, agencias de viajes y alquiler de vehículos).

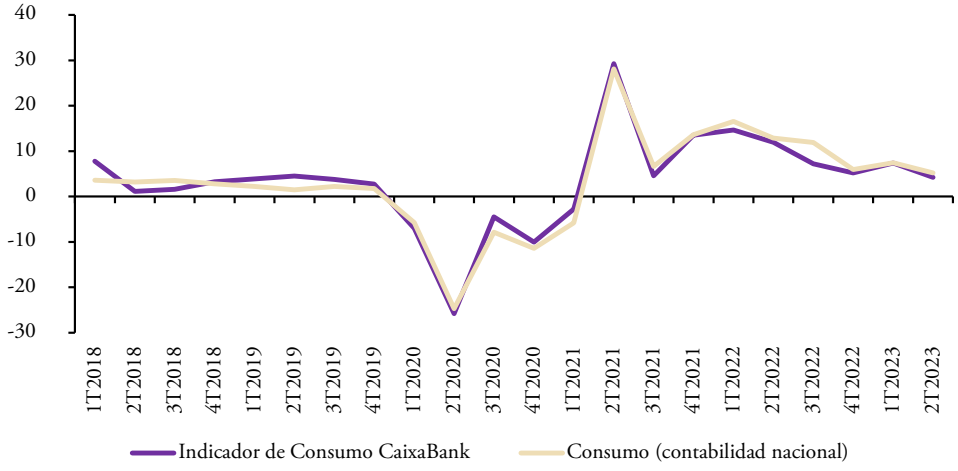
El comercio electrónico doméstico incluye el gasto con tarjetas de crédito y débito emitidas por CaixaBank en un TPV virtual. Solo se consideran personas físicas y se muestra la media móvil de los últimos dos meses.

La comparativa del indicador de consumo con una referencia externa (figura 1) muestra cómo la suma de consumo presencial doméstico y comercio electrónico doméstico puede considerarse una buena aproximación de la serie de gasto en consumo final de los hogares (nominal y no ajustada de estacionalidad y calendario) publicada por el INE en la contabilidad nacional trimestral de España. La figura 1 muestra la comparativa en términos interanuales y puede apreciarse una correlación entre ambas series muy elevada, del 0,98.

Figura 1.

Comparativa del indicador de consumo con una referencia externa

Variación interanual (%)



Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank y del INE.

3.2. Acceso a la vivienda

En la sección de vivienda, el indicador clave es la *ratio de accesibilidad a la vivienda* en propiedad. Esta ratio mide el número de años de ingresos laborales netos que el hogar mediano debe destinar a la compra de una vivienda en la zona geográfica *i*. Un valor más alto indica mayor dificultad para acceder a una vivienda en propiedad. Siguiendo la metodología del Banco de España, se calcula con la siguiente expresión:

$$Ratio\ Accesibilidad_i \frac{PMC_i * 93,75}{mediana\ ingresos_i * 1,6} \quad [1]$$

Donde

- PMC_i = precio de la vivienda por metro cuadrado en la zona geográfica *i* (media móvil de 12 meses).
- 93,75 m^2 es la superficie de referencia que usa el Banco de España para calcular la ratio de accesibilidad a nivel nacional.
- La mediana de los ingresos laborales (nóminas y subsidio de paro) por persona en la zona geográfica *i* (media móvil de 12 meses para reducir la estacionalidad por las pagas extras).
- 1,6 es el número promedio de perceptores de ingresos en un hogar (*EPA*).

Los datos del precio de la vivienda (PMC) provienen del Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA) y tienen frecuencia trimestral. Para pasarlos a frecuencia mensual se replica el mismo dato de PMC para todos los meses del trimestre, y se calcula una media móvil de 12 meses del PMC para suavizar la serie. En consecuencia, los datos de la ratio de accesibilidad se publican en el portal de Economía en tiempo real en bloques de tres meses tras la publicación del dato del MITMA. Las categorías disponibles incluyen el total nacional, CC. AA. y capitales de provincia (68 series).

La comparativa del indicador de accesibilidad a la vivienda con una referencia externa se realiza con la ratio de accesibilidad a la vivienda que publica del Banco de España (BdE) a nivel nacional (figura 2)[5]. El BdE usa el precio de la vivienda del MITMA hasta 2005; luego se enlaza con la variación interanual de Tinsa para los años 2005 y 2006, y a partir de 2007 se enlaza con la variación interanual del IPV del INE. El indicador CaixaBank usa el precio del MITMA por la disponibilidad de datos de PMC por capitales de provincia.

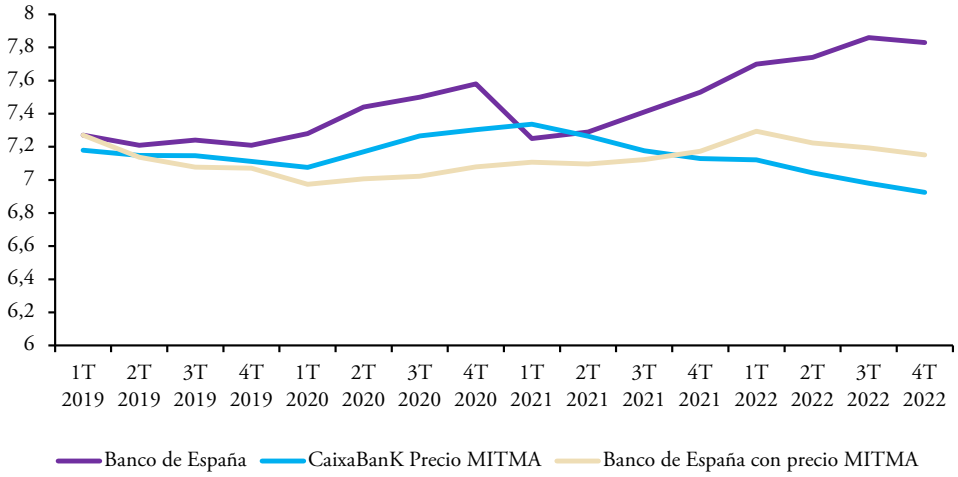
Para los ingresos, el BdE usa la renta bruta disponible del hogar mediano procedente de la *Encuesta Financiera de las Familias* (trianual) y extrapolando el resto de los años a partir de otras fuentes estadísticas. El indicador CaixaBank usa la mediana de los ingresos (nóminas y subsidios de paro) por persona y en términos netos (ingreso a cuenta del trabajador).

[5] El Banco de España publica trimestralmente la ratio de accesibilidad en la Síntesis de Indicadores, cuadro 1.5: https://www.bde.es/webbde/es/estadis/infoest/sindi.html

Los datos del BdE tienen frecuencia trimestral. Para la comparación con el indicador creado a partir de datos internos de CaixaBank, escogemos el último mes del trimestre (serie suavizada). La siguiente figura muestra tres series de datos: 1) la ratio de accesibilidad CaixaBank calculada a partir de la mediana de los ingresos (datos internos) y el precio de la vivienda del INE; 2) el indicador CaixaBank de la ratio de accesibilidad, calculado a partir de la mediana de los ingresos (datos internos) y el precio de la vivienda del MITMA; y (3) la ratio de accesibilidad del BdE. El primer indicador CaixaBank es el directamente comparable con el BdE por usar el mismo precio de la vivienda en los años recientes (INE). Su correlación es de 0,97. Sin embargo, el indicador CaixaBank publicado en el portal usa el precio de la vivienda del MITMA por tener mayor detalle geográfico (municipios). En este caso, la correlación con la ratio de accesibilidad del BdE es positiva, pero más débil (0,13) debido al menor crecimiento del precio del MITMA en comparación con el precio del INE en el periodo considerado.

Figura 2.

Comparativa de la ratio de accesibilidad con una referencia externa (Banco de España)



Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank, del Banco de España y del MITMA.

3.3. Salarios

Los salarios se miden en términos nominales (no deflactados) a partir de los ingresos salariales mensuales de personas físicas desde los movimientos en la cuenta corriente identificados específicamente como nómina. Se incluyen todos los ingresos de nómina acumulados en el mes para el cálculo de un salario mensual. Se consideran todos los clientes con nómina presente en el mes anterior como criterio de estabilización. El indicador de salarios corresponde a la mediana de la variación interanual de la nómina mensual, que se calcula cliente a cliente. El indicador corresponde a la media móvil de dos meses.

Las categorías disponibles corresponden a ámbitos geográficos (total nacional y CC. AA.), franja de edad, género, sector público/privado y cinco sectores de actividad:

a. Agricultura (CNAE A)

b. Industria (CNAE B-E)

c. Construcción (CNAE F)

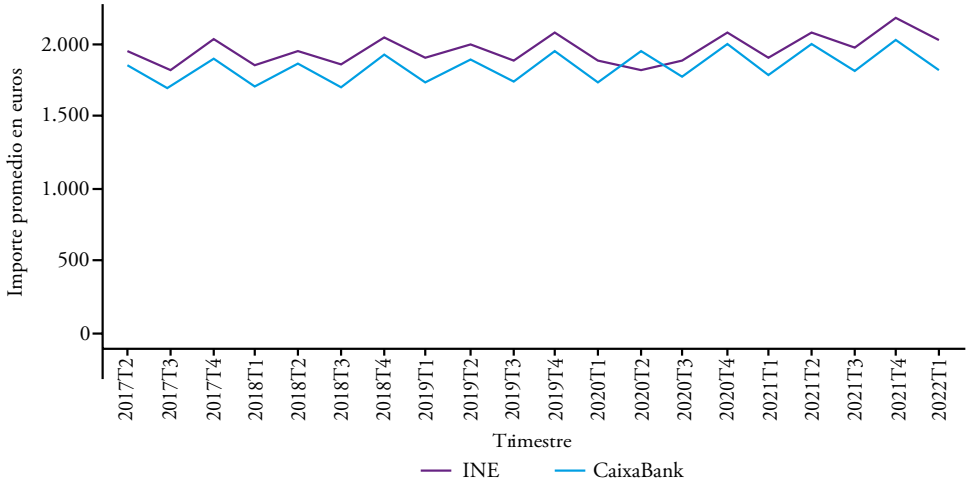
d. Servicios turísticos (CNAE a dos dígitos 49-51, 55, 56, 77, 79, 90, 91, 93)

e. Servicios no turísticos (CNAE G a J y L a U, excluyendo a aquellos incluidos en los servicios turísticos).

La primera referencia externa para comparar el indicador interno de salarios corresponde al coste salarial total por trabajador de la *Encuesta Trimestral de Coste Laboral (ETCL)* del INE (figura 3). Este coste salarial es el que soportan los empleadores para emplear a su plantilla (excluye otros costes laborales como las cotizaciones sociales, etc.) y está expresado en términos brutos (al contrario que el indicador interno de salarios, que está expresado en

Figura 3.

Comparativa del indicador de salario con una referencia externa (ETCL)



Nota: El coste salarial total por trabajador de la *Encuesta Trimestral de Coste Laboral (ETCL)* del INE corresponde al coste salarial promedio que los empleadores soportan para emplear a su plantilla (excluye otros costes laborales como las cotizaciones sociales, etc.) y está expresado en términos brutos. El indicador interno de salarios está expresado en términos netos y la comparativa con el INE realizada a partir del promedio.

Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank y de la *Encuesta Trimestral de Coste Laboral (ETCL)* del INE.

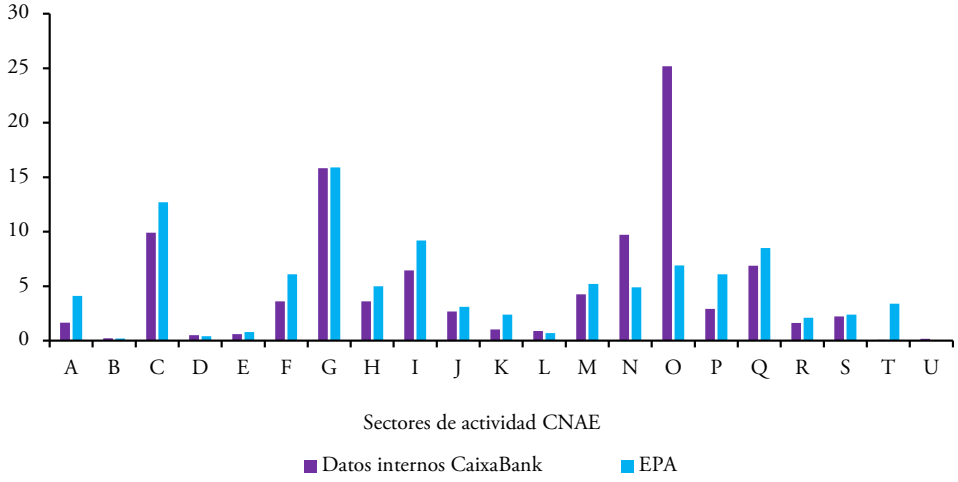
términos netos). En la figura se puede observar la similitud entre las series en el agregado para España, tanto en el nivel como en la estacionalidad (con la excepción del segundo trimestre de 2020).

Asimismo, se puede comprobar la representatividad de los datos internos de nóminas de CaixaBank con la *Encuesta de Estructura Salarial (EES)* en la sección de desigualdad (sección 3.4.). Otra comparativa realizada corresponde a la representatividad sectorial a nivel de sección CNAE entre los datos internos y la *Encuesta de Población Activa*. La figura 4 muestra que la distribución es muy similar en la mayoría de los sectores de actividad. Las dos únicas excepciones corresponden al sector público (grupo CNAE O), en el que la muestra de CaixaBank está sobrerrepresentada por su mayor cuota de mercado, y el sector de empleadas del hogar (grupo CNAE T), en el que la muestra interna está infrarrepresentada puesto que a menudo la retribución se paga en efectivo.

Figura 4.

Comparativa de la distribución sectorial de los asalariados entre la muestra interna y la *EPA*

Peso del número de asalariados en cada sector sobre el total (%)



Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank y de la *Encuesta de Población Activa (EPA)* del INE correspondientes al 3T 2022.

3.4. Turismo

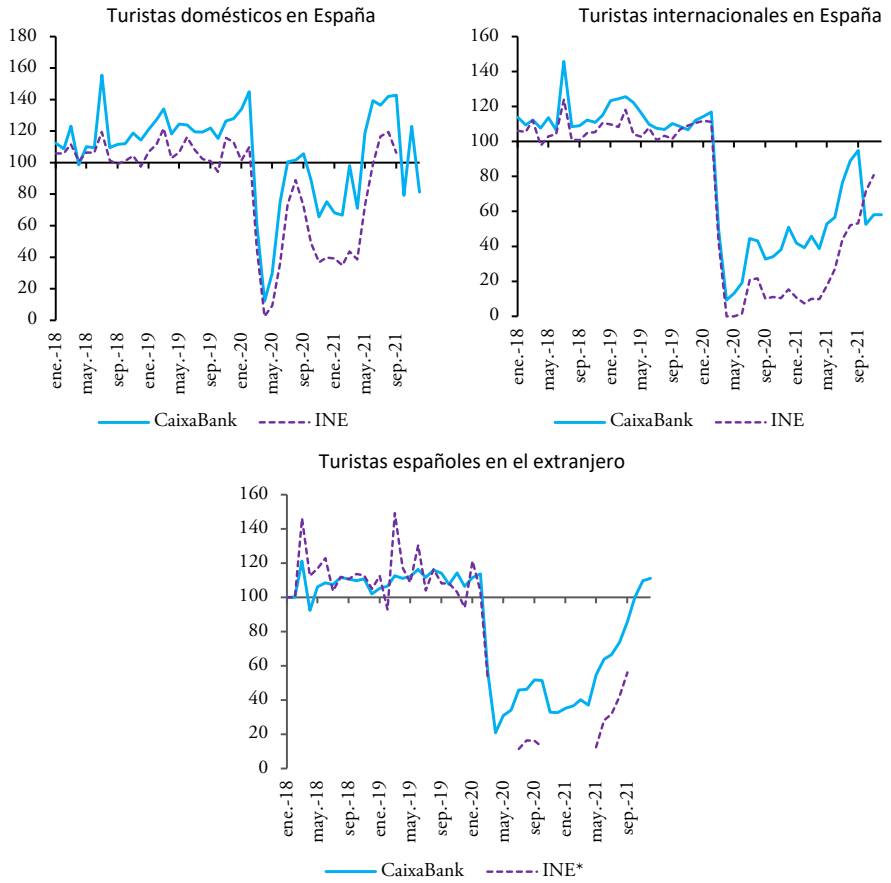
El bloque de turismo contiene tres indicadores de gasto nominal (no deflactado): turismo doméstico, turismo internacional y gasto exterior (gasto turístico de españoles en el extranjero). Se aplica una media móvil de tres meses sobre la serie de gasto y, a continuación, se calcula la tasa de variación interanual de la serie suavizada.

El turismo doméstico (gasto turístico doméstico) se identifica mediante los pagos presenciales y retiradas de efectivo realizados por los turistas españoles en España. Consideramos como gasto de los turistas domésticos los pagos en TPV de CaixaBank con tarjetas emitidas por entidades españolas efectuados fuera de sus áreas habituales de consumo (área de residencia, trabajo, consumo rutinario, etc.). Las categorías disponibles corresponden a distintos ámbitos geográficos (total nacional y CC. AA.) y distintos sectores de actividad (alojamiento, comercio y ocio y restauración).

Figura 5.

Comparativa de los niveles de gasto según destino y origen

Índice (100 = mismo mes de 2017)



Nota: El INE registra gasto nulo para turistas internacionales en España en los meses de abril y mayo de 2020, y no publica resultados de gasto de españoles en el extranjero para 8 meses de 2020 y 2021, debido al reducido número de encuestas realizadas.

Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank y del INE.

El turismo internacional (gasto turístico internacional en España) corresponde a los pagos presenciales y retiradas de efectivo realizados por los turistas internacionales en España. Consideramos como gasto de turistas internacionales los pagos efectuados en España en TPV de CaixaBank con las tarjetas emitidas por entidades extranjeras. Las categorías disponibles corresponden a ámbitos geográficos (total nacional y CC. AA.), país de origen (Alemania, Estados Unidos, Francia, Reino Unido, otros de Europa y otros países) y sectores de actividad (alojamiento, comercio y ocio y restauración).

El gasto exterior (Gasto turístico español en el extranjero) se identifica mediante los pagos presenciales y retiradas de efectivo con tarjetas de crédito y débito emitidas por CaixaBank realizados en el extranjero. Las categorías disponibles corresponden a ámbitos geográficos (total nacional y CC. AA. de residencia del turista), franja de edad, franja de ingresos y país/región de destino.

Las referencias externas para comparar los indicadores de CaixaBank son los indicadores de gasto turístico publicados por el INE en las encuestas *EGATUR* (encuesta de gasto de turistas internacionales) y *FAMILITUR* (encuesta de turismo doméstico). En la figura 5 puede apreciarse la similar dinámica de las series del INE y los indicadores calculados con datos internos de CaixaBank, si bien se observa un sesgo positivo en el nivel capturado por estos últimos. Este sesgo puede deberse a multitud de factores, si bien estimamos que los principales son el uso cada vez más predominante de los pagos con tarjeta, la inclusión de tarjetas de extranjeros residentes en España y la mejor representación del turismo doméstico de cercanía (excursiones o estancias muy cortas) y el turismo de frontera. Este turismo se mostró considerablemente más resiliente a partir de marzo de 2020, con el inicio de la pandemia.

3.5. Desigualdad

El bloque de desigualdad contiene cuatro indicadores que miden la desigualdad de ingresos salariales: el índice de Gini, los percentiles de ingresos, las distribuciones de ingresos y las curvas de Lorenz.

El indicador del índice de Gini corresponde a la variación del índice de Gini[6], en puntos porcentuales, respecto a febrero de 2020, corregida por la variación estacional promedio experimentada durante el mismo periodo en 2018 y 2019. Se presentan dos series del índice de Gini, que se calculan a partir de la distribución mensual de los ingresos salariales antes y después de las transferencias públicas. La diferencia entre los índices antes y después de las transferencias públicas refleja en qué medida los esquemas de soportes de rentas (como pueden ser las prestaciones por desempleo o las que reciben los trabajadores que se encuentran en un ERTE) reducen la desigualdad salarial. Las categorías disponibles corresponden a ámbitos geográficos (total nacional y CC. AA.), franja de edad, género y país de nacimiento (España, extranjero).

[6] El índice de Gini es el índice de referencia para analizar el nivel de desigualdad salarial. Puede tomar valores entre 0 y 100. Cifras más elevadas reflejan niveles de desigualdad más altos, y viceversa.

Los percentiles de ingresos se definen como la ratio entre dos percentiles de la distribución de ingresos después de las transferencias del sector público: 1) ratio entre los percentiles 80 y 20 (p80/p20); 2) ratio entre los percentiles 80 y 50 (p80/p50), y 3) ratio entre los percentiles 50 y 20 (p50/p20). En este caso, no se presentan las ratios de percentiles antes de las transferencias públicas. Las categorías disponibles corresponden a ámbitos geográficos (total nacional y CC. AA.) y género.

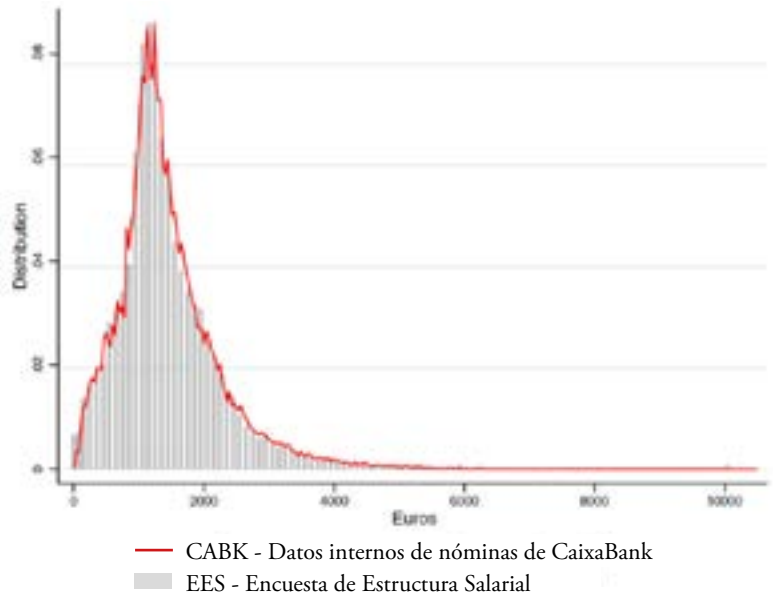
Las distribuciones de ingresos corresponden a la distribución mensual de los ingresos salariales antes y después de las transferencias del sector público. Las franjas representan el porcentaje de población con "ingresos bajos", que se refiere a salarios o transferencias inferiores a 1.000 euros/mes; "ingresos medios", entre 1.000 y 2.000 euros/mes; "ingresos altos", más de 2.000 euros/mes; y "sin ingresos". Las categorías disponibles corresponden a ámbitos geográficos (total nacional y CC. AA.), franja de edad, género y país de nacimiento (España, extranjero).

Las curvas de Lorenz corresponden a la distribución cumulativa de los ingresos salariales a lo largo de la población. Cada punto de la curva de Lorenz muestra qué porcentaje

Figura 6.

Distribución de salarios mensuales netos en España

Frecuencia (%)



Nota: Para facilitar la comparación entre muestras, ajustamos la distribución salarial de la *ESS* de 2018 por el aumento salarial promedio entre 2018 y 2019.

Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank y de la *Encuesta de Estructura Salarial* (INE).

acumulado de los ingresos tiene un determinado porcentaje acumulado de la población. La curva empieza en el origen (0,0) y termina en el punto (100,100). Si los ingresos salariales estuvieran distribuidos de manera igual en toda la población, la curva correspondería con la línea de 45 grados. Si una persona dispusiera de todos los ingresos, la curva correspondería al eje horizontal hasta el punto (100,0), donde pasaría al punto (100,100). Cuanto más cerca está la curva de la perfecta igualdad, más equitativa es la distribución de ingresos (es decir, menor desigualdad existe), y viceversa.

Para confirmar la representatividad de los datos internos de CaixaBank con datos externos, utilizamos la *Encuesta de Estructura Salarial (EES),* y restringimos la muestra a los asalariados para poder comparar la distribución de las nóminas. Como se puede observar en la figura 6, las dos distribuciones son extremadamente parecidas.

Para corroborar este hecho, comparamos también las ratios de los percentiles de cada distribución y las características de las personas con una nómina en CaixaBank con los de la *ESS* en el cuadro 2. Así, podemos comprobar que el peso de distintos colectivos en los datos de CaixaBank es muy parecido al de la *EES*, y, por tanto, igual de representativo del conjunto de la población asalariada (aunque con un mayor tamaño muestral).

Cuadro 2.

Comparativa de los ratios de percentiles de la distribución de salarios netos y de la distribución por edad y género

	CaixaBank (2018)	*Encuesta de Estructura Salarial (2018)*
Ratio de percentiles		
P90/P10	4,23	4,04
P90/P50	1,95	1,86
P10/P50	0,46	0,46
P75/P25	1,87	1,82
Género (%)		
Hombre	53	52
Mujer	47	48
Edad (%)		
15-19	1	0
20-29	19	12
30-39	25	25
40-49	28	32
50-59	21	24
60+	6	7
Tamaño muestral	2.325.908	216.726

Nota: La tabla muestra las ratios de los percentiles de la distribución de salarios netos y la distribución por edad y género separadamente para la muestra interna de CaixaBank y para la *Encuesta de Estructura Salarial (ESS)* en el año 2018.

Fuentes: CaixaBank Research, a partir de datos internos de CaixaBank y de la *Encuesta de Estructura Salarial* (INE).

Tener unos datos que sean representativos del conjunto de la población nos permite construir indicadores de desigualdad, como el índice de Gini o las curvas de Lorenz, para el conjunto de la población y por los distintos subgrupos de población, y analizar su evolución a lo largo del tiempo.

4. TOMANDO EL PULSO DE LA ECONOMÍA ESPAÑOLA A PARTIR DE LOS INDICADORES DEL PORTAL DE ECONOMÍA EN TIEMPO REAL

El portal de Economía en Tiempo Real permite analizar la coyuntura económica con indicadores representativos del conjunto de la economía española, como vimos en el apartado anterior. En esta sección se describen algunos de los resultados que se pueden obtener con los indicadores. Con todo, al actualizar el portal mensualmente y con poca latencia, le recomendamos al lector ir directamente a la página web https://realtimeeconomics.CaixaBankresearch.com/#/home para analizar los últimos indicadores. Parte del esfuerzo de la creación del portal consistió en crear distintos niveles de entrada, con un apartado "Descubrir" para todos los públicos en el que la información ya está analizada y se proporcionan directamente los resultados principales; otro más interactivo para un público, como periodistas o *policy makers,* más interesado en cuestiones concretas y donde se pueden seleccionar los indicadores y el nivel de detalle deseado; hasta el nivel en el que se pueden descargar datos y gráficos con mayor desagregación para realizar análisis e investigaciones académicas propios.

El indicador de consumo muestra el dinamismo del consumo en la segunda mitad del año 2023, con un ritmo de crecimiento alrededor del 7,0 % en promedio, aunque algo inferior a la primera mitad del año (9,4 % interanual). El consumo presencial doméstico en España, medido a partir del gasto presencial y reintegros con tarjetas de crédito y débito emitidas por CaixaBank, cerró el año 2023 creciendo un 3,3 % interanual en diciembre (3,7 % en 2S 2023), con una desaceleración respecto al primer semestre (5,8 %), pero que no fue a más en el segundo semestre del año. Por su parte, el comercio electrónico, medido a partir del gasto con tarjetas de crédito y débito emitidas por CaixaBank en un TPV virtual, mantiene su ritmo de crecimiento de doble dígito, al avanzar un 13,3 % interanual en diciembre de 2023 (11,4 % en 2S 2023).

El indicador de salarios muestra que en 2023 aumentaron las retribuciones de los trabajadores, pero sin producirse efectos de segunda ronda significativos. Los ingresos salariales del sector privado español aumentaron un 4,3 % interanual en diciembre de 2023, con lo que se mantiene el avance alrededor del 4 % a lo largo de 2023 en las nóminas del sector privado (4,1 % en promedio). Con todo, este ritmo es 1,7 p.p. superior al que experimentaron un año antes (+2,6 % interanual en diciembre de 2022). Por sectores, los ingresos salariales avanzaron un 4,3 % en 2023 en el sector de servicios turísticos, un ritmo algo superior al del sector de servicios no turísticos (3,9 %) y de la industria (3,8 %). Por su parte, los ingresos salariales en 2023 avanzaron algo menos en la construcción (3,1 %) y en la agricultura (2,5 %).

El indicador de accesibilidad de la vivienda permite observar cómo un hogar típico destina 6,8 años de renta íntegra para comprar una vivienda en España, aunque existen marcadas

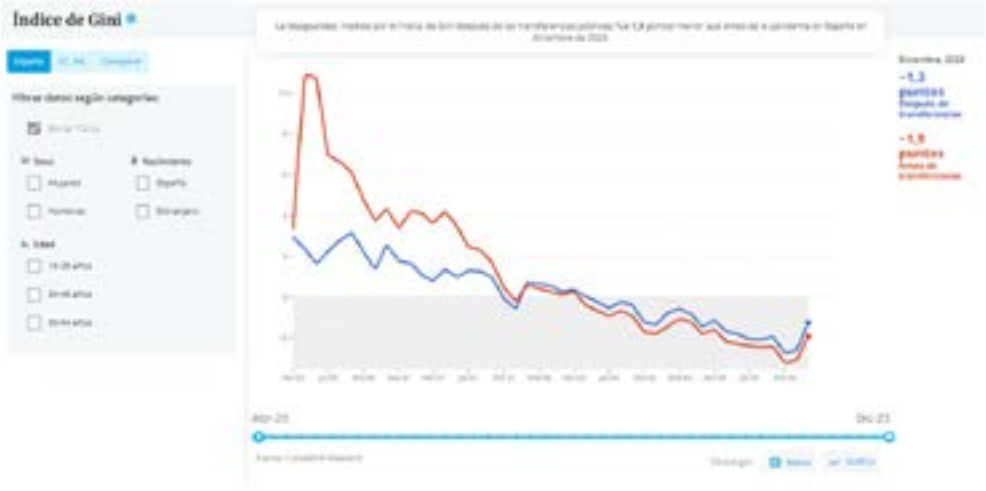
diferencias entre regiones. Madrid y Baleares son las CC. AA. donde más cuesta acceder a la vivienda (11 y 10 años de renta); mientras que Extremadura y Castilla-La Mancha son las CC. AA. donde menos (3,6 años).

Los indicadores de turismo muestran un fuerte crecimiento del turismo internacional en España en 2023. En la segunda mitad del año, el gasto turístico de los extranjeros en España pasó de crecer un 8,3 % en agosto a un 13,2 % en diciembre. El turismo interior de los españoles se mantuvo en niveles elevados al inicio del año (9,2 % interanual en enero), aunque ha ido perdiendo fuelle a lo largo del año (1,2 % en diciembre). Por otro lado, los indicadores de gasto en el extranjero muestran que los españoles han salido al extranjero mucho más en 2023 que en el año anterior, con crecimientos del gasto por encima del 6 % interanual en todo el año.

Los indicadores disponibles en el apartado de desigualdad nos muestran que se ha deshecho el impacto de la pandemia en la desigualdad y que esta ya alcanza un nivel inferior al que existía antes de la pandemia (figura 7). Concretamente, la desigualdad medida por el índice de Gini antes de las transferencias públicas fue 1,9 puntos menor que antes de la pandemia en España en diciembre de 2023 (-1,4 puntos en diciembre de 2022). Alcanzar una situación de desigualdad salarial menor a la de febrero de 2020 es un hito remarcable, porque la pandemia afectó fuertemente a las nóminas de los trabajadores. La desigualdad en España, medida por el índice de Gini, aumentó en 10,8 puntos en tan solo dos meses, de febrero a abril de 2020 antes de las transferencias públicas. A lo largo de los siguientes meses, y en particular a partir de la primavera de 2021, con una mayor reactivación económica, se aceleró la

Figura 7.

Índice de Gini



Fuentes:

reducción de la desigualdad. Las transferencias públicas jugaron un papel clave para contener ese aumento de la desigualdad, dado que las ayudas permitieron, en el momento álgido de la crisis (abril de 2020), reducir la desigualdad en un 80 % aproximadamente. Dos años y medio después, en diciembre de 2023, el índice de Gini, una vez incluidas las transferencias públicas era ya 1,3 puntos menor que antes de la pandemia en España.

5. CONCLUSIÓN

Los movimientos de cuentas bancarias, de alta frecuencia y con un nivel de granularidad elevado, permiten obtener información económica de gran calidad y precisión de manera casi instantánea y con un enorme potencial para la investigación económica. En este trabajo hemos descrito la experiencia de CaixaBank Research en el análisis de datos masivos de transacciones financieras a partir de los proyectos llevados a cabo, y en particular, mediante la creación de un portal de Economía en Tiempo Real. Este proyecto monitoriza la evolución de la economía española a través de 12 indicadores construidos con datos internos de CaixaBank, agregados mediante técnicas de *big data*. Los indicadores se agrupan en cinco ámbitos: el consumo, la vivienda, el turismo, los salarios y la desigualdad. Cada uno de estos indicadores se puede consultar por distintas dimensiones, lo cual permite identificar diferencias entre colectivos (edad, sexo, ingresos, sector de actividad), regiones (CC. AA., municipios), etc. En total, actualmente publicamos más de 800 series de datos que actualizamos mensualmente. El proyecto está en curso de mejora constante, y se añadirán al portal nuevos indicadores y dimensiones a medida que estén disponibles.

Referencias

Aspachs, O., Durante, R., Graziano, A., Mestres, J., G. Montalvo, J. y Reynal-Querol, M. (2021). Tracking the impact of COVID-19 on economic inequality at high frequency. *PLoS One, 16*(3), e0249121. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249121

Aspachs, O., Durante, R., Graziano, A., Mestres, J., G. Montalvo, J. y Reynal-Querol, M. (2022). Real-time inequality and the welfare state in motion: evidence from COVID-19 in Spain.

Bisquert Perles, M., Mestres Domènech, J., Sánchez Palomino, J. A. y Vidal Martínez, A. (2023).“Patrones de consumo y ahorro tras la jubilación. *Informe Mensual Junio 2023.* CaixaBank Research. https://www.caixabankresearch.com/es/economia-y-mercados/mercado-laboral-y-demografia/patrones-consumo-y-ahorro-tras-jubilacion

Campos, A. y Montoriol, J. (2019). ¿Cómo puede el big data potenciar la sostenibilidad del sector turístico? *Informe Sectorial Turismo,* enero de 2019. CaixaBank Research. https://www.caixabankresearch.com/sites/default/files/content/file/2022/04/05/91184/is_turismo1_2022_esp.pdf

Chetty, R., Friedman, J. N. y Stepner, M. (2020). The economic impacts of COVID-19: Evidence from a new public database built using private sector data (No. w27431). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w27431/w27431.pdf

Graziano, A. y Mestres, J. (2022). Los precios de la luz están por las nubes, ¿y el importe de su recibo? *Informe Mensual Enero 2022.* CaixaBank Research. https://www.caixabankresearch.com/es/economia-y-mercados/actividad-y-crecimiento/precios-luz-estan-nubes-y-importe-su-recibo?1150

Montoriol, J. (2020). ¿Comprar o alquilar? Una cuestión de ingresos, pero sobre todo de capacidad de ahorro. *Informe Sectorial Inmobiliario Julio 2020.* CaixaBank Research. https://www.caixabankresearch.com/es/analisis-sectorial/inmobiliario/comprar-o-alquilar-cuestion-ingresos-sobre-todo-capacidad-ahorro?1150

CAPÍTULO V

Predicción de la volatilidad: una comparación entre métodos paramétricos y semiparamétricos

Isabel Casas*
J. Miguel Marín
Helena Veiga

En este capítulo se estudia la eficacia de distintos modelos (paramétricos, semiparamétricos y semiparamétricos con aprendizaje automático) en la predicción de la varianza diaria realizada, utilizando datos intradiarios de Bitcoin, NASDAQ y S&P500 que representan distintos grupos de mercados: criptomonedas, tecnología y el mercado de acciones estadounidense, respectivamente. La disponibilidad de datos a frecuencias muy altas, característica del *big data,* constituye una gran ventaja, ya que posibilita la estimación de la volatilidad de forma consistente y, por ende, su predicción con mayor precisión.

Se ha llevado a cabo una comparación de los modelos en términos de predicción de varianza realizada, mediante test de habilidad predictiva, tanto incondicionales como condicionales, así como en la predicción del valor en riesgo y se ha observado que el modelo autorregresivo heterogéneo de cuarticidad (o variación cuártica) sobresale en la predicción de la varianza realizada.

En cuanto al valor en riesgo, se han empleado trece contrastes condicionales para evaluar el rendimiento de los diversos modelos. Se obtiene que, en general, no presentan un buen desempeño durante el periodo correspondiente a la pandemia global del coronavirus. Los mejores resultados se observan en Bitcoin y NASDAQ, usando los modelos heterogéneos autorregresivos con cuarticidad realizada, *Random Forests* y el modelo heterogéneo autorregresivo con parámetros variables en el tiempo.

Palabras clave: modelos de aprendizaje automático, modelos HAR, varianza realizada, valor en riesgo.

* Los autores desean expresar su profundo agradecimiento al profesor Daniel Peña, así como a las profesoras Pilar Poncela y Eva Senra, por la generosa invitación a participar en estas jornadas. Además, queremos destacar nuestro agradecimiento a la Fundación Funcas por su apoyo financiero. También agradecemos la ayuda financiera de los proyectos de investigación PID2021-122919NB-I00 y PID2022-139614NB-C22, y a la Fundação para a Ciência e a Tecnologia, proyecto UIDB/00315/2020.

1. INTRODUCCIÓN

La predicción de la varianza realizada en los rendimientos de activos financieros importantes, como el Bitcoin (BTC), NASDAQ y S&P 500 (SP500), es esencial debido a su influencia significativa en los mercados financieros globales. Estos activos no solo son indicativos de la salud económica y financiera, sino que también su volatilidad es un área de interés crítico para un amplio espectro de participantes del mercado, incluidos inversores, analistas y gestores de carteras.

Entender y predecir esta volatilidad no solo facilita la toma de decisiones informadas en inversión y gestión de carteras, sino que también es vital en la asignación de activos y la gestión de riesgos. Esto es particularmente importante, ya que permite a los participantes del mercado anticipar y mitigar los efectos adversos de eventos inesperados y fluctuaciones bruscas en los precios.

Además, la capacidad de predecir con precisión la varianza realizada es un hecho muy valioso en la gestión de riesgos. Los modelos predictivos eficaces permiten identificar períodos de alta volatilidad, lo cual es crucial para la implementación de medidas preventivas, como el ajuste de las carteras de inversión y la aplicación de estrategias de cobertura. Estas acciones son fundamentales para mitigar los riesgos asociados a la volatilidad en los mercados financieros.

En la era de automatización, el seguimiento detallado de los precios de los activos financieros de forma continua facilita el cálculo de la varianza realizada (RV) diaria a partir de los rendimientos financieros intradiarios. Esta medida ofrece una representación más precisa de la volatilidad y, por ende, del riesgo financiero. La predicción de esta varianza y, en consecuencia, del riesgo del activo es crucial, por ejemplo, para tomar decisiones con conocimiento de causa en la gestión del riesgo, tanto para inversores como para instituciones financieras.

En este contexto, sugerimos la adopción de diversos métodos para predecir la varianza realizada. Estos incluyen enfoques paramétricos clásicos, así como métodos semiparamétricos y de aprendizaje automático, que se aplican a series financieras como Bitcoin (BTC), NASDAQ y S&P 500 (SP500), que representan diferentes grupos de mercados: criptomonedas, tecnología y el mercado de acciones estadounidense, respectivamente. Nuestro objetivo es realizar una comparativa exhaustiva entre las metodologías empleadas en el ámbito de predicción de la volatilidad y en el cálculo del valor en riesgo (VaR).

Dentro de los enfoques paramétricos, estimamos modelos heterogéneos autorregresivos (HAR) que se utilizan para predecir la volatilidad de los mercados financieros. Los modelos HAR utilizan información histórica de los precios para estimar cómo la volatilidad de un activo financiero cambia a lo largo del tiempo en diferentes intervalos, desde momentos cercanos en el tiempo hasta momentos más distantes en el futuro. Aunque los modelos HAR no se clasifican formalmente como modelos de memoria larga, se ha identificado que muestran

una alta persistencia y reflejan las características de memoria larga de ciertas series financieras. Por ello, son apropiados para modelar la varianza realizada.

Para poder predecir la volatilidad con mayor flexibilidad, añadimos coeficientes variantes en el tiempo a los modelos HAR para crear los TVHAR que estimaremos semiparamétricamente, lo cual permite capturar las no linealidades potenciales presentes en la varianza realizada.

Es importante destacar que los *shocks* negativos pueden influir más en la volatilidad que los positivos de igual magnitud. La razón es que los inversores suelen ser más sensibles a la incertidumbre que generan los *shocks* negativos, dada la mayor probabilidad de incurrir en pérdidas significativas. Por lo tanto, incorporar la asimetría de la volatilidad en los modelos de varianza realizada podría mejorar la precisión de las predicciones y el cálculo del valor en riesgo.

Los métodos de aprendizaje automático que utilizaremos son el *Deep Learning (DL)* y *Random Forests* (RF). Los primeros permiten ajustar redes neuronales de arquitectura *feedforward* y con capas ocultas, que posibilitan predecir observaciones futuras basadas en la estructura de los datos previos. Por otro lado, en los RF se emplean múltiples árboles de decisión en su entrenamiento, generando resultados basados en el consenso de estos árboles, o en su promedio para la predicción de nuevas observaciones.

Para analizar los resultados de la predicción de la varianza realizada, hemos utilizado contrastes de habilidad predictiva incondicionales y condicionales, llegando a la conclusión de que, para el BTC condicional a sus rendimientos, los mejores modelos son los modelos HAR que incluyen la cuarticidad realizada[1], el modelo HAR cuyos coeficientes cambian en el tiempo y RF.

Usamos como variable condicional los rendimientos, dada la ley financiera de compensación de "a mayor riesgo, mayor rendimiento". Para el NASDAQ y SP500, utilizamos como variable condicional la volatilidad de mercado medida por el VIX, que es un índice de volatilidad en tiempo real creado por el Chicago Board Options Exchange (CBOE). Este índice fue el primero en cuantificar las expectativas del mercado respecto a la volatilidad. Los resultados indican que los mejores modelos para el NASDAQ son los modelos HAR simétrico y asimétrico que incluyen la cuarticidad realizada, y los RF. Lo mismo para el SP500, aunque el RF no ha sido calificado como un modelo dominante para la predicción de la varianza realizada. Estos resultados se basan en una función de pérdida robusta.

[1] La cuarticidad realizada es un concepto utilizado en econometría financiera, especialmente en la modelización de datos financieros de alta frecuencia. Forma parte de una familia más amplia de medidas conocidas como momentos realizados, que se utilizan para estimar la volatilidad y otros momentos superiores de los rendimientos de los activos durante un periodo de tiempo específico. Mientras que la varianza realizada (segundo momento) y la asimetría realizada (tercer momento) se discuten a menudo, la cuarticidad realizada (cuarto momento) es crucial por varias razones, en particular en la estimación de la curtosis.

En cuanto al cálculo del valor en riesgo, los mejores modelos, considerando 13 pruebas condicionales de *backtesting*, son para el Bitcoin, los modelos HAR simétrico y asimétrico que incluyen la cuarticidad realizada, y para el NASDAQ, el modelo HAR con coeficientes que varían en el tiempo. En relación al SP500, parece no haber ningún modelo que sea validado por las pruebas de *backtesting*. La razón puede ser que el periodo de análisis corresponde al de la pandemia mundial del COVID-19, y las empresas tradicionales incluidas en el SP500 fueron las que registraron más pérdidas y tardaron en recuperar sus cotizaciones de antes de la pandemia. En cuanto al BTC, se ha verificado que durante la pandemia ha tenido un buen desempeño en términos de rentabilidad. A su vez, el NASDAQ, que incluye empresas tecnológicas, como Netflix, ha registrado menos pérdidas y se ha recuperado más rápidamente que el SP500.

El trabajo se organiza de la siguiente manera: en la sección segunda se presentan los modelos que se utilizarán para predecir la varianza realizada y para el cálculo del valor en riesgo. En la sección tercera se presentan los datos y los resultados empíricos, y en la sección cuarta se presentan las principales conclusiones del estudio.

2. MODELACIÓN DE LA VARIANZA REALIZADA

Si bien los modelos heterogéneos autorregresivos (HAR) no se clasifican formalmente como modelos de memoria larga, se ha observado que exhiben una alta persistencia y capturan las características de memoria larga de ciertas series temporales financieras.

El modelo HAR es un tipo de modelo utilizado en análisis financiero para predecir la volatilidad en los mercados. Este modelo toma en cuenta diferentes horizontes de tiempo, como corto plazo, medio plazo y largo plazo, para capturar mejor la complejidad de los movimientos de los precios de los activos financieros.

La varianza realizada (RV, en inglés) del día t se define mediante

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2, t = 1, \ldots, N, \qquad [1]$$

donde $r_{t,i}$ es el rendimiento del precio en el momento i del día t y M es el número total de valores de tiempo intradiario.

El modelo HAR introducido por Corsi (2009) se utiliza ampliamente para la previsión de la RV.

Sin embargo, este modelo no tiene en cuenta la respuesta asimétrica de la volatilidad a los *shocks* positivos y negativos, como se analiza en la literatura sobre los efectos del apalancamiento Christie (1982); Campbell y Hentschel (1992); Bollerslev *et al.* (2006)[2]. De hecho,

[2] Aunque la asimetría y el apalancamiento no son lo mismo, los usamos indistintamente a continuación. El apalancamiento se considera un caso especial de asimetría; ver McAleer (2014).

los *shocks* negativos suelen tener un mayor impacto en la volatilidad que los *shocks* positivos de la misma magnitud. En otras palabras, los inversores suelen ser reacios a la incertidumbre asociada con los *shocks* negativos porque aumentan la probabilidad de pérdidas significativas. En consecuencia, incorporar la asimetría de la volatilidad en los modelos HAR puede mejorar la predicción de la RV.

El modelo HAR tampoco tiene en cuenta las no-linealidades en el comportamiento de la RV. Para cubrir esta necesidad, algunos autores han incluido un término extra en los modelos de la familia HAR que es una interacción entre la RV pasada y su varianza. También es posible modelar la no-linealidad usando modelos semiparamétricos que se adaptan a las condiciones del mercado. Otra forma es usar modelos de aprendizaje supervisado como las redes neuronales y los *Random Forests.* A continuación, presentamos en detalle los modelos mencionados.

2.1. Modelos de la familia HAR

Antes de presentar los modelos paramétricos sugeridos en este análisis, necesitamos definir las variables predictoras. Primero, las semivarianzas realizadas positiva y negativa como las presenta Barndorff-Nielsen *et al.* (2010):

$$RV_t^+ = \sum_{i=1}^{M} r_{t,i}^2, I\left(r_{t,i} \geq 0\right),$$

$$RV_t^- = \sum_{i=1}^{M} r_{t,i}^2, I\left(r_{t,i} \leq 0\right),$$

Como en la definición de la RV, M es el número de observaciones intradiarias, $r_{t,i}$ es el rendimiento del día t en el momento i. Aquí $I(\cdot)$ es una función indicadora que toma el valor 1 si el argumento es verdadero y cero en caso contrario.

Además, consideramos los siguientes retardos de la RV:

$$RV_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} RV_{t-i},$$

con $j \leq h$.

Por lo tanto, RV_{t-1}, $RV_{t-1|t-5}$ y $RV_{t-1|t-22}$ corresponden a los retardos diarios, semanales y mensuales, respectivamente.

Utilizando estas variables, los modelos de la familia HAR que proponemos están resumidos en el cuadro 1. El modelo original propuesto por Corsi (2009) y su extensión propuesta por Patton y Sheppard (2015) que incluye las semivarianzas realizadas para modelar la asimetría. Además, Bollerslev *et al.* (2016) muestran que la RV, si bien es un estimador consistente de la volatilidad integrada bajo ciertas condiciones, es propenso a errores de medición en

muestras finitas. Estos errores pueden provocar sesgos en la estimación del modelo HAR. Para disminuir estos sesgos, proponen una nueva familia de modelos conocida como HARQ. Esta familia depende de la cuarticidad realizada, dada por:

$$RQ_t \equiv \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4.$$

Cuadro 1.

Modelos HAR paramétricos

Modelo	*Especificación*	*Autores*
HAR	$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1\|t-5} + \beta_3 RV_{t-1\|t-22} + u_t$	Corsi (2009)
SHAR	$RV_t = \beta_0 + \beta_1^+ RV_{t-1}^+ + \beta_1^- RV_{t-1}^- + \beta_2 RV_{t-1\|t-5} + \beta_3 RV_{t-1\|t-22} + u_t$	Patton y Sheppard (2015)
HARQ	$RV_t = \beta_0 + \left(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}\right) RV_{t-1} + \beta_2 RV_{t-1\|t-5} + \beta_3 RV_{t-1\|t-22} + u_t$	Bollerslev *et al.* (2016)
SHARQ	$RV_t = \beta_0 + \left(\beta_1^+ \beta_{1Q}^+ RQ_{t-1}^{1/2}\right) RV_{t-1}^+ + \left(\beta_1^- + \beta_{1Q}^- RQ_{t-1}^{1/2}\right) RV_{t-1}^- + \beta_2 RV_{t-1\|t-5} + \beta_3 RV_{t-1\|t-22} + u_t$	Bollerslev *et al.* (2016)

Fuente: Candel *et al.* (2020).

En nuestro estudio no consideramos la presencia de saltos en los datos. Según Patton y Sheppard (2015), los modelos que incluyen saltos, como el modelo HAR-J, tienden a comportarse de manera deficiente en comparación con el modelo SHAR. El modelo HAR-J incorpora un componente de salto al modelo básco HAR para considerar movimientos bruscos y significativos en los precios de los activos que no son capturados por las medidas estándar de volatilidad realizada. Por lo tanto, este capítulo se enfoca en modelos que excluyen la consideración de saltos.

2.2. Modelo HAR con coeficientes variables

Es necesario extender el enfoque de los modelos HAR, mediante la incorporación de coeficientes variables en el tiempo, para mejorar la predicción de la varianza realizada en los mercados financieros porque nos permite predecir cómo evoluciona el riesgo con el tiempo.

Al añadir coeficientes que cambian dinámicamente, podemos capturar la naturaleza no lineal de la volatilidad en los mercados, adaptándose mejor a los cambios en las condiciones económicas y financieras.

El modelo TVHAR ha sido propuesto por Chen *et al.* (2018) donde se explica el algoritmo para predecir futuros valores de la RV. Su expresión es:

$$RV_t = \gamma_0(\tau) + \gamma_1(\tau)\, RV_{t-1} + \gamma 2(\tau)\, RV_{t-1|t-5} + \gamma 3(\tau)\, RV_{t-1|t-22} + ut \qquad [2]$$

para $t = 1, 2,..., N$ y $\tau = t/N$.

Los valores estimados de los coeficientes $\gamma(\tau) = (\gamma_0(\tau), \gamma_1(\tau), \gamma_2(\tau), \gamma_3(\tau))^t$ se obtienen mediante técnicas no paramétricas detalladas en las referencias Robinson (1989), Fan y Gijbels (1996) y Cai (2007). En particular, empleamos el enfoque del estimador local lineal.

En esquema, para cada valor de τ, calculamos los coeficientes de [2] mediante un proceso de regresión local que otorga un mayor peso a los datos cercanos a un punto temporal *t*. Este peso se basa en un *kernel* (núcleo).

En la siguiente ecuación [3] se muestra la expresión del estimador local lineal para γ y su primera derivada, donde el vector de predictores es $X_i = (1, RV_{t-1}, RV_{t-1|t-5}, RV_{t-1|t-22})^t$.

$$\begin{pmatrix} \hat{\gamma}_i(\tau) \\ \hat{\gamma}_i^{(1)}(\tau) \end{pmatrix} = \begin{pmatrix} S_{N,0}(\tau) & S_{N,1}^{\top}(\tau) \\ S_{N,1}(\tau) & S_{N,2}(\tau) \end{pmatrix}^{-1} \begin{pmatrix} T_{N,0}(\tau) \\ T_{N,1}(\tau) \end{pmatrix}, \qquad [3]$$

donde las funciones $S_{N,s}(\tau)$ y $T_{N,s}(\tau)$ para $s = 0, 1, 2$ se definen como:

$$S_{N,s}(\tau) = \frac{1}{N}\sum_{i=1}^{N} X_i^{\top} X_i (\tau_i - \tau)^s K\left(\frac{\tau_i - \tau}{b}\right),$$

$$T_{N,s}(\tau) = \frac{1}{N}\sum_{i=1}^{N} X_i^{\top} (\tau_i - \tau)^s K\left(\frac{\tau_i - \tau}{b}\right) RV_i.$$

En estos estimadores, el *bandwidth* o ancho de ventana del *kernel* denotado como *b*, es un parámetro crítico. La elección de una ventana demasiado amplia no lograría capturar con precisión las variaciones del coeficiente, mientras que una ventana muy estrecha produciría un estimador excesivamente inestable.

Dado que los datos que analizamos tienen dependencia temporal, utilizamos el método de *leave-k-out cross-validation* con un tamaño de bloque igual a $k = \lfloor N/3 \rfloor$ para seleccionar *b*. De esta manera, garantizamos la independencia de las submuestras y su selección robusta.

2.2.1. Deep Learning

Existen varios marcos teóricos para Deep Learning (Hastie *et al.,* 2009), y aquí resumimos la arquitectura *feedforward* utilizada por librería H20 de R (ver *e.g.* Landry, 2016; LeDell *et al.,* 2018; Candel *et al.,* 2020).

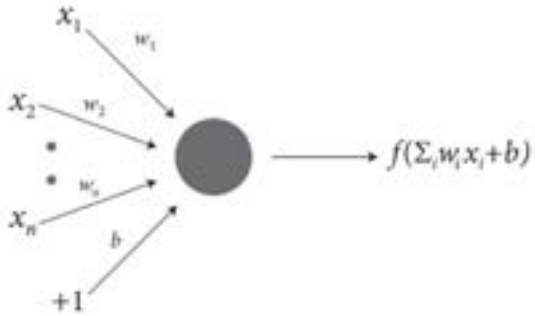
La unidad básica en el modelo es la *neurona* (figura 1), un modelo inspirado en la neurona humana biológica.

En los humanos, las señales de salida de las neuronas de diferente intensidad viajan a lo largo de las uniones sinápticas y luego se agregan como entrada para la activación de una neurona conectada.

En el modelo, se agrega la combinación ponderada de señales de entrada $\alpha = \sum_{i=1}^{n} w_i x_i + b$ y luego se transmite una señal de salida $f(\alpha)$ por la neurona conectada. La función f repre-

Figura 1.

Esquema neurona básica



Fuente: Candel *et al.* (2020.

senta la función de activación no lineal utilizada en toda la red, y el sesgo *b* representa el umbral de activación de la neurona.

Las redes neuronales de múltiples capas predictivas (figura 2) constan de muchas capas de unidades neuronales interconectadas: comenzando con una capa de entrada para que coincida con el espacio de entrada; seguido de múltiples capas no lineales; y terminan con una regresión lineal o capa de clasificación para que coincida con el espacio de salida. Las entradas y salidas de las unidades del modelo siguen la lógica básica de la neurona única descrita anteriormente. Las constantes se incluyen en cada capa de la red que no es de salida. Los pesos que vinculan las neuronas con otras neuronas determinan completamente la salida de toda la red, y el aprendizaje ocurre cuando estos pesos se adaptan para minimizar el error en los datos de entrenamiento.

Figura 2.

Esquema de red neuronal multicapa

Fuente: Candel *et al.* (2020).

Este marco básico de redes neuronales multicapa se puede utilizar para realizar tareas de *Deep Learning*. Las arquitecturas de *Deep Learning* son modelos de extracción de características jerárquicas, que normalmente implican múltiples capas con niveles de no linealidad. Dichos modelos pueden modelizar con alto rendimiento datos complejos.

Específicamente, usamos el procedimiento h2o.deeplearning de la librería H20 de R, que utiliza un protocolo de entrenamiento puramente supervisado. El esquema de inicialización predeterminado es la opción de adaptación uniforme, que es una inicialización optimizada basada en el tamaño de la red.

Usamos las opciones por defecto del comando, así, la función de activación *Rectifier*: $f(\alpha) = \text{máx}(0, \alpha)$, y como función de pérdida el error cuadrático medio (*ECM*). A su vez, usamos dos capas ocultas con 200 neuronas en cada una.

2.2.2. Random Forests

Los *Random Forests* (RF) presentados por primera vez en Breiman (2001), son una modificación sustancial de las técnicas de *bagging* que generan una numerosa colección de árboles no correlacionados y luego se promedian. La idea esencial de *bagging* es promediar muchos modelos ruidosos, pero aproximadamente insesgados y, por lo tanto, reducir la varianza. Los árboles son candidatos ideales para el *bagging,* ya que pueden capturar estructuras de interacción complejas en los datos y, si crecen lo suficientemente en profundidad, tienen un sesgo relativamente bajo. Dado que los árboles son habitualmente ruidosos, se benefician enormemente de hacer promedios.

En muchos problemas, el rendimiento de los *Random Forests* es muy similar al *boosting* y son más sencillos de entrenar y ajustar. Como consecuencia, los *Random Forests* son bastante populares y se implementan en una gran variedad de paquetes estadísticos.

El algoritmo general de *Random Forests* para usar con variables continuas es:

1. Desde $b = 1$ a B:

   (a) Obtener una muestra *bootstrap* $\mathbf{Z}^*$ de tamaño N a partir de los datos de entrenamiento.

   (b) Ajustar un árbol del *Random Forest* T_b con los datos remuestreados, repitiendo de modo recursivo los siguientes pasos para cada nodo terminal del árbol, hasta alcanzar el tamaño mínimo de nodo $n_{mín}$.

      i. Seleccionar m variables al azar de las p variables.

      ii. Elegir la mejor variable o punto de división entre las m.

      iii. Dividir el nodo en dos nodos *hijos*.

2. Generar el conjunto de árboles $\{T_b\}_1^B$.

Para hacer una predicción en un nuevo punto *x*:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

En este estudio usamos la librería H20 de R. Específicamente, usamos el procedimiento

h2o.randomForest.

Después de realizar un estudio previo para optimizar los valores de los parámetros del modelo, fijamos algunos valores del procedimiento para optimizar tanto el rendimiento, como la velocidad de computación.

Usamos una profundidad máxima de los árboles (max_depth) igual a 10, dado que valores más altos pueden llevar a sobreajustes.

Para simplificar la interpretabilidad, fijamos en 5 el número mínimo de observaciones para especificar una hoja del árbol (min_rows).

Finalmente, el número de árboles por defecto (ntrees) se fija en 30.

3. ANÁLISIS EMPÍRICO

Vamos a predecir la varianza realizada de los rendimientos del BTC, NASDAQ y SP500 que son muy relevantes en los mercados financieros globales.

3.1. Datos

En predicción, dividimos los datos en un conjunto de entrenamiento en el que ajustamos el modelo y un conjunto de prueba donde evaluamos su poder de predicción. En nuestro análisis, empleamos los últimos 1.000 valores de cada serie, como conjunto de prueba para determinar la efectividad de cada modelo y para comparar su rendimiento prediciendo la varianza realizada. El cuadro 2 muestra las fechas exactas de las series financieras en nuestro análisis.

Una vez ajustado el modelo con el conjunto de entrenamiento, cuyos predictores tienen índices 1,...,*T*, se predice el valor de la *RV* para $T + 1$ que denotamos con $RV^{*}_{T+1|T}$.

Cuadro 2.

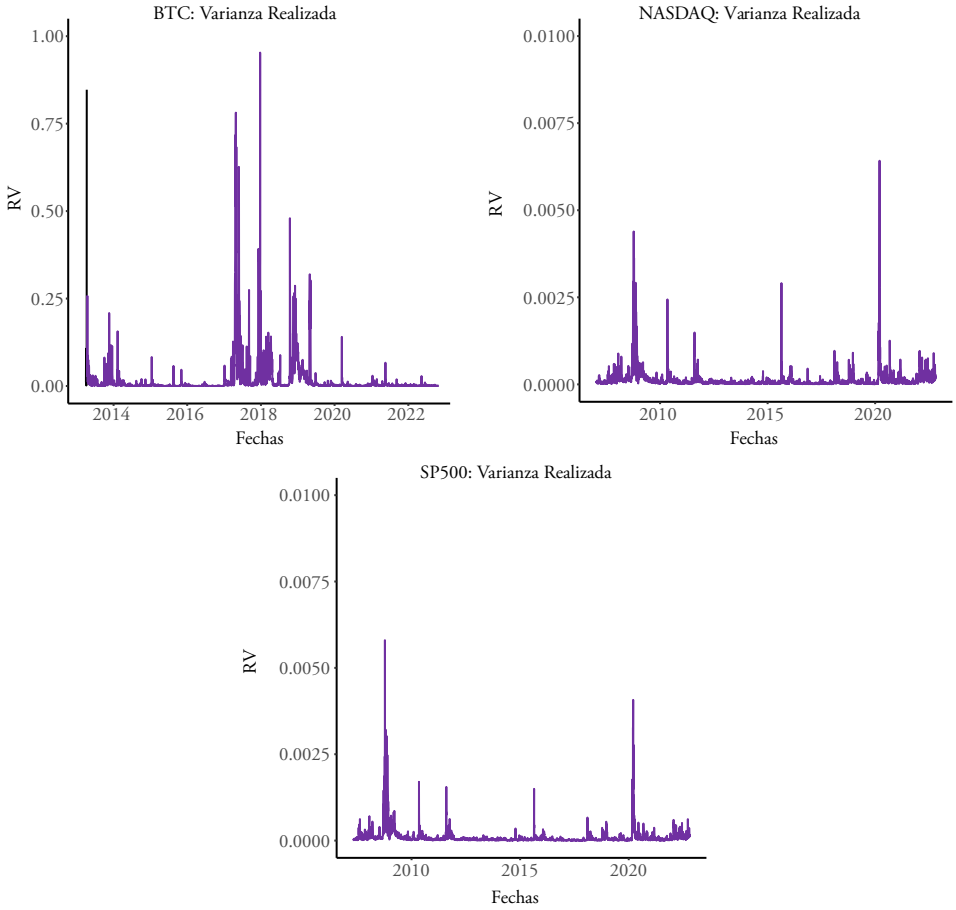
Datos

	Muestra completa	*Muestra de predicción (1.000 valores)*	*Frecuencia*
BTC	23/04/2013-28/10/2022	Desde 02/02/2020	Diaria Continuo
NASDAQ	16/02/2007-21/10/2022	Desde 01/11/2018	Diaria Lunes a viernes
SP500	31/05/2007-21/10/2022	Desde 01/11/2018	Diaria Lunes a viernes

Fuentes:

Figura 3.

Varianzas realizadas



Fuente: Candel *et al.* (2020).

Movemos la ventana de entrenamiento un día, es decir los predictores tienen índices 2,...,T +1, y calculamos $RV^{*}_{T+2|T+1}$. Este proceso se repite 1.000 veces, siempre entrenando el modelo con conjuntos de tamaño T y moviendo la ventana un dato cada vez, para obtener la predicción diaria de la RV para las tres series en nuestro análisis.

Calculamos el error de predicción en cada paso como:

$$E_h = RV_{T+h} - RV^{*}_{T+h|T+h-1},$$

para h = 1,..., 1000 y RV_{T+h} calculada con la ecuación [1].

La figura 3 muestra las varianzas realizadas, y se puede apreciar que, durante el periodo evaluado para las predicciones de volatilidad, la volatilidad del BTC es inferior en comparación con la volatilidad observada en el periodo de entrenamiento.

3.2. Evaluación del funcionamiento de los modelos de predicción

La evaluación del desempeño de los modelos de predicción es esencial en cualquier proceso de modelización de datos. No solo se trata de comprender cómo es de efectivo un modelo con datos conocidos, sino también de su capacidad para predecir datos desconocidos. La medición del desempeño en la predicción se convierte en el indicador principal para elegir un modelo que explique los patrones ocultos en los datos que nos sea útil en aplicaciones prácticas.

La evaluación comparativa de varios modelos nos desvela cuál de ellos es el mejor en nuestra tarea de predicción. Cuando decimos mejor, no nos referimos al modelo con el menor error de predicción promedio en nuestra muestra. Buscamos una respuesta estadística sólida que nos permita extrapolar los resultados a cualquier conjunto de datos con características o distribución similares.

Para ello, introducimos el procedimiento del "Model Confidence Set" y los contrastes de habilidad predictiva superior condicional y utilizamos 3 funciones de pérdida:

- El error cuadrático (EC),
- El error absoluto (EA), y
- La "Quasi-Likelihood Exponential" (QLIKE).

Patton (2011) propuso la QLIKE como medida de error de predicción de la volatilidad realizada en el contexto de modelos de volatilidad financiera. Esta medida se deriva de la función de verosimilitud, que compara la probabilidad de que los datos observados provengan de

un modelo de volatilidad específico en comparación con la volatilidad realizada. Un modelo cuyas predicciones de volatilidad son muy cercanas a la volatilidad realizada observada, el valor de QLIKE será bajo.

La figura 4 muestra las QLIKE medias de ventana deslizante para varios modelos de predicción de la RV y para las tres series de rendimientos. Esta función de pérdida de ventana móvil se calcula para una ventana fija de 90 días. Se observa que el modelo TVHAR presenta valores elevados de QLIKE durante 2020, coincidiendo con el confinamiento y el periodo más restrictivo de la pandemia, seguido de un aumento en los errores de predicción hacia finales de 2021 que se prolonga durante 2023. Los modelos HAR, HARQ, SHARQ y RF muestran los valores más bajos de QLIKE.

En relación a la RV del NASDAQ, el patrón es similar, siendo los modelos mencionados previamente los que presentan valores más bajos de QLIKE. En cambio, para el SP500, la evidencia difiere: casi todos los modelos presentan un mal desempeño durante el confinamiento. En ese periodo, el modelo con mayores QLIKE es el SHARQ, seguido por DL, HARQ y SHAR. Sin embargo, tras ese periodo, las QLIKE de la mayoría de los modelos son similares, excepto las del DL, que son relativamente mayores que las demás.

3.2.1. Model confidence set

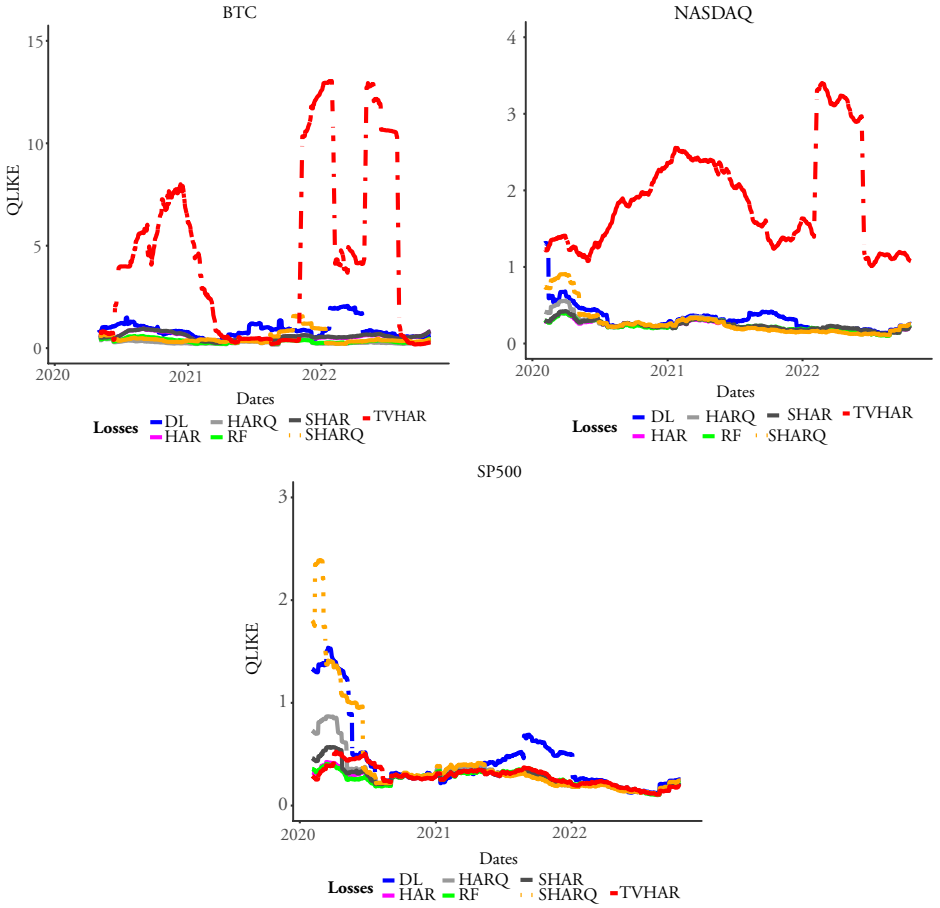
El procedimiento del *Model Confidence Set (MCS)* propuesto en Hansen *et al.* (2011) es una herramienta estadística que se utiliza para comparar la confianza en las predicciones de varios modelos. El *MCS* es más preciso que las métricas comunes como el *error cuadrático medio (ECM)* y el *error absoluto medio (EAM),* ya que no solo se fija en cuánto se equivoca cada modelo en promedio, sino que también considera cuán variables son los *errores cuadráticos (EC)* y los *errores absolutos (EA)* de las predicciones. Además de con los *EC* y *EA*, este procedimiento se puede usar con otras funciones de pérdida como el QLIKE.

De manera más detallada, el *MCS* es un procedimiento que consiste en una serie de pruebas de hipótesis basadas en una función de pérdida. En la primera prueba estadística, el *MCS* selecciona el modelo con el mayor error promedio y compara el rango que abarca sus errores de predicción con los de los otros modelos. Si este rango no se superpone con ninguno de los otros modelos, se descarta y se pasa a hacer lo mismo con el siguiente modelo. Si hay superposición, se agrupan los modelos que tienen errores estadísticamente similares y se les asigna un p-valor. Al final de este proceso, el procedimiento *MCS* ha asignado un p-valor a cada modelo.

Los modelos cuyos p-valores superen cierto valor de significación (α) se consideran parte del llamado "Conjunto Superior de Modelos" (*Superior Set of Models* o *SSM*) y son los modelos que se eligen para realizar las predicciones.

Figura 4.

Predicción de la volatilidad: ventana deslizante de QLIKE



Fuente: Candel *et al.* (2020).

El cuadro 3 muestra el *ECM, EAM* y *QLIKE* medio de los errores de predicción de los modelos propuestos usados para predecir los últimos valores de la varianza realizada de nuestras muestras del BTC, NASDAQ y SP500. También incluye los *p*-valores obtenidos con el procedimiento MCS para cada modelo.

La interpretación de los resultados del cuadro 3 es:

- *ECM*: En la serie NASDAQ, todos los modelos tienen *p*-valores relativamente altos (por encima de 0.05). Esto significa que todos pertenecen al 95 % SSM y no hay diferencia a nivel de significación del 5 % para considerar que un modelo es mejor que los demás en términos de errores de predicción cuadráticos. Sin embargo, si se reduce el

umbral al 90 % *SSM*, solo el modelo SHARQ y DL pertenecen al conjunto superior de modelos del 90 %. Esto sugiere que para un nivel de significación del 10 %, SHARQ y DL son estadísticamente superiores al resto para predecir la RV. En las series SP500 y BTC, ningún modelo tiene un *p*-valor menor que 0.05, lo que significa que todos los modelos pertenecen al 95 % *SSM*.

- *EAM*: en la serie BTC, los modelos HARQ y TVHAR pertenecen al 95 % *SSM*, mientras que el resto de los modelos tienen errores de predicción absolutos significativamente más grandes, por lo que el HARQ y TVHAR son estadísticamente superiores en términos de *EAM* para predecir la RV de BTC. En las series NASDAQ y SP500, todos los modelos pertenecen al 95 % *SSM* para predecir la RV ya que sus *p*-valores son mucho más grandes que 0.05. El modelo con el menor *EAM* es el HAR.
- QLIKE medio: muchos de los modelos son estadísticamente igualmente buenos para predecir la RV de las series NASDAQ y SP500. Sin embargo, el HARQ es el único modelo que pertenece al 95 % *SSM* y el preferido para predecir la RV del BTC.

Cuadro 3.

Errores de predicción medios y p-valores del procedimiento *MCS* de los diferentes modelos

	HAR	*SHAR*	*HARQ*	*SHARQ*	*TVHAR*	*DL*	*RF*
				ECM $\times 10^6$			
BTC	35.505	46.399	39.063	36.616	38.695	47.838	39.181
p-valor	1.000	0.569	0.705	0.780	0.832	0.549	0.666
NASDAQ	0.053	0.060	0.076	0.131	0.067	0.196	0.051
p-valor	0.508	0.316	0.562	0.585	0.518	0.289	1.000
SP500	0.027	0.031	0.026	0.032	0.038	0.039	0.027
p-valor	0.615	0.525	1.000	0.522	0.336	0.319	0.245
				EAM $\times 10^3$			
BTC	2.768	2.887	1.691	1.929	1.686	3.202	1.836
p-valor	< 0.001	<0.001	0.954	0.042	1.000	<0.001	0.003
NASDAQ	0.079	0.082	0.084	0.087	0.083	0.102	0.080
p-valor	1.000	0.469	0.580	0.641	0.573	0.104	0.798
SP500	0.054	0.055	0.054	0.057	0.059	0.064	0.054
p-valor	1.000	0.782	0.965	0.284	0.439	0.030	0.951
				QLIKE medio			
BTC	0.571	0.587	0.291	0.481	4.114	0.889	0.366
p-valor	<0.001	<0.001	1.000	0.039	0.006	0.002	<0.001
NASDAQ	0.223	0.232	0.243	0.282	1.769	0.595	0.226
p-valor	1.000	0.054	0.196	0.153	<0.001	0.060	0.529
SP500	0.256	0.269	0.308	0.958	0.265	0.547	0.256
p-valor	1.000	0.778	0.178	0.067	0.677	0.028	0.947

Fuente: Candel *et al.* (2020).

3.2.2. Habilidad predictiva superior condicional

En esta sección vamos a analizar la capacidad predictiva superior condicional de los modelos para predecir la volatilidad con el contraste propuesto por Li *et al.* (2022).

Sea $\{RV_t\}_{t>T}$ la serie temporal de volatilidad que se va a predecir; $\{\widehat{RV}_{0,t}\}_{T<t\leq N}$ es una serie de pronóstico de referencia y $\{\widehat{RV}_{i,t}\}_{T<t\leq N}$, $1 \leq i \leq S$, series competidoras.

El desempeño del modelo de referencia en comparación con la i-ésima alternativa competidora se evalúa mediante la diferencia de las funciones de pérdida, tal que:

$$e_{i,t} \equiv L(\widehat{RV}_t, RV_{i,t}) - L(RV_t, \widehat{RV}_{0,t}).$$

La hipótesis nula del test HPSC establece que:

$$H_0 : E(e_{i,t}|X_t = x) \geq 0, x \in \chi, \forall 1 \leq i \leq S.$$

Aquí, X_t representa una variable de estado condicional elegida por el evaluador, y χ denota la región condicional como un subconjunto del dominio de X. En este caso particular, las variables de estado condicional son el VIX que es una proxy de la volatilidad del mercado para NASDAQ y SP500, y los rendimientos de BTC para la variable BTC. En este último caso buscamos validar la teoría financiera de compensación del riesgo. El objetivo es descubrir para qué valores (bajos o altos) de la variable condicional el modelo de referencia falla o tiene éxito en predecir volatilidad futura en comparación con el modelo alternativo. La hipótesis nula establece que el modelo de referencia debe dominar débilmente a todos los modelos competidores en la región condicional χ, porque su valor de función de pérdida es estadísticamente el más pequeño. Si la hipótesis nula no puede ser rechazada, esto indica que el modelo de referencia tiene propiedades predictivas deseables.

Se pueden realizar dos tipos de test condicionales. El primero se llama "Uno-contra-Uno", en el que cada referencia se compara con una alternativa para todos los pares de modelos. El segundo tipo se llama "Uno-contra-Todos", que compara cada modelo de referencia con los seis modelos competidores juntos[3].

El cuadro 4 muestra los resultados de los contrastes para tres funciones de pérdida: errores cuadráticos, errores absolutos y QLIKE. Los contrastes se realizan para un nivel de confianza elegido del 95 %.

Basándonos en los contrastes (Uno-contra-Todos) para la RV de la BTC, se observa que los modelos que dominan débilmente a sus competidores en la predicción de la RV son el HAR, HARQ, SHARQ, TVHAR y RF en términos del EC. Para el EA solo el HARQ parece dominar débilmente a sus competidores. Sin embargo, para la QLIKE, parece que ningún modelo domina a sus competidores, aunque en los contrastes Uno-contra-Uno los modelos menos rechazados son el HARQ, TVHAR, DL y RF.

[3] Usamos el código proporcionado por los autores disponible en https://zenodo.org/record/4884813

Cuadro 4.

Habilidad predictiva superior condicional para predicción de volatilidad usando tres funciones de pérdida

	HAR	*SHAR*	*HARQ*	*SHARQ*	*TVHAR*	*DL*	*RF*	*HAR*	*SHAR*	*HARQ*	*SHARQ*	*TVHAR*	*DL*	*RF*	*HAR*	*SHAR*	*HARQ*	*SHARQ*	*TVHAR*	*DL*	*RF*
	EC							*EA*							*QLIKE*						
BTC																					
Panel A: Contrastes HPSC Uno-contra-Uno																					
HAR		0	0	0	0	1	0		0	0	0	1	1	0		0	0	0	0	0	0
SHAR	0		0	0	0	1	0	0		0	0	1	0	0	0		0	0	0	0	0
HARQ	1	1		0	0	1	0	1	1		1	1	1	1	1	1		1	0	0	0
SHARQ	0	0	0		0	1	0	1	1	0		1	1	0	1	1	0		0	0	0
TVHAR	1	1	0	1		1	0	1	1	0	1		1	1	1	0	0	0		0	0
DeepL	0	0	0	0	0		0	0	0	0	0	0		0	0	0	0	0	0		0
RF	0	0	0	0	0	1		1	1	0	0	1	1		1	1	0	0	0	0	
Panel B: Contrastes HPSC Uno-contra-Todos																					
	0	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1
NASDAQ																					
Panel A: Contrastes HPSC Uno-contra-Uno																					
HAR		1	1	0	1	0	1		0	0	0	0	1	0		0	0	0	1	0	0
SHAR	1		1	0	1	0	1	0		0	0	0	1	0	1		0	0	1	0	0
HARQ	1	1		0	1	0	1	0	1		0	1	1	1	0	0		0	1	0	0
SHARQ	0	0	0		0	0	0	0	0	0		0	1	0	1	0	0		1	0	0
TVHAR	1	1	1	0		0	1	0	0	0	0		1	0	0	0	0	0		0	0
DeepL	0	0	0	0	0		0	0	0	0	0	0		0	0	0	0	0	1		0
RF	1	1	1	0	1	0		0	0	0	0	0	1		0	0	0	0	1	1	
Panel B: Contrastes HPSC Uno-contra-Todos																					
	1	1	1	1	1	1	1	0	1	0	0	1	1	0	0	1	0	0	1	1	0
SP500																					
Panel A: Contrastes HPSC Uno-contra-Uno																					
HAR		1	1	1	1	1	1		0	0	0	0	1	1		0	0	0	0	0	0
SHAR	1		1	1	1	1	1	1		0	1	0	1	1	1		0	0	0	0	0
HARQ	1	1		1	1	1	1	1	1		1	0	1	1	0	0		0	0	0	0
SHARQ	1	1	1		1	1	1	0	0	0		0	1	0	1	0	0		0	0	0
TVHAR	1	1	1	1		1	1	0	0	0	0		1	0	0	0	0	0		0	1
DeepL	1	1	1	1	1		1	0	0	0	0	0		0	0	0	0	0	0		0
RF	1	1	1	1	1	1		0	0	0	1	0	1		0	0	0	0	0	0	
Panel B: Contrastes HPSC Uno-contra-Todos																					
	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	0	1	1	1

Fuente: Candel *et al.* (2020).

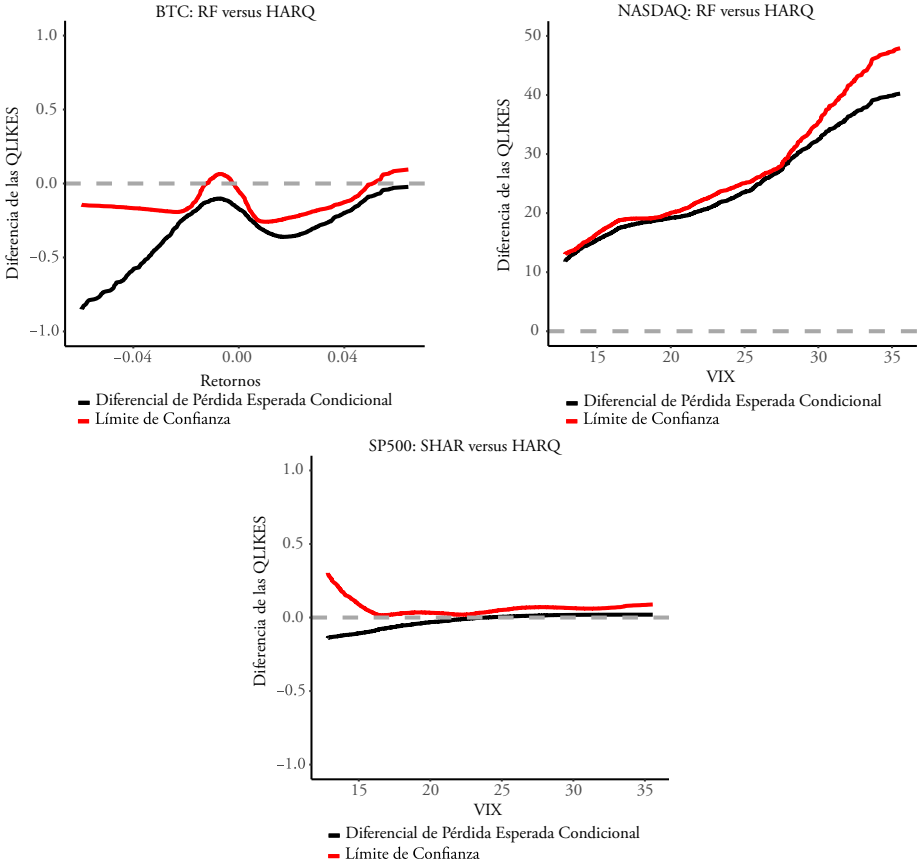
Para el NASDAQ, observamos que cuando usamos el EA o la QLIKE, los modelos HAR, HARQ, SHARQ y RF dominan débilmente a sus competidores, mientras que para la función de pérdida EC ningún modelo domina débilmente a los demás.

Finalmente, para el SP500, y también para el EC, parece no haber ningún modelo que domine débilmente en la predicción de la RV dada la volatilidad del mercado. Aunque, usando el EA, hay dos modelos que dominan, que son el HARQ y el TVHAR, y según la QLIKE, los modelos que dominan débilmente a las alternativas son el HAR, el HARQ y el SHARQ.

En resumen, los resultados de los contrastes de habilidad predictiva superior condicional parecen apuntar al HARQ como el modelo que más frecuentemente domina a sus competidores en la predicción de la RV.

Figura 5.

Predicción de la volatilidad: contrastes HPSC uno-versus-uno. Función de pérdida QLIKE



Fuente: Candel *et al.* (2020).

La figura 5 muestra la diferencia estimada entre las funciones de pérdida esperadas condicionales, junto con los límites de confianza superior del 95 % (ver Li *et al.,* 2022) para obtener más detalles sobre los límites de confianza superior).

La función de pérdida utilizada para representar la figura 5 es la QLIKE. Por ejemplo, la "Diferencia de las QLIKE" en el panel izquierdo se define como la diferencia entre las QLIKE del modelo RF versus HARQ. La referencia aquí es el modelo HARQ.

La prueba HPSC rechaza la hipótesis nula (es decir, HARQ es débilmente superior) si el límite de confianza está por debajo de cero en alguna región del espacio de estados condicionantes.

Comparando el RF con el QLIKE en la predicción de la RV de la BTC, no habiendo ningún modelo dominante, observamos que el RF parece tener un desempeño superior al desempeño del HARQ condicionando a los rendimientos de la BTC. También se observa que el desempeño del RF es mejor para rendimientos más negativos y empeora para valores más positivos de los rendimientos, aunque este es no es estadísticamente superior para rendimientos alrededor de 0 y superiores a 0.04.

En cuanto al NASDAQ, ambos modelos son débilmente dominantes, pero el desempeño del HARQ es estadísticamente superior al del RF.

Finalmente, para el SP500 observamos que, aunque los modelos SHAR y HARQ son ambos débilmente dominantes, para volatilidades de mercado bajas, el modelo SHAR parece tener un mejor desempeño que el modelo HARQ, aunque nunca rechazamos que el modelo HARQ sea débilmente dominante.

3.2.3. Valor en riesgo

Una posibilidad es usar las predicciones del RV de nuestras series para calcular su Valor en Riesgo (VaR). Debemos recordar que en las recomendaciones del Acuerdo de Basilea III, que establece normas internacionales para regular la cantidad de capital que los bancos deben tener en reserva para protegerse contra riesgos financieros y operativos, es calcular el VaR para el nivel de confianza del 99 % ($\alpha = 0.01$).

Sea los rendimientos en el tiempo t dados por

$$y_t = \mu + \sqrt{RV_t}\,\in_t, \quad t = 1,...,T$$

donde RV_t viene dado por las especificaciones en el cuadro 1 y ϵ_t es un proceso i.i.d.

La expresión matemática del VaR se define como:

$$VaR_{T+h} = -\mu_{T+h-1} - z\alpha\sqrt{RV^*_{T+h|T+h-1}},$$

donde μ_{T+h-1} es el promedio de los rendimientos de la muestra de entrenamiento, z_α es el valor crítico de una distribución normal estándar, que corresponde al nivel de confianza α, y el resto es la raíz cuadrada de la predicción de la varianza realizada.

Andersen *et al.* (2000, 2001b, 2003), Andersen *et al.* (2001a) y Andersen *et al.* (2007) encuentran que las distribuciones de los rendimientos diarios de las acciones tipificados por la correspondiente RV diaria son aproximadamente Gaussianos, especialmente considerando saltos y asimetría de la volatilidad.

Decimos que tenemos una violación en el momento $T+h$, si el VaR obtenido es menor que el rendimiento cambiado de signo:

$$\text{violación}_{T+h} = I(VaR_{T+h} < -r_{T+h}).$$

Un modelo que predice adecuadamente, tendrá un promedio de violaciones cercano a α.

El cuadro 5 muestra el promedio de violaciones en las 1.000 predicciones de RV para cada modelo y para α = 0,01; 0,05.

En líneas generales vemos que la tasa de fallo no está próxima a la tasa de fallo teórica. Esto puede ocurrir porque nuestro periodo de predicción del VaR corresponde a la época de la pandemia mundial de coronavirus (COVID-19), la cual provocó que todos los principales índices del mercado de valores experimentaran una caída abrupta en marzo de 2020. Sin embargo, tanto la magnitud de esa caída como la forma de la recuperación subsiguiente variaron enormemente.

A modo de ejemplo, el 15 de marzo de 2020, los principales mercados europeos y las acciones tradicionales en Estados Unidos habían perdido aproximadamente el 40 % de su valor en comparación con el 5 de enero de 2020. En cambio, los mercados asiáticos y el índice compuesto NASDAQ solo perdieron alrededor del 20 al 25 % de su valor. Una situación similar se observa en la recuperación posterior al coronavirus. A partir del 14 de noviembre de 2021, el valor del índice compuesto NASDAQ era aproximadamente un 65 % más alto que en enero de 2020, mientras que la mayoría de los otros mercados solo habían recuperado entre un 20 y un 40 %. Esto sugiere que ha habido menos turbulencias en el NASDAQ en comparación con el SP500, que incluye empresas más tradicionales.

En cuanto al BTC, hay estudios que indican que la pandemia de COVID-19 no influyó en la volatilidad de las criptomonedas, Sifat (2021), mientras que otros concluyeron que los precios de BTC aumentaron en el período de la pandemia. Así, Goodell y Goutte (2021) y Corbet *et al.* (2020) sugirieron que las grandes criptomonedas actuaron como una reserva de valor durante este tiempo. Lo que sugiere que este periodo fue favorable a las principales criptomonedas.

Además, en el marco de nuestro análisis, realizamos 13 contrastes condicionales. Los contrastes condicionales evalúan si la proporción de violaciones es consistente con el

nivel de confianza del VaR, teniendo en cuenta la dependencia temporal de las violaciones. Las pruebas condicionales son más generales que las pruebas incondicionales y pueden incluir pruebas de independencia como parte de su evaluación. Los contrastes incondicionales contrastan si la tasa de fallo es similar a la tasa de fallo esperada, mientras que los contrastes de independencia chequean si los fallos son independientes entre sí.

La independencia de los fallos es un requisito importante para asegurar que el modelo de VaR es válido. Si los fallos están correlacionados, podría indicar que el modelo no está capturando adecuadamente el riesgo en diferentes momentos del tiempo.

Cuando se hace *backtesting* el principal desafío y problema asociado con la validación del pronóstico del VaR es la baja potencia de las pruebas de *backtesting.* Se hace necesario, así, una metodología de *backtesting* libre de modelos y tener en cuenta el error de estimación.

Cuadro 5.

Promedio de violaciones del VaR (en %) de los diferentes modelos de predicción de la RV

	HAR	*SHAR*	*HARQ*	*SHARQ*	*TVHAR*	*DL*	*RF*
				α = 1 %			
BTC	0,4	0,4	**1,1**	**1,5**	6,9	2,5	**1,2**
NASDAQ	3,9	4,2	3,7	3,3	**1,6**	4,1	3,5
SP500	4,1	3,3	3,8	3,2	3,7	4,4	3,6
				α = 5 %			
BTC	1,4	1,5	3,5	3,8	10,6	**4,7**	2,4
NASDAQ	10,0	10,3	9,6	8,6	2,8	10,2	9,6
SP500	9,3	8,5	8,9	8,2	9,0	9,3	8,8

Fuente: Candel *et al.* (2020).

Se han propuesto varias pruebas de *backtesting* para abordar estos problemas, incluida la prueba de Cuantil Dinámico (DQ) propuesta por Engle y Manganelli (2004), Escanciano y Olmo (2010, 2011) y, más recientemente, la propuesta de Barendse *et al.* (2021).

Barendse *et al.* (2021) también proponen versiones robustas para las pruebas de *Expected shortfall* producidas por error de estimación y muestran vía simulaciones de Monte Carlo que las pruebas estándar pueden sufrir distorsiones de tamaño debido al error de estimación, con frecuencias de rechazo empíricas que exceden los niveles de significación nominal. La robustificación de las pruebas de *backtesting* generalmente corrige este problema con bastante éxito, aunque puede llevar a alguna pérdida de potencia y no son de fácil implementación.

Para evaluar los modelos en el contexto del VaR además de los contrastes propuestos en Christoffersen (1998) y Engle y Manganelli (2004) usaremos una herramienta para el *backtesting* de pronósticos del VaR que se basa en un modelo no lineal, llamado modelo de regresión Binaria Dinámica (DB), propuesto por Dumitrescu *et al.* (2012). Esta propuesta

mejora la propuesta de Engle y Manganelli (2004) que implementa un contraste basado en una regresión simple de las violaciones en las violaciones pasadas. Se sabe que este modelo no es adecuado para variables dependientes binarias dado que los residuos son heterocedásticos por construcción y su distribución es discreta, Gourieroux y Jasiak (2001).

El modelo de regresión binaria dinámica, además de ser adecuado para variables dependientes binarias, tiene en cuenta la posible correlación entre violaciones (*clusters*) en la estimación y exhibe una potencia más alta que las pruebas de *backtesting* usuales en la literatura.

Usaremos siete especificaciones diferentes para el modelo DB, denotadas por DB_1 a DB_7, que están inspiradas en las especificaciones CAViaR propuestas por Engle y Manganelli (2004). Estos contrastes son más robustos frente al error de estimación y tienen buenas propiedades en muestras finitas. Además, tienen en cuenta la posible correlación entre violaciones, que no es abordada por algunas de las pruebas existentes. Por otro lado, pueden implementarse fácilmente y permiten pruebas separadas de las hipótesis de cobertura incondicional, independencia y cobertura condicional.

El modelo de respuesta binaria dinámica considera que la probabilidad condicional de violación en el tiempo t viene dada por:

$$Pr(I_t(\alpha) = 1|F_{t-1}) = E[I_t(\alpha)|F_{t-1}] = F(\pi_t),$$

donde F (.) denota una función de distribución acumulada y F_{t-1} es el conjunto de información disponible en $t - 1$.

Se asume que el índice π_t satisface la siguiente representación autorregresiva:

$$\pi_t = c + \sum_{j=1}^{q_1} \beta_j \pi_{t-j} + \sum_{j=1}^{q_2} \delta_j I_{t-j}(\alpha) + \sum_{j=1}^{q_3} \psi_j l(x_{t-j}, \phi) + \sum_{j=1}^{q_4} \gamma_j l(x_{t-j}, \phi) I_{t-j}, \qquad [4]$$

donde l(.) es una función de un número finito de valores retardados observables, y x_t es un vector de variables explicativas. El papel de l(.) es vincular el índice t con las variables observables que pertenecen al conjunto de información. Una elección natural para x_{t-j} podría ser los rendimientos retardados o los retardos del VaR.

Dada la ecuación [4], Dumitrescu *et al.* (2012) sugieren usar siete especificaciones. Las primeras cuatro especificaciones (DB_1 a DB_4) corresponden a un modelo de respuesta binaria dinámica que incluye el índice π_t retardado como variable explicativa y alguna información adicional a través de los valores pasados observados del proceso de violación.

Los modelos quinto y sexto (DB_5 y DB_6) se derivan de las especificaciones autorregresivas de cuantiles CAViaR utilizadas por Engle y Manganelli (2004).

La séptima especificación (DB_7) incluye una interacción entre el VaR retardado y la variable binaria de violación para incluir una respuesta asimétrica del índice a valores pasados del VaR en caso de una violación.

$$DB_1: \quad \pi_t = c + \beta_1 \pi_{t-1}, \qquad [5]$$

$$DB_2: \quad \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha) \qquad [6]$$

$$DB_3: \quad \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha), + \delta_2 I_{t-2}(\alpha), \qquad [7]$$

$$DB_4: \quad \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha), + \delta_2 I_{t-2}(\alpha), + \delta_3 I_{t-3}(\alpha), \qquad [8]$$

$$DB_5: \quad \pi_t = c + \beta_1 \pi_{t-1} + \psi_1 \mathrm{VaR}_{t-1}, \qquad [9]$$

$$DB_6: \quad \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha) + \psi_1 \mathrm{VaR}_{t-1} \qquad [10]$$

$$DB_7: \quad \pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}(\alpha) + \psi_1 \mathrm{VaR}_{t-1} + \gamma_1 \mathrm{VaR}_{t-1} I_{t-1}. \qquad [11]$$

Cada una de estas especificaciones tiene diferentes variables explicativas y, por lo tanto, captura diferentes aspectos del proceso de violación.

Para la estimación de cada uno de los modelos mencionados anteriormente, Dumitrescu *et al.*(2012) han usado el método propuesto por Kauppi y Saikkonen (2008) que se basa en la estimación de máxima verosimilitud con restricciones. Este método requiere que las variables explicativas sean estacionarias y que las variables aleatorias sigan una distribución normal para cumplir con las condiciones de regularidad necesarias.

De manera formal, sea π_t proveniente de cualquiera de las especificaciones [5] a [11], el logaritmo de la función de verosimilitud viene dado por:

$$\ln L\left(\theta, c; I(\alpha), Z\right) = \sum_{t=1}^{T}\left[I_t(\alpha) \ln F\left(\pi_t(\theta, c, Z_t)\right) + \left(1 - I_t(\alpha)\right) \ln\left(1 - F\left(\pi_t(\theta, c, Z_t)\right)\right)\right],$$

donde θ es el vector de parámetros de las especificaciones anteriores excepto la constante, *i.e.*, $\theta = [\beta_1, \delta_1, \delta_2, \delta_3, \psi_1, \gamma_1]$ y Z_t es el vector de variables explicativas en el tiempo t. La estimación por máxima verosimilitud lleva a estimadores con buenas propiedades estadísticas en muestras grandes.

Contrastar la hipótesis nula de cobertura condicional ($E[I_t(\alpha)|F_{t-1}] = \alpha$) utilizando el modelo binario dinámico es similar al caso de la prueba DQ de Engle y Manganelli (2004), ya que se basa en el contraste estándar sobre los coeficientes de la regresión:

$H_0 : \beta_1 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0, \psi_1 = 0, \gamma_1 = 0$ y $c = F^{-1}(\alpha)$.

Bajo la hipótesis nula, la variable de violaciones es ortogonal a cualquier variable explicativa que pertenezca al conjunto de información F_{t-1} y la probabilidad condicional de una violación es igual a α. Se pueden usar contrastes de razón de verosimilitudes (LR), *Wald* y multiplicadores de *Lagrange* (LM).

De manera similar, la prueba de independencia se basa en $H_0 : \beta_1 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0, \psi_1 = 0, \gamma_1 = 0$, es decir, la nulidad de los coeficientes, pero esta vez en un modelo donde la constante ya no es una función de α. Bajo la hipótesis nula de independencia, no hay dependencia temporal de las violaciones, pero la probabilidad de violación de VaR podría ser diferente de la tasa de cobertura α.

Cuadro 6.

Valor en riesgo: nivel de confianza del 99 %

Variable	*Modelo*	*Promedio VaR*	*Contrastes de* "backtesting" *de cobertura condicional*													
			DB1	*DB2*	*DB3*	*DB4*	*DB5*	*DB6*	*DB7*	*LR*	*DQ1*	*DQ2*	*DQ3*	*DQV aR*$_1$	*DQV aR*$_2$	*DQV aR*$_3$
BTC																
	HAR	0.140	15.98	15.98	15.98	15.98	16.89	16.90	16.71	6.79	3.62	3.62	3.63	3.68	3.75	3.80
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.03)	(0.16)	(0.31)	(0.46)	(0.30)	(0.59)	(0.80)
	SHAR	0.141	15.98	15.98	15.98	15.98	16.96	16.96	16.86	6.79	3.62	3.62	3.63	3.72	3.74	3.77
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.03)	(0.16)	(0.31)	(0.46)	(0.29)	(0.59)	(0.81)
	HARQ	0.101	9.31	9.95	12.71	14.02	12.65	12.62	12.34	0.28	7.35	13.60	14.16	7.67	14.48	15.80
			(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.87)	(0.03)	(0.00)	(0.01)	(0.05)	(0.01)	(0.03)
	SHARQ	0.106	10.68	10.70	12.76	13.23	15.23	15.24	15.22	1.90	2.91	6.98	7.26	5.16	10.03	10.76
			(0.00)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	(0.00)	(0.39)	(0.23)	(0.07)	(0.12)	(0.16)	(0.07)	(0.15)
	TVHAR	0.076	153.92	192.92	193.20	194.86	218.51	218.15	217.92	160.25	507.83	628.71	648.47	649.37	735.61	768.87
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	DL	0.119	20.15	24.66	23.14	23.16	25.92	26.16	23.78	11.33	26.22	34.21	40.28	27.28	37.13	45.66
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	RF	0.111	9.21	12.16	12.80	13.02	13.76	13.78	11.31	0.22	6.52	12.07	12.61	7.01	12.74	13.46
			(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.89)	(0.04)	(0.01)	(0.01)	(0.07)	(0.03)	(0.06)
NASDAQ																
	HAR	0.024	49.27	50.84	50.91	53.49	51.48	51.64	51.02	49.40	85.99	86.75	103.44	95.75	102.06	116.42
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	SHAR	0.024	57.88	65.05	65.14	65.31	65.74	68.98	59.61	58.64	107.74	107.84	120.20	116.63	116.66	126.76
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	HARQ	0.024	43.80	45.92	46.23	48.08	50.54	50.93	50.41	44.05	75.14	76.21	95.47	105.39	108.63	128.74
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	SHARQ	0.024	33.55	36.50	36.61	37.47	40.70	40.91	40.12	34.18	56.34	58.81	69.75	86.30	87.30	94.92
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	TVHAR	0.078	3.15	12.71	12.78	13.04	33.27	33.26	33.27	4.43	7.23	10.45	13.40	27.14	30.80	34.06
			(0.21)	(0.01)	(0.01)	(0.02)	(0.00)	(0.00)	(0.00)	(0.11)	(0.03)	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)
	DL	0.024	54.96	56.84	57.59	57.81	58.77	58.79	57.87	54.99	97.80	102.22	106.53	105.25	116.37	119.11
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	RF	0.025	38.56	40.34	40.64	41.52	44.56	44.73	44.49	38.98	65.25	66.95	75.98	89.99	93.28	116.48
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
SP500																
	HAR	0.019	54.96	58.13	58.88	58.94	59.34	59.42	58.03	55.85	101.99	115.24	118.75	116.07	125.29	140.27
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	SHAR	0.019	33.55	34.88	35.08	39.03	35.80	35.82	34.81	34.18	56.34	58.81	61.14	59.94	64.98	66.74
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	HARQ	0.019	46.51	48.99	49.23	49.24	51.60	51.66	50.48	47.89	86.25	92.37	92.87	103.15	106.40	118.35
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	SHARQ	0.019	31.13	31.13	31.14	31.14	36.42	36.52	35.62	31.12	49.17	49.17	49.18	63.91	71.35	73.07
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	TVHAR	0.019	43.80	44.19	44.25	44.39	44.68	44.68	47.20	44.05	75.14	76.21	76.65	77.21	80.69	82.94
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	DL	0.019	63.86	66.78	67.35	67.37	67.43	67.54	66.41	64.39	120.02	129.93	132.14	130.87	139.66	142.46
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	RF	0.020	41.15	41.22	41.23	42.13	45.43	45.36	45.25	41.21	68.91	69.17	70.58	88.38	94.56	104.81
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Nota: Se utilizaron trece contrastes de cobertura condicional, a saber, siete especificaciones binarias dinámicas *DB* Dumitrescu *et al.* (2012), seis contrastes *DQ* Engle y Manganelli (2004) incluyendo varios retardos de la variable de violaciones y VaR, así como el contraste LR_{cc} Christoffersen (1998). Los valores *p* correspondientes se presentan entre paréntesis.

Fuente: Candel *et al.* (2020).

Cuadro 7.

Valor en riesgo: nivel de confianza del 95 %

Variable	*Modelo*	*Promedio VaR*	*Contrastes de* "backtesting" *de cobertura condicional*													
			DB1	*DB2*	*DB3*	*DB4*	*DB5*	*DB6*	*DB7*	*LR*	*DQ1*	*DQ2*	*DQ3*	*DQV aR*$_1$	*DQV aR*$_2$	*DQV aR*$_3$
BTC																
	HAR	0.098	48.93 (0.00)	53.61 (0.00)	54.34 (0.00)	54.35 (0.00)	53.75 (0.00)	53.84 (0.00)	48.96 (0.00)	43.25 (0.00)	29.69 (0.00)	30.60 (0.00)	31.41 (0.00)	29.72 (0.00)	30.63 (0.00)	31.44 (0.00)
	SHAR	0.099	46.08 (0.00)	49.86 (0.00)	50.60 (0.00)	50.61 (0.00)	50.24 (0.00)	50.27 (0.00)	46.08 (0.00)	40.46 (0.00)	28.03 (0.00)	28.83 (0.00)	29.55 (0.00)	28.03 (0.00)	28.83 (0.00)	29.59 (0.00)
	HARQ	0.073	12.73 (0.00)	17.09 (0.00)	18.92 (0.00)	19.28 (0.00)	17.72 (0.00)	24.49 (0.00)	15.80 (0.00)	6.75 (0.03)	5.71 (0.06)	10.56 (0.01)	10.85 (0.03)	7.00 (0.07)	11.72 (0.04)	12.00 (0.10)
	SHARQ	0.074	10.43 (0.01)	13.80 (0.00)	14.87 (0.00)	14.80 (0.01)	13.05 (0.01)	13.04 (0.02)	14.62 (0.00)	4.54 (0.10)	3.69 (0.16)	5.13 (0.16)	6.50 (0.16)	5.39 (0.15)	6.78 (0.24)	8.12 (0.32)
	TVHAR	0.053	53.92 (0.00)	97.48 (0.00)	97.57 (0.00)	97.84 (0.00)	142.86 (0.00)	142.89 (0.00)	142.83 (0.00)	59.37 (0.00)	95.36 (0.00)	115.18 (0.00)	117.89 (0.00)	165.46 (0.00)	174.92 (0.00)	190.28 (0.00)
	DL	0.084	6.73 (0.03)	14.69 (0.00)	17.97 (0.00)	18.37 (0.00)	20.06 (0.00)	20.31 (0.00)	15.58 (0.00)	4.39 (0.11)	7.88 (0.02)	18.01 (0.00)	18.01 (0.00)	8.85 (0.03)	22.74 (0.00)	22.74 (0.00)
	RF	0.078	26.45 (0.00)	29.06 (0.00)	31.50 (0.00)	31.63 (0.00)	30.34 (0.00)	30.32 (0.00)	28.25 (0.00)	21.50 (0.00)	15.40 (0.00)	17.26 (0.00)	17.40 (0.00)	15.89 (0.00)	17.70 (0.00)	17.84 (0.01)
NASDAQ																
	HAR	0.017	41.85 (0.00)	48.82 (0.00)	49.70 (0.00)	49.82 (0.00)	48.97 (0.00)	54.53 (0.00)	46.94 (0.00)	41.74 (0.00)	53.11 (0.00)	56.80 (0.00)	57.71 (0.00)	67.37 (0.00)	70.61 (0.00)	71.35 (0.00)
	SHAR	0.017	46.45 (0.00)	59.57 (0.00)	61.17 (0.00)	61.26 (0.00)	60.29 (0.00)	64.51 (0.00)	52.16 (0.00)	46.40 (0.00)	59.74 (0.00)	64.07 (0.00)	66.74 (0.00)	74.78 (0.00)	77.01 (0.00)	78.04 (0.00)
	HARQ	0.017	36.01 (0.00)	37.34 (0.00)	41.54 (0.00)	42.02 (0.00)	41.95 (0.00)	47.03 (0.00)	41.87 (0.00)	36.13 (0.00)	45.35 (0.00)	48.74 (0.00)	48.75 (0.00)	63.73 (0.00)	69.85 (0.00)	70.57 (0.00)
	SHARQ	0.017	23.07 (0.00)	24.52 (0.00)	30.73 (0.00)	30.76 (0.00)	31.77 (0.00)	31.80 (0.00)	29.35 (0.00)	24.02 (0.00)	29.16 (0.00)	29.77 (0.00)	29.89 (0.00)	44.48 (0.00)	44.70 (0.00)	44.86 (0.00)
	TVHAR	0.055	11.96 (0.00)	30.01 (0.00)	31.05 (0.00)	33.80 (0.00)	55.06 (0.00)	55.10 (0.00)	54.69 (0.00)	15.83 (0.00)	13.87 (0.00)	17.09 (0.00)	28.89 (0.00)	23.40 (0.00)	27.03 (0.00)	36.93 (0.00)
	DL	0.017	44.90 (0.00)	45.30 (0.00)	47.78 (0.00)	47.53 (0.00)	48.88 (0.00)	49.95 (0.00)	48.72 (0.00)	45.05 (0.00)	57.90 (0.00)	58.45 (0.00)	59.05 (0.00)	61.52 (0.00)	65.47 (0.00)	66.45 (0.00)
	RF	0.017	36.01 (0.00)	37.15 (0.00)	43.86 (0.00)	43.87 (0.00)	43.47 (0.00)	43.77 (0.00)	40.99 (0.00)	36.63 (0.00)	46.19 (0.00)	46.90 (0.00)	47.07 (0.00)	60.49 (0.00)	60.50 (0.00)	63.12 (0.00)
SP500																
	HAR	0.014	31.88 (0.00)	38.92 (0.00)	39.13 (0.00)	40.16 (0.00)	39.11 (0.00)	41.27 (0.00)	34.94 (0.00)	32.01 (0.00)	39.75 (0.00)	46.78 (0.00)	47.12 (0.00)	46.79 (0.00)	51.83 (0.00)	54.84 (0.00)
	SHAR	0.014	21.92 (0.00)	28.35 (0.00)	29.11 (0.00)	31.07 (0.00)	28.35 (0.00)	29.43 (0.00)	25.91 (0.00)	21.82 (0.00)	26.11 (0.00)	35.07 (0.00)	35.23 (0.00)	34.02 (0.00)	42.33 (0.00)	43.09 (0.00)
	HARQ	0.013	26.70 (0.00)	30.52 (0.00)	31.68 (0.00)	31.67 (0.00)	30.05 (0.00)	30.02 (0.00)	29.69 (0.00)	26.74 (0.00)	32.60 (0.00)	42.09 (0.00)	42.46 (0.00)	40.66 (0.00)	47.17 (0.00)	49.62 (0.00)
	SHARQ	0.013	18.60 (0.00)	20.92 (0.00)	22.17 (0.00)	22.29 (0.00)	20.50 (0.00)	20.92 (0.00)	20.50 (0.00)	18.51 (0.00)	21.84 (0.00)	26.91 (0.00)	27.79 (0.00)	24.89 (0.00)	30.31 (0.00)	32.89 (0.00)
	TVHAR	0.014	27.96 (0.00)	35.75 (0.00)	32.35 (0.00)	37.67 (0.00)	36.31 (0.00)	40.01 (0.00)	33.62 (0.00)	27.96 (0.00)	34.24 (0.00)	43.09 (0.00)	43.10 (0.00)	40.59 (0.00)	47.10 (0.00)	47.36 (0.00)
	DL	0.013	31.88 (0.00)	35.03 (0.00)	35.36 (0.00)	35.35 (0.00)	40.05 (0.00)	40.21 (0.00)	40.03 (0.00)	31.79 (0.00)	39.34 (0.00)	46.42 (0.00)	46.44 (0.00)	55.95 (0.00)	59.94 (0.00)	60.42 (0.00)
	RF	0.014	25.47 (0.00)	31.88 (0.00)	34.23 (0.00)	34.29 (0.00)	32.80 (0.00)	33.00 (0.00)	30.84 (0.00)	25.59 (0.00)	31.13 (0.00)	48.84 (0.00)	49.11 (0.00)	41.75 (0.00)	55.74 (0.00)	59.46 (0.00)

Nota: Se utilizaron trece contrastes de cobertura condicional, a saber, siete especificaciones binarias dinámicas *DB* Dumitrescu *et al.* (2012), seis contrastes *DQ* Engle y Manganelli (2004) incluyendo varios retardos de la variable de violaciones y VaR, así como el contraste LR_{cc} Christoffersen (1998). Los valores *p* correspondientes se presentan entre paréntesis.

Fuente: Candel *et al.* (2020).

El cuadro 6 muestra los resultados de las pruebas de *backtesting* para los niveles de confianza del 99 % y 95 %, respectivamente, para las tres series de rendimientos financieros. Teniendo en cuenta el nivel de confianza del 99 % y los rendimientos de BTC, observamos que los modelos HARQ y RF son los que mejor predicen el VaR, dado que las hipótesis nulas de cobertura condicionales son menos rechazadas.

Los contrastes de Engle y Manganelli (2004) y de Christoffersen (1998) no son rechazados casi nunca, y los contrastes en los modelos dinámicos binarios no rechazan la cobertura condicional al nivel de significación del 1 %.

En cuanto a los rendimientos de NASDAQ, los modelos en general predicen muy mal el VaR, excepto el modelo TVHAR, que genera coberturas condicionales más cercanas a las teóricas. En cuanto a los rendimientos de SP500, los modelos predicen VaR que no superan las pruebas de *backtesting*.

Los resultados del *backtesting* al nivel de confianza del 5 % son un poco peores (véase cuadro 7). Solo para el BTC, el modelo SHARQ logra predecir el valor en riesgo de forma más o menos adecuada.

4. CONCLUSIONES

El propósito de nuestro estudio se centra en profundizar en la predicción de la volatilidad medida por la varianza realizada que es calculada a partir de datos intradiarios, particularmente de datos de cinco minutos. Se ha demostrado que esta medida es una medida consistente de la volatilidad diaria, una variable que no es observable. El objetivo es analizar qué modelos son los más adecuados para prever las fluctuaciones en el riesgo, con el fin de optimizar las estrategias de inversores y gestores de carteras.

En el capítulo que presentamos, se llevaron a cabo predicciones utilizando diversos modelos paramétricos, semiparamétricos y de aprendizaje automático. La evaluación de su desempeño se hizo mediante funciones de pérdida específicas y contrastes de habilidad predictiva superior, tanto incondicionales como condicionales. Además, se contrastaron los modelos en el contexto de la gestión del riesgo mediante el cálculo del valor en riesgo.

Los datos analizados incluyen los rendimientos de Bitcoin y dos índices bursátiles: NASDAQ y S&P 500, que representan diferentes grupos de mercados: criptomonedas, tecnología y el mercado de acciones estadounidense, respectivamente.

Se observó que algunos modelos tenían un desempeño superior en función de la serie de rendimientos financieros utilizada. Sin embargo, en términos generales, se encontró que el modelo heterogéneo autorregresivo y los modelos heterogéneos autorregresivos con cuarticidad simétrico y asimétrico y *Random Forests* presentaban los valores más bajos de QLIKE, indicando que son los más adecuados para predecir la varianza realizada (a un día) en las tres series de rendimientos analizadas.

En cuanto al cálculo del valor en riesgo, se aplicaron trece pruebas condicionales de *backtesting,* que tienen la ventaja de una mayor potencia en comparación con las pruebas convencionales y tienen en cuenta el error de estimación. Las pruebas muestran evidencia de que los modelos heterogéneos autorregresivos con cuarticidad, tanto simétrico como asimétrico, y modelos semiparamétricos como *Random Forests,* parecen adecuados para el cálculo del valor en riesgo para Bitcoin. Mientras que el modelo heterogéneo autorregresivo con parámetros que cambian con el tiempo, es el único modelo avalado por las pruebas de *backtesting* para el NASDAQ.

En relación al S&P 500, los modelos muestran un bajo desempeño para el cálculo del valor en riesgo en el periodo considerado. Es relevante señalar que, durante periodos de alta volatilidad, como el confinamiento por la pandemia, que afectó en mayor medida a las empresas convencionales incluidas en el índice S&P 500, se observó un peor desempeño de los modelos en términos del cálculo del valor en riesgo.

En conclusión, el estudio de la predicción de la varianza realizada en los mercados financieros es un campo de gran importancia, que tiene un impacto directo en la toma de decisiones de inversión y en la gestión del riesgo. Contar con modelos precisos y confiables es esencial para navegar en los dinámicos y competitivos mercados financieros actuales.

Referencias

Andersen, T. G., Bollerslev, T., Diebold, F. X. y Ebens, H. (2001a). The distribution of realized stock return volatility. *Journal of Financial Economics,* 61, pp. 43–76. doi: https://doi.org/10.1016/S0304-405X(01)00055-1

Andersen, T. G., Bollerslev, T., Diebold, F. X. y Labys, W. C. (2000). Great realizations in the exchange rate: Exchange rates and inflation in industrialized countries. *Journal of International Economics,* 50, pp. 87–108.

Andersen, T. G., Bollerslev, T., Diebold, F. X. y Labys, W. C. (2001b). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association,* 96, pp. 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. y Labys, W. C. (2003). Modeling and forecasting realized volatility. *Econometrica* 71, pp. 579–625.

Andersen, T. G., Bollerslev, T. y Dobrev, D. (2007). No-arbitrage semimartingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications. *Journal of Econometrics,* 138, pp. 125–180.

Barendse, S., Kole, E. y van Dijk, D. (2021). Backtesting Value-at-Risk and Expected Shortfall in the Presence of Estimation Error. *Journal of Financial Econometrics,* 21, pp. 528–568. doi:10. 1093/jjfinec/nbab008

Barndorff-Nielsen, O. E., Kinnebrock, S. y Shephard, N. (2010). Measuring downside risk? Realised semivariance. En T. Bollerslev, J. Russell y M. Watson, M. (Eds.), *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle.* Oxford University Press, New York.

Bollerslev, T., Litvinova, J. y Tauchen, G. (2006). Leverage and volatility feedback effects in high- frequency data. *Journal of Financial Econometrics,* 4, pp. 353–384.

Bollerslev, T., Patton, A. J. y Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics,* 192, pp. 1–18.

Breiman, L. (2001). Random forests. *Machine learning,* 45, pp. 5–32.

Cai, Z. (2007). Trending time-varying coefficient time series with serially correlated errors. *Journal of Econometrics,* 136, pp. 163–188.

Campbell, J. Y. y Hentschel, L. (1992). No news is good news: A asymmetric model of changing volatility in stock returns. *Journal of Financial Economics,* 31, pp. 281–318.

Candel, A., Parmar, V., LeDell, E. y Arora, A. (2020). Deep learning with h2o. H_{2O}. ai Inc , pp. 1–21.

Chen, X. B., Gao, J., Li, D. y Silvapulle, P. (2018). Nonparametric estimation and forecasting for time-varying coefficient realized volatility models. *Journal of Business & Economic Statistics,* 36, pp. 88–100.

Christie, A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics,* 10, pp. 407–432.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review,* 39, pp. 841–862.

Corbet, S., Larkin, C. y Lucey, B. (2020). The contagion effects of the COVID-19 pandemic: Evidence from gold and cryptocurrencies. *Finance Research Letters,* 35, 101554. doi: https://doi.org/10.1016/j.frl.2020.101554.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics,* 7, pp. 174–196.

Dumitrescu, E. I., Hurlin, C. y Pham, V. (2012). Backtesting Value-at-Risk: From Dynamic Quantile to Dynamic Binary Test. *Revue de l'association française de finance,* 33, pp. 79–112.

Engle, R. F. y Manganelli, S. (2004). CAViaR. *Journal of Business and Economic Statistics,* 22, pp. 367–381. doi:10.1198/073500104000000370

Escanciano, J. y Olmo, J. (2010). Backtesting parametric Value-at-Risk with estimation risk. *Journal of Business and Economic Statistics,* 28, pp. 36–51. doi:10.1198/jbes.2009.07063

Escanciano, J. C. y Olmo, J. (2011). Robust backtesting test for value-at-risk models. *Journal of Financial Econometrics,* 9, pp. 132–161. doi:10.1093/jjfinec/nbq021

Fan, J. y Gijbels, I. (1996). Local Polynomial Modeling and Its Applications. Hapman and Hall, London.

Goodell, J. W. y Goutte, S. (2021). Co-movement of COVID-19 and Bitcoin: Evidence from wavelet coherence analysis. *Finance Research Letters,* 38, 101625. doi: https://doi.org/10.1016/j.frl.2020.101625

Gourieroux, C. y Jasiak, J. (2001). Financial Econometrics: Problems, Models, and Methods. Princeton University Press.

Hansen, P. R., Lunde, A. y Nason, J. M. (2011). The model confidence set. *Econometrica,* 79, pp. 453–497.

Hastie, T., Tibshirani, R., Friedman, J. H. y Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Volume 2. Springer.

Kauppi, H. y Saikkonen, P. (2008). Predicting US recessions with dynamic binary response models. *Review of Economics and Statistics,* 90, pp. 777–791.

Landry, M. (2016). Machine learning with r and h2o. H_{2O}. ai: Mountain View, CA, USA .

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., *et al.*,(2018). *Package h2o. R* Foundation for Statistical Computing: Vienna, Austria , 17.

Li, J., Liao, Z. y Quaedvlieg, R. (2022). Conditional superior predictive ability. *The Review of Economic Studies,* 89, pp. 843–875.

McAleer, M. (2014). Oil shocks and the macroeconomy: The role of price variability. *Econometrics,* 2, pp. 145–150.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160, 246–256.

Patton, A. J. y Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility full access. *Review of Economics and Statistics,* 97, pp. 683–697.

Robinson, P. (1989). Nonparametric estimation of time-varying parameters. En P. Hackl (Ed.), *Statistical Analysis and Forecasting of Economic Structural Change*. Springer, Berlin.

Sifat, I. (2021). On cryptocurrencies as an independent asset class: Long-horizon and COVID- 19 pandemic era decoupling from global sentiments. *Finance Research Letters,* 43, 102013. doi: https://doi.org/10.1016/j.frl.2021.102013

CAPÍTULO VI

Selección de activos para construir carteras de inversión en base a su asimetría y curtosis

M. Angeles Carnero
Ángel León
Trino-Manuel Ñíguez

El objetivo de este trabajo es evaluar la rentabilidad de carteras de inversión construidas con acciones seleccionadas según algunas medidas de rendimiento basadas en los cuatro primeros momentos de la distribución de rentabilidades (media, varianza, asimetría y curtosis). Dichos momentos se estiman a partir de un modelo de volatilidad condicional con innovaciones potencialmente asimétricas. Usando los activos del índice Russell 1000 observados diariamente durante los últimos 22 años, comparamos los rendimientos de las distintas carteras obtenidas a partir de medidas de selección tales como extensiones de la ratio de Sharpe que tienen en cuenta la asimetría y curtosis de las rentabilidades. Además, analizamos la diferencia entre estimar los momentos marginales de la distribución de las rentabilidades asumiendo que éstas siguen un modelo TGARCH con innovaciones posiblemente asimétricas y usar los correspondientes estimadores muestrales.

Palabras clave: TGARCH, volatilidad condicional, ratio de Sharpe, VaR, Cornish-Fisher.

L. INTRODUCCIÓN

En finanzas, la selección de activos para construir carteras de inversión es una tarea crucial. El uso de grandes bancos de datos puede ayudar a anticipar tendencias en el mercado y comportamientos de diferentes activos, permitiendo a los inversores ajustar sus carteras de manera eficiente. Además, la capacidad de acceder y analizar datos en tiempo real proporciona a los inversores una ventaja competitiva al tomar decisiones rápidas y fundamentadas en la información más reciente del mercado. En este sentido, este capítulo ilustra cómo el uso de grandes conjuntos de datos puede ayudar a predecir la rentabilidad que tendrá una determinada cartera de inversión y cómo usar dichas predicciones para seleccionar los activos con los que construir una cartera de inversión.

Los inversores buscan maximizar sus rendimientos teniendo en cuenta el riesgo. Algunos métodos usados en la práctica son: i) evaluar los activos en función de fundamentos económicos y financieros (por ejemplo, ingresos, ganancias, deuda, perspectivas de crecimiento, etc.) y seleccionar aquellos activos con mejores resultados; ii) predecir el comportamiento futuro de los activos a través de modelos estadísticos usando toda la información disponible (por ejemplo, precios históricos de los activos, volúmenes de negociación, etc.); iii) reducir el riesgo de la cartera seleccionando activos de diferentes clases, sectores o regiones buscando que las posibles pérdidas en un activo sean compensadas por las ganancias en otros. La elección de un método de selección de activos dependerá de factores como los objetivos de inversión, el horizonte temporal, la tolerancia al riesgo, etc., por lo que, con frecuencia, se combinan varios métodos.

Una vez seleccionados los activos que formarán parte de la cartera, otro problema importante es determinar qué peso dar a cada activo. Supongamos que se dispone de un euro para invertir en una cartera de cinco activos. Podría construirse una cartera en la que todos los activos tengan el mismo peso (0,2 euros para cada activo) o se podrían determinar los pesos siguiendo otro criterio, por ejemplo, seleccionar los pesos de manera tal que la cartera tenga la menor varianza posible. En esta línea, Brownlees *et al.* (2021) analiza el problema de selección de carteras de inversión con un gran número de activos financieros destacando la importancia que tiene la estimación de la matriz de varianzas y covarianzas de la cartera. Dicho trabajo presenta nueva metodología para modelizar covarianzas y correlaciones dinámicas de elevada dimensionalidad e ilustra sus ventajas a través de un ejercicio empírico usando los 100 activos con mayor capitalización bursátil del S&P 500.

El objetivo de este trabajo es analizar fundamentalmente el problema de cómo seleccionar los activos que formarán parte de la cartera y, en menor medida, cómo seleccionar los pesos de cada activo. Como ilustración usaremos dos carteras, la cartera equiponderada y la cartera de mínima varianza global. En la primera, los pesos son iguales para todas las acciones, mientras en la segunda se seleccionan los pesos tal que minimizan la varianza global de la cartera de rentabilidades que se encuentra en la frontera eficiente de Markowitz (espacio

media-varianza) construida con posiciones largas en todos los activos seleccionados, véase para más información Bodie *et al.* (2023).

Para elegir los activos de nuestras carteras utilizamos varias métricas del rendimiento, que indican la rentabilidad del activo ajustada a su riesgo, donde cada medida se construye a partir de diferentes candidatos de medición de rentabilidad y riesgo como veremos a continuación. En concreto, consideramos los siguientes indicadores: la ratio de Sharpe (SR) (Sharpe, 1966), la ratio de Sharpe modificada (mSR), propuesta por Favre y Galeano (2002) y Gregoriou y Gueyie (2003), así como la ratio de asimetría-curtosis (SKR) de Watanabe (2006). Por un lado, la SR está basada en el análisis tradicional media-varianza, mientras que la mSR y la SKR tienen en cuenta cuantiles y momentos de orden superior, como asimetría y curtosis, de la distribución de la rentabilidad. Obtenemos cada uno de estos indicadores mediante momentos muestrales, por un lado, y, por otro, mediante los momentos incondicionales dados por densidades, basadas en expansiones polinómicas (PA) con estructura TGARCH en la varianza condicional (TGARCH-PA), véase Carnero *et al.* (2023), y media condicional constante. En particular, utilizamos densidades PA construidas a partir de: i) la densidad secante hiperbólica (HS) dando lugar a la distribución PAHS, y de ii) la densidad normal (N), dando lugar a la distribución PAN, o más conocida en la literatura como Gram-Charlier (GC). El modelo TGARCH-PA permite implementar momentos de tercer y cuarto orden incondicionales útiles para evaluar el riesgo y el rendimiento de los activos. Para más información sobre características de este tipo de densidades PA véase Bagnato *et al.* (2015) y las extensiones en León y Ñíguez (2022). El interés del trabajo se centra en la comparativa del rendimiento de carteras obtenidas mediante indicadores basados en momentos muestrales, y aquellos basados en los momentos incondicionales del modelo TGARCH-PA. Para este fin, implementamos un análisis fuera de muestra (*out-of-sample*) para carteras compuestas a partir de la selección de acciones que constituyen el índice Russell 1000 utilizando varias estrategias alternativas. Las rentabilidades acumuladas de la cartera se obtienen durante el período *out-of-sample* para cada estrategia elegida. Nuestros resultados empíricos muestran evidencia de ganancias considerables en los rendimientos acumulados de las carteras obtenidas mediante indicadores basados en los momentos implícitos del modelo TGARCH-PA.

La metodología usada a lo largo del trabajo se describe en la sección segunda. A continuación, la sección tercera contiene una aplicación empírica usando datos del índice Russell 1000. Por último, la sección cuarta presenta las conclusiones.

2. METODOLOGÍA

Esta sección describe algunos de los indicadores usados en finanzas para medir el rendimiento de una inversión, así como dos métodos alternativos para calcularlos. Además se describen dos estrategias para construir una cartera de inversión a partir de varios activos.

2.1. Indicadores de rendimiento de una inversión

La ratio de Sharpe (SR) es una medida propuesta por Sharpe (1966) y ampliamente utilizada para evaluar el rendimiento de una cartera de inversión en relación con su nivel de riesgo. Dicha ratio viene dada por:

$$SR = \frac{\mu - r_0}{\sigma} \qquad [1]$$

donde μ y σ son, respectivamente la media y desviación típica de la rentabilidad de la cartera y r_0 es la rentabilidad que se podría obtener con una inversión libre de riesgo (en este caso, por simplicidad, asumiremos que $r_0 = 0$). Cuanto mayor sea la ratio de Sharpe mejor será el rendimiento de la inversión ajustado al riesgo. Esta medida tiene la ventaja de que permite comparar de manera fácil y rápida varias inversiones aunque también tiene algunas limitaciones. Entre ellas está que su cálculo requiere estimar r_0, un valor que depende de las condiciones del mercado y/o los tipos de interés y que, por tanto, no es constante en el tiempo. Otra limitación es que, al usar únicamente los dos primeros momentos de la distribución de la rentabilidad de la cartera, implícitamente se está asumiendo que dicha rentabilidad sigue una distribución normal. Con el objetivo de solventar esta última desventaja, Favre y Galeano (2002), y Gregoriou y Gueyie (2003) proponen usar la ratio de Sharpe modificada (mSR), que viene dada por:

$$mSR = \frac{\mu - r_0}{|mVaR(\alpha)|} \qquad [2]$$

siendo

$$|mVaR(\alpha)| = -\mu - \sigma \times z_{CF}(\alpha)$$
$$= -\mu - \sigma\left[z_\alpha + \frac{1}{6}\left(z_\alpha^2 - 1\right)SK + \frac{1}{24}\left(z_\alpha^3 - 3z_\alpha\right)(K-3) - \frac{1}{36}\left(2z_\alpha^3 - 5z_\alpha\right)SK^2\right]$$

donde z_α es el cuantil $1 - \alpha$ de la distribución normal estándar (es decir, $P(Z \leq z_\alpha) = 1 - \alpha$ donde $Z \sim N(0, 1)$) y Sk y K son, respectivamente, los coeficientes de asimetría y curtosis de la rentabilidad de la cartera. Como puede verse, mVaR(α) es el valor en riesgo (estimación de la pérdida máxima esperada en una cartera de inversión en un período de tiempo específico y bajo cierto nivel de confianza) modificado para no tener que asumir que la distribución de la rentabilidad de la cartera es gaussiana.

Por último, vamos a considerar la ratio de asimetría-curtosis (SKR), propuesta por Watanabe (2006) y Bacon (2008), que proporciona una medida compuesta del riesgo y el rendimiento de una cartera de inversión. Cuanto mayor sea el valor de esta ratio, mayor será el rendimiento potencial de la cartera en relación con su riesgo. Dicha ratio viene dada por:

$$SKR = \frac{asimetría}{curtosis} \qquad [3]$$

Un problema con estos indicadores surge cuando el numerador (media o asimetría) es negativo. En este caso usaremos la metodología propuesta por Israelsen (2005) y extendida en León y Ñíguez (2020) que consiste en elevar el denominador de la ratio al signo del numerador.

2.1.1. Cálculo de los indicadores usando los estimadores muestrales

Denotemos por $r_1, r_2, ..., r_T$ la serie temporal de los rendimientos diarios de una cartera. Lo habitual es estimar su media, desviación típica y los coeficientes de asimetría y curtosis mediante sus respectivos valores muestrales:

$$\bar{r} = \frac{1}{T}\Sigma_{t=1}^{T} r_t \qquad s = \sqrt{\frac{1}{T-1}\Sigma_{t=1}^{T}\left(r_t - \bar{r}\right)^2}$$

$$\widehat{Sk} = \frac{\frac{1}{T-1}\Sigma_{t=1}^{T}\left(r_t - \bar{r}\right)^3}{s^3} \qquad \hat{K} = \frac{\frac{1}{T-1}\Sigma_{t=1}^{T}\left(r_t - \bar{r}\right)^4}{s^4}$$

De este modo tendríamos los correpondientes estimadores muestrales de los tres indicadores:

$\widehat{SR} = \frac{\bar{r}}{s}$; $\widehat{mSR} = \frac{\bar{r}}{\left|-\bar{r} - s\hat{z}_{CF}(\alpha)\right|}$ donde

$$\hat{z}_{CF}(\alpha) = z_\alpha + \frac{1}{6}\left(z_\alpha^2 - 1\right)\widehat{Sk} + \frac{1}{24}\left(z_\alpha^3 - 3z_\alpha\right)\left(\hat{K} - 3\right) - \frac{1}{36}\left(2z_\alpha^3 - 5z_\alpha\right)\widehat{SK}^2$$

y finalmente, $\widehat{SKR} = \frac{\widehat{Sk}}{\hat{K}}$.

2.1.2. Cálculo de los indicadores usando los estimadores implícitos obtenidos a partir de un modelo de volatilidad condicional

Supongamos que los rendimientos diarios de una inversión siguen un proceso r_t dado por

$$r_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t, \qquad [4]$$

donde μ_t y σ_t^2 denotan, respectivamente, la media y varianza de r_t condicionadas a la información disponible en el instante $t - 1$ y las innovaciones, z_t, son independientes e idénticamente distribuidas, con media 0 y varianza 1 (*i.e.* $\{z_t\} \sim iid\ (0, 1)$). En este caso, el tercer y cuarto momento de z_t son exactamente su asimetría y curtosis que denotaremos por sk_z y k_z, respectivamente. Consideremos $\psi_k = E(z_t^k)$, entonces $\psi_1 = 0$, $\psi_2 = 1$, $\psi_3 = sk_z$ y $\psi_4 = k_z$. Asu-

mimos que el error $\{\varepsilon_t\}$ en [4] sigue un proceso TGARCH(1,1) propuesto por Zakoian (1994) que modeliza directamente la volatilidad σ_t y que, como se muestra en Rodríguez y Ruiz (2012), es un modelo tipo GARCH apropiado y flexible que representa bien las propiedades de los rendimientos financieros, en particular, el exceso de curtosis, la heterocedasticidad condicional y el efecto apalancamiento o *leverage*. En este modelo, σ_t viene dado por:

$$\begin{aligned}\sigma_t &= \omega + \beta\sigma_{t-1} + \alpha^+\varepsilon_{t-1}^+ - \alpha^-\varepsilon_{t-1}^- \\ &= \omega + \sigma_{t-1}\left(\beta + \alpha^+ z_{t-1}^+ - \alpha^- z_{t-1}^-\right),\end{aligned} \qquad [5]$$

donde $\omega > 0$, $\beta \geq 0$, $\alpha^+ \geq 0$, y $\alpha^- \geq 0$. Se usa la notación $y_t^+ = \text{máx}\,(y_t, 0)$ y $y_t^- = \text{mín}\,(y_t, 0)$ donde y_t puede ser ε_t o z_t.

Este modelo permite una respuesta asimétrica de la volatilidad a rendimientos pasados positivos y negativos. En particular, la volatilidad tiende a ser mayor después de impactos negativos en los rendimientos que después de impactos positivos de la misma magnitud. Esto generalmente conduce a correlaciones cruzadas negativas entre rendimientos pasados y volatilidad. Como se puede ver en [5], cuando ε_{t-1} es positivo, la respuesta de la volatilidad es lineal en ε_{t-1} con pendiente α^+, pero si ε_{t-1} es negativo, la pendiente es α^-, y se espera que $\alpha^+ < \alpha^-$. Nótese que cuando $\alpha^+ = \alpha^-$, la volatilidad responde de manera simétrica a rendimientos pasados positivos y negativos y el modelo se reduce al modelo *Absolute Value GARCH* (AVGARCH) propuesto por Taylor (1986) y Schwert (1989). Como puede verse en Francq y Zakoian (2010), el modelo TGARCH es estrictamente estacionario si,

$$E\left[\ln\left(\beta + \alpha^+ z_t^+ - \alpha^- z_t^-\right)\right] < 0,$$

y es estacionario en covarianza si,

$$E\left[\left(\beta + \alpha^+ z_t^+ - \alpha^- z_t^-\right)^2\right] < 1$$

En Carnero *et al.* (2023) se muestran las expresiones analíticas para los cuatro primeros momentos del modelo TGARCH con innovaciones potencialmente asimétricas. Supongamos que $\{r_t\}$ viene dado por el modelo [4], estrictamente estacionario con momento de orden 4 finito. En el caso más sencillo en que $\mu_t = \mu$, puede demostrarse que los coeficientes de asimetría y curtosis de la rentabilidad r_t vienen dados por:

$$sk_r = sk_z \frac{E\left(\sigma_t^3\right)}{E\left(\sigma_t^2\right)^{3/2}},$$

$$k_r = k_z \frac{E\left(\sigma_t^4\right)}{E\left(\sigma_t^2\right)^2}.$$

donde,

$$E\left(\sigma_t^k\right)=\frac{\omega^k f_k}{\Pi_{j=1}^k\left(1-a_j\right)},\quad k=1,2,3,4$$

$f_1 = 1,$
$f_2 = 1 + a_1,$
$f_3 = 1 + 2a_1 + 2a_2 + a_1a_2,$
$f_4 = 1 + 3a_1 + 5a_2 + 3a_3 + 3a_1a_2 + 5a_1a_3 + 3a_2a_3 + a_1a_2a_3.$

$$a_k=\beta^k+\sum_{j=1}^{k}\frac{k!}{j!(k-j)!}\beta^{k-j}\left\{\left(\alpha^+\right)^j\psi_j+\left[t_jI\left(j\text{ es par}\right)-g_jI\left(j\text{ es impar}\right)\right]\psi_j^-\right\} \quad [6]$$

$$\psi_j=E\left[z_t^j\right],\psi_j^-=E\left[\left(z_t^-\right)^j\right],g_j=\left(\alpha^-\right)^j+\left(\alpha^+\right)^j,t_j=\left(\alpha^-\right)^j-\left(\alpha^+\right)^j$$

Como se deduce de las expresiones anteriores, los coeficientes de asimetría y curtosis de la rentabilidad r_t van a depender de los parámetros del TGARCH (*i.e.* ω, β, α^+ y α^-) pero también de los valores ψ_j y ψ_j^- que vienen dados por la distribución de las innovaciones z_t. Por ejemplo, si z_t tiene una distribución normal estándar, $\psi_1 = 0$; $\psi_2 = 1$; $\psi_3 = 0$, y $\psi_4 = 3$, y $\psi_1^- = -\frac{1}{\sqrt{2\pi}}; \psi_2^- = \frac{1}{2}; \psi_3^- = -\sqrt{\frac{2}{\pi}}; \psi_4^- = \frac{3}{2}.$

En Carnero *et al.* (2023) se asume que la función de densidad de las innovaciones z_t pertenece a la familia de densidades PA, basadas en expansiones polinómicas ortogonales de una densidad principal seleccionada. Por ejemplo, la densidad Gram-Charlier (GC) expande la normal estándar como densidad principal mediante polinomios de Hermite. Expansiones similares con la secante hiperbólica estandarizada (PAHS) y la logística (PAL) como densidades principales también se han propuesto en Bagnato *et al.* (2015) y Vacca *et al.* (2022) para modelizar los rendimientos de activos financieros. En este capítulo asumimos que los rendimientos r_t vienen dados por el modelo [4], con $\mu_t = \mu$, y este tipo de densidades PA (en concreto, GC y PAHS) para las innovaciones z_t. Bajo dicho supuesto, en lugar de usar los estimadores muestrales para calcular las tres ratios descritas anteriormente, proponemos estimar la media, desviación típica y los coeficientes de asimetría y curtosis de los rendimientos diarios a partir del modelo TGARCH-PA. Para ello, estimamos el modelo para las rentabilidades y usando los parámetros estimados y las expresiones obtenidas en Carnero *et al.* (2023), podemos calcular los valores de μ, σ, *Sk* y *K* que implica el modelo.

2.2. Estrategias de inversión

Suponiendo que disponemos de las series de rendimientos diarios de una muestra de activos, la estrategia de inversión que vamos a seguir, asumiendo por simplicidad que los costes de transacción son cero, es la siguiente:

1. Dividimos la muestra total, de tamaño *T*, en dos muestras, una que llamaremos *in-sample* y que usaremos para elegir los cinco activos que inicialmente formarán nuestra cartera y otra que llamaremos *out-of-sample* en donde evaluamos la rentabilidad acumulada de nuestra cartera de inversión.

2. Elegimos uno de los tres indicadores descritos anteriormente, por ejemplo, la ratio de Sharpe.

3. Con la muestra *in-sample* calculamos dicho indicador para todas las acciones en las que podríamos invertir, las ordenamos de mayor a menor valor y seleccionamos las cinco primeras.

4. Movemos nuestra muestra 25 días descartando las primeras observaciones de la muestra anterior e incluyendo las observaciones siguientes. Con esta nueva muestra repetimos el paso anterior. En este punto, podría suceder que las cinco primeras acciones sigan siendo las que ya forman parte de nuestra cartera, en cuyo caso la composición de mi cartera no cambiaría. Sin embargo, podría suceder que haya acciones en nuestra cartera que no estén entre las cinco primeras con el nuevo orden, lo que nos llevaría a venderlas y comprar las nuevas acciones que aparecen entre las cinco primeras.

5. Repetimos el proceso 100 veces, lo que nos lleva a una muestra *out-of-sample* de 2.500 observaciones en la que podemos evaluar nuestra estrategia observando el valor de nuestra cartera el último día (observación *T*) después de haber invertido un euro al inicio de la muestra *out-of-sample.*

Un problema muy relevante, que no abordamos aquí, es cómo seleccionar los pesos para cada una de las cinco acciones que formarán parte de nuestra cartera. En este trabajo vamos a hacerlo de dos maneras bien conocidas y estudiadas. En primer lugar vamos a construir una cartera equiponderada, aquella en la que se asigna el mismo peso o proporción a cada una de las acciones seleccionadas. La cartera equiponderada es una estrategia simple que busca diversificar entre todos los activos, sin favorecer a ninguno en particular. Como alternativa a la cartera equiponderada, vamos a construir también la cartera de mínima varianza, seleccionando los pesos de cada activo de manera que se minimice la varianza total de la cartera, teniendo en cuenta las correlaciones entre los activos.

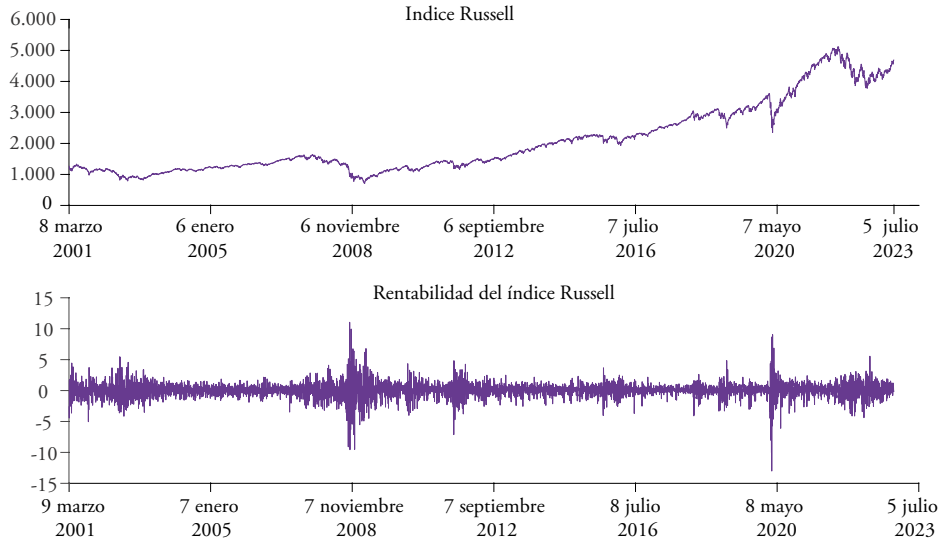
3. APLICACIÓN EMPÍRICA

En esta sección vamos a comparar los rendimientos de las distintas estrategias de inversión, descritas anteriormente, usando los activos del índice Russell 1000 observados diariamente desde el 8 de marzo de 2001 hasta el 5 de julio de 2023. El Russell 1000 es un índice bursátil que proporciona una visión general del rendimiento de las 1.000 empresas más grandes, de diversos sectores, que cotizan en bolsa en los Estados Unidos. Se actualiza anualmente para reflejar cambios en la capitalización de mercado de las empresas y por tanto, refleja

continuamente las 1.000 empresas más grandes en el mercado de valores estadounidense. Este índice está administrado por Russell Investments, una empresa de gestión de activos. La figura 1 muestra la evolución del índice y su rentabilidad a lo largo de nuestra muestra.

Figura 1.

Índice Russell 1000



Fuente: Elaboración propia.

Nuestro objetivo es seleccionar cinco empresas, de entre las 1.000, con las que construir una cartera de inversión. Disponemos por tanto de una muestra total de tamaño $T = 5.825$. Seguimos nuestra estrategia eligiendo primero la muestra *in-sample,* observada desde el 8 de marzo de 2001 hasta el 4 de diciembre de 2013 dando lugar a un tamaño muestral de 3.325 observaciones. Por otra parte, la muestra *out-of-sample* tendrá un tamaño muestral de 2.500 observaciones y se observa desde el 5 de diciembre de 2013 hasta el 5 de julio de 2023.

Cuadro 1.

Estadísticos descriptivos

Rentabilidad	*Índice*	*Russell*
	Total	*In-sample*
Media	0,0222	0,0120
Desviación típica	1,2232	1,2925
CV	55. 10	107. 71
Asimetría	-0,4580	-0,2354
Curtosis	14,319	11,684

Fuente: Elaboración propia.

En el cuadro 1 se muestran los estadísticos descriptivos del índice en la muestra total y en la muestra *in-sample* donde se observan algunas diferencias, especialmente en el coeficiente de variación.

Cuadro 2.

Códigos SIC de los 417 activos del índice Russell en los que poder invertir

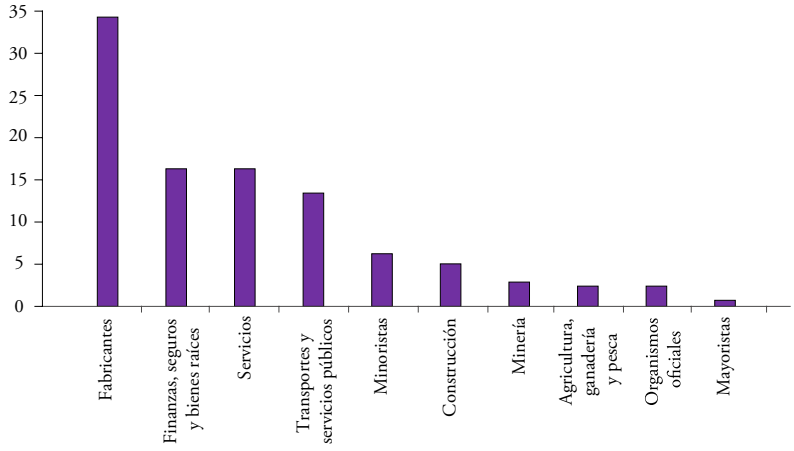
Categoría	*No activos*	*Porcentaje (%)*
Agricultura, ganadería y pesca	10	2,40
Minería	12	2,88
Construcción	21	5,04
Fabricantes	143	34,29
Transportes y servicios públicos	56	13,43
Mayoristas	3	0,72
Minoristas	26	6,24
Finanzas, seguros y bienes raíces	68	16,3
Servicios	68	16,3
Organismos oficiales	10	2,40
	417	100

Fuente: Elaboración propia.

Antes de elegir las cinco acciones que formarán parte de nuestra primera cartera, es conveniente destacar que hay un número considerable de empresas que no permanecen en el índice durante largos períodos de tiempo. Con el objetivo de que todas las empresas puedan formar parte de la cartera, en una primera selección nos quedamos con 577 activos que han

Figura 2.

Códigos SIC de los 417 activos del índice Russell en los que poder invertir



Fuente: Elaboración propia.

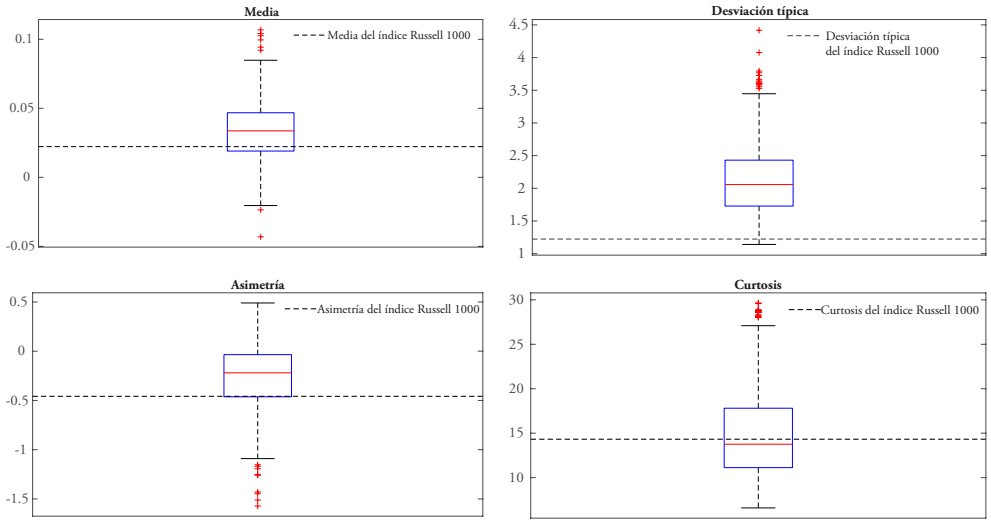
formado parte del índice durante todo el período muestral, esto es, desde el 8 de marzo de 2001 hasta el 5 de julio de 2023. Dichos activos aparecen detallados en el Apéndice 1. Por otra parte, para poder estimar μ, σ, Sk y K a partir del modelo TGARCH-PA es necesario que existan los cuatro primero momentos de las rentabilidades, y para ello se ha de cumplir que a_k para $k \leq 4$ dado por [6] sea menor que 1. Seleccionamos, por tanto, aquellos 417 activos en los que, una vez estimado el modelo, se cumple dicha condición. Podemos clasificar los 417 activos de acuerdo a los códigos SIC (*Standard Industrial Classification*) en 10 categorías como muestran tanto el cuadro 2 como la figura 2.

La figura 3 presenta la distribución de los cuatro primeros momentos de los 417 activos seleccionados en comparación con el índice Russell. Como podemos ver, tanto la media como los coeficientes de asimetría y curtosis de los activos, se distribuyen, en media, como el índice. Por otra parte, como era de esperar, la desviación típica del índice es menor que la de sus activos componentes.

Con los 417 activos seleccionados procedemos entonces con nuestra estrategia: i) con la muestra *in-sample,* es decir, con la información que tenemos el 4 de diciembre de 2013, para cada activo, calculamos los tres indicadores, dados por [1], [2] y [3] respectivamente, usando los estimadores muestrales y los implícitos según el modelo TGARCH-PA estimado. A continuación, ordenamos las 417 acciones de mayor a menor valor y seleccionamos las cinco primeras, obteniendo así seis carteras, una por cada indicador y cada estimador; ii) movemos la

Figura 3.

Diagramas de caja de los cuatro primeros momentos de los 417 activos del índice Russell



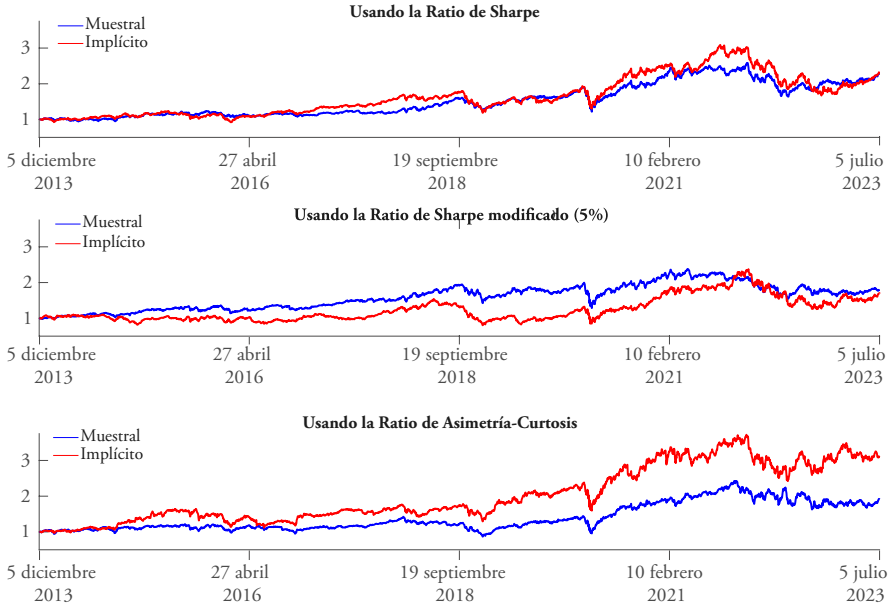
Fuente: Elaboración propia.

muestra 25 días descartando las primeras observaciones de la muestra anterior e incluyendo las observaciones siguientes (es decir, la nueva muestra irá desde el día 12 de abril de 2001 hasta el 8 de enero de 2014). Con esta nueva muestra repetimos el proceso 100 veces, y evaluamos nuestra estrategia observando el valor de nuestra cartera el día 5 de julio de 2023 después de haber invertido un euro el 4 de diciembre de 2013.

En el Apéndice 2 se muestran los activos que han sido seleccionados para cada una de las seis carteras descritas anteriormente, así como la frecuencia con la que dicho activo forma parte de la cartera. Por ejemplo, podemos ver que el activo número 30, que corresponde a APPLE (véase Apéndice 1), ha resultado seleccionado siguiendo la ratio de Sharpe obtenido con ambos estimadores. Además, en las 100 veces que se ha repetido el proceso, esta acción ha sido seleccionada en 81 ocasiones cuando la SR se ha obtenido con el estimador muestral, y en las 100 veces cuando se ha usado el estimador implícito. Sin embargo, APPLE no ha formado parte de ninguna cartera cuando el indicador elegido es la mSR obtenida con el estimador muestral ni cuando seguimos el criterio dado por la SKR. Cabe señalar que únicamente 126 de los 417 activos posibles han formado parte de alguna cartera y, por tanto, hay 291 acciones que, de acuerdo a los indicadores usados, no han quedado nunca entre las cinco primeras y en consecuencia no han formado parte de ninguna cartera. Un ejemplo de este último caso es el activo número 20, que corresponde a American Express o el número 97 correspondiente a

Figura 4.

Evolución de las distintas carteras equiponderadas



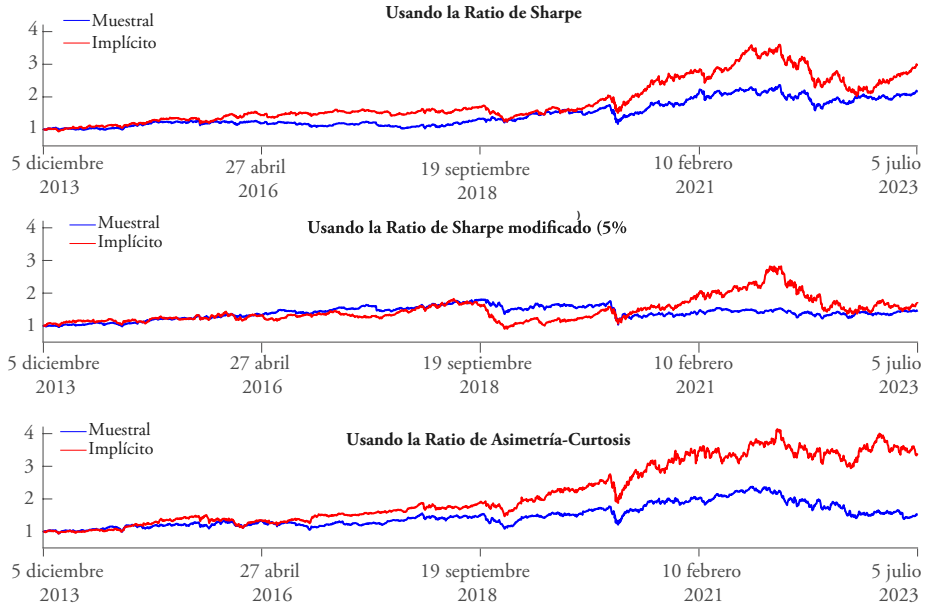
Fuente: Elaboración propia.

Coca Cola. El Apéndice 3 contiene una tabla en la que pueden verse aquellos activos comunes en grupos de dos carteras cada una asociada a un indicador diferente. Como se puede observar, el activo número 478, que corresponde a Amazon.com, es el que más se repite cuando comparamos todas las carteras. Dicho activo ha formado parte de todas las carteras excepto de la formada siguiendo la mSR obtenida con el estimador implícito. En cambio, el activo número 372, que corresponde a Starbucks, únicamente ha formado parte de las carteras formadas siguiendo la SR obtenida con ambos estimadores.

La figura 5 muestra la evolución de una cartera equiponderada (de mínima varianza) a lo largo de la muestra *out-of-sample*. Se observan carteras que han evolucionado mejor que otras. El gráfico de arriba de la figura 4 compara la evolución de las carteras equiponderadas elegidas de acuerdo a la SR estimada usando los estimadores muestrales (línea azul) y usando los estimadores implícitos por el modelo TGARCH-PA (línea roja). Como puede verse, las diferencias son muy pequeñas en los primeros años, alternándose luego períodos en los que una mejora ligeramente a la otra. Sin embargo, no parece que una cartera evolucione sistemáticamente mejor que la otra, lo cual no es sorprendente puesto que ambas carteras tienen muchos activos en común (véase Apéndice 3). Esto cambia si nos fijamos en el gráfico de arriba de la figura 5, donde podemos hacer la misma comparación pero con la cartera de mínima varianza. En este caso puede verse que la evolución de la cartera elegida según la SR

Figura 5.

Evolución de las distintas carteras de mínima varianza



Fuente: Elaboración propia.

estimada usando los estimadores implícitos por el modelo TGARCH-PA es sistemáticamente mejor que aquella construida usando los estimadores muestrales.

En los gráficos centrales de ambas figuras se compara la evolución de las carteras elegidas según la mSR. Se observa que, tanto para las carteras equiponderadas como de mínima varianza, la diferencia debida al estimador elegido es marginal, especialmente al final del horizonte de inversión. Por último, como puede verse en ambas figuras, la rentabilidad acumulada de la cartera construida según la SKR implícita es sistemáticamente mejor que la correspondiente a la cartera seleccionada con la SKR muestral.

4. CONCLUSIONES

En este trabajo se ha llevado a cabo un ejercicio para evaluar la rentabilidad de carteras de inversión construidas con acciones seleccionadas según tres medidas de rendimiento basadas en los cuatro primeros momentos de la distribución de rentabilidades (media, varianza, asimetría y curtosis); en concreto, la ratio de Sharpe, la ratio de Sharpe modificada y la ratio de asimetría-curtosis. Usando los activos del índice Russell 1000 observados diariamente desde el 5 de diciembre de 2013 hasta el 5 de julio de 2023, se han comparado los rendimientos de las distintas carteras obtenidas a partir de dichas medidas de selección. Los resultados muestran la diferencia entre estimar los momentos marginales de la distribución de las rentabilidades asumiendo que éstas siguen un modelo TGARCH con innovaciones posiblemente asimétricas y usar los correspondientes estimadores muestrales. Cuando la cartera elegida es la equiponderada, de acuerdo con los indicadores SR y mSR, parece no haber grandes diferencias entre el uso de los estimadores muestrales y los implícitos. Sin embargo, las diferencias son notables cuando se elige la cartera equiponderada o de mínima varianza, de acuerdo con la ratio de asimetría-curtosis, en donde usar los estimadores implícitos mejora la rentabilidad acumulada.

Referencias

Bacon, C. R. (2008). *Practical portfolio performance measurement and attribution.* John Wiley & Sons.

Bagnato, L., Potl, V. y Zoia, M. G. (2015). The role of orthogonal polynomials in adjusting hyperbolic secant and logistic distributions to analyse financial asset returns. *Statistical Papers,* 56, pp. 1205-1234.

Bodie, Z., Kane, A. y Marcus, A. J. (2023). *Investments.* McGraw-Hill Education, edición n.º 13.

Brownlees, C., Llorens, J. y Senar, N. (2021). Modelos de selección de carteras con muchos activos. En D. Peña, P. Poncela y E. Ruiz, *Nuevos métodos de predicción económica con datos masivos* (pp. 33-60). Madrid: Funcas.

Carnero, M. A., León, A. y Ñíguez, T.-M. (2023). *Analytic moments of TGARCH models with polynomially adjusted densities.* SSRN 3973456.

Favre, L. y Galeano, J. A. (2002). Mean-modified value-at-risk optimization with hedge funds. *Journal of Alternative Investments, 5*(2), pp. 21-25.

Francq, C. y Zakoian, J. M. (2010). G*ARCH Models: Structure, Statistical Inference and Financial Applications.* Wiley.

Gregoriou, G. N. y Gueyie, J. P. (2003). Risk-adjusted performance of funds of hedge funds using a modified Sharpe ratio. *Journal of Wealth Management, 6*(3), pp. 77-83.

Israelsen, C. L. (2005). A refinement of the Sharpe ratio and information ratio. *Journal of Asset Management, 5*(6), pp. 423-427.

León, Á. y Ñíguez, T.-M. (2020). Modeling asset returns under time-varying seminonparametric distributions. *Journal of Banking and Finance,* 118, 105870.

León, Á. y Ñiguez, T.-M. (2022). Polynomial adjusted Student-t densities for modeling asset returns. *European Journal of Finance, 28*(9), pp. 907-929.

Rodríguez, M. J. y Ruiz, E. (2012). GARCH models with leverage effect: differences and similarities. *Journal of Financial Econometrics,* 10, pp. 637-668.

Schwert, G. W. (1989). Why does stock market volatility change over time? *Journal of Finance,* 45, pp. 1129-1155.

Sharpe, W. F. (1966). Mutual Fund Performance. *The Journal of Business, 39*(1), pp. 119-138.

Taylor, S. J. (1986). *Modelling Financial Time Series.* Wiley.

Vacca, G., Zoia, M. G. y Bagnato, L. (2022). Forecasting in GARCH models with polynomially modified innovations. *International Journal of Forecasting,* 38, pp. 117-141.

Watanabe, Y. (2006). Is Sharpe ratio still effective? *Journal of Performance Measurement, 11*(1), p. 55.

Zakoian, J. M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control,* 18, pp. 931-955.

APÉNDICE 1

Acciones seleccionadas del índice Russell l000 durante el período 8/03/200l hasta 5/07/2023

1	AFLAC	46	BAKER HUGHES A	91	CINCINNATI FINL.
2	AGCO	47	BALL	92	CIRRUS LOGIC
3	AES	48	BAXTER INTL.	93	CISCO SYSTEMS
4	ABBOTT LABORATORIES	49	BECTON DICKINSON	94	CINTAS
5	ACTIVISION BLIZZARD	50	VERIZON COMMUNICATIONS	95	CLEVELAND CLIFFS
6	ADOBE (NAS)	51	W R BERKLEY	96	CLOROX
7	ADVANCED MICRO DEVICES	52	BERKSHIRE HATHAWAY 'B'	97	COCA COLA
8	AFFILIATED MANAGERS	53	BEST BUY	98	COGNEX
9	AIR PRDS.& CHEMS.	54	BIO-RAD LABORATORIES 'A'	99	COLGATE-PALM.
10	ALASKA AIR GROUP	55	H&R BLOCK	100	COMERICA
11	ALBEMARLE	56	BOEING	101	COMMERCE BCSH.
12	HONEYWELL INTL.	57	BORGWARNER	102	NNN REIT
13	HANOVER INSURANCE GROUP	58	BOSTON BEER 'A'	103	DXC TECHNOLOGY
14	ALLSTATE ORD SHS	59	BOSTON SCIENTIFIC	104	CONAGRA BRANDS
15	HOWMET AEROSPACE	60	BOYD GAMING	105	CONSOLIDATED EDISON
16	HESS	61	BRISTOL MYERS SQUIBB	106	COOPER COS.
17	U HAUL HOLDING	62	BROWN-FORMAN 'A'	107	MOLSON COORS BEVERAGE CO. B
18	AMEREN	63	BROWN-FORMAN 'B'	108	COPART
19	AMER.ELEC.PWR.	64	BRUNSWICK	109	INGREDION
20	AMERICAN EXPRESS	65	SIRIUS XM HOLDINGS	110	CORNING
21	AMERICAN FINL.GP.OHIO	66	CIGNA	111	COUSINS PROPS.
22	AMERICAN INTL.GP.	67	CMS ENERGY	112	CREDIT ACCEP.
23	AMERISOURCEBERGEN	68	CNA FINANCIAL	113	WOLFSPEED
24	AMETEK	69	CSX	114	CROWN HDG.
25	AMGEN	70	CVS HEALTH	115	CULLEN FO.BANKERS
26	AMPHENOL 'A'	71	COTERRA ENERGY	116	CUMMINS
27	ANALOG DEVICES	72	CACI INTERNATIONAL 'A'	117	CURTISS WRIGHT
28	AON CLASS A	73	CADENCE DESIGN SYS.	118	D R HORTON
29	APA	74	CAMDEN PROPERTY TST.	119	DTE ENERGY
30	APPLE	75	CAMPBELL SOUP	120	DANAHER
31	APPLIED MATS.	76	CONSTELLATION BRANDS'A'	121	DARDEN RESTAURANTS
32	APTARGROUP	77	CAPITAL ONE FINL.	122	DARLING INGREDIENTS
33	ARCHER DANIELS MIDLAND	78	CARDINAL HEALTH	123	TARGET
34	ARROW ELECTRONICS	79	CARLISLE COS.	124	DECKERS OUTDOOR
35	ASHLAND	80	CARNIVAL	125	DEERE
36	ASPEN TECHNOLOGY	81	CASEY'S GENERAL STORES	126	DENTSPLY SIRONA
37	ATMOS ENERGY	82	CATERPILLAR	127	WALT DISNEY
38	AUTODESK	83	AVIS BUDGET GROUP	128	DOLLAR TREE
39	AUTOMATIC DATA PROC.	84	EAGLE MATERIALS	129	DOMINION ENERGY
40	AUTOZONE	85	JP MORGAN CHASE & CO.	130	DONALDSON CO.
41	AVALONBAY COMMNS.	86	CHEMED	131	DOVER
42	AVERY DENNISON	87	CHEVRON	132	DUKE ENERGY
43	AVNET	88	CHURCH & DWIGHT CO.	133	EASTMAN CHEMICAL
44	TRUIST FINANCIAL	89	CHURCHILL DOWNS	134	EATON
45	BOK FINL.	90	CIENA	135	ECOLAB

Acciones seleccionadas del índice Russell l000 durante el período 8/03/200l hasta 5/07/2023

(continuación)

136	EDISON INTL.	180	HARTFORD FINL.SVS.GP.	224	KOHL'S
137	ELECTRONIC ARTS	181	HASBRO	225	KROGER
138	EMERSON ELECTRIC	182	HAWAIIAN ELECTRIC INDS.	226	LAB. CORP.OF AM. HDG.
139	EOG RES.	183	HEALTHPEAK PROPERTIES	227	LAM RESEARCH
140	ENTERGY	184	WELLTOWER	228	LANDSTAR SYSTEM
141	EQUIFAX	185	JACK HENRY AND ASSOC.	229	LATTICE SEMICONDUCTOR
142	EQT	186	HERSHEY	230	ESTEE LAUDER COS.'A'
143	EQUITY RESD.TST.PROPS. SHBI	187	HP	231	LEGGETT&PLATT
144	ERIE INDEMNITY 'A'	188	HEXCEL	232	LENNAR 'A'
145	NEWMARKET	189	HIGHWOODS PROPERTIES	233	JEFFERIES FINANC. GROUP
146	EVEREST RE GP.	190	HF SINCLAIR	234	ELI LILLY
147	EXPEDITOR INTL.OF WASH.	191	HOLOGIC	235	BATH AND BODY WORKS
148	EXXON MOBIL	192	HOME DEPOT	236	LINCOLN ELECTRIC HDG.
149	FMC	193	HORMEL FOODS	237	LINCOLN NATIONAL
150	FNB	194	HOST HOTELS & RESORTS R	238	LITTELFUSE
151	NEXTERA ENERGY	195	HUBBELL	239	LOCKHEED MARTIN
152	FAIR ISAAC	196	HUMANA	240	LOEWS
153	FASTENAL	197	HUNT JB TRANSPORT SVS.	241	RANGE RES.
154	FEDEX	198	HUNTINGTON BCSH.	242	LOUISIANA PACIFIC
155	FEDERAL REALTY INV.TST.	199	ICU MEDICAL	243	LOWE'S COMPANIES
156	MACY'S	200	MOSAIC	244	M&T BANK
157	FIFTH THIRD BANCORP	201	ITT	245	MDU RESOURCES GROUP
158	FIRST CTZN.BCSH.A	202	IDACORP	246	MGIC INVESTMENT
159	FIRST INDL.REALTY TST.	203	IDEX	247	MGM RESORTS INTL.
160	FIRST HORIZON	204	IDEXX LABORATORIES	248	MSC INDL.DIRECT 'A'
161	FISERV	205	ILLINOIS TOOL WORKS	249	MANPOWERGROUP
162	FIRSTENERGY	206	TRANE TECHNOLOGIES	250	EQUITY LIFESTYLE PROPS.
163	FLOWSERVE	207	INTEGRA LFSC.HDG.	251	MARKEL GROUP
164	FRANKLIN RESOURCES	208	INTEL	252	MARSH & MCLENNAN
165	FREEPORT-MCMORAN	209	INTERNATIONAL BUS.MCHS.	253	MARRIOTT INTL.'A'
166	WHITE MOUNTAINS IN.GP.	210	INTL.FLAVORS & FRAG.	254	MARTIN MRTA.MATS.
167	ARTHUR J GALLAGHER	211	INTERNATIONAL PAPER	255	MASCO
168	GARTNER 'A'	212	INTERPUBLIC GROUP	256	MASTEC
169	GENERAL DYNAMICS	213	INTUIT	257	MATTEL
170	GENERAL ELECTRIC	214	IONIS PHARMACEUTICALS	258	MCCORMICK & COMPANY NV.
171	GENERAL MILLS	215	JABIL	259	MCDONALDS
172	GENTEX	216	JACOBS SOLUTIONS	260	S&P GLOBAL
173	GENUINE PARTS	217	JOHNSON & JOHNSON	261	MCKESSON
174	GILEAD SCIENCES	218	KLA	262	MEDTRONIC
175	GRACO	219	KELLOGG	263	BANK OF NEW YORK MELLON
176	WW GRAINGER	220	KEYCORP	264	MICROSOFT
177	HALLIBURTON	221	KIMBERLY-CLARK	265	MICROCHIP TECH.
178	HARLEY-DAVIDSON	222	KIMCO REALTY	266	MICRON TECHNOLOGY
179	L3HARRIS TECHNOLOGIES	223	KIRBY	267	MID-AMER.APT COMMUNITIES

Acciones seleccionadas del índice Russell l000 durante el período 8/03/200l hasta 5/07/2023

(continuación)

268	MSA SAFETY	312	PENTAIR	356	SCOTTS MIRACLE-GRO
269	3M	313	PEPSICO	357	SEABOARD
270	MOHAWK INDUSTRIES	314	PERRIGO	358	SEALED AIR
271	MORGAN STANLEY	315	PFIZER	359	SEMPRA
272	MOTOROLA SOLUTIONS	316	ESSENTIAL UTILITIES	360	SERVICE CORP.INTL.
273	VIATRIS	317	ALTRIA GROUP	361	SHERWIN-WILLIAMS
274	NCR	318	CONOCOPHILLIPS	362	SIMON PROPERTY GROUP
275	NVR	319	PVH	363	SMITH (AO)
276	NATIONAL FUEL GAS	320	PINNACLE WEST CAP.	364	SNAP-ON
277	NATIONAL INSTS.	321	PIONEER NTRL.RES.	365	SONOCO PRODUCTS
278	NETAPP	322	POLARIS INDUSTRIES	366	SOUTHERN
279	NEUROCRINE BIOSCIENCES	323	POPULAR	367	SOUTHWEST AIRLINES
280	NEW YORK TIMES 'A'	324	T ROWE PRICE GROUP	368	SOUTHWESTERN ENERGY
281	NEWMONT	325	PROCTER & GAMBLE	369	LIFE STORAGE
282	NIKE 'B'	326	PROGRESSIVE OHIO	370	STANLEY BLACK & DECKER
283	NORDSON	327	PUB.SER.ENTER.GP.	371	US BANCORP
284	NORDSTROM	328	PUBLIC STORAGE	372	STARBUCKS
285	NORFOLK SOUTHERN	329	PULTEGROUP	373	STATE STREET
286	EVERSOURCE ENERGY	330	QUANTA SERVICES	374	STIFEL FINANCIAL
287	XCEL ENERGY	331	NEW YORK COMMUNITY BANC.	375	STRYKER
288	NORTHERN TRUST	332	RPM INTERNATIONAL	376	SUN COMMUNITIES
289	NORTHROP GRUMMAN	333	RAYMOND JAMES FINL.	377	GEN DIGITAL
290	WELLS FARGO & CO	334	RAYONIER	378	SYNOPSYS
291	NUCOR	335	REALTY INCOME	379	SYNOVUS FINANCIAL
292	OGE ENERGY	336	REGAL REXNORD	380	SYSCO
293	OCCIDENTAL PTL.	337	REGENERON PHARMS.	381	TJX
294	OLD DOMINION FGT.LINES	338	REGENCY CENTERS	382	BIO-TECHNE
295	OLD REPUBLIC INTL.	339	RELIANCE STEEL AND ALMN	383	TELEFLEX
296	OLIN	340	REPLIGEN	384	TENET HEALTHCARE
297	OMNICOM GROUP	341	RESMED	385	TERADYNE (XSC)
298	OMEGA HLTHCR.INVRS.	342	ARCH CAP.GP.	386	TETRA TECH
299	ONEOK	343	ROBERT HALF INTERNAT.	387	TEXAS INSTRUMENTS
300	ORACLE	344	ROCKWELL AUTOMATION	388	TEXTRON
301	O REILLY AUTOMOTIVE	345	ROLLINS	389	THERMO FISHER SCIENT.
302	EXELON	346	ROPER TECHNOLOGIES	390	THOR INDUSTRIES
303	PG&E	347	ROSS STORES	391	TIMKEN
304	PNC FINL.SVS.GP.	348	ROYAL GOLD	392	TOLL BROTHERS
305	PPL	349	RYDER SYSTEM	393	GLOBE LIFE
306	PPG INDUSTRIES	350	AT&T	394	TORO
307	PACCAR	351	SEI INVESTMENTS	395	DAVITA
308	BANK OF HAWAII	352	TRAVELERS COS.	396	TRACTOR SUPPLY
309	PTC	353	HENRY SCHEIN	397	CITIGROUP
310	PARKER-HANNIFIN	354	SCHLUMBERGER	398	WENDY'S CLASS A
311	PAYCHEX	355	CHARLES SCHWAB	399	YUM! BRANDS

Acciones seleccionadas del índice Russell l000 durante el período 8/03/200l hasta 5/07/2023

(continuación)

400	TRIMBLE	444	MOODY'S	488	COGNIZANT TECH.SLTN.'A'
401	TYSON FOODS 'A'	445	QUEST DIAGNOSTICS	489	CROWN CASTLE
402	UGI	446	STEEL DYNAMICS	490	EAST WEST BANCORP
403	MARATHON OIL	447	STERICYCLE	491	EBAY
404	WASTE MANAGEMENT	448	SOUTHERN COPPER	492	GOLDMAN SACHS GP.
405	IAC	449	FACTSET RESEARCH SYS.	493	LITHIA MOTORS
406	UNITED STATES STEEL	450	PROLOGIS REIT	494	MKS INSTRUMENTS
407	UNION PACIFIC	451	UNITED RENTALS	495	MERCURY SYSTEMS
408	UDR	452	CENTERPOINT EN.	496	NVIDIA
409	UNITEDHEALTH GROUP	453	NEWELL BRANDS (XSC)	497	BOOKING HOLDINGS
410	RAYTHEON TECHNOLOGIES	454	BOSTON PROPERTIES	498	REPUBLIC SVS.'A'
411	KEMPER	455	ESSEX PROPERTY TST.	499	POOL
412	UNIVERSAL HEALTH SVS.'B'	456	ALEXANDRIA RLST.EQTIES.	500	TREX
413	V F	457	CH ROBINSON WWD.	501	WESCO INTL.
414	VAIL RESORTS	458	CHOICE HOTELS INTL.	502	UNUM GROUP
415	VALERO ENERGY	459	COLUMBIA BKG.SYS.	503	COSTAR GP.
416	VALMONT INDUSTRIES	460	COLUMBIA SPORTSWEAR	504	COSTCO WHOLESALE
417	VENTAS	461	EASTGROUP PROPS.	505	DEVON ENERGY
418	VERISIGN	462	DISH NETWORK 'A'	506	REVVITY
419	VERTEX PHARMS.	463	EPR PROPERTIES	507	TAKE TWO INTACT.SFTW.
420	VORNADO REALTY TRUST	464	HEICO	508	TELEDYNE TECHS.
421	VULCAN MATERIALS	465	KILROY REALTY	509	UNITED PARCEL SER.'B'
422	WALMART	466	MANHATTAN ASSOCS.	510	JUNIPER NETWORKS
423	WALGREENS BOOTS ALLIANCE	467	METTLER TOLEDO INTL.	511	LENNOX INTL.
424	WATERS	468	PEGASYSTEMS	512	AMDOCS
425	WATSCO	469	IRON MOUNTAIN	513	WORLD WRESTLING ENTM'A'
426	WEBSTER FINANCIAL	470	SL GREEN REALTY	514	RENAISSANCERE HDG.
427	WESTERN DIGITAL	471	SILGAN HOLDINGS	515	RB GLOBAL (NYS)
428	EVERGY	472	PENSKE AUTOMOTIVE GP.	516	WP CAREY
429	WABTEC	473	WEST PHARM.SVS.	517	HEICO NEW 'A'
430	WEYERHAEUSER	474	WOODWARD	518	WINTRUST FINANCIAL
431	WHIRLPOOL	475	NISOURCE	519	ANNALY CAPITAL MAN.
432	WILLIAMS	476	SKYWORKS SOLUTIONS	520	EDWARDS LIFESCIENCES
433	WILLIAMS-SONOMA	477	BANK OF AMERICA	521	AGILENT TECHS.
434	WEC ENERGY GROUP	478	AMAZON.COM	522	VIASAT
435	SPECTRUM BRANDS HOLDINGS	479	AUTONATION	523	BIOMARIN PHARM.
436	ZEBRA TECHNOLOGIES 'A'	480	BROWN & BROWN	524	BLACKROCK
437	ZIONS BANCORP.	481	JONES LANG LASALLE	525	EXELIXIS
438	CHUBB	482	LAMAR ADVERTISING 'A'	526	F5
439	QIAGEN	483	NOV	527	COHERENT
440	JOHNSON CONTROLS INTL.	484	RALPH LAUREN CL.A	528	METLIFE
441	ROYAL CARIBBEAN GROUP	485	ALLIANT ENERGY (XSC)	529	PACKAGING CORP.OF AM.
442	AMERICAN TOWER	486	LUMEN TECHNOLOGIES	530	PENN ENTERTAINMENT
443	AZENTA	487	TYLER TECHNOLOGIES	531	PLUG POWER

Acciones seleccionadas del índice Russell l000 durante el período 8/03/200l hasta 5/07/2023

(continuación)

532	SBA COMMS.	548	EXACT SCIS.	563	QUIDELORTHO
533	ON SEMICONDUCTOR	549	FTI CONSULTING	564	SSR MINING
534	UNITED THERAPEUTICS	550	ILLUMINA	565	TEXAS PACIFIC LAND TRUST
535	UNIVERSAL DISPLAY	551	INTUITIVE SURGICAL	566	J M SMUCKER
536	FORD MOTOR	552	PROSPERITY BCSH.	567	COMCAST A
537	PDC ENERGY	553	AMEDISYS	568	GRAPHIC PACKAGING HLDG.
538	BANK OZK	554	CLEAN HARBORS	569	REGIONS FINL.NEW
539	SLM	555	EURONET WWD.	570	PARAMOUNT GLOBAL B
540	GLOBAL PAYMENTS	556	EQUINIX REIT	571	PARAMOUNT GLOBAL A
541	TAPESTRY	557	PACWEST BANCORP	572	REINSURANCE GROUP OF AM.
542	ALIGN TECHNOLOGY	558	GARMIN	573	MERCK & COMPANY
543	ANSYS	559	MONSTER BEVERAGE	574	QORVO
544	BRUKER	560	MIDDLEBY	575	STERIS
545	CHAS.RVR.LABS.INTL.	561	MARVELL TECHNOLOGY	576	KNIGHT-SWIFT TRSP.HDG. 'A'
546	CARMAX	562	PINNACLE FINANCIAL PTNS.	577	LINDE
547	ENTEGRIS				

Nota: Se han seleccionado inicialmente un total de 577 acciones. A partir de ese conjunto inicial se realiza el filtrado que se detalla en la sección tercera, siendo 417 el número final de activos con el que se lleva a cabo el análisis.

Fuente: Elaboración propia.

APÉNDICE 2

Activos seleccionados de acuerdo al indicador considerado y frecuencia con la que forman parte de la cartera

Ratio de Sharpe (SR)				*Ratio de Sharpe modificado (mSR)*				*Ratio Asimetría-Curtosis (SKR)*			
Muestral		*Implícito*		*Muestral*		*Implícito*		*Muestral*		*Implícito*	
Activo	Frecuencia	Activo	Frecuencia	Activo	Frecuencia	Activo	Frecuencia	Activo	Frecuencia	Activo	Frecuencia
30	0 .81	30	1	76	1	451	0 .61	25	0 .57	340	0 .91
487	0 .74	478	0 .90	550	0 .74	190	0 .50	201	0 .57	89	0 .57
88	0 .58	451	0 .50	409	0 .67	415	0 .46	54	0 .48	229	0 .57
478	0 .41	108	0 .47	340	0 .45	533	0 .37	229	0 .45	523	0 .43
301	0 .37	445	0 .29	468	0 .41	476	0 .36	270	0 .37	60	0 .36
40	0 .21	467	0 .26	81	0 .37	448	0 .34	421	0 .37	414	0 .27
204	0 .20	473	0 .24	259	0 .20	500	0 .34	396	0 .29	54	0 .24
560	0 .20	487	0 .21	298	0 .20	266	0 .32	478	0 .28	214	0 .19
294	0 .16	124	0 .20	471	0 .16	92	0 .30	400	0 .20	52	0 .16
347	0 .16	193	0 .12	478	0 .16	496	0 .25	137	0 .19	421	0 .15
473	0 .16	340	0 .08	317	0 .14	30	0 .18	241	0 .19	372	0 .13
409	0 .12	560	0 .08	40	0 .08	356	0 .16	368	0 .14	396	0 .13
467	0 .12	318	0 .07	566	0 .07	503	0 .11	192	0 .13	315	0 .11
317	0 .11	499	0 .07	152	0 .06	124	0 .09	230	0 .13	26	0 .10
396	0 .11	239	0 .06	108	0 .05	537	0 .09	468	0 .12	524	0 .10
94	0 .08	496	0 .06	434	0 .04	95	0 .08	473	0 .11	535	0 .10
108	0 .07	23	0 .05	548	0 .04	112	0 .08	112	0 .10	270	0 .09
174	0 .07	396	0 .04	86	0 .03	10	0 .07	5	0 .08	112	0 .07
405	0 .07	41	0 .03	88	0 .02	108	0 .05	476	0 .06	273	0 .06
504	0 .06	116	0 .03	301	0 .02	109	0 .04	51	0 .05	555	0 .06
538	0 .06	562	0 .03	380	0 .02	227	0 .03	142	0 .02	478	0 .04
193	0 .04	252	0 .02	39	0 .01	488	0 .03	471	0 .02	254	0 .03
259	0 .03	260	0 .02	201	0 .01	348	0 .02	547	0 .02	275	0 .03
120	0 .02	317	0 .02	251	0 .01	26	0 .01	234	0 .01	364	0 .02
314	0 .01	353	0 .02	368	0 .01	93	0 .01	343	0 .01	562	0 .02
361	0 .01	503	0 .02	428	0 .01	191	0 .01	372	0 .01	93	0 .01
488	0 .01	504	0 .02	458	0 .01	241	0 .01	437	0 .01	357	0 .01
496	0 .01	540	0 .02	499	0 .01	261	0 .01	472	0 .01	413	0 .01
		93	0 .01			427	0 .01	535	0 .01	481	0 .01
		168	0 .01			495	0 .01			491	0 .01
		184	0 .01			508	0 .01			496	0 .01
		294	0 .01			547	0 .01				
		362	0 .01			548	0 .01				
		448	0 .01			550	0 .01				
		480	0 .01			562	0 .01				

Nota: La acción correspondiente a cada número de la columna Activo puede verse en el Apéndice 1. El número de acciones diferentes seleccionadas según el indicador en los 100 períodos correspondientes donde se cambia la composición de la cartera son: 28 (SR muestral), 35 (SR implícito), 28 (mSR muestral), 35 (mSR implícito), 29 (SKR muestral), 31 (SKR implícito).

Fuente: Elaboración propia.

APÉNDICE 3

Activos comunes en grupos de dos carteras cada una asociada a un indicador diferente

Indicador 1	*Indicador 2*	*Total activos (Indicador 2)*	*Total activos (Indicador 2)*	*N. de activos comunes*	*Activos comunes*
SR muestral	SR implícito	28	35	13	30 108 193 294 317 396 467 473 478 487 496 504 560
SR muestral	mSR muestral	28	28	8	40 88 108 259 301 317 409 478
SR muestral	mSR implícito	28	35	4	30 108 488 496
SR muestral	SKR muestral	28	29	3	396 473 478
SR muestral	SKR implícito	28	31	3	396 478 496
SR implícito	mSR muestral	35	28	5	108 317 340 478 499
SR implícito	mSR implícito	35	35	9	30 93 108 124 448 451 496 503 562
SR implícito	SKR muestral	35	29	3	396 473 478
SR implícito	SKR implícito	35	31	6	93 340 396 478 496 562
mSR muestral	mSR implícito	28	35	3	108 548 550
mSR muestral	SKR muestral	28	29	5	201 368 468 471 478
mSR muestral	SKR implícito	28	31	2	340 478
mSR implícito	SKR muestral	35	29	4	112 241 476 547
mSR implícito	SKR implícito	35	31	5	26 93 112 496 562
SKR muestral	SKR implícito	29	31	9	54 112 229 270 372 396 421 478 535

Fuente: Elaboración propia.

CAPÍTULO VII

Clasificación de conjuntos de ofertas de electricidad en el mercado diario español

Jorge Arias Martí*
Andrés M. Alonso Fernández

El mercado eléctrico en España permite a los productores de electricidad ofrecer bloques de energía a diferentes precios, generalmente relacionados con sus costes marginales, en momentos concretos del día. El operador del sistema reúne las ofertas (bloques de energía) y sus correspondientes precios de todos los participantes para formar la curva de oferta con la que se obtendrá el precio marginal de cada hora. En este trabajo se estudian los conjuntos de oferta mediante la distancia de Hausdorff y se realiza la clasificación no supervisada de estos conjuntos. Adicionalmente, se caracterizan los grupos obtenidos mediante variables de producción de energía por las distintas tecnologías y variables temporales como hora, día de la semana y mes.

Palabras clave: clasificación no supervisada, distancia de Hausdorff, conjuntos de oferta.

* Este trabajo ha sido parcialmente financiado por la Agencia Estatal de Investigación (PID2019-108311GB-I00 / AEI / 10.13039/501100011033 / PID2022-138114NB-I00). Los autores agradecen los comentarios y mejoras sugeridas por el equipo editorial del libro, Daniel Peña, Pilar Poncela y Eva Senra.

1. INTRODUCCIÓN

El precio de la electricidad es un tema de gran interés actual. Casi a diario aparecen noticias relacionadas con los precios y la influencia de las diferentes tecnologías de producción. No hay duda de que los precios de la electricidad tienen un profundo impacto tanto en la economía nacional como en la empresarial. Por ejemplo, en Khobai *et al.* (2017) se encontró que "un aumento del 4 % en los precios de la electricidad provoca que el crecimiento económico disminuya en un 0,036 % en Sudáfrica". Por supuesto, ese impacto es común a cualquier economía.

Agosti *et al.* (2007) ofrecen información sobre cómo se fijan estos precios en España. Los autores explican que el componente más importante es el mercado diario, en el que se negocia la mayor parte de la energía para cada hora del día siguiente. En este mercado los productores de energía hacen ofertas de venta mientras que los consumidores hacen ofertas de compra, especificando tanto cantidad como precio. Para cada hora del día siguiente se obtienen dos curvas:

- Curva de oferta: se obtienen mediante la agregación de las cantidades de energía ofrecidas por los productores ordenadas de forma ascendente según su precio.
- Curva de demanda: se obtienen mediante la agregación de las cantidades de energía demandadas por los consumidores ordenadas en orden descendente según su precio.

El resultado son curvas formadas por puntos [p, q], donde p hace referencia al precio y q representa la cantidad de energía que se puede comprar al precio p. Al considerar simultáneamente ambas curvas, el precio de la electricidad (precio de casación) para esa hora viene dado por el punto donde se cruzan ambas curvas (Aggarwal *et al.,* 2009). A modo de ilustración, en la figura 1 el precio para la fecha y hora indicadas rondaría los 52 euros/MWh. Los compradores con demandas ubicadas a la izquierda del punto de cruce no podrán adquirir energía en ese período (pues su oferta está por debajo del precio de mercado), y la misma situación ocurre con los productores. Si hicieron una oferta ubicada a la derecha del punto de cruce no producirán energía en ese periodo (porque la ofrecen a un precio superior al precio de mercado).

En este trabajo se estudiarán los conjuntos horarios de oferta del mercado diario entre 2017 y 2021 mediante métodos de clasificación no supervisada (*clustering*). Se intentará descubrir, por ejemplo, cuáles son los periodos en los que se ofrece menor cantidad de energía, y si corresponden a horarios nocturnos o, por ejemplo, si hay una diferencia notable entre días laborables y festivos. También se tratará de descubrir cómo cambia la estructura de generación en función de la cantidad total de energía ofrecida. El conjunto de los datos a analizar se clasifica en la categoría de datos masivos (*big data*) por dos motivos: i) su volumen, pues tenemos 43.800 conjuntos de ofertas, y ii) su variabilidad o complejidad, pues cada conjunto de ofertas tiene dimensiones diferentes.

El resto del capítulo está organizado en seis secciones. La sección segunda presenta los elementos del mercado eléctrico necesarios para entender la obtención de las curvas de oferta;

Figura 1.

Ejemplo del proceso de obtención del precio de casación

Fuente: Elaboración propia a partir de datos disponibles en https://www.omie.es/es/market-results/daily/daily-market/aggragate-suply-curves

la sección tercera presenta una medida de disimilitud entre conjuntos (distancia de Hausdorff) y su procedimiento de cálculo; en la sección cuarta se recogen los métodos de agrupamiento que se han utilizado; la sección quinta desarrolla la metodología para caracterizar los diferentes clústers; en la sección sexta se presentan los resultados obtenidos, y finalmente en la sección séptima se presentan las conclusiones y posibles extensiones.

2. OBTENCIÓN DE LAS OFERTAS Y CURVAS DE OFERTA EN EL MERCADO ELÉCTRICO ESPAÑOL

2.1. Legislación del mercado eléctrico español

En el apartado anterior se ilustró el procedimiento para calcular los precios de la electricidad que se obtienen como consecuencia de la legislación de la Comisión Nacional de Mercados y la Competencia. En BOE/CNMC (2021) se establece cómo funciona el mercado eléctrico en España. En particular, existen algunas reglas de funcionamiento que se mencionan a continuación para una mejor comprensión del mercado de suministro eléctrico. La Regla 1 establece que en el mercado diario las operaciones de compra y venta se realizan para el día siguiente para cada una de las 24 horas o períodos naturales (este número también puede ser 23 ó 25 en los días correspondientes a días de cambio de hora oficial).

Además, la Regla 30.1 dice que el precio de cada período será el resultante de la compensación realizada por el algoritmo Euphemia. Este algoritmo utiliza curvas agregadas (Regla 30.2) que se obtienen de la siguiente manera:

- Curvas de oferta: Se obtienen sumando las cantidades de energía ofertadas para la venta en orden ascendente por el precio de las mismas (Regla 30.2.1).

- Curvas de demanda: Se obtienen sumando las cantidades de energía ofrecidas para comprar en orden decreciente por el precio de las mismas (Regla 30.2.2).

Como se indicó anteriormente, se obtienen como resultado las curvas formadas por los puntos [*p, q*] y que se han ilustrado en la figura 1.

En este trabajo, se obtienen todas las curvas de oferta en el periodo de 2017 a 2021 y, para ello, se necesita la información de las ofertas para cada hora. Esta información la proporciona el operador del mercado eléctrico de la península ibérica (OMIE) y se recoge en dos tipos de ficheros, obtenidos de OMIE (b) y OMIE (a). El tipo de información contenida en cada tipo de archivo se muestra en OMIE (c) y se resume a continuación:

- Los archivos de tipo cab_aaaammdd.1, contienen el número de identificación (código) de cada oferta y la clase de la misma (compra o venta).
- Los archivos de tipo det_aaaammdd.1, contienen información detallada de las ofertas. Son de interés el número o código de identificación, que vuelve a aparecer, la cantidad de energía ofrecida y el precio de la misma.

Figura 2.

Ejemplo de un archivo cab_.

4271539	2ICUVD04IBCUR	RE	4	VNO	0.000	0.000	0.0	0.0	0.000	0.000	2500.0	0.0	0.0	120170112112326
4271540	2ICUVD05IBCUR	RE	5	VNO	0.000	0.000	0.0	0.0	0.000	0.000	1000.0	0.0	0.0	120170112112326
4271541	2ICUVD06IBCUR	RE	6	VNO	0.000	0.000	0.0	0.0	0.000	0.000	1000.0	0.0	0.0	120170112112326
4271542	2ICUVD08IBCUR	RE	8	VNO	0.000	0.000	0.0	0.0	0.000	0.000	1000.0	0.0	0.0	120170112112326

Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

Figura 3.

Ejemplo de un archivo det_.

1717319	311	1	0.000	0.000	1.0SS
1717319	312	1	0.000	0.000	1.0SS
1717319	313	1	0.000	0.000	1.0SS
1717319	314	1	0.000	0.000	1.0SS
1717319	315	1	0.000	0.000	1.0SS
1717319	316	1	0.000	0.000	1.0SS
1717319	317	1	0.000	0.000	1.0SS
1717319	318	1	0.000	0.000	1.0SS
1717319	319	1	0.000	0.000	1.0SS
1717319	320	1	0.000	0.000	1.0SS
1717319	321	1	0.000	0.000	1.0SS
1717319	322	1	0.000	0.000	1.0SS

Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

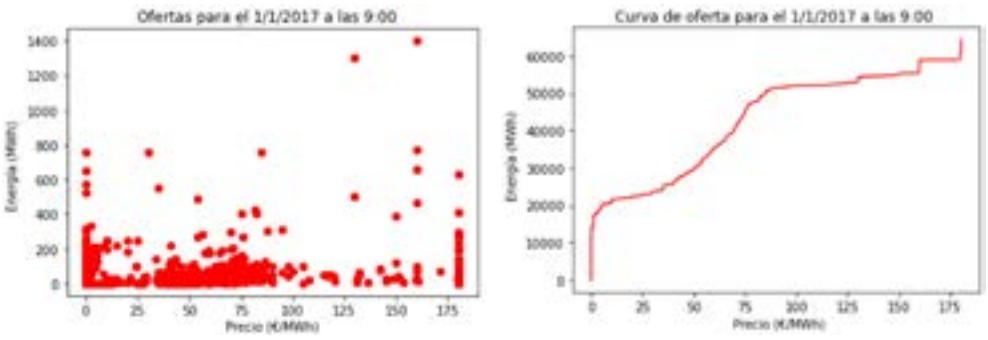
Las figuras 2 y 3 muestran un ejemplo de cada tipo de archivo.

Estos datos se ofrecen en conjuntos mensuales en archivos .zip, por lo que es necesario descargarlos y descomprimirlos.

Una vez descargados los archivos solo es necesario tomar el código y el tipo de oferta del archivo cab_aaaammdd.1 y cruzarlos con las variables del archivo det_aaaammdd.1. Iterando este procedimiento en los 1.826 archivos se obtienen las ofertas para cada periodo de cada día desde el 1 de enero de 2017 al 31 de diciembre de 2021. A partir de estas ofertas se obtienen las curvas de oferta clasificando las ofertas en orden creciente por su precio y haciendo la suma acumulada de las cantidades de energía (Regla 30.2.1). La figura 4 muestra un ejemplo de esta transformación, la cual es útil porque permite distinguir entre ofertas con el mismo precio y cantidad de energía. Sin embargo, se debe tener en cuenta que dos curvas diferentes no necesariamente contienen el mismo número de puntos. En un mismo escalón de una curva de oferta puede haber más de un punto y esto es debido a que existen varias ofertas al mismo precio.

Figura 4.

Transformación del conjunto de ofertas (a) en su curva de oferta correspondiente (b)



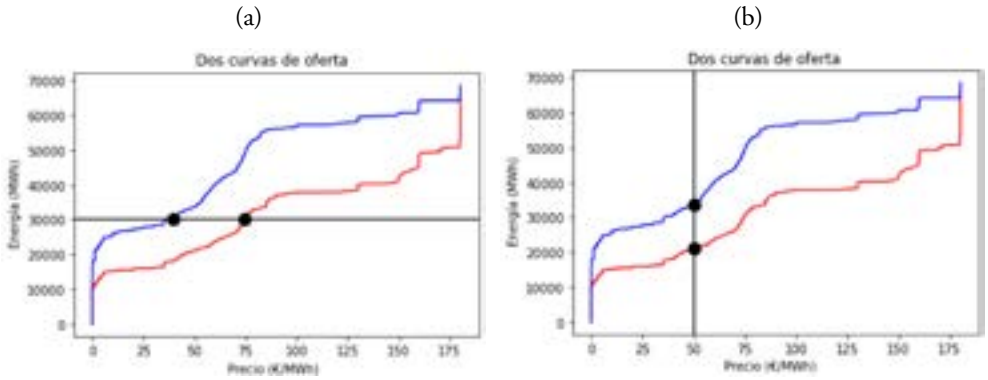
Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

Una vez obtenidas, se pueden comparar las curvas de oferta. La figura 5 muestra dos ejemplos de curvas. Para una determinada cantidad de energía (a) la curva en rojo ofrece esa cantidad a un precio más alto, por lo que la curva en azul es más interesante desde la perspectiva del comprador. Para un precio determinado (b), la curva en azul hace referencia a una mayor cantidad de energía ofertada.

Ahora que se tienen las curvas, el siguiente paso es normalizarlas y seleccionar una medida de disimilitud entre conjuntos de puntos para poder realizar un método de agrupamiento. En este estudio se ha elegido la distancia de Hausdorff, que se presentará en la siguiente sección.

Figura 5.

Ejemplo de dos curvas de oferta



Nota: La figura (a) muestra los precios para una cantidad fija de energía, mientras que la figura (b) muestra las cantidades a un precio fijo.

Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

3. DISTANCIA DE HAUSDORFF

Dados dos conjuntos de puntos de datos A y B, para cada punto $x \in A$ e $y \in B$ se define la distancia de Hausdorff dirigida de la siguiente manera (Taha y Hanbury, 2015):

$$\widehat{\mathrm{H}}(A,B) = \max_{x\in A}\left\{\min_{y\in B}\{\|x,y\|\}\right\}, \qquad [1]$$

donde $||\cdot, \cdot||$ puede ser cualquier norma en el espacio de los datos. En este trabajo se utiliza la norma euclidiana.

Se debe tener en cuenta que $\widehat{H}(A, B) \neq \widehat{H}(B, A)$, por lo que no se usa la distancia de Hausdorff dirigida, sino la no dirigida, que se define de la siguiente manera (Taha y Hanbury, 2015):

$$H(A, B) = \max\{\widehat{H}(A, B), \widehat{H}(B, A)\}. \qquad [2]$$

En este trabajo, A y B se corresponden con los puntos en dos curvas de oferta diferentes. La idea original era utilizar la distancia de Hausdorff aplicada a las nubes de puntos de oferta como las que se muestran en la figura 4(a), pero éstas no son propiamente un conjunto porque puede haber ofertas repetidas, es decir, ofertas con el mismo precio y cantidad. Al pasar de las ofertas a las curvas de oferta, se elimina el inconveniente de las repeticiones.

Para obtener la distancia de Hausdorff, se debe seguir el algoritmo 1:

Algoritmo 1.

Algoritmo para el cálculo de la distancia de Hausdorff

Input: Dos conjuntos de puntos, A y B.

Output: Distancia de Hausdorff.

1. Para un punto en A se calcula la distancia euclídea a todos los puntos en B. Se selecciona la menor distancia.
2. Se repite el paso anterior para todos los puntos de A.
3. De todas las distancias menores obtenidas en el paso 1, se selecciona la mayor, $\widehat{H}(A, B)$.
4. Se repiten los tres pasos anteriores intercambiando A y B para obtener $\widehat{H}(B, A)$.
5. Finalmente, se obtiene máx $\{\widehat{H}(A, B), \widehat{H}(B, A)\}$.

Fuente: Elaboración propia.

3.1. Computación de las distancias de Hausdorff

Tomando n como el número total de curvas de oferta, el objetivo es construir una matriz de distancias $n \times n$ donde cada elemento (i, j) sea la distancia de Hausdorff entre la curva i y la curva j. En este caso, $n = 43.800$ correspondientes a las curvas desde el 1 de enero de 2017 al 31 de diciembre de 2021[1]. Esto implica el cálculo de casi un billón de distancias.

Entre varios métodos de cálculo disponibles en Python, se encuentra que el más rápido para calcular la distancia de Hausdorff es usar la función *directed_hausdorff* ubicada en el módulo scipy.spatial.distance para encontrar $\widehat{H}(A, B)$ y $\widehat{H}(B, A)$ y tomar el máximo de estas dos distancias. También se consideraron las implementaciones en el módulo cuspatial. Sin embargo, a pesar de que este método era el mejor en términos de velocidad, el cálculo de toda la matriz habría llevado alrededor de tres meses, por lo que se modificó el objetivo inicial y se consideró el cálculo de las siguientes matrices de distancia:

- La primera es la matriz correspondiente a todas las curvas de 2019. Se eligió este año porque es el último antes de la pandemia del COVID-19.

- Para calcular la segunda matriz se seleccionó una hora valle (5:00) y una hora pico (12:00) para cada día. La matriz contiene todas las distancias entre todas estas horas para todos los años.

Con estas dos matrices se realizaron cuatro análisis: un análisis completo para todas las curvas de oferta de 2019; un segundo considerando solo una hora pico y una hora valle para cada día, y otros dos resultantes del estudio de estos dos tipos de curvas por separado. En la sección sexta se muestran los resultados de los dos primeros análisis, mientras que los resultados obtenidos de los dos restantes se han omitido por disponibilidad de espacio. No obstante, están disponibles mediante solicitud a los autores.

[1] En https://www.omie.es no estaban disponibles los archivos cab y det para el 1 de noviembre de 2021.

4. CLASIFICACIÓN NO SUPERVISADA

Después de calcular la matriz de distancias, es posible proceder con los métodos de agrupamiento o clasificación no supervisada. Se han considerado los procedimientos de partición alrededor de medoides (PAM) y agrupación jerárquica aglomerativa.

En la descripción del algoritmo PAM se utiliza el trabajo de Park y Jun (2009). Además del método mostrado para elegir los mediodes iniciales, los autores también proponen otras técnicas para seleccionarlos, pero en este capítulo se utilizó el enfoque heurístico. Este algoritmo se muestra a continuación:

Algoritmo 2.

Algoritmo para la obtención de la partición alrededor de medoides

Input: Un conjunto de datos o una matriz de disimilitud, y el número de grupos deseados = K.

Output: Grupo de cada observación y medoides.

1. Seleccione medoides iniciales.
   (a) Calcule la matriz de disimilitud utilizando una distancia predefinida.
   (b) Calcule v_j para la observación j:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \qquad [3]$$

   (c) Ordene v_j en orden ascendente y seleccione las primeras K observaciones con el valor más bajo. Estos objetos serán los medoides iniciales.
   (d) Asigne cada observación a su medoide más cercano para tener grupos iniciales.
   (e) Calcule la suma de las distancias entre cada observación y su medoide correspondiente.
2. Reemplace cada medoide con la observación que minimice la distancia total a las otras observaciones en el grupo.
3. Asigne cada observación a su medoide más cercano.
4. Calcule la suma de distancias de todos los objetos a su medioide correspondiente. Si esta suma es igual a la encontrada antes, detenga el algoritmo. De lo contrario, regrese al paso 2.

Fuente: Park y Jun (2009).

Notése que los *medoides* resultantes del algoritmo anterior son observaciones de cada grupo cuya distancia global es mínima al resto de las observaciones en el grupo, por tanto, pueden ser considerados como los representantes de los grupos.

- Por otro lado, una descripción del agrupamiento jerárquico aglomerativo puede consultarse, por ejemplo, en Nielsen (2016). El procedimiento "comienza desde los datos individuales [...] y va fusionando (iterativamente) de dos en dos los subconjuntos más cercanos" hasta que todos los datos hayan sido agrupados. Un elemento determinante en el agrupamiento jerárquico es la distancia de enlace o "distancia entre dos subconjuntos" que se denotará por $\Delta(X_i, X_j)$. Cuando los subconjuntos están compuestos sólo por un elemento, esta distancia es igual a la distancia entre las dos observaciones, pero cuando alguno o ambos subconjuntos tienen más de una observación entonces

se pueden utilizar varias definiciones de la distancia de enlace. Las siguientes son las más usadas:

- Enlace simple:

$$\Delta\left(X_i X_j\right) = \min_{x_i \in X_i, x_j \in X_j} D\left(x_i, x_j\right). \qquad [4]$$

- Enlace completo:

$$\Delta\left(X_i X_j\right) = \min_{x_i \in X_i, x_j \in X_j} D\left(x_i, x_j\right). \qquad [5]$$

- Enlace promedio

$$\Delta\left(X_i X_j\right) = \frac{1}{\left|X_i\right|\left|X_j\right|} \sum_{x_i \in X_i} \sum_{x_j \in X_j} D\left(x_i, x_j\right). \qquad [6]$$

Debido a que son los más interpretables, en la sección sexta se muestran los resultados obtenidos por el enlace promedio.

4.1. Evaluación del desempeño de la clasificación no supervisada

Para evaluar la agrupación obtenida mediante un procedimiento de clasificación no supervisada se han tenido en cuenta dos tipos de métricas:

- La primera métrica que se utiliza es el estadístico Silueta, cuya expresión es la siguiente (Kaufman y Rousseeuw, 1990):

$$s(i) = \frac{b(i) - a(i)}{\max\left\{a(i), b(i)\right\}}, \qquad [7]$$

donde $a(i)$ es la disimilitud promedio del objeto i con todos los demás objetos que pertenecen a su mismo grupo, mientras que $b(i)$ es el mínimo de todas las disimilitudes promedio entre el objeto i y el resto de los grupos, de los cuales no forma parte. El valor de $s(i)$ está entre -1 y 1. Cuando es cercana a 1 significa que la clasificación del objeto es adecuada. Si es próxima a 0, no está claro que la observación i deba asignarse a ese grupo o al grupo más cercano del resto de los grupos, y finalmente, si $s(i)$ es negativa significa que el objeto ha sido clasificado erróneamente.

Como medida global, se utiliza la media de las $s(i)$.

- La segunda métrica que se utiliza es el índice de separación, el cual está definido en Akhanli y Hennig (2020). Primero se define K como el número total de grupos. Para un objeto dado i en el grupo C_k, donde $k = \{1, 2, ..., K\}$, se calcula $d_{k:i}$ que es la distancia mínima entre el objeto i y todas las observaciones que no forman parte del grupo C_k.

Se repite el proceso para todas las observaciones en ese grupo y se ordenan todas las distancias resultantes en orden creciente. Luego, se toma una fracción igual a $[pn_k]$ de los valores más bajos de ellos, donde se sugiere que $p = 0.1$, n_k es el tamaño del grupo C_k y $[pn_k]$ representa el mayor número entero menor o igual a pn_k. El índice de separación tiene la siguiente expresión:

$$SI = \frac{1}{\sum_{k=1}^{K} [pn_k]} \sum_{k=1}^{K} \sum_{i=1}^{[pn_k]} d_{k:i}. \qquad [8]$$

A mayor índice de separación, mejor es la agrupación obtenida (Akhanli y Hennig, 2020). En la práctica, se han utilizado los procedimientos PAM y algoritmos jerárquicos para $K = 2$ hasta $K = 10$ y se ha seleccionado el número de grupos según la media del estadístico Silueta y, como segundo criterio, el índice de separación.

4.2. Evaluación de la similitud entre agrupamientos

Para evaluar la similitud de dos métodos de agrupamiento se utiliza el índice de Rand. La definición de esta métrica se puede encontrar, por ejemplo, en Warrens y van der Hoef (2020). Sean $\mathcal{A} = \{A_1, A_2, ..., A_I\}$ y $B = \{B_1, B_2, ..., B_J\}$ dos particiones diferentes de las observaciones donde I y J es el número de grupos de cada partición. El índice de Rand viene dado por la siguiente expresión:

$$R = \frac{a + d}{N}, \qquad [9]$$

donde $N = \frac{n(n-1)}{2}$ es el número total de pares de observaciones, n es el número de observaciones, a es el número de pares que pertenecen al mismo clúster en ambas particiones y d es el número de pares que no pertenecen al mismo clúster en ambas particiones. El índice de Rand está entre 0 y 1, y si $R = 1$, las particiones son idénticas.

5. DESCRIPCIÓN DE LAS CLASIFICACIONES OBTENIDAS

Una vez que se tienen las curvas divididas en grupos, el siguiente paso es caracterizar cada grupo. Para ello, se ha elaborado un conjunto de variables que contiene información relacionada con la producción eléctrica para cada hora desde el 1 de enero de 2017 al 31 de diciembre de 2021. Se distingue entre dos tipos de variables: las relacionadas con la estructura de generación eléctrica y las variables temporales (por ejemplo, si la hora es diurna o nocturna):

- Variables relacionadas con la estructura de generación: han sido recogidas del sitio web de ESIOS (Sistema de Información del Operador del Sistema) y corresponden a

los programas de generación de cada periodo en el sistema eléctrico peninsular español (BOE/SEE, 2012).

Se ha elegido la energía programada en lugar de la energía generada real porque la primera se acerca más que la segunda al concepto de curvas de oferta que se está analizando.

Las variables de este tipo que se han seleccionado son la energía programada (en MWh) para cada hora por las siguientes fuentes: biogás, biomasa, carbón, ciclo combinado (CCGT), derivados del carbón y del petróleo, hidroeléctrica, cogeneración a gas natural (NGcog), nuclear, eólica, solar fotovoltaica y solar térmica. Además, se ha incluido la variable Generación Horaria Operativa Total, que es, para cada hora, la suma de la energía programada producida por todas las fuentes.

Las fuentes que no se han tenido en cuenta son las siguientes: genéricos, geotérmicos y oceánicos, residuos domésticos y afines, subproductos mineros, energía residual y bombeo por turbinas. Se ha comprobado que, por ejemplo, en 2019 estos tipos de generación representaron en su conjunto tan sólo el 1,22 % del total de energía programada para este año.

- Variables temporales:

  - Mes: mes correspondiente a la curva.
  - Día: día de la semana correspondiente a la curva.
  - Noche: variable binaria cuyo valor es 1 si la hora es entre las 23:00 y las 6:00 horas, incluidas ambas, y 0 en caso contrario.
  - Festivo: variable binaria cuyo valor es 1 si la oferta corresponde a sábado o domingo o si corresponde a un día festivo nacional en España. Su valor es 0 en caso contrario.
  - Festivo & Verano: variable binaria cuyo valor es 1 si corresponde a un festivo o el mes correspondiente de la oferta es julio o agosto. Su valor es 0 en caso contrario.

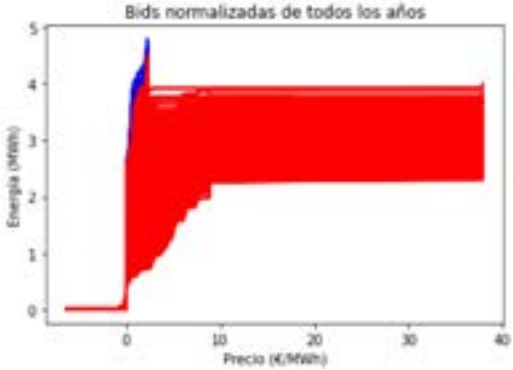
Una vez creado el conjunto de datos, se dividen todas las variables relacionadas con la estructura de generación entre la Generación Horaria Operativa Total, para obtener el porcentaje de energía generada por cada fuente para cada hora.

6. RESULTADOS Y DISCUSIÓN

Finalmente, se decidió omitir en el análisis las curvas de 2021 porque eran sustancialmente diferentes a las demás. El motivo es la gran diferencia entre los precios máximos de este año y el resto por un cambio en la legislación. La figura 6 ilustra este hecho. En la figura 6 y siguientes, se representan las curvas normalizadas, es decir, tanto los precios como las cantidades de energía se dividirán por la desviación estándar de todos los precios y todas

Figura 6.

Curvas de oferta normalizadas para el período de 2017 a 2021



Nota: Las curvas rojas corresponden a 2021 y las curvas azules corresponden al período 2017–2020.
Fuente: Elaboración propia.

las cantidades de energía, respectivamente. El objetivo de dicha normalización es que la distancia de Hausdorff no dependa de las unidades de medida.

6.1. Resultados de clasificación de las curvas de 2019 usando partición alrededor de medoides

En el cuadro 1 se muestra la media del estadístico Silueta y el índice de separación para diferentes valores de K utilizando PAM como procedimiento de clasificación. Los mejores resultados se obtienen para $K = 2$ y 3. Las clasificaciones obtenidas se describen en las siguientes subsecciones.

Cuadro 1.

Media del estadístico Silueta e Indice de separación en clasificaciones obtenidas mediante partición alrededor de medoides

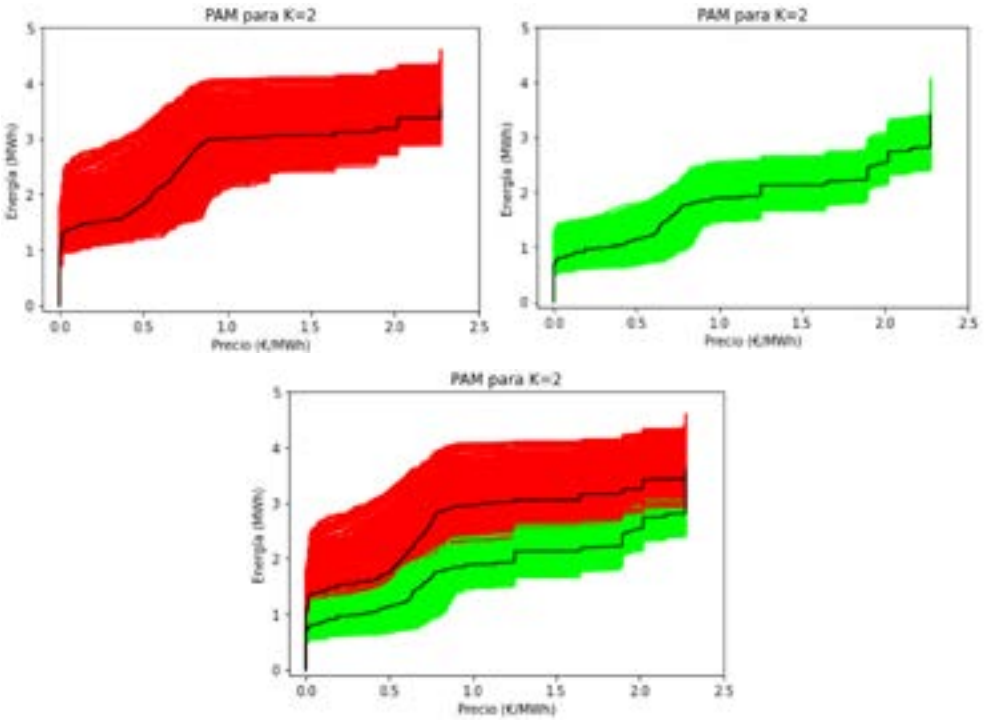
K	*Media del estadístico Silueta*	*Índice de separación*
2	0,57	0,131
3	0,43	0,069
4	0,35	0,061
5	0,33	0,047
6	0,31	0,046
7	0,28	0,044
8	0,26	0,042
9	0,23	0,041
10	0,22	0,040

Fuente: Elaboración propia.

6.1.1. Resultados utilizando PAM para K=2

Figura 7.

Curvas de oferta agrupadas mediante partición alrededor de medoides para K = 2



Nota: Las líneas en negro representan los medoides.
Fuente: Elaboración propia.

En la figura 7 se representan todas las curvas de oferta de 2019 de los grupos obtenidos con PAM para $K = 2$. Se observa que hay una cierta superposición entre las curvas verdes y rojas, lo que significa que la distancia de Hausdorff no solamente tiene en cuenta la proximidad entre dos curvas, sino también su forma. En general, las curvas verdes se refieren a ofertas más caras que las curvas rojas.

Para estos grupos, se analiza la proporción (en porcentaje) de curvas que pertenecen a cada mes. Hay que tener en cuenta que para un grupo determinado, una proporción uniforme para cada mes sería igual a 100·1/12=8,33 %, por lo que se esperan valores en torno a este número. Los resultados se muestran en el cuadro 2.

Como se ve, las curvas están distribuidas de forma más o menos uniforme en meses para cada grupo, aunque existen ligeras diferencias en los meses de enero y diciembre (mayor

Cuadro 2.

Para cada grupo definido en la figura 7, porcentaje de ofertas que corresponden a cada mes

	En.	*Feb.*	*Mar.*	*Abr.*	*May*	*Jun.*	*Jul.*	*Ag.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
■	9,05	7,89	8,41	7,95	8,11	7,85	8,21	8,21	8,07	8,10	8,70	9,39
■	7,19	7,16	8,71	8,82	9,35	9,05	9,13	9,13	8,56	9,35	7,08	6,40

Fuente: Elaboración propia.

porcentaje en las curvas rojas respecto a las verdes). Se analiza ahora si ocurre lo mismo con los días de la semana, pero ahora teniendo en cuenta que en una distribución uniforme para cada día correspondería a 100·1/7=14,28 %. Como se puede comprobar, ambos grupos parecen tener distribuciones uniformes respecto a los días de la semana.

Cuadro 3.

Para cada grupo definido en la figura 7, porcentaje de ofertas que corresponden a cada día de la semana

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	14,18	14,37	14,31	14,49	14,42	14,24	13,96
■	14,43	14,81	14,09	13,64	13,83	14,24	14,92

Fuente: Elaboración propia.

En lo que sigue se consideran las variables binarias: Noche (1 si la curva se fija para una hora entre las 23:00 y las 6:00), Festivo (1 si la curva se fija para un sábado, domingo o festivo Nacional) y Festivo & Verano (1 si la variable Festivo es 1 o el mes correspondiente de la curva es julio o agosto). En el cuadro 4 se presenta el porcentaje de curvas cuya respectiva variable es igual a 1 para cada grupo. Se observa que el 83 % de las curvas que pertenecen al clúster verde son nocturnas, mientras que sólo el 11 % de las curvas que forman el clúster rojo lo son.

Cuadro 4.

Para cada grupo definido en la figura 7, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a día Festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	30,31	42,05	11,73
■	31,56	44,33	83,44

Fuente: Elaboración propia.

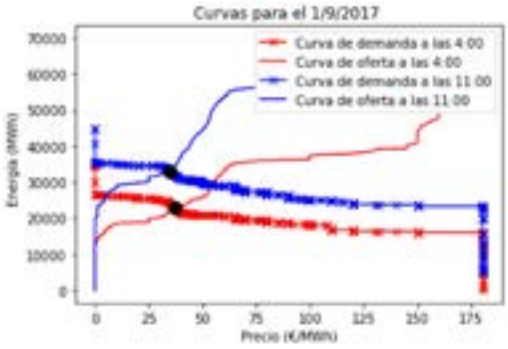
Si se mira nuevamente la figura 7, se ve que el grupo verde corresponde a curvas donde la energía es más cara, lo que puede parecer contradictorio porque se espera que la energía sea más barata en la noche. Antes de explicar este hecho, se debe analizar la figura 8. En esta figura se trazan, sin normalizar, las curvas de oferta y demanda para una hora nocturna y lo mismo

para una hora diurna. En primer lugar, se observa que la curva de demanda correspondiente a las 4 AM está por debajo (la demanda es menor) que la curva de demanda correspondiente a las 11 AM, lo que provoca que ambos precios sean similares, aunque el nocturno sigue siendo más caro (debe recordarse que el precio está dado por la intersección entre oferta y demanda). Sin embargo, se debe tener en cuenta que estas curvas corresponden a todas las ofertas realizadas para estos dos periodos, no las que corresponden únicamente a las ofertas que entran en la casación.

Se debe considerar que las ofertas que se han realizado pueden estar sujetas a una condicionalidad fijada por el productor como la que permite la Regla 40.3 de BOE/CNMC (2021), y cuando esta condicionalidad no se cumple, se hacen algunas correcciones, que al final provocan que la parte de la curva de oferta que ha casado se mueva hacia la izquierda (más barata) respecto a la original.

Figura 8.

Curvas de oferta y demanda para horas nocturnas y diurnas



Fuente: Elaboración propia.

En el cuadro 5 se recoge el porcentaje promedio de energía programada por fuente para cada grupo y el promedio total de energía:

Cuadro 5.

Para cada grupo definido en la figura 7, porcentaje promedio de energía programada por fuente para cada grupo y el promedio total de energía

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	30.202,30	21,06	10,44	10,30	21,38	20,81	4,75	2,70
■	24.044,08	19,85	8,42	12,65	26,87	22,50	0,05	0,70

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,31	1,09	4,19	1,05
■	0,37	1,34	4,44	1,22

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

En el cuadro 5 se observa que obviamente, el porcentaje promedio de energía solar generada (fotovoltaica y térmica) es notablemente mayor en el grupo rojo (diurno) que en el verde (nocturno). El porcentaje casi nulo de energía solar parece ser sustituido en la horas nocturnas principalmente por la nuclear (diferencia de unos 6 puntos), pero también por la cogeneración. Además, la energía media total es notablemente mayor en el grupo diurno rojo que en el verde, lo cual tiene sentido.

Finalmente, se realiza una clasificación supervisada mediante el algoritmo de bosque aleatorio (*random forest*), propuesto por Breiman (2001), para predecir la etiqueta del grupo usando las variables explicativas. Se dividen los datos en conjunto de entrenamiento (80 % de los datos) y de prueba (20 % de los datos). Se obtuvo una precisión del 94,51 %. La matriz de confusión correspondiente se muestra en el cuadro 6 y las 15 variables más importantes en la clasificación se muestran en la figura 9.

Cuadro 6.

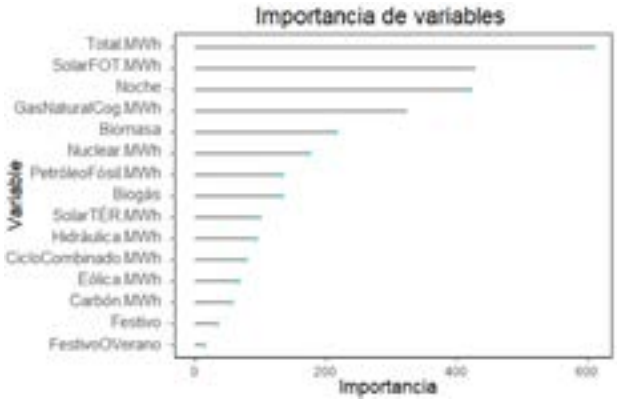
Matriz de confusión en el conjunto de prueba

	Etiqueta real	
Predicción	■	■
■	1.139	11
■	85	516

Fuente: Elaboración propia.

Figura 9.

Variables más importantes en la clasificación supervisada



Fuente: Elaboración propia.

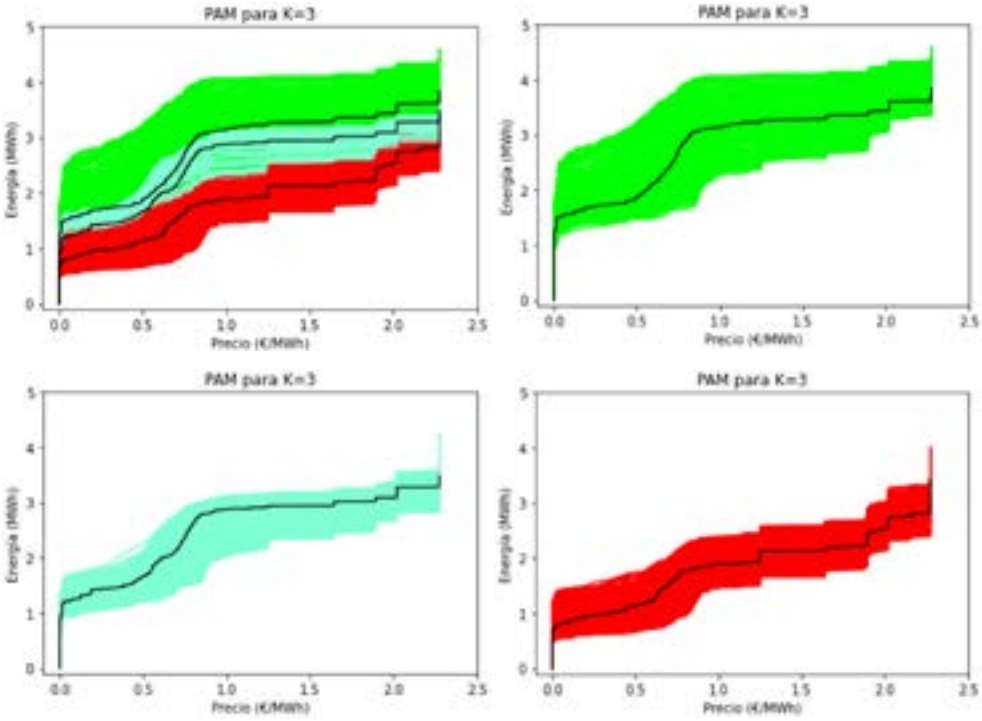
Como se observa, las variables Solar fotovoltaica y Noche son dos de las más importantes, confirmando las conclusiones anteriores.

6.1.2. Resultados utilizando PAM para K=3

Los grupos obtenidos mediante PAM para $K = 3$ se muestran en la figura 10, donde se observa nuevamente algunos solapamientos, especialmente entre las curvas verde y azul. Además, las medias de los estadísticos Silueta promedio y los índices de separación son algo más bajos en este caso, como muestra el cuadro 1.

Figura 10.

Curvas de oferta agrupadas mediante partición alrededor de medoides para $K = 3$



Nota: Las líneas en negro representan los medoides.

Fuente: Elaboración propia.

En el cuadro 7 se presenta la proporción de cada mes para cada grupo.

Un aspecto destacable que se observa en el cuadro 7 es que la proporción de curvas verdes asociadas a los meses comprendidos entre noviembre y abril es mayor que en el resto. Posteriormente se encontrará que curvas superiores, como las situadas en la parte más elevada del clúster verde, están relacionadas con una alta producción de energía eólica, pero esto se puede anticipar observando la distribución mensual de la producción con este tipo de energía. En la figura 11 se representa la energía eólica promedio programada mensualmente

Cuadro 7.

Para cada grupo definido en la figura 10, porcentaje de ofertas que corresponden a cada mes

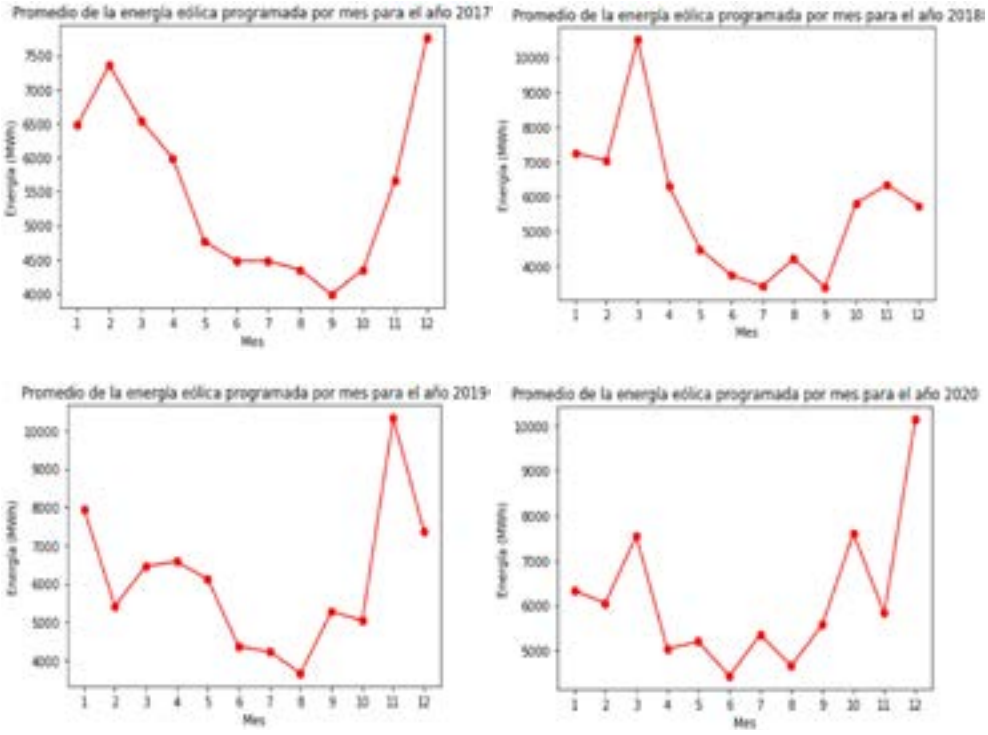
	En.	*Feb.*	*Mar.*	*Abr.*	*May*	*Jun.*	*Jul.*	*Ag.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
■	16,20	11,85	12,00	10,70	7,12	3,65	1,07	3,15	5,50	2,00	12,54	14,16
■	3,96	5,06	5,86	5,94	8,92	10,81	13,30	11,74	9,77	12,59	6,02	6,02
■	9,09	7,13	8,73	8,97	9,29	9,17	9,17	9,29	8,77	9,21	6,93	6,22

Fuente: Elaboración propia.

para cada año. Nótese que este promedio es menor en los meses de verano más septiembre y octubre, por lo que si las curvas de un grupo no se ubican en estos meses, se espera que su producción eólica asociada sea alta. Por lo tanto, se espera que las curvas verdes tengan un alto porcentaje de energía eólica.

Figura 11.

Energía eólica promedio programada por mes para los años 2017–2020



Fuente: Elaboración propia.

Además, para cada grupo se ha observado nuevamente una proporción igual para todos los días de la semana (ver cuadro 29 en el Apéndice). Sin embargo, la distinción entre curvas de día y de noche vuelve a aparecer, como muestra el cuadro 8.

Cuadro 8.

Para cada grupo definido en la figura 10, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a Festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	29,75	32,95	6,74
■	30,54	48,44	16,15
■	31,86	44,58	85,93

Fuente: Elaboración propia.

Como se observa, el grupo azul tiene un porcentaje del 16 % de curvas que son nocturnas, mientras que en el rojo este porcentaje es del 85,3 %. Además, el porcentaje de días no laborables es similar para los tres grupos mientras que el grupo verde tiene una menor proporción de curvas Festivo & Verano, lo que se debe a la baja proporción de conjuntos de ofertas en los meses estivales.

En el cuadro 9, se presenta la estructura de generación para los tres grupos obtenidos.

Cuadro 9.

Para cada grupo definido en la figura 10, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	31.180,03	13,65	11,03	10,01	20,55	28,13	5,10	2,45
■	29.379,62	26,34	10,04	10,55	22,06	15,70	4,36	2,80
■	23.907,41	19,78	8,30	12,71	27,02	22,44	0,03	0,72

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,30	1,04	4,82	0,99
■	0,31	1,13	3,74	1,09
■	0,37	1,35	4,47	1,22

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Observando el cuadro 9 se comprueba que el grupo verde tiene un porcentaje de energía eólica notablemente superior a los demás, y cómo su energía total programada es la más alta, se puede concluir que en términos absolutos la cantidad de energía eólica que ofrecen estas curvas es también la más alta. El hecho de disponer de una cantidad tan grande de energía eólica es una posible hipótesis que explica por qué estas curvas hacen referencia a las ofertas más baratas. Además, las curvas azules tienen un mayor porcentaje de energía de ciclo

combinado, lo cual tiene sentido, porque están relacionadas con horas diurnas con un bajo porcentaje de energía eólica, por lo que se necesita algún tipo de energía de respaldo.

Finalmente, se muestran los resultados obtenidos con el algoritmo de bosque aleatorio (dividiendo nuevamente los datos en un conjunto de entrenamiento y uno de prueba). En este caso, la matriz de confusión (ver cuadro 10) ilustra que la clasificación errónea entre los grupos superior e inferior (verde y rojo, respectivamente) es casi nula, mientras que el error es mayor cuando el algoritmo intenta distinguir entre los grupos superiores y medio (azul) o entre los grupos medio e inferior. La precisión obtenida en este caso es del 91,94 %.

Cuadro 10.

Matriz de confusión en el conjunto de prueba

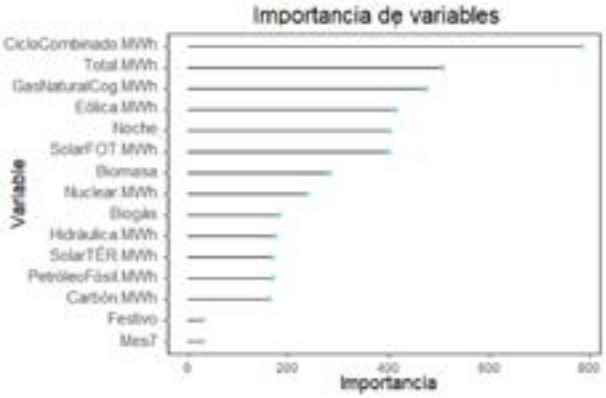
	Etiqueta real		
Predicción	■	■	■
■	469	34	1
■	48	656	16
■	2	40	484

Fuente: Elaboración propia.

El diagrama de la importancia de las variables en el procedimiento de clasificación se muestra en la figura 12. Las variables Noche y Solar fotovoltaica vuelven a ser importantes porque ayudan a distinguir el grupo rojo del resto. Además, la variable Ciclo Combinado es la más importante, lo que parece lógico porque el porcentaje medio de energía generada con ella es diferente para cada grupo. Sin embargo, ha sorprendido la importancia de la Cogeneración de Gas Natural, pero quizás también sea importante para distinguir entre el grupo rojo y el resto de los grupos. Se observa también que la energía eólica es la cuarta más importante.

Figura 12.

Variables más importantes en la clasificación supervisada



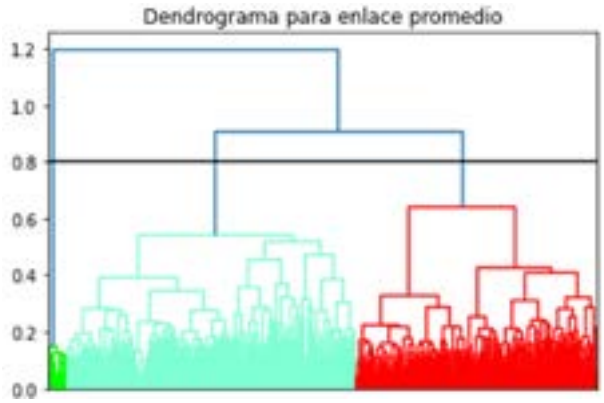
Fuente: Elaboración propia.

6.2. Resultados de clasificación de las curvas de 2019 usando un procedimiento aglomerativo

En la figura 13 se muestra el dendrograma de un agrupamiento aglomerativo usando enlace promedio (*average linkage*) aplicado a las curvas de 2019. Además, en el cuadro 11 se muestra la media del estadístico Silueta y el índice de separación para entre 2 y 10 grupos. Nuevamente son mejores los valores bajos de K. Se decide eligir K=3 puesto que K=2 conduce a una solución con un grupo muy pequeño y otro con el resto de las curvas.

Figura 13.

Dendrograma para las curvas de oferta de 2019



Nota: La línea negra corta las líneas verticales correspondientes a los tres grupos elegidos.
Fuente: Elaboración propia.

Cuadro 11.

Media del estadístico Silueta e índice de separación en clasificaciones obtenidas mediante un procedimiento aglomerativo con enlace promedio

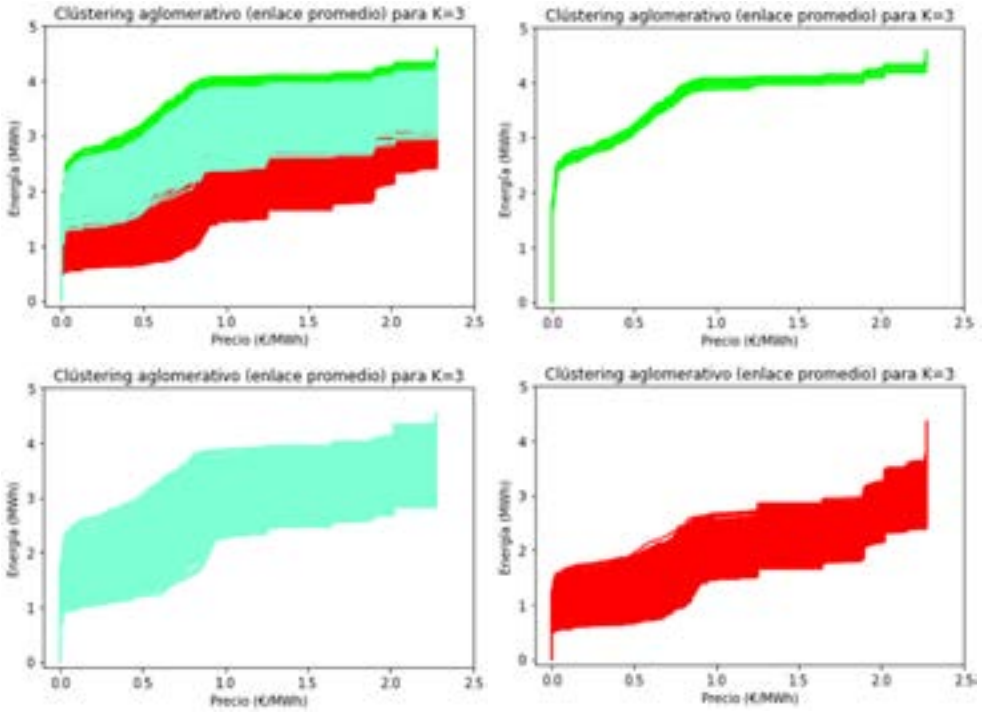
K	*Media del estadístico Silueta*	*Índice de separación*
2	0,46	0,553
3	0,51	0,145
4	0,40	0,142
5	0,40	0,084
6	0,40	0,084
7	0,39	0,084
8	0,37	0,084
9	0,27	0,084
10	0,29	0,074

Fuente: Elaboración propia.

En la figura 14 se muestran las curvas de los grupos obtenidos por agrupamiento aglomerativo para K=3. Se observa que no están tan superpuestos como los grupos obtenidos con PAM.

Figura 14.

Curvas de oferta agrupadas mediante procedimiento aglomerativo con enlace promedio para K=3



Fuente: Elaboración propia.

Se comprueba que el clúster verde está formado por 21 curvas correspondientes a las fechas y horas comprendidas entre las 10:00 y las 22:00 horas del 21 de diciembre de 2019 (Sábado) y entre las 10:00 y las 17:00 horas del 22 de diciembre de 2019 (Domingo). Para las curvas roja y azul no se ha observado ninguna distribución especial, ni en meses ni en días (ver cuadros 30 y 31 en el Apéndice).

En el cuadro 12 se presentan los porcentajes de las variables binarias en estos grupos. Nuevamente se tiene una fuerte separación entre las curvas diurnas y nocturnas.

Por otra parte, en el cuadro 13, se observa que las curvas verdes tienen un alto porcentaje de energía generada tanto por fuentes eólicas como hidráulicas, y un porcentaje bajo de

Cuadro 12.

Para cada grupo definido en la figura 14, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a Festivo o Verano (Festivo & Verano) o que corresponde a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	100,00	100,00	0,00
■	32,24	42,24	10,25
■	31,68	43,63	86,22

Fuente: Elaboración propia.

energía de ciclo combinado (CCGT), lo que puede explicar lo baratas que son estas ofertas. Además, los principales componentes de las curvas azules son el ciclo combinado, la nuclear y la cogeneración. En conclusión, se tienen tres grupos: uno que se refiere a las ofertas nocturnas y dos que se refieren a las diurnas. Entre ambos grupos diurnos, un grupo está relacionado con una gran cantidad de energía a bajos precios y tiene un alto porcentaje de energía generada por fuentes renovables.

Cuadro 13.

Para cada grupo definido en la figura 14, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	30.620,12	3,58	25,08	7,41	14,32	39,73	3,76	0,27
■	30.157,27	21,50	8,37	10,33	26,76	20,23	4,78	0,69
■	24.167,51	18,97	10,42	12,59	21,44	23,65	0,03	2,72

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,29	0,91	0,00	0,33
■	0,31	1,09	4,20	1,05
■	0,37	1,34	4,47	1,21

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Finalmente, se muestran los resultados de realizar una clasificación supervisada mediante un procedimiento de bosque aleatorio. Para este caso, se presentan dos situaciones diferentes: en la primera no se han dividido los datos en conjunto de entrenamiento y prueba debido al tamaño reducido del grupo más pequeño (el verde) y en la segunda se ha dividido el conjunto de datos ignorando el grupo verde. La matriz de confusión y el diagrama con la importancia de las variables para cada caso se muestran en el cuadro 14 y en el gráfico 15, respectivamente. Las precisiones son del 99,60 % para el conjunto completo de tres grupos, y del 96,56 % para el caso con dos grupos. Curiosamente, la energía eólica y la hidroeléctrica casi no son importantes, ni siquiera en el caso en que se toma en consideración el grupo verde. Ambos diagramas son muy similares y Noche y SolarFOT siguen siendo dos de las variables más relevantes.

Cuadro 14.

Matrices de confusión en el conjunto completo de datos y en el conjunto de prueba con dos grupos

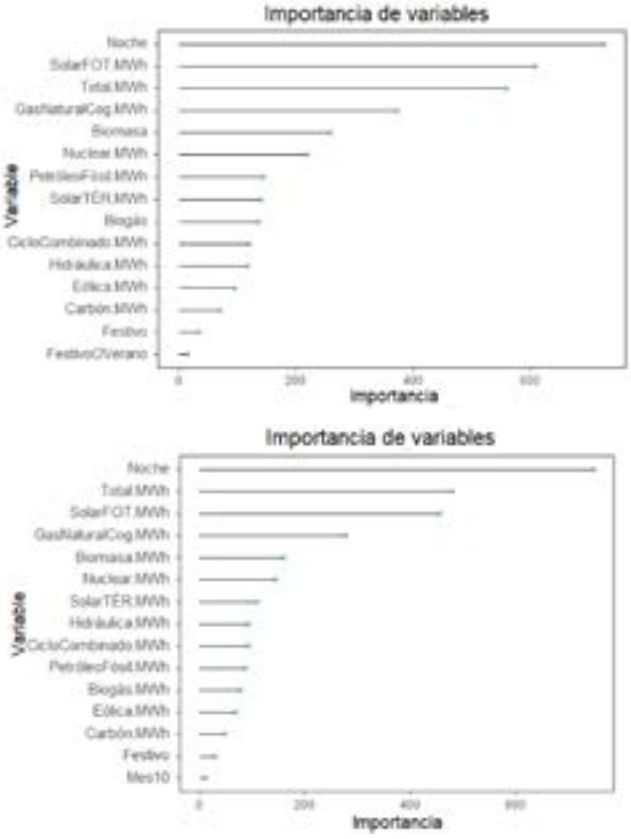
	Etiqueta real		
Predicción	■	■	■
■	21	10	0
■	0	6.055	0
■	0	25	2.649

	Etiqueta real	
Predicción	■	■
■	1.200	42
■	18	487

Fuente: Elaboración propia.

Figura 15.

Variables más importantes en la clasificación supervisada



Nota: El gráfico superior corresponde a la clasificación del conjunto completo con tres grupos y el gráfico inferior corresponde a la clasificación de los dos grupos mayoritarios.

Fuente: Elaboración propia.

6.3. Resultados de clasificación de curvas de horas pico y valle usando partición alrededor de medoides

En esta sección se toman las curvas de una hora pico (12 a.m) y una hora valle (5 a.m.) de cada día para todo el período (2017-2020). La idea es comprobar si los resultados obtenidos al analizar las curvas de 2019 son generalizables o no.

En el cuadro 15 se presentan los valores de la media del estadístico Silueta y el índice de separación.

Cuadro 15.

Media del estadístico Silueta e índice de separación en clasificaciones obtenidas mediante partición alrededor de medoides

K	*Media del estadístico Silueta*	*Índice de separación*
2	0,60	0,206
3	0,45	0,115
4	0,35	0,070
5	0,31	0,066
6	0,27	0,063
7	0,24	0,061
8	0,22	0,061
9	0,21	0,062
10	0,20	0,061

Fuente: Elaboración propia.

Nuevamente, se obtienen los valores más altos de los índices cuando K es bajo, por lo que se selecciona $K = 2$ y 3. Las clasificaciones obtenidas se describen en las siguientes subsecciones.

6.3.1. Resultados utilizando PAM para K = *2*

En la figura 16 se presentan los grupos obtenidos para $K = 2$ usando PAM. Debe tenerse en cuenta que en este caso los dos grupos no parecen estar demasiado superpuestos.

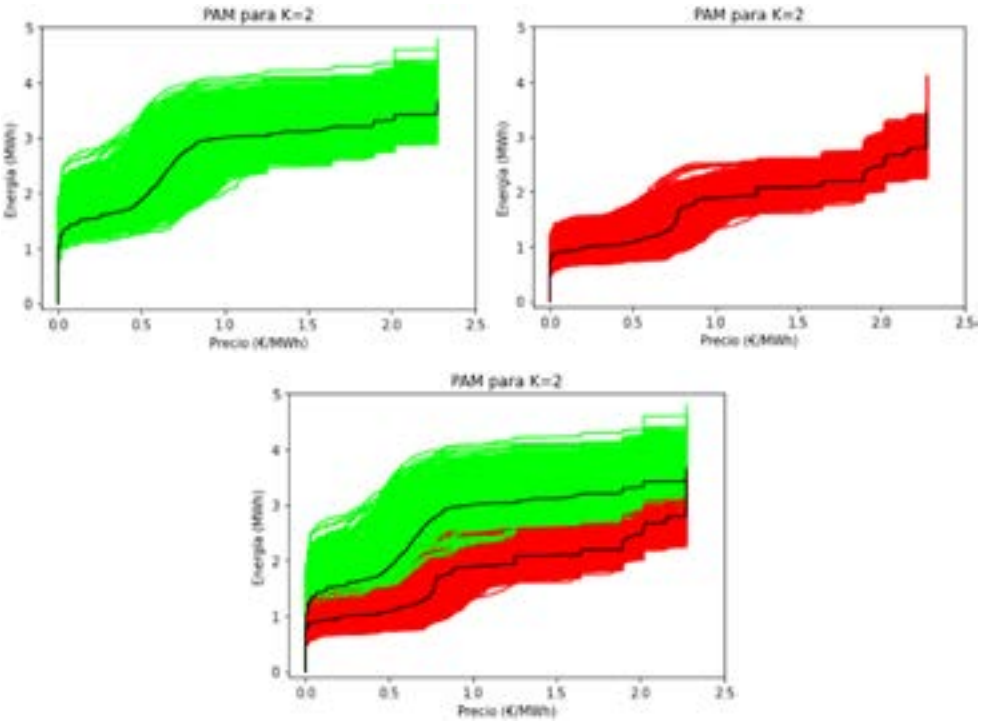
Una vez más, para cada grupo se ha observado una distribución igual en meses y en días (cuadros 32 y 33 en el Apéndice). Además, aunque no hay diferencia en la proporción de días festivos o de verano, se observa nuevamente la separación entre las curvas nocturnas y diurnas, como ilustra el cuadro 16.

Este es un buen resultado porque implica que los resultados obtenidos analizando las curvas correspondientes a 2019 pueden generalizarse a todo el período. De hecho, se puede realizar la siguiente comparación:

- Para todas estas horas punta y valle tomar únicamente las correspondientes al año 2019.
- Para todas las curvas de 2019, se toman solo las correspondientes a las 5:00 y a las 12:00.

Figura 16.

Curvas de oferta agrupadas mediante partición alrededor de medoides para K = 2



Nota: Las líneas en negro representan los medoides.
Fuente: Elaboración propia.

Cuadro 16.

Para cada grupo definido en la figura 16, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a un festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	31,15	43,38	4,59
■	30,85	43,38	99,57

Fuente: Elaboración propia.

Tomando la partición dada por PAM para $K = 2$ para ambos conjuntos y comparándolas, se obtiene un índice de Rand de 0.99 lo que implica que hay una gran similitud entre ambas clasificaciones.

El cuadro 17 muestra la estructura de generación para cada grupo. De nuevo hay una diferencia notable en los porcentajes de energía solar, lo cual es coherente con el conjunto de datos analizado.

Cuadro 17.

Para cada grupo definido en la figura 16, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	31.354,80	14,52	10,98	10,45	20,30	17,87	10,17	3,77
■	23.107,13	13,53	8,74	13,79	27,43	22,27	0,06	0,50

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,29	1,13	8,68	1,02
■	0,39	1,52	10,17	1,29

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Finalmente, se estudia la importancia de las variables en un procedimiento de bosque aleatorio. Dividiendo entre conjuntos de entrenamiento y de prueba se obtiene que la clasificación de las curvas del conjunto de prueba ha sido correcta en el 99,31 % de los casos. El cuadro 18 muestra la matriz de confusión, y la figura 17 muestra el diagrama de importancia de las variables.

Cuadro 18.

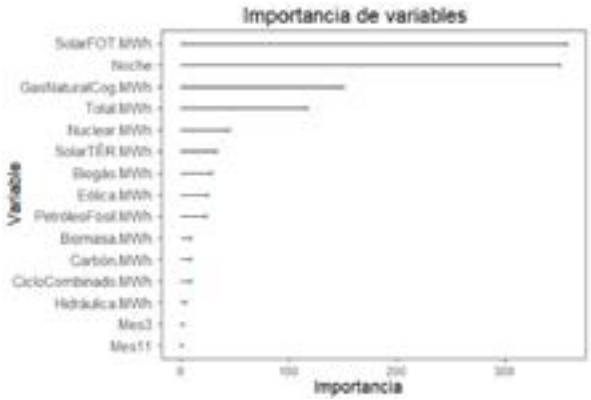
Matriz de confusión en el conjunto de prueba

	Etiqueta real	
Predicción	■	■
■	303	2
■	2	277

Fuente: Elaboración propia.

Figura 17.

Variables más importantes en la clasificación supervisada



Fuente: Elaboración propia.

La figura 17 muestra que SolarFOT y Noche son, con diferencia, las variables más relevantes para distinguir entre los dos grupos.

6.3.2. Resultados utilizando PAM para K = 3

La figura 18 representa la clasificación obtenida usando PAM para $K = 3$. En este caso, en el cuadro 19 se observa que el clúster verde tiene una mayor proporción de curvas que pertenecen a los meses comprendidos entre diciembre a abril, por lo que se espera un alto porcentaje de energía eólica. El clúster azul, por su parte, tiene una mayor proporción de curvas en julio, agosto y octubre. Por otro lado, se observa una distribución uniforme para los tres grupos al estudiar los días de la semana (ver cuadro 34 en el Apéndice).

Observando las variables binarias (cuadro 20), mientras el grupo verde muestra una menor proporción de curvas con la variable Festivo & Verano igual a 1, su proporción de días

Cuadro 19.

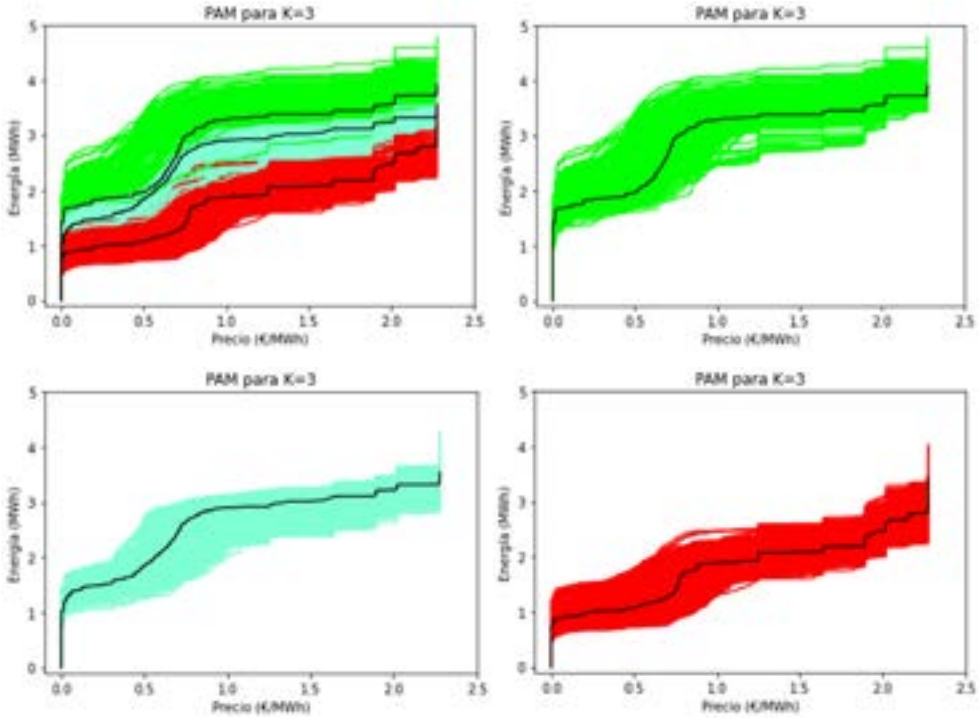
Para cada grupo definido en la figura 18, porcentaje de ofertas que corresponden a cada mes

	En.	*Feb.*	*Mar.*	*Abr.*	*May*	*Jun.*	*Jul.*	*Ag.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
■	12,05	11,73	17,10	11,40	7,98	6,03	3,26	3,09	3,42	3,75	6,84	13,36
■	7,21	5,06	4,63	5,71	8,18	8,93	11,19	11,30	1,76	11,09	8,83	7,10
■	7,76	7,76	7,25	8,48	8,92	8,70	8,99	8,99	8,63	8,85	8,41	7,25

Fuente: Elaboración propia.

Figura 18.

Curvas de oferta agrupadas mediante partición alrededor de medoides para K = 3



Nota: Las líneas en negro representan los medoides.
Fuente: Elaboración propia.

festivos es bastante similar a la de los otros dos grupos. Nuevamente, el grupo inferior (rojo, en este caso) es principalmente nocturno, mientras que los otros dos se refieren principalmente a curvas diurnas.

Cuadro 20.

Para cada grupo definido en la figura 18, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a un festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	29,32	34,04	2,77
■	32,29	48,01	7,32
■	30,89	43,58	99,78

Fuente: Elaboración propia.

El cuadro 21 es similar al cuadro 9, donde se muestra la estructura de generación obtenida también con PAM para $K = 3$ para las curvas de 2019. Por ejemplo, en esta última, los porcentajes de energía eólica eran, del grupo superior al inferior, 28,13, 15,70 y 22,44, mientras que ahora son 25,77, 12,99 y 22,10. En general, todas las variables muestran valores similares. En conclusión, se vuelve a obtener tres grupos con las siguientes características: en primer lugar, una clara separación entre las curvas nocturnas y diurnas, y dentro de las diurnas, un grupo de ellas ofrece más energía a un precio más barato como consecuencia de la alta proporción de energía eólica.

Cuadro 21.

Para cada grupo definido en la figura 18, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	31.877,30	9,84	11,17	10,22	19,86	25,77	9,83	3,08
■	30.946,21	17,50	10,80	10,62	20,65	12,99	10,23	4,15
■	23.041,93	13,59	8,75	13,82	27,49	22,10	0,03	0,50

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,29	1,10	6,97	0,98
■	0,30	1,15	9,78	1,04
■	0,39	1,52	10,21	1,29

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Finalmente, el cuadro 22 muestra la matriz de confusión (la precisión obtenida es 91,92 %) resultante de aplicar el algoritmo de bosque aleatorio para predecir las etiquetas de los clústers usando las variables explicativas, y la figura 19 muestra el correspondiente diagrama de las 15 variables más importantes.

Cuadro 22.

Matriz de confusión en el conjunto de prueba

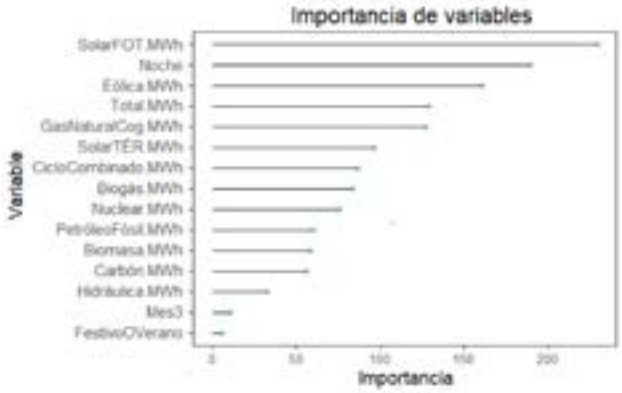
	Etiqueta real		
Predicción	■	■	■
■	98	14	0
■	23	165	3
■	1	6	272

Fuente: Elaboración propia.

La figura 19 es coherente con los hechos que se han descrito. SolarFOT y Noche vuelven a ser las variables más importantes, mientras que Eólica es la tercera.

Figura 19.

Variables más importantes en la clasificación supervisada



Fuente: Elaboración propia.

6.4. Resultados de clasificación de las curvas de horas pico y valle usando un procedimiento aglomerativo

En el cuadro 23 se muestran la media del estadístico Silueta y el índice de separación cuando se utiliza agrupamiento jerárquico aglomerativo (con enlace promedio). Nuevamente se dan valores más altos de Silueta e índice de separación cuando K es bajo. Cuando, $K = 2$ da un resultado casi igual al obtenido por PAM (índice Rand de 0.911). Para $K = 4$ se verifica la presencia de un grupo formado únicamente por cinco curvas situadas entre diciembre de 2019 y marzo de 2020, todas ellas horas valle y con un porcentaje de energía eólica entre el 30 % y el 56 %. Si $K = 5$ aparece otro clúster de cinco observaciones, siendo estas horas punta con

Cuadro 23.

Media del estadístico Silueta e índice de separación en clasificaciones obtenidas mediante un procedimiento aglomerativo con enlace promedio

K	*Media del estadístico Silueta*	*Índice de separación*
2	0,59	0,263
3	0,51	0,228
4	0,47	0,229
5	0,34	0,232
6	0,33	0,139
7	0,32	0,088
8	0,29	0,089
9	0,26	0,086
10	0,23	0,086

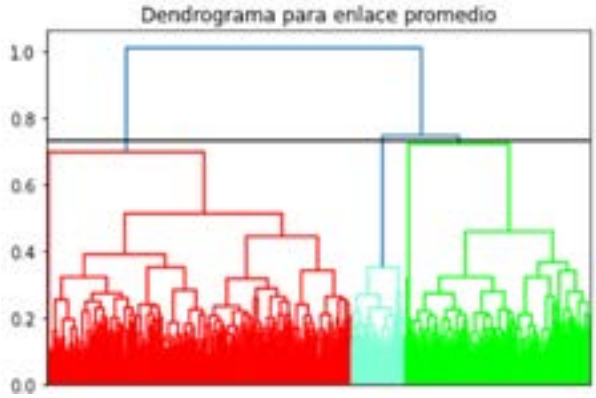
Fuente: Elaboración propia.

un bajo porcentaje de energía eólica (entre el 3 y el 12 %) cuatro de ellas en noviembre y la última en septiembre de 2020. En lo que sigue, se muestra el resultado para $K = 3$ con lo que se evitan grupos tan pequeños que podrían corresponder a días atípicos.

En la figura 20 se muestra el dendrograma y la separación correspondiente a K = 3.

Figura 20.

Dendrograma para las curvas de oferta de horas pico y valle



Nota: La línea negra corta las líneas verticales correspondientes a los tres grupos elegidos.
Fuente: Elaboración propia.

Para $K = 3$ los grupos obtenidos se muestran en la figura 21. Entre esta clasificación y la obtenida con PAM (claro, con K=3) el índice de Rand es 0.847, por lo que a pesar de algunas diferencias se esperan resultados similares en el análisis.

Figura 21.

Curvas de oferta agrupadas mediante procedimiento aglomerativo con enlace promedio para $K = 3$

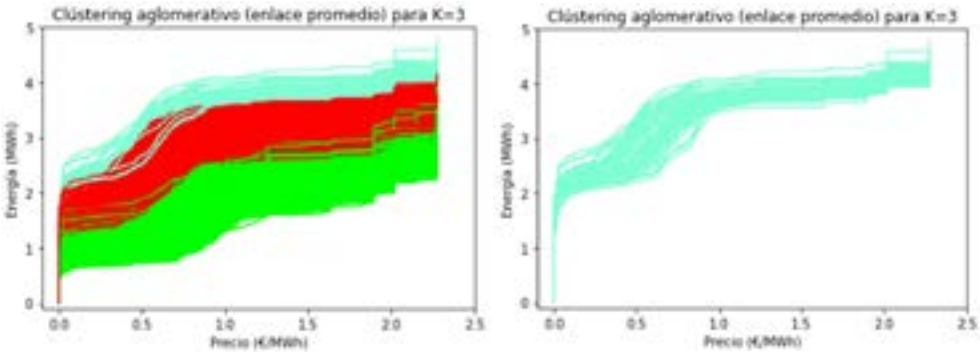
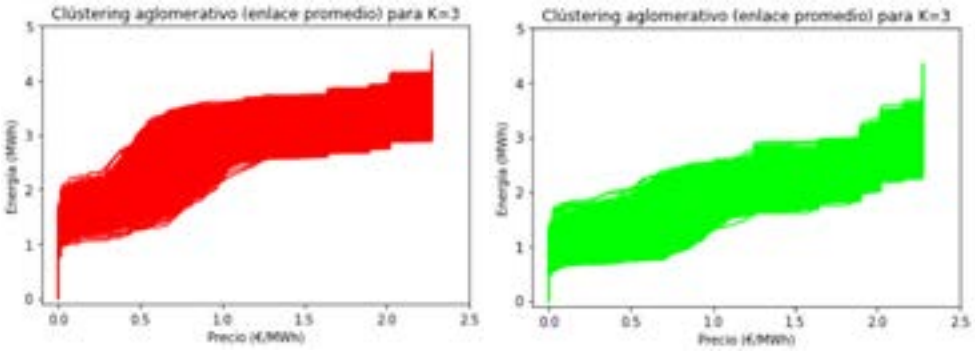


Figura 21. *(continuación)*

Curvas de oferta agrupadas mediante procedimiento aglomerativo con enlace promedio para $K = 3$



Fuente: Elaboración propia.

Primero, para el clúster azul, en su distribución en meses se observa una alta proporción de curvas correspondientes a marzo y un porcentaje nulo o muy pequeño de curvas relacionadas con los meses entre junio y octubre (ver cuadro 24). En segundo lugar, para este grupo se observa que hay una alta proporción de curvas en Lunes (cuadro 25), pero debe notarse que este grupo consta solo de 63 observaciones.

Cuadro 24.

Para cada grupo definido en la figura 21, porcentaje de ofertas que corresponden a cada mes

	En,	*Feb,*	*Mar,*	*Abr,*	*May*	*Jun,*	*Jul,*	*Ag,*	*Sep,*	*Oct,*	*Nov,*	*Dic,*
■	12,70	7,94	39,68	11,11	4,17	1,59	0,00	0,00	1,59	0,00	4,76	17,46
■	8,30	7,80	7,15	8,08	8,73	8,51	8,87	8,87	8,44	8,87	8,08	8,30
■	8,49	7,67	8,42	8,21	8,49	8,21	8,49	8,49	8,28	8,49	8,49	8,28

Fuente: Elaboración propia.

Cuadro 25.

Para cada clúster definido en la figura 21, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	23,81	14,29	12,70	11,11	14,29	9,52	14,29
■	13,88	14,23	14,31	14,38	14,23	14,59	14,38
■	14,31	14,37	14,37	14,37	14,24	14,10	14,24

Fuente: Elaboración propia.

Al observar las variables binarias en el cuadro 26, se constata que la distinción entre día y noche es casi perfecta. El grupo verde se refiere a los períodos nocturnos, mientras que los grupos rojo y azul se refieren a horas diurnas.

Cuadro 26.

Para cada grupo definido en la figura 21, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a un festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	30,16	30,16	0,00
■	31,19	43,71	0,36
■	30,87	42,85	99,66

Fuente: Elaboración propia.

La estructura de generación se presenta en el cuadro 27. Nuevamente, el grupo superior (azul) es el que tiene la mayor proporción promedio de energía eólica. Además, para este grupo la proporción media total de energía generada por fuentes renovables (solar, eólica e hidroeléctrica) es del 61,71 %. Para el grupo rojo este porcentaje es del 37,33 % y para el verde del 23,81 %. Además, el bajo porcentaje de energía generada por ciclo combinado también diferencia al clúster azul de los otros dos.

Cuadro 27.

Para cada grupo definido en la figura 21, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	33.631,12	4,62	12,31	9,04	18,37	38,26	9,28	1,86
■	31.420,41	15,43	10,96	10,50	20,26	15,69	10,68	4,03
■	23.307,43	13,13	8,80	13,66	27,24	23,28	0,05	0,48

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,27	1,02	3,02	0,78
■	0,29	1,13	9,16	1,03
■	0,38	1,51	9,88	1,27

Nota: La primera columna es la energía promedio programada por grupo.
Fuente: Elaboración propia.

Para la clasificación supervisada, nuevamente, se muestran los resultados para dos situaciones: la primera sin dividir los datos en conjunto de entrenamiento y prueba, y la segunda dividiendo los datos, pero sin considerar el grupo más pequeño, en este caso el azul. En el primer caso, la precisión es del 100 %, en el segundo es del 99,82 % (ver cuadro 28). Los diagramas de importancia de las variables se muestran en la figura 22.

Las variables Solar fotovoltaica y Noche vuelven a ser las más importantes en ambos diagramas, que son muy similares.

Cuadro 28.

Matrices de confusión en el conjunto completo de datos y en el conjunto de prueba con dos grupos

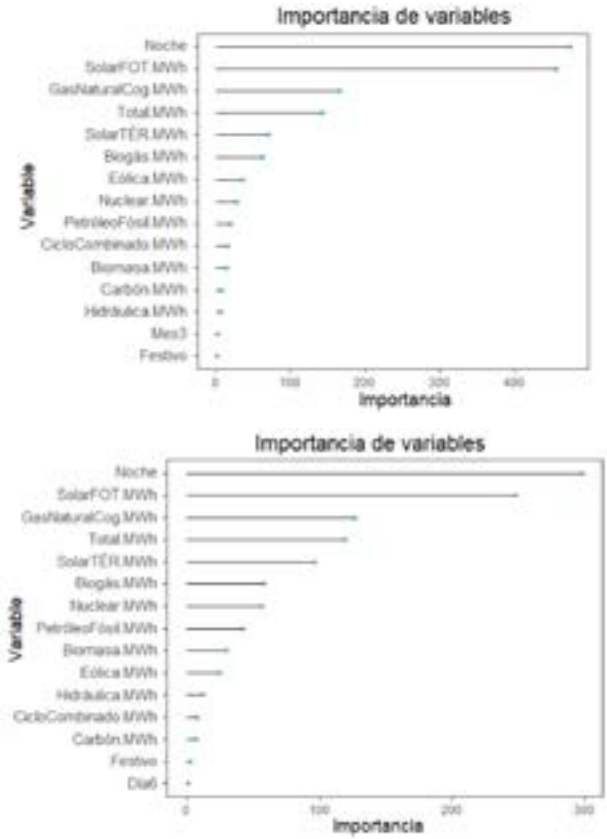
	Etiqueta real		
Predicción	■	■	■
■	63	0	0
■	0	1.398	0
■	0	0	1.461

	Etiqueta real	
Predicción	■	■
■	279	1
■	0	291

Fuente: Elaboración propia.

Figura 22.

Variables más importantes en la clasificación supervisada



Nota: El gráfico superior corresponde a la clasificación del conjunto completo con tres grupos y el gráfico inferior corresponde a la clasificación de los dos grupos mayoritarios.

Fuente: Elaboración propia.

7. CONCLUSIONES

En este proyecto, se ha encontrado una dificultad computacional notable al calcular la distancia de Hausdorff debido a la gran cantidad de cálculos iterativos que necesita, lo cual es costoso especialmente con grandes conjuntos de datos. Por ese motivo no se analizaron todas las curvas del período 2017 a 2020, y se optó por realizar un análisis para un año y para horas seleccionadas de tipo pico y valle de todo el período. Sin embargo, aun así, el estudio que se ha realizado es consistente y permite obtener conclusiones relevantes.

- En primer lugar, se han transformado los conjuntos de ofertas en curvas para poder calcular la distancia de Hausdorff entre dichos conjuntos.
- Usando esta medida de similitud para agrupar las curvas se ha observado que, en general, es preferible un valor bajo del número de grupos para obtener una buena separación entre ellos.
- No se han encontrado diferencias en las curvas de oferta respecto a los días de la semana o los días de baja actividad laboral. Si se consideran los meses del año, la distribución de las curvas cambia debido a la energía eólica.
- Este tipo de generación tiene un fuerte impacto en las curvas. Existe una clara distinción entre curvas en función de su porcentaje de energía eólica. Las curvas con un porcentaje alto hacen referencia a una gran cantidad de energía ofrecida a bajo precio.
- Además, los métodos de agrupamiento que se han utilizado también dan una separación natural entre las curvas diurnas y nocturnas.

Finalmente, una posible vía de extender esta investigación sería realizar el mismo análisis teniendo en cuenta las ofertas de demanda. Este estudio sería complementario a este proyecto y daría como resultado una comprensión más profunda de los factores que son relevantes en la fijación del precio de la energía.

Referencias

Aggarwal, S. K., Saini, L. M. y Kumar, A. (2009). Day-ahead electricity price forecasting in Victoria Electricity Market using Support Vector Machine based Model. *Power Research,* 5, pp. 37–45.

Agosti, L., Padilla, A. J. y Requejo, A. (2007). El mercado de generación eléctrica en España: Estructura, funcionamiento y resultados. *Economía Industrial,* 364, pp. 21–37.

Akhanli, S. E. y Hennig, CH. (2020). Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Statistics and Computing,* 30, pp. 1523– 1544.

BOE/CNMC (2021). Corrección de errores de la resolución de 6 de mayo de 2021, de la Comisión Nacional de los Mercados y la Competencia, por la que se aprueban las reglas de funcionamiento de los mercados diario e intradiario de energía eléctrica para su adaptación de los límites de oferta a los límites de casación europeos. *Boletín Oficial del Estado,* 131, pp. 67380–67543.

BOE/SEE. (2012). Resolución de 24 de julio de 2012, de la Secretaría de Estado de Energía, por la que se aprueba la modificación de los procedimientos de operación del Sistema Eléctrico Peninsular (SEP) p.o.-3.1; p.o.-3.2; p.o.-9 y p.o.-14.4 y los procedimientos de operación de los sistemas eléctricos insulares y extrapeninsulares (seie) p.o. seie-1 p.o. seie-2.2; p.o. seie-3.1; p.o. seie-7.1; p.o. seie-7.2; p.o. seie-8.2; p.o. seie-9 y p.o. seie-2.3 para su adaptación a la nueva normativa eléctrica. *Boletín Oficial del Estado,* 120, pp. 57263–57496.

Breiman, L. (2001). Random forests. *Machine Learning,* 45, pp. 5–32.

Kaufman, L. y Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis.* New York: Wiley.

Khobai, H., Mugano, G. y Le Roux, P. (2017). The impact of electricity price on economic growth in South Africa. *International Journal of Energy Economics and Policy,* 7, pp. 108–116.

Nielsen, F. (2016). Hierarchical clustering. En *Introduction to HPC with MPI for Data Science* (pp. 195–211). Springer Cham.

OMIE. Day-ahead Market bids detail. Fecha de acceso: 2022-04-10.

OMIE. Header of bids for Day-ahead Market. Fecha de acceso: 2022-04-10.

OMIE. *Modelo de Ficheros para la distribución pública de Información del mercado de electricidad. Versión 1.33.* Fecha de acceso: 2021-03-11.

Park, H.-S. y Jun, C-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems and Applications,* 36, pp. 3336–3341.

Taha, A. A. y Hanbury, A. (2015). An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 37, pp. 2153–2163.

Warrens, M. J. y van der Hoef, H. (2020). Understanding the Rand index. En *Advanced Studies in Classification and Data Science* (pp. 301–313). Springer.

APÉNDICE

Cuadro 29.

Para cada grupo definido en la figura 10, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	14,24	13,09	14,63	15,43	15,63	13,90	13,09
■	14,23	15,38	14,04	13,93	13,46	14,26	14,70
■	14,31	14,71	14,15	13,44	13,96	14,59	14,83

Fuente: Elaboración propia.

Cuadro 30.

Para cada grupo definido en la figura 14, porcentaje de ofertas que corresponden a cada mes

	En,	*Feb,*	*Mar,*	*Abr,*	*May*	*Jun,*	*Jul,*	*Ag,*	*Sep,*	*Oct,*	*Nov,*	*Dic,*
■	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00
■	8,41	7,80	8,51	8,13	8,39	8,03	8,39	8,39	8,18	8,42	8,52	8,83
■	8,76	7,44	8,57	8,49	8,80	8,72	8,80	8,80	8,38	8,68	7,59	6,98

Fuente: Elaboración propia.

Cuadro 31.

Para cada grupo definido en la figura 14, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	0,00	0,00	0,00	0,00	0,00	61,90	38,10
■	14,33	14,58	14,35	14,33	14,53	13,97	13,89
■	14,19	14,46	14,12	14,12	13,70	14,50	14,91

Fuente: Elaboración propia.

Cuadro 32.

Para cada grupo definido en la figura 16, porcentaje de ofertas que corresponden a cada mes

	En,	*Feb,*	*Mar,*	*Abr,*	*May*	*Jun,*	*Jul,*	*Ag,*	*Sep,*	*Oct,*	*Nov,*	*Dic,*
■	8,85	7,74	9,57	8,07	8,20	7,87	8,13	8,13	7,87	8,20	7,87	9,51
■	8,09	7,73	7,30	8,38	8,80	8,59	8,88	8,88	8,59	8,80	8,59	7,37

Fuente: Elaboración propia.

Cuadro 33.

Para cada grupo definido en la figura 16, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	14,36	14,23	13,97	14,36	14,43	14,36	14,30
■	14,24	14,39	14,67	14,24	14,03	14,10	14,32

Fuente: Elaboración propia.

Cuadro 34.

Para cada grupo definido en la figura 18, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	16,29	13,84	13,52	14,50	14,66	14,66	12,54
■	13,24	14,42	14,10	14,42	14,42	14,99	15,39
■	14,14	14,43	14,79	14,14	13,92	14,21	13,36

Fuente: Elaboración propia.

CAPÍTULO VIII

Análisis y predicción de curvas agregadas de oferta y demanda en el mercado eléctrico europeo

Antonio Muñoz
José Portela
Eugenio Fco. Sánchez-Úbeda
Guillermo Mestre

La predicción de las curvas de oferta en los mercados eléctricos es una herramienta fundamental para el diseño de estrategias de oferta y la planificación de los recursos de generación. Las técnicas de análisis de datos funcionales son idóneas para modelar y predecir estas curvas. Estos métodos incluyen modelos de series temporales funcionales que integran enfoques de reducción de dimensión, junto con métodos no paramétricos y paramétricos. Los modelos resultantes capturan eficazmente las dinámicas complejas y estacionales de las curvas de oferta, haciendo posible la optimización de las estrategias de oferta de los agentes del mercado con un enfoque probabilístico.

Palabras clave: mercados eléctricos, predicción de series temporales, análisis de datos funcionales, estrategias de oferta.

1. INTRODUCCIÓN

En las últimas décadas muchos países han pasado por un proceso de desregulación que ha dado lugar a mercados de electricidad liberalizados que permiten a las empresas comerciar energía en subastas organizadas (Joskow, 2008).

Muchos mercados de electricidad se basan en subastas donde los agentes presentan sus ofertas de venta y compra de energía al operador del mercado, quien luego determina los precios de casación y el conjunto de ofertas aceptadas en cada período de tiempo. Este proceso de liberalización ha introducido desafíos para los agentes del mercado en cuanto a la predicción de los precios de la electricidad, dado que es un factor muy relevante para la toma de decisiones de estas compañías. Por ello, es crítico disponer de técnicas precisas que sean capaces de modelar el proceso generador de precios del mercado.

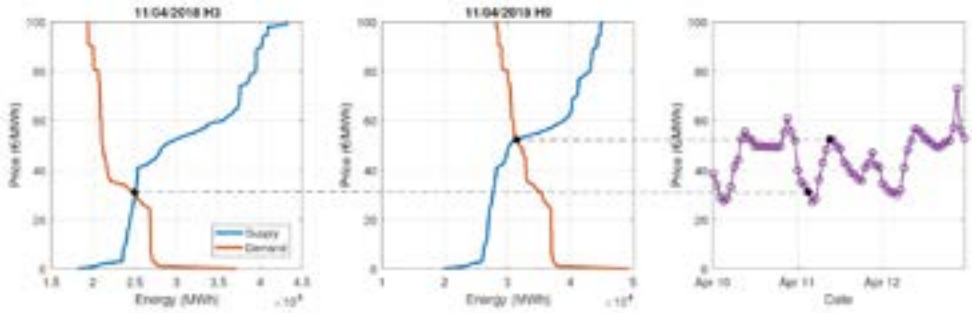
Para entender el funcionamiento de estos mercados de electricidad, consideremos un mercado horario marginalista de oferta simple, es decir, donde hay una subasta de energía para cada hora h del día. En este mercado, cada oferta de venta (o compra) de electricidad se define por un precio p y una cantidad q, que se refiere a la cantidad de energía que el agente está dispuesto a vender (o comprar) a ese precio p. Ordenando las ofertas de venta (compra) en precios crecientes (decrecientes), se construye la curva de oferta horaria de venta (compra) agregada para ese agente. Una vez que todos los agentes han presentado sus ofertas, la suma de todas las curvas de oferta resulta en la curva de oferta de venta del sistema $S_h(p)$, y la suma de las curvas de oferta de compra de cada empresa resulta en la curva de compra (o demanda) del sistema $D_h(p)$. El precio de casación p_h^* se calcula para cada hora como la intersección de las curvas de oferta de compra y venta agregadas del sistema, de ahí que $D_h\left(p_h^*\right) - S_h\left(p_h^*\right) = 0$. Este proceso se ilustra en la figura 1, donde se muestran las curvas de oferta y demanda del mercado diario español para dos horas y los precios de casación resultantes.

Para una empresa generadora que participa en el mercado, es de suma importancia planificar con anticipación para gestionar los recursos disponibles de la manera más eficiente posible. Las previsiones de diferentes indicadores y variables significativas del mercado pueden proporcionar información útil para los agentes. Por ejemplo, la previsión de la demanda permite una gestión eficiente de los recursos, una programación óptima y una planificación de la producción para minimizar los costes de generación (Bunn y Farmer, 1985). Las estimaciones del precio permiten estimar si se van a cubrir los costes de operación y ayuda a protegerse contra los movimientos de precios (Weron, 2014). Como consecuencia, los modelos de previsión de la demanda (Chen *et al.,* 2019) o del precio de casación (Monteiro *et al.,* 2018; Zhang *et al.,* 2019) son ampliamente estudiados y se mejoran continuamente.

Además, todo agente del mercado suele estar interesado en optimizar su estrategia de oferta (Liu *et al.,* 2012). Esto se puede hacer mediante el análisis y la predicción de sus cur-

Figura 1.

Curvas de oferta y demanda para el mercado diario español de electricidad



Nota: La intersección de las curvas determina el precio de casación para cada hora, como se ilustra en la serie de precios horarios mostrados en el panel derecho.

Fuente: Mestre (2021).

vas de demanda residual (*Residual Demand Curves* en inglés, o *RDC*) (Baillo *et al.,* 2004; Campos *et al.,* 2016; Prete y Hobbs, 2015; Xu y Baldick, 2007). La *RDC* de una empresa generadora *i* se puede definir para cada hora como la función que modela el comportamiento agregado de compra y venta de todos los competidores y se puede calcular como:

$$R_h(p) = D_h(p) - S_h^{-i}(p), \qquad [1]$$

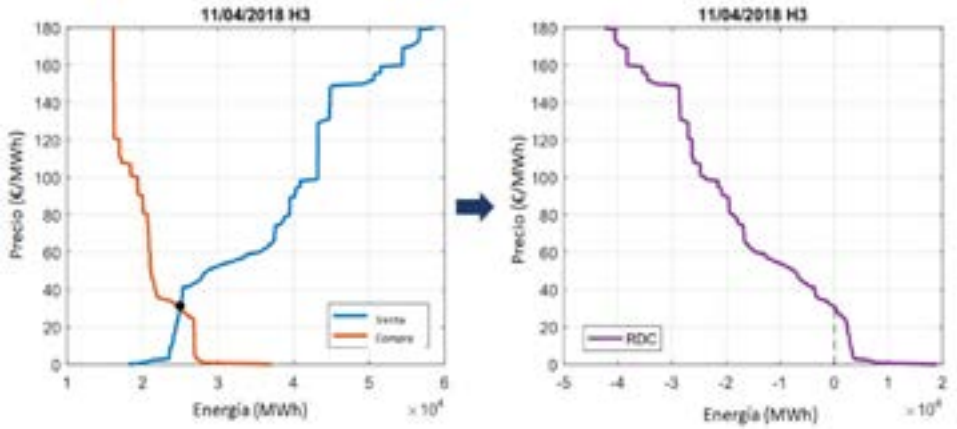
donde $S_h^{-i}(p)$ es la curva de oferta de venta de los competidores de la empresa *i*, que se calcula como la función de oferta del sistema $S_h(p)$ menos la función de oferta de la empresa $S_h^i(p)$, es decir, $S_h^{-i}(p) = S_h(p) - S_h^i(p)$. Para un valor de precio dado *p*, $R_h(p)$ da la cantidad máxima de energía q_i que la empresa generadora *i* puede vender en el mercado en la hora *h*. En Mestre *et al.* (2022) se propone una base matemática para el cálculo formal y eficiente de estas curvas. La figura 2 ilustra un ejemplo de construcción de curva de demanda residual a partir de una curva de compra y de venta.

Además cabe destacar que, en muchos mercados, el sistema de formación de precios se complica dado que existen factores adicionales que afectan a la aceptación de las ofertas. Algunos de estos factores son las limitaciones derivadas de los intercambios de energía entre zonas (que pueden saturar, causando no linealidades en la formación de precio), o las condiciones complejas del mercado ibérico, que permiten a los agentes "salirse" del proceso de casación si no consiguen unos ingresos mínimos diarios para las centrales deseadas. Esto hace que se puedan considerar curvas de demanda residual más complejas que modelen estos efectos, como se describe en Portela *et al.* (2017).

El estudio de las curvas de demanda residual permite caracterizar el comportamiento de los agentes en el mercado e identificar comportamientos anómalos en las estrategias de oferta de los mismos.

Figura 2.

Curvas de oferta y demanda para el mercado eléctrico español (izquierda) y construcción de la curva de demanda residual para esa hora (derecha)



Fuente: Elaboración propia.

Algunos estudios abordan esta tarea utilizando técnicas de agrupamiento como Ugedo *et al.* (2003) y Sánchez-Úbeda *et al.* (2006). Para que las curvas sean fácilmente tratables desde el punto de vista de análisis y aplicación de técnicas de aprendizaje automático se puede muestrear la curva original en unos puntos previamente establecidos. Por ejemplo todas las curvas representadas en la figura 3 han sido previamente muestreadas en 200 puntos uniformemente distribuidos entre dos precios dados. Aparte de esta codificación básica, en la literatura se han propuesto otros mecanismos de codificación más sofisticados basados en modelos que permiten no solamente comprimir los datos, sino también extraer conocimiento importante sobre las curvas de oferta. Por ejemplo, en Sánchez-Úbeda y García-González (2000) se propone utilizar el modelo de bisagras lineales LHM (Sánchez-Úbeda, 1999; Sánchez-Úbeda y Wehenkel, 1998), que resume la curva mediante un conjunto de rectas conectadas, capturando la forma principal de las curvas de oferta y filtrando los pequeños escalones consecutivos.

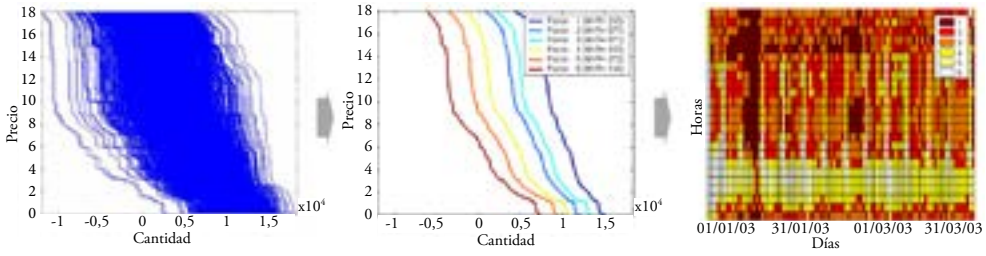
Para analizar el comportamiento temporal de las *RDC,* se pueden agrupar las curvas según patrones tipo suficientemente representativos. En Villar *et al.* (2001) y Collado *et al.* (2004) se propone utilizar técnicas de clustering y de estimación de funciones de densidad para identificar los perfiles. En la figura 3 se muestran los seis patrones obtenidos para

las curvas de demanda residual representadas, indicándose para cada patrón el número de curvas reales que representa (NVR).

Como resultado de este análisis se obtienen los patrones característicos de oferta, que son una descripción simplificada y fácilmente interpretable del comportamiento típico de los agentes en el período analizado. El análisis de la activación temporal de estos patrones y de su relación con otras variables son una fuente de conocimiento del mercado muy importante que permite, por ejemplo, analizar cambios temporales de estrategia de los agentes. La figura 3 representa la secuencia temporal de activación de los patrones obtenidos utilizando la representación mapa día-hora. En esta figura aparece claramente el efecto de las horas de llano, punta y valle, así como el efecto de los días festivos. Los patrones 4, 5 y 6 aparecen básicamente en los valles de los laborables y en los festivos salvo en las horas de punta de la noche.

Figura 3.

Curvas de demanda residual en el mercado eléctrico español (izquierda). Patrones de curvas obtenidos con un modelo de clustering (centro). Activación de los patrones en una matriz día-hora (derecha)



Nota: Cada fila representa una hora y cada columna representa un día.

Fuente: Elaboración propia a partir de Sánchez-Úbeda *et al.* (2006).

Aunque la representación de la figura 3 es muy ilustrativa, es posible utilizar modelos supervisados de aprendizaje automático para explicar no solamente la evolución temporal del patrón activado, sino la dependencia de dicha variable con otras variables explicativas apropiadas para el estudio del mercado que se quiera realizar.

Por último, es relevante destacar que el análisis de las ofertas es una tarea crítica para optimizar la estrategia de oferta de un agente en el mercado. Esta tarea implica analizar una gran cantidad de ofertas en el mercado, ya que diariamente se publican, por ejemplo en el mercado diario español, más de 100.000 ofertas de compra y venta, tanto casadas como ofertadas, resultando en más de 2 millones de registros solo en un mes. Las técnicas de *big data* ofrecen soluciones especialmente orientadas al tratamiento de grandes volúmenes de información.

Dado que las curvas de demanda residual modelan el comportamiento de oferta de los competidores, mediante una estimación de dicha curva se pueden identificar los recursos óptimos para enfrentarse a esa demanda residual de forma que se obtenga un beneficio máximo. Esta propuesta está detallada en Campos *et al.* (2016), donde se desarrolla el modelo de optimización que permite generar una escalera de oferta óptima en base a un conjunto de escenarios estimados de curvas de demanda residual. De esta forma, la oferta del agente se protege contra posibles situaciones de incertidumbre en el mercado.

Como consecuencia, es de gran relevancia para las empresas de electricidad obtener previsiones a corto plazo de las curvas de oferta y las *RDC,* ya que proporcionan una descripción precisa de la estrategia de sus competidores. El objetivo principal de este capítulo es presentar distintas metodologías para la predicción de curvas en el mercado eléctrico e ilustrar su aplicación en un caso real en el mercado italiano.

2. PREDICCIÓN DE CURVAS DE OFERTA

La estadística es una rama de las matemáticas que se ocupa de la recopilación, análisis, interpretación, presentación y organización de datos observados. Estas observaciones pueden ser en forma de escalares, vectores u otros objetos. En particular, los datos que motivan este capítulo se observan en forma de curvas, es decir, cada observación es una función de valor real que toma valores en un conjunto infinito. Por lo tanto, una variable aleatoria funcional X se define como la función $X = \{X(v);\ v \in V\}$. Restringimos el análisis en este capítulo a variables aleatorias funcionales asociadas con un intervalo $V \subset \mathbb{R}$, sin embargo, la noción de variable funcional abarca un área más grande, por ejemplo, superficies aleatorias cuando $V \subset \mathbb{R}^2$. La observación de procesos de funciones continuas aparece en la naturaleza, las ciencias sociales o los sistemas industriales, evidenciando la diversidad de datos funcionales que se pueden encontrar.

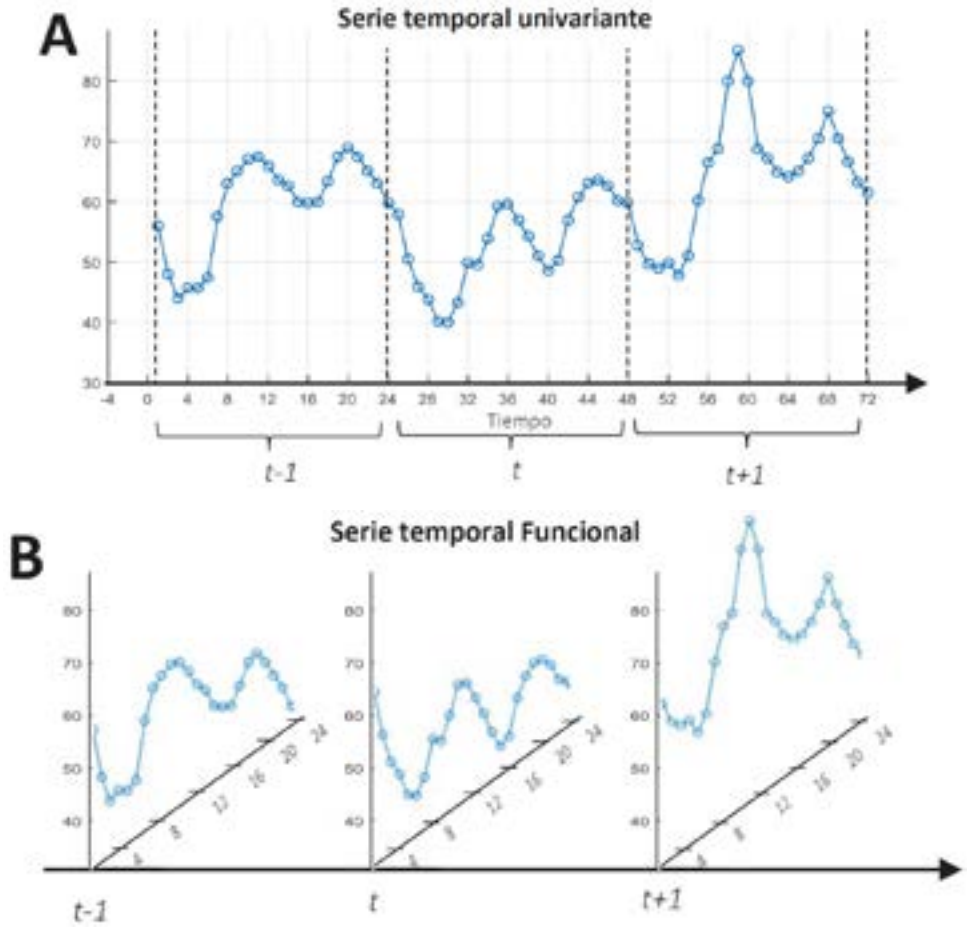
El análisis de datos funcionales (*Functional data analyses*, en inglés, o *FDA*) es el marco estadístico que proporciona las herramientas necesarias para analizar variables funcionales, donde cada observación es una función continua. Entre las diferentes ramas de *FDA,* nos vamos a enfocar en las series temporales funcionales (*Functional time series,* en inglés, o *FTS*), que son secuencias de observaciones funcionales (curvas) observadas en distintos instantes de tiempo.

Se pueden considerar dos tipos de *FTS.* Por un lado, las *FTS* pueden originarse a partir de un proceso continuo en el tiempo que se divide en segmentos de igual longitud, obteniendo una secuencia temporal de segmentos, es decir, una serie temporal funcional. Un ejemplo de este tipo de proceso sería la evolución de la temperatura en un lugar determinado como función del tiempo, que se divide en segmentos diarios, obteniendo así una secuencia de perfiles de temperatura diarios. La figura 4 ilustra esta transformación del proceso continuo (Parte A) en el proceso funcional (Parte B). El otro tipo de *FTS* se

da cuando las observaciones son funciones *per se* cuyo dominio no necesariamente tiene que ser el tiempo.

Figura 4.

Series temporales escalares vs. series temporales funcionales



Notas: Parte A. Un ejemplo de una serie temporal escalar univariante. Cuando se divide en segmentos de igual longitud, se obtiene una serie temporal funcional. Parte B. Series temporales funcionales. Para cada instante de tiempo *t*, se observa una función continua.

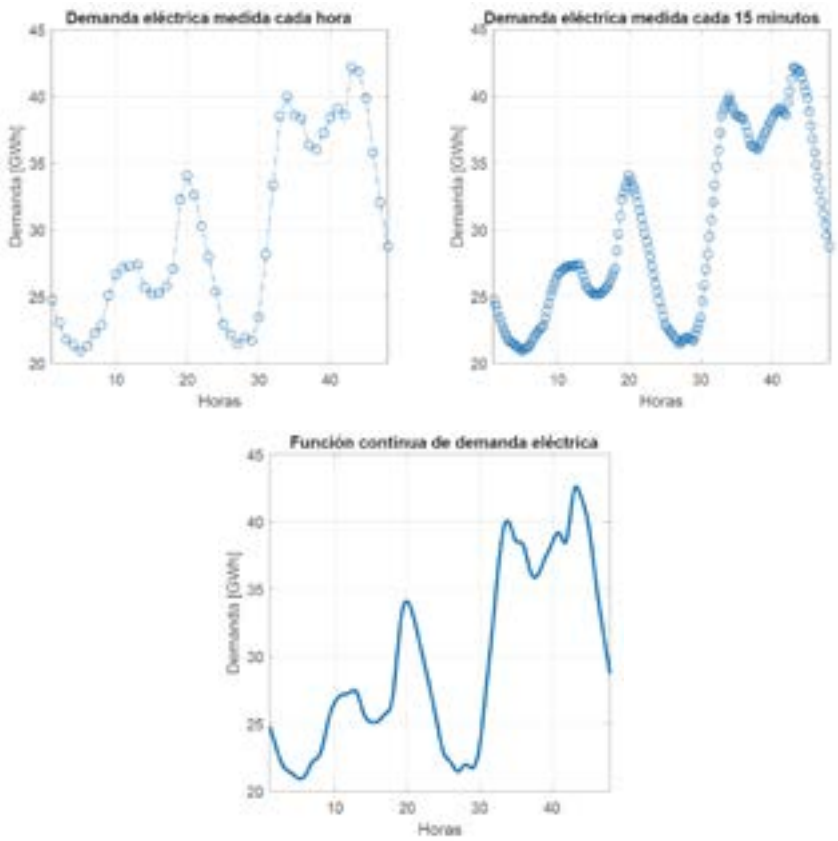
Fuente: Portela (2017).

Las series temporales funcionales también se pueden encontrar en numerosas aplicaciones de predicción en los mercados de electricidad. A continuación, se detallan algunos ejemplos:

- *Demanda de energía eléctrica.* El consumo de energía es un proceso continuo en el tiempo que puede ser analizado con métodos de análisis de datos funcionales. La predicción a corto plazo de la demanda es crucial para los agentes del mercado y los operadores del sistema. En los últimos años, la medición inteligente está aumentando la frecuencia de muestreo del consumo de electricidad de valores horarios a datos cada media hora o cada pocos minutos, como se representa en la figura 5. Por lo tanto, la predicción de la demanda está evolucionando hacia una predicción continua. Algunas aplicaciones de *FDA* a la predicción de la demanda eléctrica se pueden ver en Paparoditis y Sapatinas (2013).

Figura 5.

El efecto de la medición inteligente en la demanda de electricidad



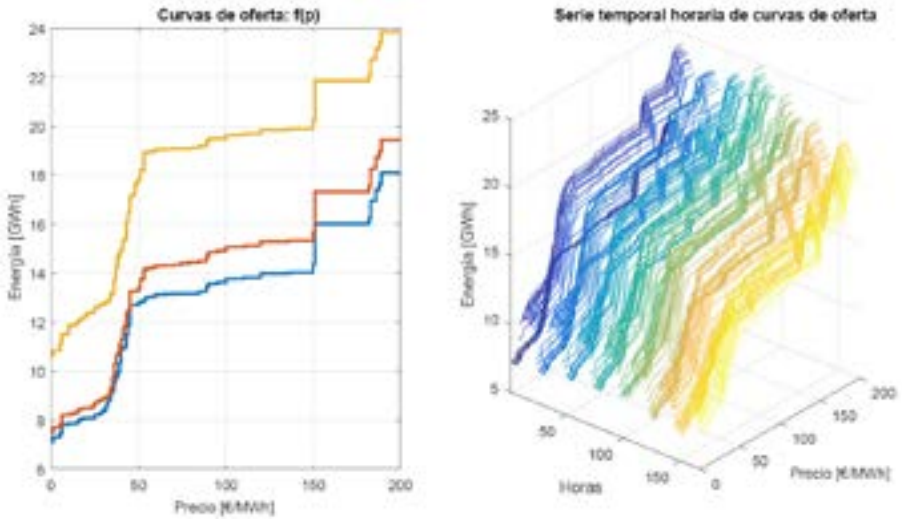
Nota: Tradicionalmente, se disponía de mediciones horarias. Con los medidores inteligentes, la frecuencia de muestreo se acerca a la función continua.

Fuente: Portela (2017).

- *Precios de la electricidad.* Hemos mencionado anteriormente la relevancia de obtener predicciones de precios de casación de los mercados eléctricos. Aunque la serie temporal resultante de los precios horarios sea discreta, la secuencia de perfiles de precios diarios puede analizarse como una serie temporal funcional. Se pueden encontrar algunos ejemplos de predicción de precios con un enfoque funcional en Vilar *et al.* (2012), Liebl (2013) y Galeano (2021).
- *Curvas de oferta.* En la gran mayoría de los mercados eléctricos, como es el caso del mercado ibérico o italiano, las ofertas de compra y venta de energía ofertadas por los agentes son publicadas diariamente. Esta información permite construir curvas agregadas de oferta, dando lugar a series temporales que por su propia naturaleza son funcionales. La figura 6 muestra un ejemplo visual de curvas agregadas de venta horarias, donde se puede observar que son curvas escalonadas no decrecientes.

Figura 6.

Curvas de oferta en los mercados de electricidad



Nota: A la izquierda, se muestran tres curvas de oferta definidas en el rango de precios de 0 a 200 euros/MWh. A la derecha, se traza la secuencia horaria de curvas de oferta para una semana. Para cada hora, se observa una función definida en el rango de precios de 0 a 200 euros/MWh.

Fuente: Portela (2017).

- *Curvas de demanda residual.* Una curva de demanda residual expresa el precio de casación del mercado como una función de la cantidad de energía que el agente está dispuesto a comprar o vender. De manera similar a las curvas de oferta, se obtiene una función (curva) para cada hora y los agentes pueden usar predicciones de las *RDC* para optimizar su estrategia de oferta (Campos *et al.,* 2016). El enfoque funcional

se ajusta a esta aplicación, como se muestra en Aneiros *et al.* (2013), que hace uso de métodos no paramétricos funcionales.

Estas series temporales que surgen en el ámbito de los mercados de electricidad comparten algunos factores comunes. El efecto de las actividades comerciales y cotidianas conduce a estacionalidades semanales y diarias. Además, hay variables explicativas que afectan de forma causal a muchas de estas series. Por ejemplo, el clima (velocidad del viento, precipitaciones, etc.) afecta la producción de tecnologías renovables con costes de generación más bajos, influyendo en el comportamiento de oferta de los agentes y, por tanto, al precio de casación. En consecuencia, el desarrollo de modelos de predicción funcional para los mercados de electricidad debe tener en cuenta estas propiedades y considerar el efecto de las variables exógenas y la estacionalidad, entre otros.

En la literatura sobre predicción de series temporales funcionales se pueden encontrar algunos modelos de referencia que podrían aplicarse en los mercados de electricidad. La siguiente sección se centrará en realizar una revisión de los diferentes enfoques más comunes.

2.1. Predicción de series temporales funcionales

En las últimas dos décadas hemos sido testigos de un creciente interés por los modelos de predicción de series temporales funcionales, así como del desarrollo de nuevas técnicas para modelar las dinámicas complejas que, a menudo, exhiben estas series. Esta sección está dedicada a revisar las diferentes técnicas de modelado encontradas en la literatura de datos funcionales, clasificándolas en diferentes grupos según las herramientas estadísticas que utilizan: reducción de dimensionalidad, métodos de estimación no paramétricos y métodos paramétricos.

2.1.1. Modelos basados en reducción de la dimensión

Un enfoque común para predecir series temporales funcionales es proyectar las observaciones funcionales en un espacio finito-dimensional adecuado para aplicar modelos multivariantes y así estimar valores futuros de la serie.

En la mayoría de las aplicaciones, las curvas se proyectan en sus primeras componentes principales funcionales (*Functional principal components,* en inglés, *FPC*) (ver Hall *et al.,* 2006; Yao *et al.,* 2005, para más información), reduciendo efectivamente la dimensión de los datos. De esta forma, cada observación $Y_t(v)$ de la serie temporal funcional se puede representar de forma aproximada como una combinación lineal de las K autofunciones $\phi_j(v)$, $j = 1 \ldots K$, lo que se conoce como la representación de Karhunen-Loéve (ver Yao *et al.,* 2005, para más detalle).

$$Y_t(v) \approx \sum_{j=1}^{K} \xi_{t,j} \phi_j(v), \qquad [2]$$

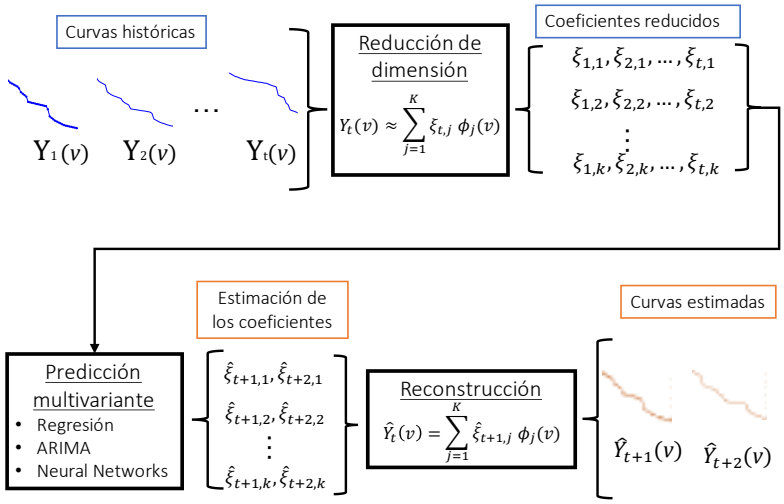
donde $\xi_{t,j}$ son las coordenadas o *scores* para cada curva t en la base de autofunciones. De esta forma, la serie funcional original se transforma en un conjunto de series temporales multivariantes en el cual se pueden aplicar modelos de predicción de serie temporal escalar para predecir los valores futuros de las coordenadas $\hat{\xi}_{t+1,j}$. Luego, se pueden reconstruir las curvas futuras estimadas como:

$$\hat{Y}_{t+1}(v) = \sum_{j=1}^{K} \hat{\xi}_{t+1,j} \phi_j(v). \qquad [3]$$

La figura 7 ilustra este proceso de descomposición en componentes principales, su predicción y su posterior reconstrucción.

Figura 7.

Proceso de predicción de series funcionales basado en técnicas de reducción de la dimensión por componentes principales funcionales



Fuente: Elaboración propia.

Los modelos de predicción funcionales basados en la reducción de la dimensionalidad se pueden clasificar según los métodos de predicción utilizados para estimar las series temporales de coeficientes. Por ejemplo, Erbas *et al.* (2007), Hyndman y Shang (2009), Hyndman y Ullah (2007), Valderrama *et al.* (2002), Wagner-Muns *et al.* (2018) proyectan las curvas en el espacio abarcado por las primeras componentes principales funcionales de la serie temporal funcional y luego ajustan modelos ARIMA univariantes a cada serie de coeficientes. También se pueden aplicar modelos autorregresivos vectoriales como en Aue *et al.* (2015), Klepsch *et al.* (2017), Sen y Klüppelberg (2019) o incluir variables explicativas para la predicción como en Aue *et al.* (2015). El trabajo de Shang (2012) resume diferentes técnicas de predicción funcional que involucran métodos de descomposición en *FPC*.

2.1.2. *Modelos no-paramétricos*

A diferencia de la metodología anterior, los modelos no paramétricos y paramétricos no reducen explícitamente la dimensión del conjunto de datos funcional original. Estas técnicas asumen que la serie temporal funcional sigue algún tipo de relación donde se utilizan operadores lineales funcionales para modelar esta dependencia temporal. El modelo de referencia para predecir series temporales funcionales es el proceso de Hilbert autorregresivo de orden 1, véase Bosq (2000), denotado por ARH(1), que se define como:

$$Y_t = \Psi(Y_{t-1}) + \varepsilon_t. \qquad [4]$$

Aquí, Ψ es el operador autorregresivo que necesita ser estimado y ε_t denota un proceso de ruido blanco funcional. Según las técnicas utilizadas para estimar estos operadores lineales, los modelos de predicción funcional pueden clasificarse como modelos paramétricos o no paramétricos.

El marco no paramétrico no asume una estructura fija para los operadores lineales, sino que realiza una ponderación local de las variables de entrada para estimar los operadores del modelo (Ferraty y Vieu, 2006). El estimador de kernel de Nadaraya-Watson (Nadaraya, 1964; Watson, 1964) se utiliza a menudo para realizar esta ponderación local de las variables, seleccionando los valores históricos más similares en el conjunto de datos disponible para predecir situaciones futuras. Como tal, el operador ARH(1) se puede estimar como:

$$\hat{\psi}\left(Y_{t-1}\right) = \sum_{j=1}^{t-2} \omega_h\left(Y_{t-1}, Y_j\right) Y_{j+1} \qquad [5]$$

siendo $\omega_h(.,.)$ la función de ponderación dada por

$$\omega_h\left(Y_{t-1}, Y_j\right) = \frac{K\left(h^{-1} d\left(Y_{t-1}, Y_j\right)\right)}{\sum_{i=1}^{t-2} K\left(h^{-1} d\left(Y_{t-1}, Y_i\right)\right)}, \qquad [6]$$

donde K denota una función kernel (como la función de densidad gaussiana estándar) y $h > 0$ se llama el parámetro de ancho de banda del kernel, que representa el nivel de suavizado que se aplicará a los datos y necesita ser estimado a partir de una muestra. La expresión [5] puede interpretarse como una media ponderada de observaciones funcionales pasadas, donde la similitud entre las observaciones se cuantifica mediante la función de distancia $d(.,.)$. La elección de dicha función de distancia es crítica, ya que se utilizará para resaltar las características relevantes de las curvas.

La monografía de Ferraty y Vieu (2006) proporciona una excelente introducción a las técnicas no paramétricas que ayudaron a popularizar estos métodos estadísticos. En aplicaciones del mundo real, la inclusión de covariables exógenas es de suma importancia, ya que a menudo son los principales impulsores de las series temporales funcionales. Aneiros y Vieu (2008) introdujeron un modelo semiparamétrico que permite la inclusión de variables exógenas en el modelo autorregresivo no paramétrico. Este modelo ha sido aplicado con éxito en la

predicción de la demanda y el precio de la electricidad (Aneiros *et al.,* 2016; Vilar *et al.,* 2018; Vilar *et al.,* 2012) y en la predicción no paramétrica de curvas de demanda residual (Aneiros *et al.,* 2013).

2.1.3. Modelos paramétricos

Según Ferraty y Vieu (2006), los modelos paramétricos para la predicción de series temporales funcionales pueden definirse como modelos que asumen que los operadores que aparecen en la formulación del modelo de predicción pertenecen a alguna familia conocida, como los operadores integrales.

Los operadores integrales aparecen, por ejemplo, en la extensión del modelo de regresión a datos funcionales. Considerando dos variables funcionales $X(u)$ y $Y(v)$ con media cero, este modelo relaciona la variable de salida funcional $Y(v)$, definida en algún intervalo $v \in V_Y$, con la variable de entrada funcional $X(u)$, definida en algún intervalo $u \in V_X$. El operador integral $\Psi(\cdot)$, por tanto, tiene la forma:

$$Y(v) = \Psi(X)(v) = \int \psi(u, v)X(u)du,$$

donde $\Psi(X(u))$ pertenece a la clase de operadores lineales en L^2 llamados operadores integrales. $\psi(u, v)$ es el núcleo del operador y puede considerarse como el parámetro funcional. La figura 8 muestra una representación visual de un operador integral funcional. En cualquier punto dado v, el valor de $Y(v)$ depende de toda la trayectoria de $X(u)$. Es una extensión directa de los modelos lineales tradicionales con respuesta multivariante y covariables vectoriales.

El enfoque más habitual en la literatura para estimar el parámetro funcional es proyectar el kernel de regresión en alguna base funcional:

$$\psi(u,v) = \sum_{l=1}^{L}\sum_{m=1}^{M} b_{l,m} e_l(u) f_m(v),$$

donde $b_{l,m}$ son coeficientes escalares y e_l y f_m son algunas funciones base que no necesitan ser ortonormales. Las funciones base e_l y f_m suelen ser elegidas por el usuario, y los parámetros $b_{l,m}$ se optimizan para minimizar el error de predicción. Las funciones base más comúnmente utilizadas son las componentes principales funcionales (Faraway, 1997), aunque también hay otras opciones disponibles como se ve en Antoch *et al.* (2010) que utiliza un estimador de B-spline.

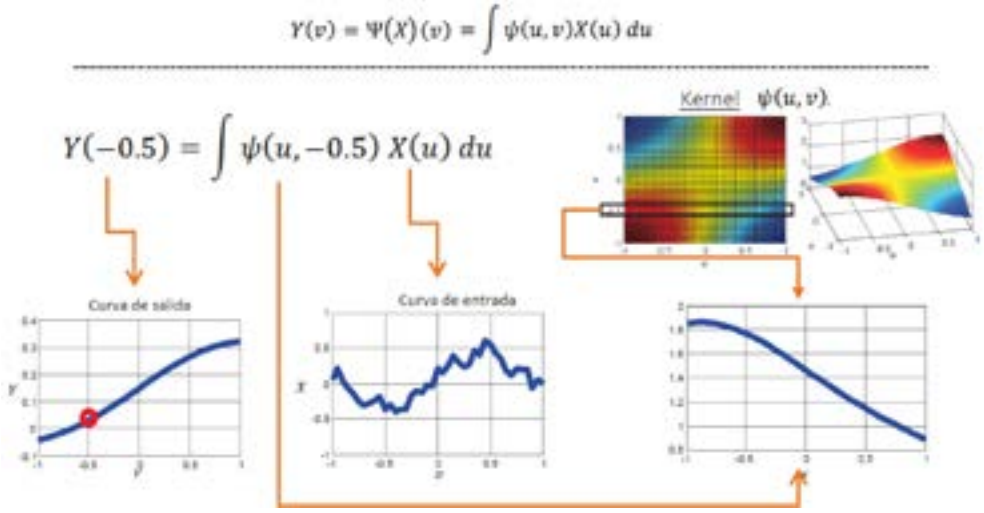
En cuanto al proceso autorregresivo de orden uno ARH(1) de la ecuación [4], Bosq (2000) propone estimar Ψ utilizando un conjunto equivalente de ecuaciones de Yule-Walker para datos funcionales. El resultado es un operador integral cuya función kernel toma la forma

$$\hat{\psi}(u,v)=\frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{j=1}^{M}\sum_{i=1}^{M}\hat{\lambda}^{-1}\left\langle Y_t,\hat{\phi}_j\right\rangle\left\langle Y_{t+1},\hat{\phi}_i\right\rangle\hat{\phi}_j(u)\hat{\phi}_i(v), \quad [7]$$

donde $\left\{\hat{\lambda}_j\right\}_{j=1}^{M}$ denota los autovalores estimados del operador de covarianza de los datos y $\left\{\hat{\phi}_j\right\}_{j=1}^{M}$ son las primeras componentes principales funcionales de la serie de curvas. Esto puede interpretarse como encontrar las coordenadas óptimas de la función de kernel $\hat{\psi}(u,v)$ cuando se proyecta en la base funcional definida por las primeras componentes principales del conjunto de datos funcional.

Figura 8.

Representación visual de un operador integral funcional



Nota: En el ejemplo, la curva de salida *Y* en el punto (–0,5) es el resultado de la integral de la multiplicación de la entrada *X* y la rebanada del operador en -0,5.
Fuente: Portela (2017).

Este modelo ha sido ampliamente estudiado en la literatura. Por ejemplo, Mas (2007) y Álvarez *et al.* (2017) analizan las propiedades asintóticas del modelo ARH(1). Otros métodos de estimación para el modelo ARH(1) se ilustran en Didericksen *et al.* (2012) o Kargin y Onatski (2008). El trabajo de Horváth *et al.* (2020) compara las capacidades predictivas de varias de estas técnicas de estimación en una aplicación financiera del mundo real.

Sin embargo, en muchas aplicaciones el modelo ARH(1) puede no ser suficiente, ya que la dependencia temporal que exhiben las curvas puede estar influenciada por distintos retardos de la serie temporal funcional. Los trabajos de Bosq (2000), Mourid (2002) proporcionan un estudio en profundidad de los procesos autorregresivos funcionales de orden $p > 1$, denominados como ARH(p).

La inclusión de covariables exógenas en el modelo ARH(p) fue analizada en Damon y Guillas (2002, 2005), donde los autores analizan las propiedades teóricas y empíricas de este modelo. Este modelo, denominado como ARHX(p), se puede formular como:

$$Y_t = \sum_{j=1}^{p} \Psi_j \left(Y_{t-j} \right) + \sum_{z=1}^{Z} \Gamma_z \left(X_t^z \right) + \varepsilon_t, \qquad [8]$$

donde Ψ_j son los operadores autorregresivos, Γ_z denota el operador de regresión asociado con la covariable funcional X_t^z y ε_t son las innovaciones i.i.d. del modelo. El modelo de media móvil funcional MAH(1) fue introducido en Turbillon *et al.* (2007). Además, Klepsch *et al.* (2017) estudian el modelo ARMA funcional, denotado por ARMAH(p, q). Los autores aproximan el modelo:

$$Y_t = \sum_{j=1}^{p} \Psi_j \left(Y_{t-j} \right) + \sum_{i=1}^{q} \Theta_i \left(\varepsilon_{t-i} \right) + \varepsilon_t \qquad [9]$$

por un modelo ARMA vectorial, obtenido después de proyectar ambas curvas, innovaciones y operadores en el espacio abarcado por el FPC de las series temporales funcionales originales.

Sin embargo, para modelar las dinámicas estacionales presentes en las series funcionales del mercado diario, Portela (2017) y Portela *et al.* (2018) proponen el modelo SARMAHX, un modelo paramétrico que es una generalización del modelo ARMA estacional escalar con covariables exógenas. Dado que ha demostrado ser un modelo competitivo en comparación con otros modelos de predicción de referencia gracias a su capacidad para capturar dinámicas complejas, la siguiente sección presentará este modelo junto con sus principales propiedades y suposiciones.

2.2. El modelo SARMAHX

El modelo funcional SARMAHX se define siguiendo el enfoque estándar de modelado de series temporales propuesto en Box *et al.* (2008), pero extendido a series temporales funcionales utilizando operadores integrales. El modelo SARMAHX generaliza el modelo ARMAX escalar extendiendo el modelo escalar a series temporales funcionales utilizando operadores integrales definidos en el espacio de Hilbert L^2 como los parámetros del modelo. Mientras que los parámetros del modelo ARMAX clásico son valores escalares, los parámetros del modelo SARMAHX son operadores funcionales que modelan la relación entre las curvas de entrada y salida. Para estimar estos parámetros, se propuso una nueva metodología en Portela (2017), que difiere sustancialmente de los métodos de estimación paramétrica descritos en la sección anterior.

Para más detalles, pueden referirse a Portela (2017) y Mestre *et al.* (2020).

2.2.1. Formulación del modelo

El modelo SARMAHX(P_0, Q_0)×(P_1, Q_1)$_{s1}$ ×(P_2, Q_2)$_{s2}$ es un modelo funcional autorregresivo de media móvil en espacio de Hilbert con dos estacionalidades (aunque podría gene-

ralizarse a cualquier número de estacionalidades) que incluye tanto variables explicativas funcionales como escalares. La expresión completa para el modelo se define de la siguiente manera:

$$\prod_{j=0}^{2}\left(I-\sum_{i=1}^{P_j}\Psi_{j,i}B^{i\cdot s_j}\right)\left(Y_t\right)=\prod_{k=0}^{2}\left(I-\sum_{l=1}^{Q_k}\Theta_{k,l}B^{l\cdot s_k}\right)\left(\varepsilon_t\right)+\sum_{z=1}^{Z_f}\Gamma_z^f\left(X_t^z\right)+\sum_{z=1}^{Z_c}\Gamma_z^c\left(x_t^z\right), \quad [10]$$

donde:

- $\{Y_t(v);\ t = 1, 2, ..., T;\ v \in V\}$ es una serie temporal funcional estacionaria de media cero.
- $\{X_t^z(v_z);\ z = 1, 2, ..., Z_f;\ t = 1, 2, ..., T;\ v_z \in V_z\}$ un conjunto Z_f de variables exógenas funcionales.
- $\{x_t^z;\ z = 1, 2, ..., Z_c;\ t = 1, 2, ..., T\}$ un conjunto Z_c de variables exógenas escalares.
- ε_t un proceso de ruido blanco funcional.
- I es el operador identidad.
- Los parámetros P_0, P_1 y P_2 son los órdenes autorregresivos regular y estacionales, respectivamente.
- Los parámetros Q_0, Q_1 y Q_2 son los órdenes de media móvil regular y estacionales, respectivamente.
- Los parámetros s_1 y s_2 son los períodos estacionales. El parámetro s_0 es igual a 0.
- $\Psi_{0,i}$, $\Psi_{1,i}$ y $\Psi_{2,i}$ son los operadores autorregresivos regulares y estacionales.
- $\Theta_{0,l}$, $\Theta_{1,l}$ y $\Theta_{2,l}$ son los operadores de media móvil regulares y estacionales.
- Γ_z^f son los operadores relacionados con las variables explicativas Z_f.
- Γ_z^c son los operadores relacionados con las variables explicativas Z_c.
- B^n es el operador de retardo que se define como $B^nY_t = Y_{t-n}$ donde $n \in \mathbb{N}$. Si se utilizan operadores integrales definidos en el espacio L^2 para los términos ARMA, se obtiene el modelo SARMAHX completamente funcional. Esta versión del modelo, ilustrada en Portela *et al.* (2018) en un problema de predicción de precios de electricidad, es capaz de capturar la dependencia de la correlación cruzada entre diferentes puntos de las curvas.

Es importante notar que, cuando se utiliza este modelo para la predicción, los términos asociados a la media móvil (retardos de ε_t) no se observan.

Para facilitar la comprensión del modelo SARMAHX, se detalla el siguiente ejemplo. El modelo SARMAHX(1, 1) $\times(1, 0)_{24}$ se definiría de la siguiente manera:

$$Y_t = \Psi_{0,1}(Y_{t-1}) + \Psi_{1,1}(Y_{t-24}) - \Psi_{0,1}\Psi_{1,1}(Y_{t-25}) - \Theta_{0,1}(\varepsilon_{t-1}) + \varepsilon_t, \qquad [11]$$

donde los términos son, respectivamente, el autorregresivo regular, el autorregresivo estacional, la interacción estacional autorregresiva y el término de media móvil. Cabe señalar que el término $\Psi_{0,1}\Psi_{1,1}(Y_{t-25})$ denota la composición, es decir, $\Psi_{0,1}(\Psi_{1,1}(Y_{t-25}))$. Luego, sustituyendo cada operador por su expresión integral, la ecuación de predicción se convierte en:

$$\begin{aligned}\hat{Y}_t(v') = &\int \psi_{0,1}(u,v')Y_{t-1}(u)\,du + \int \psi_{1,1}(u,v')Y_{t-24}(u)\,du \\ &- \int\int \psi_{0,1}(v,v')\psi_{1,1}(u,v)Y_{t-25}(u)\,du\,dv \\ &- \int \theta_{0,1}(u,v')\hat{\varepsilon}_{t-1}(u)\,du\end{aligned} \qquad [12]$$

Como se observa, el modelo resultante admite una amplia variedad de configuraciones: términos autorregresivos y de media móvil hasta dos estacionalidades, así como la inclusión de variables explicativas escalares y funcionales. Por lo tanto, este modelo es adecuado para la mayoría de las series temporales funcionales que se encuentran en muchas aplicaciones del mundo real, como la predicción de precios y demandas en los mercados de electricidad.

2.2.2. Estimación del modelo

Para ajustar el modelo SARMAHX, cada operador integral debe ser estimado a partir de los datos observados. Como ya se mencionó en la sección anterior, estimar un operador integral Ψ implica estimar la superficie del núcleo o kernel asociado $\psi(u, v)$. El modelo SARMAHX sigue un enfoque novedoso para la estimación de parámetros funcionales: cada función kernel $\psi(u, v)$ se modela como una suma finita de funciones sigmoidales (Portela, 2017, Portela *et al.,* 2018). Las funciones sigmoidales son aproximaciones universales de funciones (Cybenko, 1989) que se utilizan comúnmente en redes neuronales debido a sus propiedades para modelar relaciones no lineales. De esta manera, cada núcleo bivariante $\psi(u, v)$ puede ser modelado como:

$$\psi(u,v) = \alpha_0 + \sum_{g=1}^{G_\psi} \alpha_g \tanh\left(w_{g0} + w_{g1}u + w_{g2}v\right), \qquad [13]$$

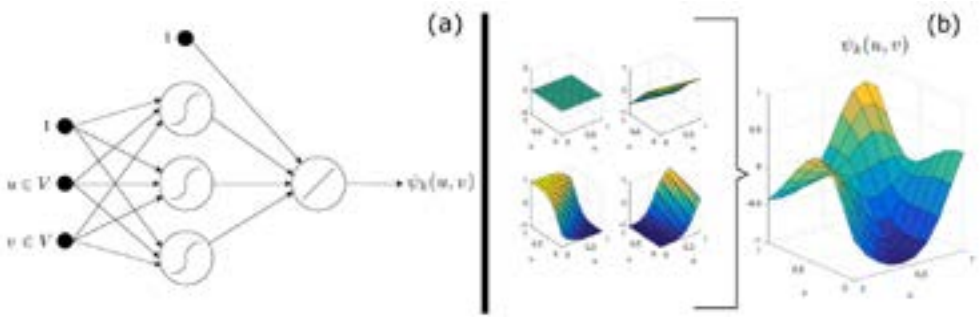
donde w_{g0}, w_{g1}, w_{g2}, α_g y α_0 son los parámetros que definen las sigmoidales y su combinación. Las variables u y v toman valores reales en los intervalos en los que se definen las variables funcionales.

En el caso de un operador concurrente (utilizado para modelar relaciones respecto de variables explicativas escalares), esta expresión se simplifica, modelando cada núcleo como una suma finita de sigmoidales:

$$\psi(v) = \alpha_0 + \sum_{g=1}^{G_\psi} \alpha_g \tanh\left(w_{g0} + w_{g1}v\right). \qquad [14]$$

Figura 9.

Diagrama de arquitectura de la red neuronal utilizada para optimizar los parámetros funcionales del modelo SARMAHX (Parte a). Función kernel estimada como suma ponderada de 3 funciones sigmoidales y una constante (Parte b)



Fuente: Mestre (2021).

Este enfoque puede verse como una red neuronal de perceptrón multicapa (*MLP*) con una configuración particular: una capa de entrada con dos variables de entrada y un sesgo; una capa oculta con un número G_ψ de unidades ocultas no lineales con la tangente hiperbólica como función de activación y w_{g0}, w_{g1}, w_{g2} como los pesos para cada entrada. Finalmente, una capa de salida con una unidad de salida lineal que tiene α_g como los pesos para la activación de las unidades ocultas. La figura 9 (Parte a) muestra el diagrama de arquitectura de la red mencionada con $G_\psi = 3$ capas ocultas. La propiedad de las superficies sigmoidales como aproximadores universales se ilustra en la Parte b, donde una superficie bastante compleja se modela como la suma de tres funciones sigmoidales y una superficie constante que modela el nivel de la superficie final.

Se ha observado empíricamente que el uso de cinco o seis funciones sigmoidales al ajustar el modelo SARMAHX es suficiente en la gran mayoría de las aplicaciones prácticas, debido a la flexibilidad de las funciones sigmoidales.

Los parámetros (w_{g0}, w_{g1}, w_{g2}, α_g, α_0) definen completamente cada núcleo bivariado en [13]. Por lo tanto, el modelo SARMAHX propuesto se estima cuando se estiman los valores para todos estos parámetros. Para lograr esto, se ha implementado un método de Quasi Newton de baja memoria con pesos iniciales aleatorios para optimizar estos parámetros reales con el fin de minimizar una función de coste. En Portela (2017), la función de coste C para estimar el modelo SARMAHX se define como la suma de los errores cuadrados L^2,

$$C = \sum_{t=1}^{T} e_t, \qquad [15]$$

donde:

$$e_t = \left\| Y_t - \hat{Y}_t \right\|_{L^2}^2 = \int \left(Y_t(u) - \hat{Y}_t(u) \right)^2 du. \qquad [16]$$

El método Quasi Newton implementado es un algoritmo que utiliza el gradiente para determinar la dirección de búsqueda, por lo que se necesitan las derivadas del término de error con respecto a los parámetros de las sigmoidales. Esta derivada de la función de error con respecto a un parámetro general *W* viene dada por:

$$\frac{\partial C}{\partial W} = \sum_{t=1}^{T} \int 2\left(Y_t(u) - \hat{Y}_t(u) \right) \left(-\frac{\partial \hat{Y}_t(u)}{\partial W} \right) du, \qquad [17]$$

donde $\frac{\partial \hat{Y}_t(u)}{\partial W}$ es la derivada parcial de la estimación con respecto al parámetro genérico *W*. Para reducir los tiempos de cálculo, se han obtenido expresiones analíticas para estas derivadas. Se remite al lector a Portela *et al.* (2018) para la formulación de las derivadas de la función de coste con respecto a los parámetros de peso y otros detalles de la implementación.

Por último, para evitar el sobreajuste, en el proceso de entrenamiento se utiliza la técnica del *early stopping,* obteniendo el conjunto de valores de los parámetros que obtiene un menor error en un conjunto de validación.

La inclusión de covariables exógenas (ya sean variables escalares o funcionales) y la capacidad de incluir hasta dos términos estacionales hace posible capturar las dinámicas complejas que exhiben las series temporales funcionales horarias, como las series de curvas de oferta o curvas de demanda residual.

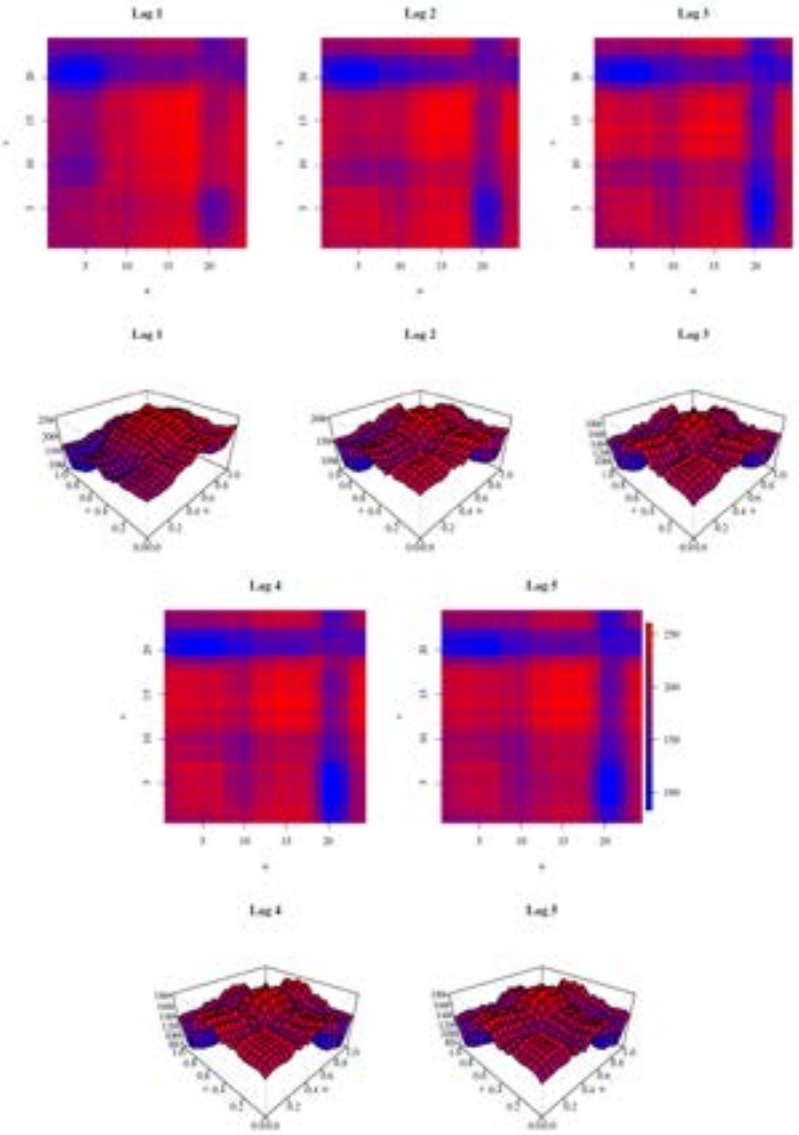
Sin embargo, antes de ajustar el modelo SARMAHX, el usuario necesita seleccionar tanto la estacionalidad presente en los datos como el orden de los términos autorregresivos y de media móvil. Por lo tanto, se presenta a continuación una técnica de identificación y diagnóstico para evaluar la adecuación de estos modelos SARMAHX funcionales.

2.3. Identificación del modelo SARMAHX

El trabajo de Mestre *et al.* (2021) desarrolla una metodología análoga a la de Box-Jenkins para la identificación de series temporales funcionales. Para ello, es fundamental caracterizar la dinámica temporal de la serie funcional, lo que suele hacerse analizando la correlación existente entre distintos retardos de la propia serie. En series temporales funcionales, se puede definir un procedimiento análogo.

Figura 10.

Primeras cinco funciones de autocovarianza retardadas para la *FTS* de los perfiles diarios de precios de la electricidad (arriba) y una representación tridimensional de las superficies (abajo)



Nota: Estas funciones proporcionan información detallada, aunque quizás difícil de interpretar, sobre la estructura de covarianza entre las curvas Y_t y Y_{t-h}.

Fuente: Mestre (2021).

En primer lugar vamos a introducir la función de autocovarianza, que mide la variación conjunta de las funciones de una *FTS* en diferentes instantes de tiempo. Se puede generalizar la estimación de la autocovarianza de una serie temporal al caso funcional como:

$$\widehat{c_h}(u,v) = \frac{1}{T}\sum_{i=1}^{T-h}\left(Y_i(u) - \hat{\mu}(u)\right)\left(Y_{i+h}(v) - \hat{\mu}(v)\right), \qquad [18]$$

donde

$$\hat{\mu}(u) = \frac{1}{T}\sum_{i=1}^{T} Y_i(u) \qquad [19]$$

se refiere a la media muestral funcional.

Por ejemplo, la figura 10 ilustra las superficies de autocovarianza para diferentes retardos de los perfiles diarios de precios de la electricidad del mercado eléctrico español en 2014. Se puede observar que los precios de la electricidad para las horas 12 a 18 están correlacionados con las curvas pasadas, y que las horas 19 a 22 están menos influenciadas por las curvas de precios anteriores.

A partir de esta función de autocovarianza, y siguiendo a Kokoszka *et al.* (2017), se puede estimar el coeficiente de autocorrelación funcional para el retardo *h* como:

$$\widehat{\rho_h} = \frac{\|\widehat{c_h}\|}{\int \widehat{c_h}(u,u)\,du}, \|\widehat{c_h}\|^2 = \int\int \widehat{c_h}^2(u,v)\,dudv. \qquad [20]$$

Dado que un proceso de ruido blanco funcional no exhibe autocorrelación entre sus términos, la norma de sus operadores de autocovarianza retardados debe ser cercana a cero para cada retardo positivo *h*. En Mestre *et al.* (2021), se indica un procedimiento para obtener los límites de confianza para los valores de $\widehat{\rho_h}$ bajo la hipótesis de que la serie es ruido blanco.

Los métodos clásicos de identificación para series temporales escalares, como la metodología de Box–Jenkins (Box *et al.,* 2008), utilizan las funciones de autocorrelación y autocorrelación parcial muestrales de la serie temporal para comprobar la hipótesis de ruido blanco del residuo y para identificar la estructura de correlación subyacente de la serie temporal. Así, por ejemplo, el patrón de decrecimiento de la autocorrelación y los retardos específicos con una fuerte autocorrelación pueden ser utilizados para seleccionar los órdenes adecuados de autorregresión y de media móvil en los modelos de series temporales ARIMA.

Una vez definida la función de autocorrelación funcional simple, se puede definir una función de autocorrelación parcial funcional de forma análoga a como se define en el caso escalar. En Mestre *et al.* (2021) se desarrolla la teoría funcional para la construcción de la fun-

ción de autocorrelación parcial funcional (*FPACF*) y la obtención de un límite de confianza bajo la hipótesis de que la serie es ruido blanco funcional.

Los métodos para estimar las funciones de autocorrelación y autocorrelación parcial presentados para datos funcionales se encuentran implementados en el paquete de R fdaACF, disponible en CRAN (http://cran.r-project.org/package=fdaACF).

Este procedimiento de identificación y diagnóstico se ilustra en la figura 11.

Figura 11.

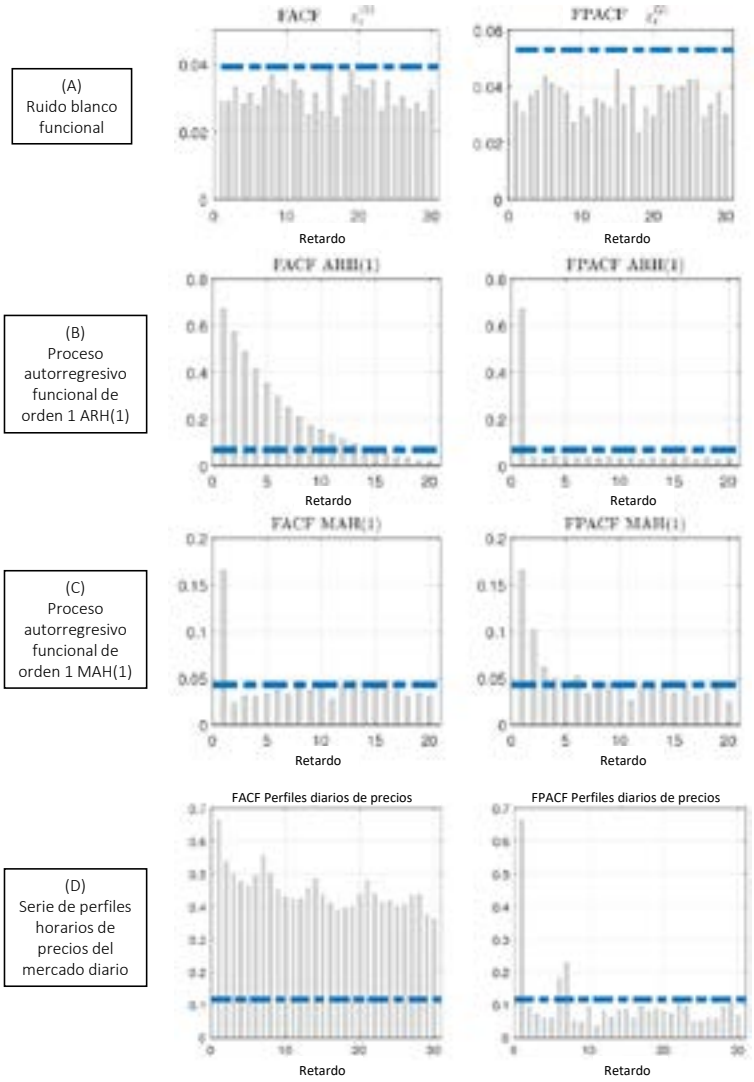
Procedimiento general para identificar y diagnosticar un modelo SARMAHX usando las *FACF* y *FPACF* propuestas

Fuente: Mestre (2021).

La figura 12 muestra las funciones de autocorrelación simple y parcial para distintas series de diferente naturaleza. La *FACF* y la *FPACF* introducidas se pueden utilizar para identificar la configuración óptima de los órdenes de autorregresión y de media móvil del modelo SARMAHX para una *FTS* dada, obteniendo predicciones más precisas de las curvas.

Figura 12.

Ejemplos de funciones de autocorrelación funcional (*FACF*) y autocorrelación funcional parcial (*FPACF*) para series de distinta naturaleza: proceso de ruido blanco funcional, proceso autorregresivo, proceso de media móvil y serie de perfiles de precios del mercado diario. Se puede observar en esta última cómo se aprecia el efecto estacional de periodicidad semanal (siete muestras)



Fuente: Elaboración propia a partir de Mestre (2021).

3. APLICACIÓN AL MERCADO ELÉCTRICO EUROPEO: PREDICCIÓN DE CURVAS DE OFERTA EN EL MERCADO ELÉCTRICO ITALIANO

Esta sección está dedicada a la predicción a corto plazo de las curvas de oferta agregadas por hora en los mercados de electricidad aplicando los modelos de predicción funcionales presentados anteriormente. En concreto, se ilustrará la aplicación a un caso de predicción dentro del mercado eléctrico italiano. En primer lugar, se presentarán las particularidades del mercado eléctrico de Italia y, posteriormente, se definirá el caso estudio. Finalmente, se comentarán los resultados y las conclusiones.

3.1. El mercado eléctrico italiano

El Mercado diario italiano (*Mercato del giorno prima – MGP*) es el mercado donde se gestionan las ofertas de compra y venta de electricidad en Italia para cada hora del día siguiente. El sistema eléctrico italiano se divide en siete zonas: Norte, Centro-Norte, Centro-Sur, Sur, Calabria, Sicilia y Cerdeña. Antes del cierre del mercado, los participantes presentan sus ofertas donde especifican la cantidad y el precio máximo y mínimo al que están dispuestos a comprar y a vender respectivamente la electricidad para cada zona y para cada subasta horaria del día siguiente. Además, las zonas italianas están interconectadas con las de países europeos cercanos, como Francia, Suiza, Austria, Eslovenia, Córcega, Montenegro, Grecia y Malta.

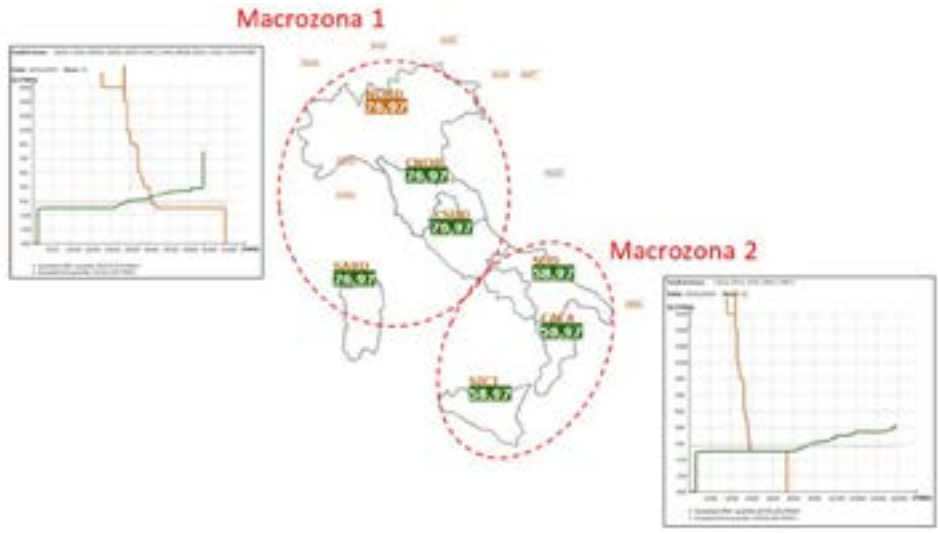
Estas zonas de mercado reflejan las diferencias entre la oferta y la demanda de electricidad en distintas áreas geográficas y tienen implicaciones directas en la formación de precios. Por ejemplo, en aquellas zonas donde la oferta supera a la demanda, los precios tienden a ser más bajos, y viceversa. Este sistema de precios diferenciales incentiva a los productores a establecer nuevas plantas en zonas con menor oferta, lo cual no solo aumenta la eficiencia del mercado sino que también mejora la oferta general de electricidad en esa zona.

Una vez que el *MGP* cierra, las ofertas de compra y de venta son aceptadas basándose en el criterio marginalista y teniendo en cuenta los límites de capacidad de transmisión entre zonas. Estas restricciones de capacidad son vinculantes. En caso de existir una saturación de la interconexión, el mercado se divide y se determinan precios de casación distintos para cada agrupación de zonas (*macrozona*). Estas *macrozonas* pueden ser distintas para cada hora del día. Las figuras 13 y 14 muestran el resultado del mercado para dos días diferentes. En el primero, se observan dos *macrozonas* de precios distintos, mientras que en el segundo caso, el resultado del mercado ha generado tres *macrozonas*.

Mientras que los generadores son remunerados al precio zonal resultante, las ofertas de demanda aceptadas que pertenecen a unidades de consumo se evalúan al precio nacional único (*PUN*) que se calcula como el promedio de los precios zonales ponderados por el consumo zonal (Gianfreda y Grossi, 2012).

Figura 13.

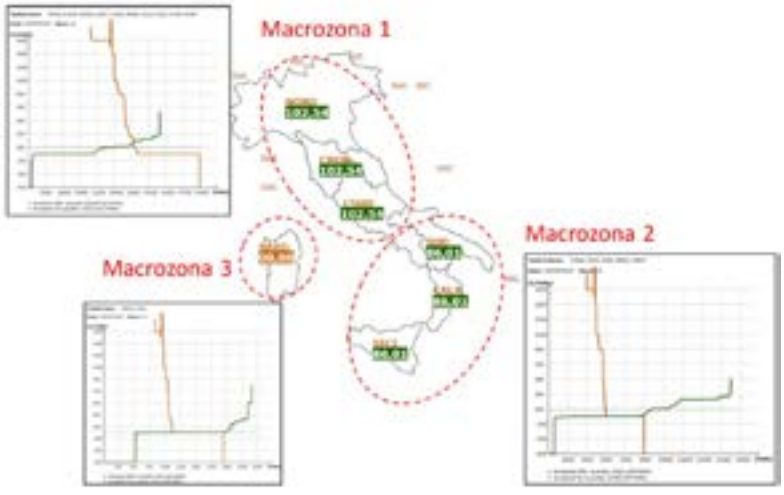
Hora 12 del día 3/11/2023 donde se han creado dos *macrozonas* con precios diferentes



Fuente: Elaboración propia.

Figura 14.

Hora 12 del día 5/9/2023 donde se han creado tres *macrozonas* con precios diferentes



Fuente: Elaboración propia.

Como se señala en Shah y Lisi (2019), debido a las políticas de confidencialidad, el Operador del Mercado Italiano (GME) no publica la información detallada relativa a las ofertas de compra y venta en un período de confidencialidad de siete días, a partir del día siguiente al cierre de la sesión del mercado a la que se refieren. Sin embargo, para los participantes del mercado, los resultados confidenciales anonimizados (incluyendo la aceptación/rechazo de las ofertas y las curvas de oferta agregadas sin información sobre la zona geográfica de cada oferta) están disponibles en https://www.IPEX.it.

3.2. Definición del caso de estudio

A pesar de que el mercado italiano consta de siete zonas interconectadas, por simplicidad y siguiendo a Pelagatti (2012), en este estudio se considera el mercado italiano como un único mercado, como si no hubiera restricciones de saturación entre zonas.

Se define como objetivo predecir las curvas de oferta para los competidores de una importante compañía eléctrica en Italia, que se obtienen agregando todas las curvas de oferta de los competidores zonales. Para este estudio, las curvas de oferta se limitan al rango de precios [0, 200] euros/MWh. Por lo tanto, cada observación funcional es la curva de oferta agregada presentada al mercado en cada hora. El rango de tiempo de los datos es desde el 01/03/2015 hasta el 29/02/2016, consistiendo así en 8784 curvas. Los datos se han obtenido del Operador del Mercado de Electricidad Italiano (https://gme.mercatoelettrico.org) y se dividen en dos conjuntos. El período de ajuste (*in-sample*) se considera desde el 01/03/2015 hasta el 31/08/2015, que se utilizará para entrenar los modelos. Los datos restantes se dejan para el período de test (*out-of-sample*).

El comportamiento de oferta de los agentes se ve afectado por las condiciones climáticas que afectan a la demanda, la generación eólica y solar, así como las circunstancias particulares de cada día. En consecuencia, se utilizan variables explicativas como una forma de tener en cuenta los factores externos que podrían influir en la decisión de los comerciantes. Las variables explicativas utilizadas en este estudio son las siguientes:

- *Demanda:* demanda total de electricidad de Italia. La demanda es de suma importancia para tener en cuenta el consumo de energía en el país.
- *Eólica:* producción total de energía eólica. El sur de Italia alberga un gran número de parques eólicos que tienen un impacto significativo en las curvas de oferta en días ventosos.
- *Solar:* producción total solar. El norte de Italia es la región con la mayor capacidad de energía solar instalada. Por lo tanto, la producción solar debería ser significativa.
- *Disponibilidad térmica:* es la suma de toda la energía ofrecida en el mercado por unidades térmicas.

- *Intercambios internacionales:* energía intercambiada con los países europeos adyacentes: Francia, Suiza, Austria, Eslovenia, Grecia y Malta. Estos intercambios juegan un papel muy importante en la comercialización de energía en Italia.

Vale la pena señalar que se utilizan valores reales de las variables explicativas, en lugar de predicciones, tanto para el entrenamiento como para la validación de todos los modelos considerados en este estudio. En la operación real, los valores actuales de estas variables son desconocidos, por lo que se deben utilizar escenarios futuros de las variables exógenas. El Operador del Mercado Italiano publica predicciones horarias de la demanda total de electricidad e intercambios internacionales para el día $D + 1$ en el día D, por lo que los modelos propuestos las pueden utilizar para obtener estimaciones de las curvas de oferta horarias antes de que se cierre la subasta para el día $D +1$. Otras variables, como la disponibilidad térmica y la producción solar y eólica, deben ser estimadas para usarlas como entradas para los modelos de predicción.

Estos modelos proporcionan indicadores útiles para cualquier agente del mercado, por lo tanto, en este capítulo se asume que los participantes del mercado ya tienen predicciones de estas variables. Se puede obtener información histórica sobre las variables mencionadas de la Plataforma de Transparencia de ENTSO-E, la Red Europea de Operadores de Sistemas de Transmisión (https://transparency.entsoe.eu/). Dado que las series temporales de salida son curvas horarias y estas variables explicativas son valores horarios, el modelo las considerará como covariables escalares y no como covariables funcionales.

3.3. Identificación del modelo SARMAHX

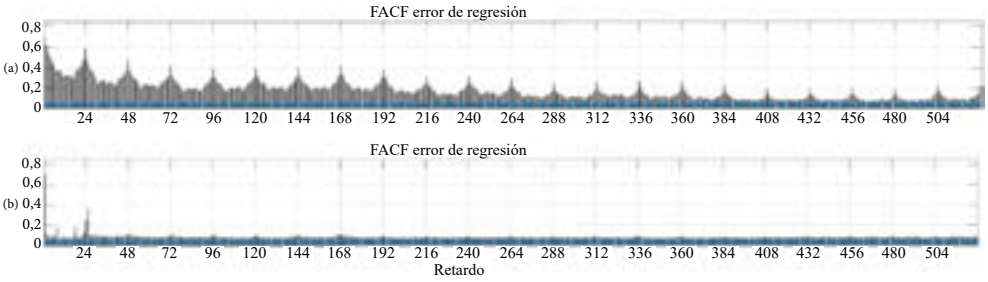
La técnica de identificación y diagnóstico presentada se utiliza para identificar el modelo funcional SARMAHX. Una vez seleccionados los órdenes de media móvil y autorregresivos regulares y estacionales del modelo, el modelo se entrena con los datos del periodo de ajuste de este caso de estudio. En primer lugar, se identifica la estructura de autocorrelación de la serie de curvas de oferta en el mercado eléctrico italiano utilizando la *FACF* y *FPACF* de la serie junto con el procedimiento de identificación propuesto mostrado en la figura 11.

En la figura 15 se muestra la estructura de autocorrelación de los errores de un modelo de regresión funcional ajustado con las variables exógenas mencionadas anteriormente. Estas variables explicativas se incluyen para capturar el efecto de las variables exógenas en las curvas de oferta. Como las series temporales de salida son curvas horarias y las variables explicativas son valores horarios, el modelo las considerará como covariables escalares.

Se observa un comportamiento estacional de la serie: una alta autocorrelación en los retardos 24 y 168 indica que la serie tiene fuertes estacionalidades diarias y semanales no modeladas con las variables explicativas. Además, el patrón de decrecimiento de la *FACF* indica la presencia de una componente autorregresiva en la serie. Por tanto, la autocorrelación observada en el error de regresión del modelo SARMAHX indica que es necesario modelar esta dinámica.

Figura 15.

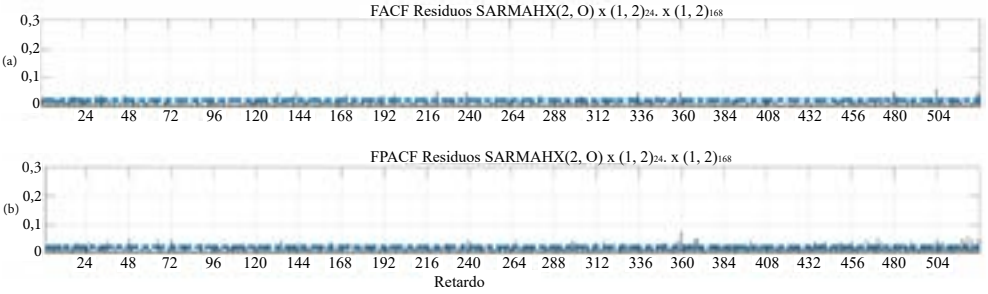
Función de autocorrelación funcional y de autocorrelación parcial funcional (*FACF/FPACF*) del error de regresión de un modelo de regresión funcional ajustado que incluye las variables exógenas



Fuente: Mestre (2021).

Figura 16.

Función de autocorrelación funcional y de autocorrelación parcial funcional (*FACF/FPACF*) de los residuos del modelo SARMAHX(2, 0) × $(1, 2)_{24}$ × $(1, 2)_{168}$ ajustado incluyendo variables exógenas. Como no se observan correlaciones significativas por encima de la banda de confianza, se pueden considerar los residuos como ruido blanco funcional



Fuente: Mestre (2021).

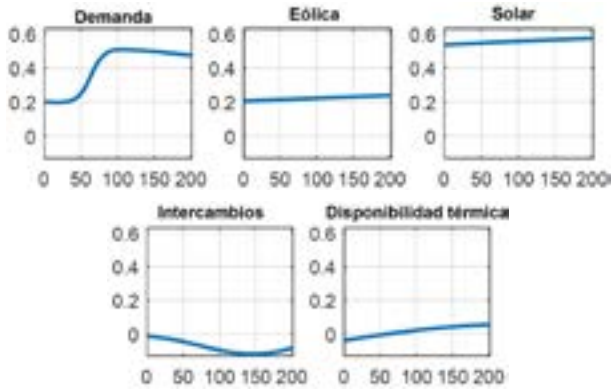
Para seleccionar los órdenes autorregresivos y de media móvil óptimos del modelo SARMAHX, se ha utilizado la metodología de identificación presentada en la sección anterior. Se han añadido términos de media móvil regular y estacional al modelo para tener en cuenta los retardos correlacionados significativos encontrados en los gráficos de *FACF/FPACF.* Este proceso iterativo se ha llevado hasta que los residuos del modelo ajustado son ruido blanco funcional. Finalmente, el modelo ajustado es un SARMAHX(2, 0) ×$(1, 2)_{24}$ ×$(1, 2)_{168}$. La figura 16 muestra las funciones de autocorrelación y autocorrelación parcial de los residuos del modelo SARMAHX ajustado con el límite superior del 95 % de las estadísticas bajo la hipótesis de ruido blanco. Como la mayoría de los valores caen por debajo de este límite superior, no se

puede rechazar la hipótesis de que los residuos sean ruido blanco, lo que valida el modelo desde un punto de vista lineal.

Vale la pena analizar las formas resultantes de los parámetros funcionales ajustados. La figura 17 muestra los operadores de regresión concurrentes. Analizando estas formas, se puede ver el efecto de cada variable en la curva de oferta. La demanda tiene valores positivos para cada precio, lo que significa que para una demanda más alta, la energía ofrecida a todos los precios aumenta. De hecho, la oferta aumenta más a precios más altos que a precios más bajos, lo que significa que el aumento de la demanda generalmente se cubre con una generación más cara. La producción de viento y solar tiene coeficientes algo planos. Por lo general, esta producción renovable se ofrece a precio 0, por lo tanto, simplemente desplaza la curva hacia arriba o hacia abajo. El intercambio internacional, por otro lado, tiene valores negativos. Esto significa que cuando se importa más energía, la curva de oferta tiene menos energía ofertada.

Figura 17.

Núcleos de los operadores para los regresores del modelo SARMAHX en el estudio de predicción de la curva de oferta



Fuente: Mestre (2021).

3.3.1. Comparación de resultados

Esta sección compara los modelos SARMAHX ajustados con algunos modelos de referencia. Se presentan dos análisis diferentes. Por un lado, como el modelo propuesto se entrena para minimizar el error de predicción a un paso, se analiza un predicción de una hora adelante. Sin embargo, en el mercado eléctrico italiano, las subastas para las 24 horas del día se liquidan al mismo tiempo. Por lo tanto, también se considera una predicción con un horizonte de 24 horas, validando el uso en una aplicación de caso real.

Todos los modelos incluidos en este capítulo se ajustaron utilizando datos del período de ajuste. Una vez que se han estimado sus parámetros, se obtienen estimaciones tanto para el

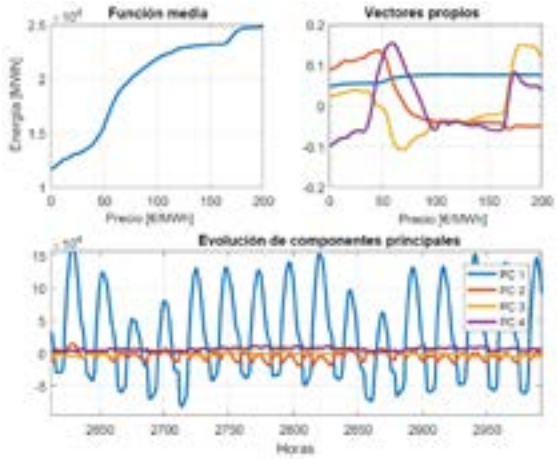
período de ajuste como para el período de test, sin recalibrar los parámetros de los modelos. Para el predicción de 24 horas adelante, se utiliza un enfoque de ventana móvil: al predecir la primera hora $h = 1$ del día D+1, la estimación $\hat{S}_{D+h}$ utiliza las curvas reales hasta el día D. Sin embargo, al predecir horas $h > 1$, las curvas reales no se han observado, por lo que las estimaciones pasadas del modelo se utilizan como entradas para el modelo.

Para enriquecer la comparativa, se ha incluido un modelo de cada grupo de los mencionados en la sección 2.1, comparando las capacidades de predicción del modelo propuesto frente a los mejores modelos funcionales encontrados en la literatura. Los modelos a comparar se describen a continuación:

- *Ingenuo.* Modelo de referencia simple para comparar los modelos ensayados. Se utilizan dos versiones, dependiendo de si la simulación es una predicción a horizonte 1 u horizonte 24. En el primer caso, la predicción es simplemente la última curva observada en los datos, es decir, la curva de la hora anterior. En el segundo caso, para los martes, miércoles, jueves y viernes, la predicción será la curva horaria del día anterior, mientras que para los sábados, domingos y lunes, la predicción será la curva horaria del mismo día de la semana anterior.
- *Enfoque de componentes principales.* Este método se utiliza en Pelagatti (2012) para la predicción de curvas de oferta. Extrae las primeras Componentes Principales Funcionales de las curvas y las correspondientes series temporales de sus valores (*scores*). Luego, estos valores se estiman mediante modelos de función de transferencia (*TF*) (Pankratz, 1991), que incluyen variables explicativas. La estimación final de

Figura 18.

Función media, vectores propios asociados a las componentes principales y evolución de las componentes principales de las curvas de oferta



Fuente: Mestre (2021).

las curvas se realiza reconstruyéndolas a partir de los valores estimados de las componentes principales. El período de ajuste se utiliza para entrenar todos los modelos. Las *FPC* se extraen para ese rango y se estiman los parámetros de la *TF*. En el período de test, se obtienen los valores de las componentes proyectando las curvas en la base definida por las *FPC* previamente extraídas.

La figura 18 representa la función media, los pesos que definen las componentes principales (vectores propios) y los valores obtenidos (*scores*) de las primeras componentes principales de las curvas de oferta. Se extraen tres y cuatro componentes principales, que explican el 98 % y el 99 % de la varianza de los datos, respectivamente. Es importante señalar que cada curva pronosticada con este método es una combinación lineal de solo las tres o cuatro componentes principales extraídas. Estos modelos se denotarán como PC_FT3 y PC_FT4, respectivamente.

- *Enfoque no paramétrico funcional.* Este enfoque es el presentado en la sección 2 de este capítulo.

  Sin embargo, en la definición original, la predicción sólo dependía de valores pasados de la serie. Como todos los modelos que se comparan con el modelo SARMAHX propuesto utilizan variables exógenas, se incluirá una versión de este modelo en el estudio de comparación. El modelo lineal parcial semifuncional (Aneiros *et al.,* 2013; Vilar *et al.,* 2018) permite la inclusión de variables escalares exógenas, generalizando el modelo no paramétrico funcional. La expresión del modelo es la siguiente:

$$Y_t(v) = x_t^T \beta(v) + \omega(Y_{t-1})(v) + \varepsilon_t(v), \qquad [21]$$

  donde $x_t^T = (x_{t,1},\ldots,x_{t,p})$ es un vector de p covariables escalares exógenas, $\beta(v) = (\beta_1(v), \ldots, \beta_p(v))$ es un vector de parámetros funcionales desconocidos a estimar y ω es el estimador no paramétrico de Nadaraya-Watson dado en [5]. Para ajustar este modelo al conjunto de datos de la curva de oferta, los parámetros se han seleccionado de la siguiente manera: la función de kernel K utilizada fue la Epanechnikov, definida como $K(u) = 3/4(1 - u^2)$; el parámetro de ancho de banda h se ha seleccionado utilizando el método de los k vecinos más cercanos propuesto en Antoniadis *et al.* (2009); y la semimétrica d seleccionada se basa en la norma L^2 de las curvas. A lo largo de esta sección, este modelo se denotará como NPARHX.

Todos los modelos se entrenan con el período de entrenamiento utilizando las mismas variables exógenas. Posteriormente, cada uno produce una predicción a horizonte 1, asumiendo que se conoce la curva de la última hora, y una predicción a horizonte 24, donde se predicen las 24 horas del día siguiente, siendo la última curva observada la hora 24 del día actual. En el caso de predicción a horizonte 24, la hora 1 siempre se estimará utilizando un horizonte de predicción de 1; la hora 2 se estimará utilizando un horizonte de predicción de 2; y finalmente, el horizonte de predicción utilizado para estimar la hora 24 será 24. De aquí en adelante, se producen predicciones a horizonte 1 y 24 para todo el rango de datos. Se calculan los errores funcionales FMAE y FRMSE, que se definen como:

$$\text{FMAE} = T^{-1}\sum_{t=1}^{T}\int\left|Y_t(u)-\hat{Y}_t(u)\right|du \qquad [22]$$

$$\text{FRMSE} = \sqrt{T^{-1}\sum_{t=1}^{T}\int\left(Y_t(u)-\hat{Y}_t(u)\right)^2 du} \qquad [23]$$

Cuando se predicen ciertos tipos de datos, se debe aplicar algún método de posprocesamiento para mejorar la salida bruta del modelo de predicción que se utiliza. Un caso de este tipo es el que nos ocupa: predecir curvas de oferta agregadas en el mercado eléctrico del día siguiente. Estas curvas son siempre monótonamente crecientes, debido a la propia definición de las funciones de oferta. La formulación del modelo SARMAHX no garantiza que las estimaciones del modelo sean funciones no decrecientes, por lo que para obtener estimaciones fieles de estas curvas con el modelo propuesto, las curvas de salida del modelo $\hat{Y}_t$ se transforman en curvas monótonamente crecientes $\hat{Y}_t^+$ obtenidas como la solución del siguiente problema de optimización:

$$\min_{\hat{Y}_t^+} \quad \frac{1}{2}\left\|\hat{Y}_t^+ - \hat{Y}_t\right\|^2 \qquad [24a]$$

$$\text{sujeto a} \quad 0 \le \hat{Y}_t^+(v_i) \le \hat{Y}_t^+(v_{i+1}), i = 1,\ldots,N-1, \qquad [24b]$$

donde $\{v_1, \ldots, v_N\}$ denota los puntos de discretización de las observaciones funcionales. La función objetivo a minimizar [24a] es la distancia entre las curvas estimadas y su transformación monótona, mientras que la restricción [24b] asegura que la nueva curva sea no decreciente. El problema de optimización se resuelve por mínimos cuadrados ordinarios.

Esta transformación garantiza que las curvas estimadas sean monótonamente crecientes, manteniendo la forma que ha sido estimada por el modelo SARMAHX. Para proporcionar una comparación justa entre el modelo SARMAHX y otros modelos de predicción encontrados en la literatura, la transformación propuesta se aplicará a las estimaciones de todos los modelos considerados en esta sección.

El cuadro 1 muestra los errores funcionales para las predicciones a horizontes 1 y 24 horas para los períodos de entrenamiento y test. El modelo SARMAHX propuesto muestra una clara ventaja sobre los modelos de referencia, obteniendo un FMAE de 288,66 MWh que es mucho menor que el FMAE del PC_FT4 (el mejor modelo de referencia), que logra un FMAE de 416,05 MWh en el período de test del estudio con horizonte de predicción de un período hacia adelante. Para el caso de predicción con horizonte de 24 horas, los resultados son similares: mientras que el modelo PC_FT4 obtiene un FMAE de 811,69 MWh, no es capaz de mejorar los resultados del modelo SARMAHX concurrente, que logra un FMAE de 709,70 MWh en el período *out-of-sample*. Los resultados de la estimación se analizan más adelante en detalle.

En primer lugar, al observar las predicciones de horizonte 1, se puede ver cómo el enfoque funcional supera a los demás métodos tanto en los períodos de entrenamiento como en el de

Cuadro 1.

Errores promedios para cada método en la predicción a horizonte 1 y horizonte 24 para la serie de curvas de oferta agregadas

		In-Sample		*Out-Of-Sample*	
Forecasting horizon	*Model*	*FMAE [MWh]*	*FRMSE [MWh]*	*FMAE [MWh]*	*FRMSE [MWh]*
1-Step Ahead	Naïve	909,39	1.258,46	818,59	1.179,16
	PC_FT3	353,85	470,51	480,54	694,15
	PC_FT4	276,65	378,83	416,05	569,36
	NPARHX	389,30	553,53	641,92	884,50
	SARMAHX	**264,62**	**363,10**	**313,11**	**429,66**
24-Step Ahead	Naïve	1.075,3	1.433,5	1.382,3	1.812,50
	PC_FT3	672,06	871,33	856,62	1.148,1
	PC_FT4	639,46	834,46	811,69	1.85,07
	NPARHX	743,47	1.010,01	987,41	1.311,51
	SARMAHX	**590,08**	**787,34**	**709,70**	**963,49**

Nota: Los errores más bajos se han resaltado en negrita.

Fuente: Mestre (2021).

test. Además, el enfoque funcional es el que ofrece los mejores resultados, con un error absoluto medio de 313 MWh en el período de test, mientras que el FMAE del modelo PC_FT4 es de 416 MWh. Esta diferencia valida la utilidad del modelo SARMAHX como un modelo competitivo para predecir curvas de oferta. Se pueden extraer varias conclusiones de estos resultados. En primer lugar, aunque el enfoque FPCA también utiliza un modelo de series temporales, el hecho de reducir la dimensionalidad afecta significativamente al rendimiento. Como las Componentes Principales se mantienen intactas, las predicciones no pueden adaptarse a los cambios en la serie temporal funcional y la reconstrucción de las curvas pierde precisión. Por el contrario, el modelo funcional propuesto no se basa en ninguna expansión de bases de la serie y tiene en cuenta los valores de toda la curva del pasado reciente. Por lo tanto, puede reflejar mejor los cambios en la serie.

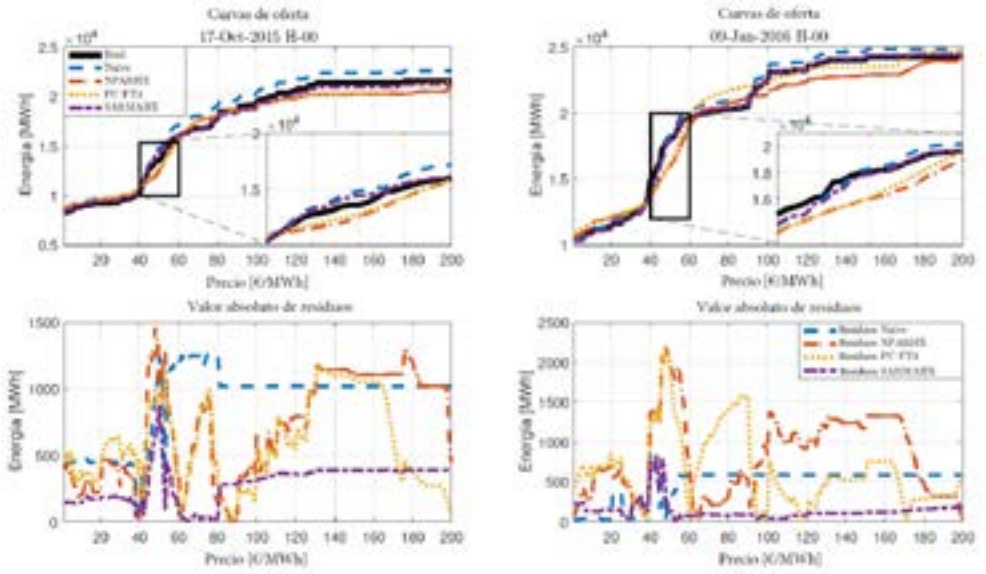
Además, como el enfoque no paramétrico no tiene en cuenta la estructura estacional de los datos, no es capaz de obtener los mismos resultados que los otros modelos de referencia. La figura 19 muestra algunos ejemplos de predicciones para el caso de predicción a horizonte 1, donde se resalta la capacidad del modelo SARMAHX de estimar adecuadamente la forma de las curvas de oferta. Mientras que el modelo propuesto es capaz de estimar con precisión las curvas de oferta, los modelos de referencia no son capaces de capturar el complejo comportamiento de oferta exhibido por las curvas, proporcionando sólo una aproximación suave a las curvas que no tiene en cuenta la naturaleza escalonada de las mismas.

En segundo lugar, las predicciones con horizonte de 24 horas muestran resultados similares al caso de horizonte 1. El modelo SARMAHX proporciona mejores errores medios con respecto a los otros modelos. Globalmente, los errores de predicción con horizonte 24 horas son mucho más altos que los de horizonte 1.

En el mercado diario italiano, el precio de casación en el período de estudio suele estar situado en el rango de precios [30, 70] euros/MWh, por lo que el volumen de las ofertas de los agentes será significativamente mayor en esa parte de la curva de oferta agregada. Por lo tanto, obtener una estimación precisa de la curva de oferta en esa región es de suma importancia para cualquier empresa. Como se puede ver en la figura 19, el modelo SARMAHX es capaz de capturar la forma de la curva de oferta en ese rango de precios, proporcionando una descripción precisa del comportamiento de oferta de los agentes en la zona de interés de la curva de oferta.

Figura 19.

Panel superior: ejemplos de predicciones para estimaciones a horizonte 1 en los períodos *out-of-sample* en el estudio de predicción de curvas de oferta. Se muestran en detalle las curvas en el rango de precios [40, 60] euros/MWh, resaltando las diferencias entre los distintos modeos de predicción. Panel inferior: valor absoluto de las curvas residuales de los modelos



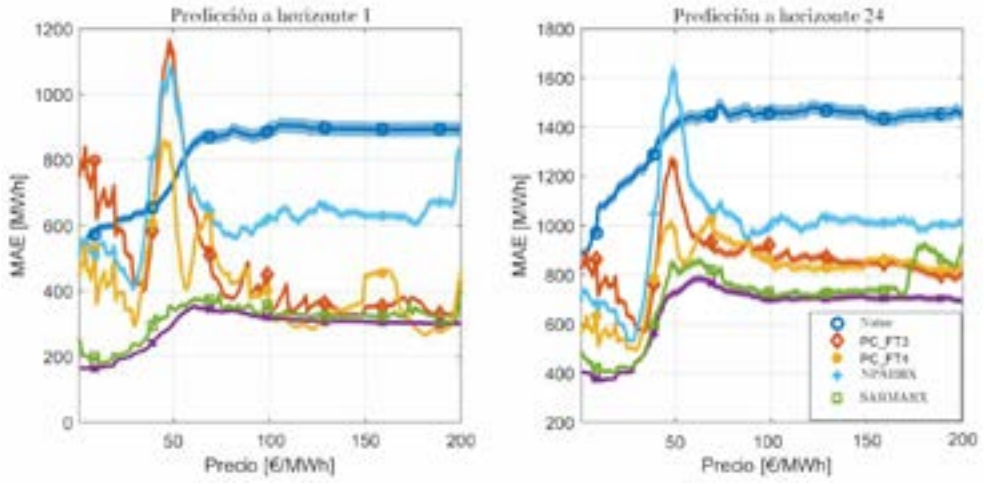
Fuente: Mestre (2021).

La forma de las curvas de oferta estimadas es esencial para analizar la oferta estratégica de los competidores de una empresa eléctrica, porque contiene información valiosa sobre el comportamiento de oferta de los agentes. Las curvas de oferta pronosticadas, junto con las estimaciones de la demanda, se pueden utilizar para definir las *RDC,* que se utilizan a menudo para calibrar modelos de equilibrio de mercado (Díaz *et al.,* 2010) así como modelos de optimización de oferta (Baillo *et al.,* 2004). En Portela *et al.* (2017), los autores analizan la importancia de obtener estimaciones fieles de las pendientes de las *RDC,* ya que indica la

capacidad de un agente para influir en los precios del mercado. Esto resalta la necesidad de un modelo de predicción que no solo capte el nivel general de las curvas de oferta, sino que también proporcione una estimación precisa de la forma de la curva.

Figura 20.

FMAE para cada precio en las predicciones de horizonte 1 y horizonte 24 en el período de test en el estudio de predicción de las curvas de oferta



Nota: Como el error FMAE del modelo Naive es significativamente mayor que el de los otros modelos, no se incluye. Las regiones sombreadas son bandas de confianza del 90 % para el FMAE.

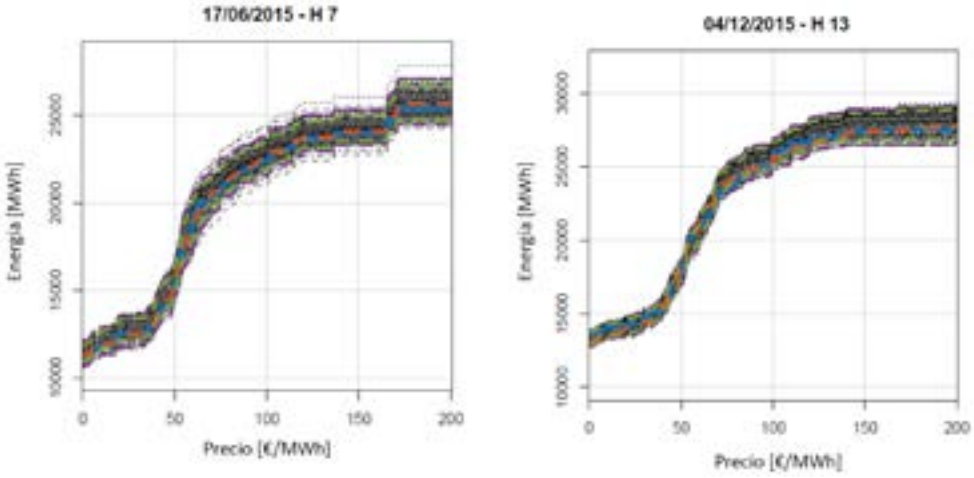
Fuente: Mestre (2021).

Como se puede ver en la figura 20, se presenta el error absoluto medio (*MAE*) para cada precio de oferta para el período de test, para las predicciones de horizonte 1 y horizonte 24, que se obtienen como el valor medio del error absoluto para cada precio de las curvas. Las figuras muestran el desempeño de cada método para diferentes rangos de precios y se puede observar cómo los modelos SARMAHX producen errores más uniformes a lo largo de los distintos precios. De nuevo, como los modelos propuestos no dependen de las *FPC*, se pueden adaptar mejor a los escalones de oferta en la curva. La mejora más significativa se observa en el rango de precios de interés mencionado anteriormente: por ejemplo, mientras que el error absoluto medio del modelo PC_FT4 en el precio de 50 euros/MWh es de 1.000 MWh, el modelo SARMAHX logra un error absoluto medio de 730 MWh en el pronóstico con horizonte de 24 horas. Esta diferencia es más pronunciada en el estudio con horizonte 1, donde los errores absolutos medios del modelo PC_FT4 en el rango de precios de interés están cerca de 800 MWh frente a los errores del modelo SARMAHX, que están cerca de 350 MWh. Además, se incluyen bandas de confianza de los errores *MAE* para facilitar la interpretación de los resultados. Como las bandas de confianza no se superponen, se puede concluir que los errores de los modelos SARMAHX son significativamente menores que los de los otros modelos considerados en este estudio.

Finalmente, cabe destacar que a partir de los residuos de estos modelos, en Mestre (2021) se propone una metodología probabilística para la generación de escenarios futuros de las curvas de oferta, especialmente indicada para la optimización de las estrategias de oferta. Un ejemplo se puede ver en la figura 21.

Figura 21.

Estimación de escenarios de curvas de oferta para dos horas seleccionadas



Fuente: Mestre (2021).

4. CONCLUSIONES

Los agentes que operan en los mercados liberalizados de electricidad se enfrentan a distintas fuentes de incertidumbre, que hacen necesario disponer de herramientas avanzadas de análisis y predicción que ayuden a la toma de decisiones para optimizar sus estrategias de oferta.

El conjunto de las ofertas históricas presentadas por los distintos agentes en el mercado son una fuente de información muy valiosa, ya que permiten caracterizar sus estrategias de oferta. Las técnicas de *big data* ofrecen una solución escalable para el procesamiento de estos grandes volúmenes de información.

La caracterización del comportamiento de los agentes se puede realizar a partir de las curvas agregadas de oferta de compra y de venta, o bien las curvas de demanda residual. Una forma de analizar y predecir dichas curvas es mediante el uso del análisis de datos funcionales. En este trabajo se propone un enfoque funcional para la predicción de curvas de oferta, utilizando un modelo funcional basado en la metodología de Box-Jenkins (Box *et al.*, 2008).

El desarrollo del modelo funcional SARMAHX, que admite la inclusión de dos estacionalidades y distintas variables explicativas, permite modelizar series temporales funcionales

horarias considerando sus estacionalidades diarias y semanales. Este modelo permite obtener predicciones más precisas de las curvas de oferta agregada, superando a otros métodos tradicionales.

Además, el procedimiento de identificación basado en la función de autocorrelación funcional simple y parcial es un elemento clave para analizar la estructura de correlación de las series temporales funcionales y para determinar los órdenes AR y MA del modelo. La eficacia del modelo se ha ilustrado con un caso real de predicción de curvas de oferta horarias en el mercado eléctrico diario de Italia.

Referencias

Álvarez, J., Bosq, D. y Ruiz-Medina, M. (2017). Asymptotic properties of a component-wise ARH(1) plug-in predictor. *Journal of Multivariate Analysis,* 155, pp. 12–34. doi: 10.1016/j.jmva.2016.11.009

Aneiros, G. y Vieu, P. (2008). Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis, 99*(5), pp. 834–857. doi: 10.1016/ j.jmva.2007.04.010

Aneiros, G., Vilar, J. y Raña, P. (2016). Short-term forecast of daily curves of electricity demand and price. *International Journal of Electrical Power & Energy Systems,* 80 , pp. 96–108. doi: 10.1016/j.ijepes.2016.01.034

Aneiros, G., Vilar, J. M., Cao, R. y Muñoz San Roque, A. (2013). Functional Prediction for the Residual Demand in Electricity Spot Markets. *IEEE Transactions on Power Systems, 28*(4), pp. 4201–4208. doi: 10.1109/ TPWRS.2013.2258690

Antoch, J., Prchal, L., Rosaria De Rosa, M. y Sarda, P. (2010). Electricity consumption prediction with functional linear regression using spline estimators. *Journal of Applied Statistics, 37*(12), pp. 2027–2041. doi: 10.1080/02664760903214395

Antoniadis, A., Paparoditis, E. y Sapatinas, T. (2009). Bandwidth selection for functional time series prediction. *Statistics & Probability Letters, 79*(6), pp. 733–740. doi: 10.1016/ j.spl.2008.10.028

Aue, A., Norinho, D. D. y Hörmann, S. (2015, enero). On the Prediction of Stationary Functional Time Series. *Journal of the American Statistical Association, 110*(509), pp. 378–392. doi: 10.1080/01621459.2014.909317

Baillo, A., Ventosa, M., Rivier, M. y Ramos, A. (2004). Optimal Offering Strategies for Generation Companies Operating in Electricity Spot Markets. *IEEE Transactions on Power Systems, 19*(2), pp. 745–753. doi: 10.1109/ TPWRS.2003.821429

Bosq, D. (2000). *Linear Processes in Function Spaces – Theory and Applications* (n.o 149). New York: Springer-Verlag.

Box, G. E. P., Jenkins, G. M. y Reinsel, G. C. (2008). Time series analysis: forecasting and control (4th ed.). Hoboken, N.J: John Wiley. (OCLC: ocn176895531).

Bunn, D. W. y Farmer, E. (1985). *Comparative Models for Electrical Load Forecasting.* New York, NY: John Wiley & Sons, Inc.

Campos, F. A., Muñoz, A., Sánchez-Úbeda, E. F. y Portela, J. (2016). Strategic Bidding in Secondary Reserve Markets. *IEEE Transactions on Power Systems , 31*(4), pp. 2847–2856. doi: 10.1109/TPWRS.2015.2453477

Chen, K., Chen, K., Wang, Q., He, Z., Hu, J. y He, J. (2019). Short-Term Load Forecasting With Deep Residual Networks. *IEEE Transactions on Smart Grid , 10*(4), pp. 3943–3952. doi: 10.1109/TSG.2018.2844307

Collado, J. V., Sánchez-Úbeda, E. F. y Muñoz San Roque, A. (2004). SGO: sistema de información para la realización de ofertas en el mercado eléctrico español. En *Anales de mecánica y electricidad, Vol. 81* (pp. 40–50). Asociación de Ingenieros del ICAI.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems, 2*(4), pp. 303–314. doi: 10.1007/BF02551274

Damon, J. y Guillas, S. (2002). The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics, 13*(7), pp. 759–774. doi: 10.1002/env.527

Damon, J. y Guillas, S. (2005). Estimation and Simulation of Autoregressive Hilbertian Processes with Exogenous Variables. *Statistical Inference for Stochastic Processes, 8*(2), pp. 185–204. doi: 10.1007/s11203-004-1031-6

Díaz, C. A., Villar, J., Campos, F. A. y Reneses, J. (2010). Electricity market equilibrium based on conjectural variations. *Electric Power Systems Research, 80*(12), pp. 1572–1579. doi: 10.1016/j.epsr.2010.07.012

Didericksen, D., Kokoszka, P. y Zhang, X. (2012). Empirical properties of forecasts with the functional autoregressive model. *Computational Statistics , 27*(2), pp. 285–298. doi: 10.1007/s00180-011-0256-2

Erbas, B., Hyndman, R. J. y Gertig, D. M. (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine, 26*(2), pp. 458–470. doi: 10.1002/sim.2306

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics, 39*(3), pp. 254–261.

Ferraty, F. y Vieu, P. (2006). *Nonparametric Functional Data Analysis – Theory and Practice.* New York: Springer-Verlag.

Galeano, P. (2021). Una aplicación del análisis de series temporales funcionales a los precios horarios de la electricidad en el mercado mibel. En D. Peña, P. Poncela y E. Ruiz (editores), *Análisis econométrico y big data* (pp. 163–190). Madrid: Funcas.

Gianfreda, A. y Grossi, L. (2012). Forecasting Italian electricity zonal prices with exogenous variables. *Energy Economics, 34*(6), pp. 2228–2239. doi: 10.1016/j.eneco.2012.06.024

Hall, P., Müller, H.-G. y Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics, 34*(3), pp. 1493–1517.

Horváth, L., Liu, Z., Rice, G. y Wang, S. (2020). A functional time series analysis of forward curves derived from commodity futures. *International Journal of Forecasting, 36*(2), pp. 646–665. doi: 10.1016/j.ijforecast.2019.08.003

Hyndman, R. J. y Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society, 38*(3), pp. 199–211. doi: 10.1016/j.jkss.2009.06.002

Hyndman, R. J. y Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis, 51*(10), pp. 4942–4956. doi: 10.1016/j.csda.2006.07.028

Joskow, P. L. (2008). Lessons Learned from Electricity Market Liberalization. *The Energy Journal, 29*(SI2). doi: 10.5547/ISSN0195-6574-EJ-Vol29-NoSI2-3

Kargin, V. y Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis, 99*(10), pp. 2508–2526. doi: 10.1016/j.jmva.2008.03.001

Klepsch, J., Klüppelberg, C. y Wei, T. (2017). Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics,* 1 , pp. 128–149. doi: 10.1016/ j.ecosta.2016.10.009

Kokoszka, P., Rice, G. y Shang, H. L. (2017). Inference for the autocovariance of a functional time series under conditional heteroscedasticity. *Journal of Multivariate Analysis,* 162, pp. 32–50. doi: 10.1016/j.jmva.2017.08.004

Liebl, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *The Annals of Applied Statistics, 7*(3), pp. 1562–1592. doi: 10.1214/13-AOAS652

Liu, Z., Yan, J., Shi, Y., Zhu, K. y Pu, G. (2012). Multi-agent based experimental analysis on bidding mechanism in electricity auction markets. *International Journal of Electrical Power & Energy Systems, 43*(1), pp. 696–702. doi: 10.1016/j.ijepes.2012.05.056

Mas, A. (2007). Weak convergence in the functional autoregressive model. *Journal of Multivariate Analysis, 98*(6), pp. 1231–1261. doi: 10.1016/j.jmva.2006.05.010

Mestre, G., Portela, J., Muñoz San Roque, A. y Alonso, E. (2020). Forecasting hoignoreurly supply curves in the Italian Day-Ahead electricity market with a double-seasonal SAR- MAHX model. *International Journal of Electrical Power & Energy Systems,* 121, 106083. doi: https://doi.org/10.1016/j.ijepes.2020.106083

Mestre, G., Portela, J., Rice, G., Muñoz San Roque, A. y Alonso, E. (2021). Functional time series model identification and diagnosis by means of auto- and partial autocorrelation analysis. *Computational Statistics & Data Analysis,* 155, 107108. doi: 10.1016/j.csda.2020.107108

Mestre, G., Sánchez-Úbeda, E. F., Muñoz San Roque, A. y Alonso, E. (2022). The arithmetic of stepwise offer curves. *Energy,* 239, 122444. doi: https://doi.org/10.1016/j.energy.2021.122444

Mestre, G. (2021). *Probabilistic forecasting of functional time series: Application to scenario-generation of residual demand curves in electricity markets* (Tesis Doctoral no publicada). Universidad Pontificia Comillas.

Monteiro, C., Ramirez-Rosado, I. J., Fernandez-Jimenez, L. A. y Ribeiro, M. (2018). New probabilistic price forecasting models: Application to the Iberian electricity market. *International Journal of Electrical Power & Energy Systems, 103,* pp. 483–496. doi: 10.1016/ j.ijepes.2018.06.005

Mourid, T. (2002). Estimation and Prediction of Functional Autoregressive Processes. *Statistics, 36*(2), pp. 125–138. doi: 10.1080/02331880212048

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability & Its Applications, 9*(1), pp. 141–142. doi: 10.1137/1109020

Pankratz, A. (1991). *Forecasting with dynamic regression models.* New York: Wiley.

Paparoditis, E. y Sapatinas, T. (2013). Short-Term Load Forecasting: The Similar Shape Functional Time-Series Predictor. *IEEE Transactions on Power Systems, 28*(4), pp. 3818–3825. doi: 10.1109/TPWRS.2013.2272326

Pelagatti, M. (2012). *Supply Function Prediction in Electricity Auctions* (Inf. Téc.).

Portela, J. (2017). *Functional time series forecasting in electricity markets: a novel parametric approach* (Tesis Doctoral no publicada). Universidad Pontificia Comillas, Madrid.

Portela, J., Muñoz, A., Sánchez-Úbeda, E. F., García-González, J. y González, R. (2017). Residual Demand Curves for Modeling the Effect of Complex Offering Conditions on Day- Ahead Electricity Markets. *IEEE Transactions on Power Systems, 32*(1), pp. 50–61. doi: 10.1109/TPWRS.2016.2552240

Portela, J., Muñoz San Roque, A. y Alonso, E. (2018). Forecasting Functional Time Series with a New Hilbertian ARMAX Model: Application to Electricity Price Forecasting. *IEEE Transactions on Power Systems, 33*(1), pp. 545–556. doi: 10.1109/TPWRS.2017.2700287

Prete, C. L. y Hobbs, B. F. (2015). Market power in power markets: an analysis of residual demand curves in California's day-ahead energy market (1998-2000). *The Energy Journal, 0*(2).

Sánchez-Úbeda, E. F. (1999). *Models for data analysis: contributions to automatic learning.* (Ph.D thesis). Madrid: Universidad Pontificia Comillas.

Sánchez-Úbeda, E. F. y García-González, J. (2000). Management of sealed-bid auction curves: Applications of the Linear Hinges Model. *IPMU Information Processing and Management of Uncertainty in Knowledge-based Systems, Madrid, 2.*

Sánchez-Úbeda, E. F., Muñoz, A. y Villar, J. (2006). Minería y visualización de datos del mercado eléctrico español. Inteligencia Artificial. *Revista Iberoamericana de Inteligencia Artificial , 10*(29), pp. 79–88.

Sánchez-Úbeda, E. F. y Wehenkel, L. (1998). The Hinges model: A one-dimensional continuous piecewise polynomial model. En *Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU* (pp. 878–885). Paris.

Sen, R. y Klüppelberg, C. (2019, julio). Time series of functional data with application to yield curves. *Applied Stochastic Models in Business and Industry , 35*(4), pp. 1028–1043. doi: 10.1002/asmb.2443

Shah, I. y Lisi, F. (2019). Forecasting of electricity price through a functional prediction of sale and purchase curves. *Journal of Forecasting*, pp. 1– 18. doi: 10.1002/for.2624

Shang, H. L. (2012). Point and interval forecasts of age-specific fertility rates: a comparison of functional principal component methods. *Journal of Population Research, 29*(3), pp. 249–267. doi: 10.1007/s12546-012-9087-4

Turbillon, C., Marion, J.-M. y Pumo, B. (2007). Estimation of the moving-average operator in a Hilbert space. En *Recent Advances in Stochastic Modeling and Data Analysis* (pp. 597–604). Chania, Greece: World Scientific Publications. doi: 10.1142/9789812709691_0070

Ugedo, A., Lobato, E., Franco, A., Rouco, L., Fernandez-Caro, J., de-Benito, J., . . . De-la-Hoz, J. (2003). Stochastic model of residual demand curves with decision trees. En *Power engineering society general meeting, 2003, ieee, Vol. 2* (pp. 979–984).

Valderrama, M. J., Ocana, F. A. y Aguilera, A. M. (2002). Forecasting PC-ARIMA models for functional data. En W. Härdle y Rönz (Eds.), *Proceedings in Computational Statistics* (pp. 25–36). Physica, Heidelberg. doi: 10.1007/978-3-642-57489-4_3

Vilar, J., Aneiros, G. y Raña, P. (2018, marzo). Prediction intervals for electricity demand and price using functional data. *International Journal of Electrical Power & Energy Systems,* 96, pp. 457–472. doi: 10.1016/j.ijepes.2017.10.010

Vilar, J. M., Cao, R. y Aneiros, G. (2012, julio). Forecasting next-day electricity demand and price using nonparametric functional methods. *International Journal of Electrical Power & Energy Systems, 39*(1), pp. 48–55. doi: 10.1016/j.ijepes.2012.01.004

Villar, J., Muñoz, A., Sánchez-Úbeda, E. F., Mateo, A., Casado, M., Campos, A., . . . Marcos, J. J. (2001). SGO: Management information system for strategic bidding in electrical markets. En *Power Tech Proceedings, 2001 IEEE Porto, Vol. 1*. IEEE.

Wagner-Muns, I. M., Guardiola, I. G., Samaranayke, V. A. y Kayani, W. I. (2018). A Functional Data Analysis Approach to Traffic Volume Forecasting. *IEEE Transactions on Intelligent Transportation Systems , 19*(3), pp. 878–888. doi: 10.1109/TITS.2017.2706143

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā*: *The Indian Journal of Statistics, 26*(4), pp. 359–372.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting, 30*(4), pp. 1030–1081. doi: 10.1016/ j.ijforecast.2014.08.008

Xu, L. y Baldick, R. (2007, noviembre). Transmission-Constrained Residual Demand Derivative in Electricity Markets. *IEEE Transactions on Power Systems, 22*(4), pp. 1563–1573. doi: 10.1109/TPWRS.2007.907511

Yao, F., Müller, H.-G. y Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association, 100*(470), pp. 577–590.

Zhang, J.-L., Zhang, Y.-J., Li, D.-Z., Tan, Z.-F. y Ji, J.-F. (2019). Forecasting day- ahead electricity prices using a new integrated model. *International Journal of Electrical Power & Energy Systems,* 105, pp. 541–548. doi: 10.1016/j.ijepes.2018.08.025

Sobre los autores

Maria Alló

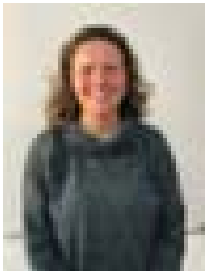
Es doctora en Economía por la Universidad de Santiago de Compostela. Actualmente, ejerce como profesora titular de Fundamentos de Análisis Económico en la Universidade da Coruña. Desde enero del 2024 es vicedecana de relaciones institucionales y movilidad, de la Facultad de Economía y Empresa de la misma universidad.

Andrés M. Alonso Fernández

Es licenciado en Matemáticas por la Universidad de La Habana (1991), máster en Epidemiología por el Instituto Pedro Kourí (1994) y doctor en Economía por la Universidad Carlos III de Madrid (2001). Ha sido profesor asociado en el Departamento de Matemáticas de la Universidad Autónoma de Madrid e investigador Juan de La Cierva en el Departamento de Estadística de la Universidad Carlos III de Madrid. Actualmente es profesor titular de Estadística en esta universidad. Sus principales intereses de investigación son: análisis de series de tiempo, técnicas de remuestreo, estadística aplicada y econometría. Ha publicado más de 70 artículos de investigación, una monografía y un libro sobre estas temáticas.

Jorge Arias Martí

Es graduado en Física por la Universidad Autónoma de Madrid (2021) y máster en Estadística para la Ciencia de Datos (2022) impartido por la Universidad Carlos III de Madrid. Actualmente trabaja como ingeniero de datos en Bluetab Solutions S.L. Sus intereses abarcan tanto el área del *Machine Learning* como el procesamiento de datos masivos en plataformas Cloud.



M. Angeles Carnero

Es profesora titular de Universidad en el Departamento de Fundamentos del Análisis Económico de la Universidad de Alicante. Es doctora en Economía desde 2003 por la Universidad Carlos III de Madrid. Su principal área de investigación es la econometría de las series temporales financieras.

Isabel Casas

Es actualmente profesora en Deusto Business School en Bilbao. Obtuvo su licenciatura en Ciencias Matemáticas de la Universidad Autónoma de Madrid y completó su doctorado en Econometría en la Universidad de Western Australia. A lo largo de su carrera ha trabajado como ingeniera de *software* en empresas privadas y ha ocupado diversas posiciones académicas en instituciones destacadas como la Universidad Carlos III, la K.U. Leuven, Aarhus University y la University of Southern Denmark. Isabel fue galardonada con una Marie Sklodowska-Curie Fellowship para llevar a cabo investigaciones en el Basque Center for Applied Mathematics en Bilbao. Sus contribuciones académicas se centran principalmente en el campo de la econometría y sus aplicaciones en finanzas y economía. Ha publicado en revistas destacadas como *Journal of Econometrics, Journal of Applied Econometrics, Journal of Financial Econometrics, Journal of Banking and Finance* y *Econometric Reviews.*

Ángel León Valle

Es profesor titular de Universidad en el Departamento de Fundamentos del Análisis Económico de la Universidad de Alicante. Es doctor en Economía Financiera desde 1998 por la Universidad de Alicante. Sus áreas de investigación son la econometría financiera, derivados y opciones reales.

María Loureiro

Es catedrática de Fundamentos de Análisis Económico en la Universidad de Santiago de Compostela y directora científica del centro interuniversitario en economía y empresa, ECOBAS. Desde el 2022 es miembro del comité científico del Real Instituto Elcano, y miembro numerario de la Real Academia Gallega de Ciencias.

Juan Miguel Marín

Es licenciado en Biología (rama Fundamental) y en Matemáticas (rama I.O.). A su vez, es doctor en Genética y doctor en Estadística e I.O. por la Universidad Complutense de Madrid. Ha sido T.E.U. en la EUE de la Universidad Complutense de Madrid, en la ESCET de la URJC y T.U. en la ESCET de la URJC. Actualmente es T.U. en la Universidad Carlos III. Trabaja en análisis multivariante, estadística bayesiana y modelos de volatilidad estocástica. Ha publicado en revistas como *Biometrics, Journal of Multivariate Analysis* y *TEST,* entre otras.

Guillermo Mestre Marcos

Obtuvo el título de doctor en Modelado de Sistemas de Ingeniería por la Universidad Pontificia Comillas en 2021. En 2016 se incorporó al área de Sistemas Inteligentes del Instituto de Investigación Tecnológica (IIT), donde desarrolló su actividad investigadora en modelos de predicción de series temporales, análisis de datos funcionales y técnicas de aprendizaje automático aplicado al análisis de los mercados de energía eléctrica. Desde 2021 forma parte del equipo de Inversiones Cuantitativas (QIS) de Santander Asset Management como analista cuantitativo, donde desarrolla metodologías de construcción de carteras aplicando técnicas de analítica avanzada de datos y aprendizaje automático. Sus áreas de interés incluyen la aplicación de técnicas de inteligencia artificial en los mercados financieros, así como en la predicción de series temporales y el uso de técnicas de aprendizaje automático en los mercados eléctricos.

Josep Mestres Domènech

Es economista sénior en la unidad de economía española de CaixaBank Research. Es doctor en Economía por el University College London y máster en Economía por la Universitat Pompeu Fabra. Antes de incorporarse a CaixaBank, ejerció de economista en el Departamento de Empleo y Políticas Sociales (DELSA) de la OCDE, en las divisiones de Migración Internacional y de Empleo. También trabajó previamente en el Centre for Research and Analysis of Migration (CReAM), en el University College London. Sus áreas de estudio comprenden la economía laboral, la economía de la inmigración y las políticas públicas en general. Actualmente coordina los proyectos de analítica de datos de CaixaBank Research (https://realtimeeconomics.caixabankresearch.com/#/home).

Twitter: @josep_mestres

Linkedn: https://www.linkedin.com/in/jmestres

Antonio Montañés Bernal

Es doctor en Ciencias Económicas y Empresariales por la Universidad de Zaragoza y en la actualidad es catedrático del departamento de Análisis Económico de la Universidad de Zaragoza. Es autor de diversas publicaciones científicas y monografías. A destacar sus artículos en revistas como *Journal of Econometrics, Econometric Theory, Econometrics Reviews, Energy Economics, Journal of Health Economics, Journal of International Money and Finance* o Social *Science & Medicine.*

Antonio Muñoz San Roque

Obtuvo el título de doctor ingeniero industrial por la Universidad Pontificia Comillas en 1996. En 1992 se incorporó al Instituto de Investigación Tecnológica (IIT) donde desarrolla su actividad investigadora en el área de Sistemas Inteligentes en estrecha colaboración con la industria. Desde 1994 imparte clases en el Departamento de Electrónica, Automática y Comunicaciones de la E.T.S. de Ingeniería ICAI. Sus áreas de interés son la aplicación de técnicas de inteligencia artificial a la monitorización y diagnóstico de procesos industriales, la predicción de series temporales y la aplicación de técnicas de aprendizaje automático en mercados eléctricos. En 2008 fue nombrado director del IIT. En 2010 pasó a ocupar el cargo de subdirector académico de la Escuela Técnica Superior de Ingeniería ICAI, de la que es director desde 2018.

Trino-Manuel Ñíguez

Es profesor en la Westminster Business School, University of Westminster, London. Es doctor en Economía desde 2004 por la Universidad de Alicante. Su principal área de investigación es la econometría financiera.

Álvaro Ortiz Vidal-Abarca

Es el responsable de análisis de *big data* en BBVA Research. Dirige la unidad que lleva a cabo el análisis de Big Data para cuestiones económicas, sociales y geopolíticas. Es profesor adjunto en IE Business School y ha sido miembro del Grupo de Expertos que ha asesorado al INE en temas relacionados con *big data.* Álvaro Ortiz es doctor en Economía por la Universidad Autónoma de Madrid y posee un Diploma de Estudios Avanzados de Economía Internacional y Política Económica por el Instituto para la Economía Mundial de Kiel (Alemania). Diploma en *Machine Learning* para Economistas (Cemfi) y Certificados profesionales en R (Harvard EdX) y Python(IBM). Ha presentado sus trabajos en numerosas conferencias organizadas por NBER,CEPR y AEA-ASSA así como en conferencias organizadas por bancos centrales (BCE, Reserva Federal, Banco de Inglaterra, Banco de Suecia, Banco de Italia, Hong Kong Monetary Authority y Banco de Turquía)... Ha sido *KeyNote Speaker* en conferencias organizadas por el Ministerio de Economía Alemana y en el Banco de Italia. Ha publicado en varias revistas académicas como *Royal Society Open Science, CEPR Discussion Papers, Cambridge Economics WP, Arxiv (Cornell), Economía the Journal of Lacea, Bank of Spain Working Papers, Opec Energy Review, The Service Studies Journal* y *Moneda y Crédito.*

José Portela

Estudió en la Universidad Pontificia Comillas donde obtuvo el título de ingeniero industrial. En 2011 se incorporó al Instituto de Investigación Tecnológica (IIT) y desde 2017 compagina la investigación con la docencia como profesor en la facultad de Ciencias Económicas y Empresariales de ICADE y en la E.T.S. de Ingeniería del ICAI. Desde 2020 es el coordinador del Área de Sistemas Inteligentes del IIT, enfocada en la investigación de técnicas de inteligencia artificial y su aplicación en el negocio en distintos sectores empresariales. Ha participado en más de 50 proyectos de investigación sobre estos temas en colaboración con empresas (ver iit.comillas.edu/personas/jportela). Sus áreas de interés incluyen modelos de predicción de series temporales, análisis de datos funcionales y técnicas de aprendizaje automático basados en redes neuronales con aplicación en los mercados de energía eléctrica y sistemas industriales.

Tomasa Rodrigo

Es economista líder de la unidad de *big data* en BBVA Research, que se encarga de aplicar la ciencia de datos al análisis económico, social y geopolítico. Tiene una gran experiencia trabajando con infraestructuras en la nube, grandes masas de datos de índole económica, financiera y social. Es licenciada en Economía por la Universidad de Granada (*Summa Cum Laude*) donde trabajó de ayudante de investigación dos años. Realizó un máster en Análisis Económico en la Universidad Carlos III de Madrid y ejerció como docente de Econometría. También realizó un máster de *data scientist* impartido por IBM, junto con numerosos cursos en ciencia de datos y *machine learning*. Es profesora asociada en la Universidad Carlos III y en la Universidad CEU San Pablo. Además, ha publicado en revistas académicas como *Royal Society Open Science, CEPR Discussion Papers, Cambridge Economics WP, Arxiv (Cornell) Economía* the Journal of Lacea, Bank of Spain Working Papers y presentado en foros relevantes en el ámbito de la ciencia de datos como Big Data Spain, Machine Learning Spain y Google Cloud Summit.

Eugenio Fco. Sánchez Úbeda

Obtuvo el título de doctor ingeniero industrial por la Universidad Pontificia Comillas en 1999. En 1991 se incorporó al Instituto de Investigación Tecnológica (IIT), donde desarrolla su actividad investigadora en el campo de las técnicas de aprendizaje automático y predicción, dentro del área de Sistemas Inteligentes. Ha participado en más de cien proyectos de investigación sobre estos temas en colaboración con empresas de diferentes sectores (ver iit.comillas.edu/personas/euge). Imparte clases de Estadística, Aprendizaje Automático y Análisis de Datos en la Escuela Técnica Superior de Ingeniería ICAI. Sus áreas de interés incluyen las técnicas de inteligencia artificial y su aplicación en los mercados eléctricos y en sector de la salud.

Helena Veiga

Es doctora en Economía por la Universidad Autónoma de Barcelona y actualmente profesora titular del Departamento de Estadística de la Universidad Carlos III de Madrid. Su investigación se centra en la econometría financiera, especialmente en la modelización de la volatilidad estocástica, métodos de estimación de estos modelos, análisis de fronteras estocásticas y finanzas empíricas. En el ámbito de las finanzas empíricas, recibió financiación del laboratorio social de la Caixa para realizar varios experimentos sobre el contagio de los mercados financieros. Helena Veiga ha sido editora, junto con Sofia Ramos, del libro The Interrelationship Between Financial and Energy Markets, publicado por Springer. Además, ha publicado sus trabajos en destacadas revistas como Journal of Banking and Finance, Journal of Empirical Finance, European Review of Economics, Energy Economics, Experimental Economics, Computational Statistics and Data Analysis, Journal of Productivity Analysis, Econometric Reviews y European Journal of Operational Research, entre otras.

Funcas

Caballero de Gracia, 28
28013 Madrid
Teléfono: 91 596 54 81
Fax: 91 596 57 96

publica@funcas.es
www.funcas.es

ISBN 978-84-17609-80-1