BIG DATA

NUEVOS MÉTODOS DE PREDICCIÓN ECONÓMICA CON DATOS MASIVOS

Daniel Peña
Pilar Poncela
Esther Ruiz
(editores)

funcas

BIG DATA

NUEVOS MÉTODOS DE PREDICCIÓN ECONÓMICA CON DATOS MASIVOS

Daniel Peña
Pilar Poncela
Esther Ruiz
(editores)

funcas

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

ISBN: 978-84-17609-48-1
Depósito legal: M-5439-2021
Imprime: Cecabank

Contenido

Colaboradores

Arratia, Argimiro
Ciencias de la Computación
Universidad Politécnica de Cataluña

Blázquez, Desamparados
Universitat Politècnica de València

Brownlees, Christian
Universitat Pompeu Fabra y Barcelona GSE

Camacho, Máximo
Universidad de Murcia

Caro, Ángela
Departamento de Estadística
Universidad Carlos III de Madrid

Debón, Ana
Universitat Politècnica de València

Doménech, Josep
Universitat Politècnica de València

Ghirelli, Corinna
Banco de España

Guillen, Montserrat
Departamento de Econometría, Estadística
y Economía Aplicada-Riskcenter-IREA
Universitat de Barcelona

Hurtado, Samuel
Banco de España

Láinez, María
Dirección Corporativa Actuarial, MAPFRE S.A.

Llorens, Jordi
Universitat Pompeu Fabra y Barcelona GSE

Colaboradores

Pascual, Lorenzo
Deloitte Consulting-Analytics & Cognitive

Peña, Daniel
Departamento de Estadística y UC3M-Santander Big Data Institute
Universidad Carlos III de Madrid

Pérez, Javier J.
Banco de España

Pérez-Marín, Ana M.
Departamento de Econometría, Estadística
y Economía Aplicada-Riskcenter-IREA
Universitat de Barcelona

Poncela, Pilar
Departamento de Análisis Económico, Economía Cuantitativa
Universidad Autónoma de Madrid

Ramallo, Salvador
Universidad de Murcia

Ruiz, Esther
Departamento de Estadística
Universidad Carlos III de Madrid

Ruiz Marín, Manuel
Universidad Politécnica de Cartagena

Sánchez, Eduardo
Dirección Corporativa Actuarial, MAPFRE S.A.

Senar, Nuria
Analytics Centre of Excellence, Sano

Urtasun, Alberto
Banco de España

Presentación

En este siglo la revolución en la velocidad de almacenamiento y computación, de transmisión de la información y la disponibilidad de grandes bases de datos masivos, el conocido en el argot como *big data,* han abierto en la última década nuevas posibilidades en la predicción de fenómenos complejos en economía y finanzas. Desde el campo de aprendizaje automático (*Machine Learning, ML* por sus siglas en inglés) e inteligencia artificial, se han desarrollado nuevos métodos de predicción que, prescindiendo de hipótesis sobre el modelo generador de los datos, tratan de explotar las relaciones (posiblemente no lineales) existentes entre muchas variables en grandes bancos de datos generados automáticamente. Estos nuevos datos han impulsado avances en Estadística y Econometría en el análisis conjunto de grandes sistemas de series temporales interdependientes, con el objetivo de mejorar las predicciones. En particular, la combinación de varios métodos y modelos está proporcionando buenos resultados en gestión de carteras, análisis de coyuntura, modelos de riesgo y otras muchas áreas. Este libro tiene por objeto presentar y evaluar algunas aplicaciones en economía y empresa de los nuevos enfoques de predicción basados en datos masivos.

En sus ocho capítulos, reconocidos expertos en este campo analizarán cómo pueden mejorarse las predicciones incorporando a los enfoques tradicionales los resultados de los modelos factoriales dinámicos, árboles de decisión, bosques aleatorios (*Random Forest*) y redes neuronales, entre otros procedimientos desarrollados en los últimos años. También, se analizará como el análisis de sentimientos y de textos, así como otras nuevas herramientas no convencionales que se basan en la abundancia de datos no estructurados pero con potencial valor predictivo, pueden aportar mejoras a los métodos de predicción habitualmente utilizados. El conjunto de los capítulos publicados en este libro representa una visión amplia de aplicaciones de procedimientos de predicción ML en el contexto de variables económicas y financieras.

En el primer capítulo, **Ángela Caro** y **Daniel Peña** describen la evolución histórica de la predicción basada en series temporales y revisan algunas de las herramientas actuales más utilizadas para la predicción de datos masivos. Utilizando *redes neuronales profundas (Deep learning)* para la predicción de varias magnitudes macroeconómicas en 35 países de la OCDE. Estas predicciones se comparan con las obtenidas mediante modelos univariantes y modelos factoriales dinámicos con el objetivo de ilustrar las dificultades con las que se encuentra el analista en la construcción de las reglas de predicción.

En el segundo capítulo, **Christian Brownless, Jordi Llorens** y **Nuria Senar** justifican la necesidad de *regularización,* o imponer restricciones sobre los parámetros en la estimación, para calcular matrices de varianzas y covarianzas dinámicas en el contexto de carteras de activos financieros de alta dimensión. Se demuestra e ilustra cómo la regularización aumenta la precisión en la estimación de dichas matrices.

En el tercer capítulo, **Máximo Camacho, Salvador Ramallo** y **Manuel Ruiz Marín** realizan una descripción de los *árboles de decisión* en el contexto de datos independientes observados en un momento del tiempo determinado. La metodología es implementada para predecir el precio de la vivienda en Madrid. Dado el fuerte componente no lineal de la relación entre el precio y sus determinantes, la utilización de árboles de decisión es ventajosa frente a metodologías más tradicionales que pueden tener dificultades representando dichas no linealidades.

El cuarto capítulo, escrito por **Montserrat Guillen, María Láinez, Ana M. Pérez-Marín** y **Eduardo Sánchez,** analiza cómo los datos masivos están afectando a la medición del riesgo y la fijación de primas en seguros de automóviles. El trabajo presenta un interesante análisis sobre cómo es posible utilizar la *información de los sensores en los vehículos* para monitorizar la conducción y detectar los factores que afectan al riesgo en el contexto de los seguros de automóviles.

El capítulo quinto presenta la contribución de **Desamparados Blázquez, Josep Doménech** y **Ana Debón.** Estos autores utilizan un repositorio público y de libre acceso que contiene *capturas de más de 400 billones de sitios web* y analizan cómo la información sobre cambios en las páginas web de una empresa puede aportar información sobre la supervivencia empresarial.

El capítulo seis, debido a **Argimiro Arratia,** utiliza la información escrita en medios digitales que contiene información no estructurada puede combinarse con información cuantitativa tradicional para mejorar las predicciones. Se describe cómo construir un modelo de predicción con *indicadores de sentimiento* derivados de datos textuales para series temporales financieras que incluyen indicadores extraídos de noticias sobre mercados financieros.

En el capítulo siete **Corinna Guirelli, Samuel Hurtado, Javier J. Pérez** y **Alberto Urtasun** analizan los desafíos a los que los bancos centrales se enfrentan a la hora de utilizar datos granulares y obtenidos con mayor frecuencia de las previamente habituales, para llevar a cabo sus funciones. Los autores señalan que, en 2019, más del 60% de los bancos centrales utilizaron este tipo de datos en sus operaciones y dos tercios de ellos, los usaron como instrumento en el diseño de sus políticas. Entre los casos de éxito descritos aparecen la *medición de la incertidumbre económica basada en artículos de prensa,* el uso de informes regulares de los bancos centrales como herramienta de comunicación sobre el estado de la economía y la predicción macroeconómica.

Finalmente, **Lorenzo Pascual** y **Esther Ruiz** realizan una *descripción de la literatura sobre aplicaciones* empíricas en las que los procedimientos ML se han utilizado con éxito en la predicción económica. En comparación con procedimientos tradicionales, ML tiene mejor comportamiento predictivo en situaciones en las que las no-linealidades son importantes (como en épocas de crisis) y cuando el horizonte de predicción se incrementa.

Estos trabajos fueron presentados en una Jornada celebrada en Funcas el 8 de octubre de este año y el lector interesado puede encontrar en la web las grabaciones de las presentaciones que allí se hicieron. Los editores queremos agradecer a Funcas su apoyo en la realización de este libro que esperamos contribuya a difundir nuevas oportunidades para mejorar la predicción incorporando datos masivos.

Daniel Peña, Pilar Poncela y **Esther Ruiz**

Diciembre, 2020

CAPÍTULO I

Predicción de series temporales económicas con datos masivos: perspectiva, avances y comparaciones

Ángela Caro*
Daniel Peña**

Este trabajo analiza cómo la predicción económica ha ido evolucionando en función de los datos disponibles y cómo la reciente disponibilidad de datos masivos está transformando los métodos utilizados para el pronóstico. Se revisan brevemente tres períodos en la evolución de los procedimientos de predicción económica y empresarial y se presentan las características de una cuarta etapa, que se ha iniciado en este siglo con la revolución del Big data. Se analizan los cambios metodológicos para construir predicciones basadas en modelos econométricos, estadísticos y de aprendizaje de máquina *(machine learning)* y se describen algunos de los más utilizados para la predicción con series temporales. Como ilustración, se comparan las predicciones de un conjunto de variables que describen el ciclo económico en los países de la OCDE obtenidas con un modelo factorial dinámico y una red neuronal recurrente.

Palabras clave: CART, combinación de predicciones, inteligencia artificial, matrices de correlación, modelos no lineales, redes neuronales.

* Con el apoyo del Ministerio de Ciencia, Educación y Universidades de España con referencia FPU15/03983.

** Con el apoyo parcial de la Agencia Nacional de Evaluación de la Calidad y Acreditación con referencia PID2019-109196GB-I00.

1. INTRODUCCIÓN

En la actualidad, una proporción creciente de fenómenos o actividades que realizamos (ocio, salud, trabajo, etc), o que observamos (en procesos climáticos o ambientales, de producción, comerciales o agrícolas), se controla con sensores (en teléfonos móviles, ordenadores, aparatos TIC, etc.) y genera automáticamente datos de forma continua y con bajo coste marginal. Esta situación crea gigantescas bases de datos, conocidas con el nombre de *big data,* que contienen muchas variables, con frecuencia recogidas con localización geográfica y temporal, y que se almacenan, transmiten y analizan, de forma digital. Estos datos incluyen medidas numéricas, pero también imágenes, vídeos o audios. Por otro lado, los avances en las tecnologías TIC permiten procesar y analizar con gran velocidad estas grandes masas de datos. El análisis de este big data está cambiando nuestro ocio, nuestro trabajo, el cuidado de nuestra salud o nuestras relaciones sociales, y transformará también, en el futuro, nuestra democracia y organización social. En particular, ha modificado cómo aprendemos de los datos y los utilizamos para hacer predicciones.

Esta situación de abundancia de datos es nueva: hasta hace pocos años, el problema principal para la predicción era su escasez. En la actualidad, el problema es cómo extraer de un exceso de información la relevante, y cómo combinar muchas variables temporales, con frecuencia medidas en series de distinta frecuencia y periodicidad, con otras creadas a partir de imágenes, vídeos o audios. Este nuevo escenario esta transformando los métodos de predicción ya que su evolución histórica ha venido precisamente condicionada por la disponibilidad de los datos.

Este trabajo presenta una breve panorámica de algunas de las herramientas que se han utilizado en la predicción con series temporales y de las actuales para datos masivos. En la sección siguiente se analiza cómo la existencia de datos ha condicionado los métodos de predicción. Se describen tres etapas en su evolución y se justifica que desde principios de este siglo hemos entrado en un nuevo período. En la sección tercera se analiza el cambio de paradigma en este nueva etapa, que ha evolucionado de los modelos causales a las reglas empíricas de predicción, y del mejor modelo a la combinación de muchos. La cuarta sección describe algunos de los métodos utilizados hoy, que tratan de aprovechar el potencial de los datos masivos. En la sección quinta se comparan los modelos factoriales con las redes neuronales en un ejemplo de previsión del ciclo económico. La sección 6 incluye unos comentarios finales.

2. LOS DATOS Y LA PREDICCIÓN ECONÓMICA

La predicción en economía, y en general en las ciencias sociales, comienza a realizarse a partir de los datos en el siglo XVII, pero no se establece como disciplina científica basada en métodos probabilistas hasta la segunda mitad del siglo XX. Podemos considerar cuatro períodos distintos en su evolución. El primero, (1649-1940), abarca desde

los trabajos pioneros de Graunt, a mediados del siglo XVII, hasta la aparición de la econometría en los años 40 del siglo pasado. Comienza con los trabajos de Graunt (1620-1674) sobre datos demográficos en Inglaterra y la predicción del sexo en los nacimientos. Continúa con los trabajos de Quetelet (1796–1874), un astrónomo belga que intenta identificar leyes naturales en los fenómenos sociales, con la esperanza de construir un marco general similar a la propuestas por Newton para el mundo físico. La teoría de Newton supuso la primera explicación coherente de la naturaleza, incluyendo el movimiento de los cuerpos en el espacio y los objetos en la tierra, y de ella se deducen predicciones contrastables de lo que ocurrirá en fenómenos físicos observables. Esta capacidad profética de la teoría de Newton estimula a los científicos para recoger datos astronómicos y físicos para contrastarla. La necesidad de ajustar ecuaciones lineales a los datos observados lleva a Legendre (1752-1833) y a Gauss (1777-1855) a descubrir la estimación de mínimos cuadrados, que es utilizada por F. Galton (1822-1911) con datos biológicos, recogidos para contrastar la teoría de su primo Darwin, para introducir el concepto de regresión. Poco después, K. Pearson (1857-1936) inventa el coeficiente de correlación. Gracias a estos autores, a principios del siglo XX las ideas básicas para relacionar variables no deterministas quedan bien establecidas, y pocos años después aparecen los primeros manuales de predicción económica como el de Morgenstern (véase Clements y Hendry, 1998), donde se señalaba ya la escasez de datos como la principal limitación para la predicción económica.

El segundo período (1940-1975) se inicia con la creación de la econometría como ciencia a partir de los trabajos de la comisión Cowles entroncando la predicción económica dentro de la estadística con un enfoque probabilista, y finaliza con la entrada del ordenador en los centros de predicción económica. En este período el objetivo es construir un modelo causal, donde la variable de interés se explica por otras variables explicativas o exógenas que influyen sobre la que queremos prever en un sentido estadístico. Los modelos utilizados son inicialmente los modelos de regresión, que se generalizan en los sistemas de ecuaciones simultáneas. Con ellos podemos hacer predicciones de escenarios condicionados por las variables exógenas. Jan Tinbergen (1903-1994), que recibió el primer premio Nobel de economía, y Haavelmo (1944) son representantes destacados de este enfoque. Tinbergen se formó en física en Holanda y trasladó con éxito las ecuaciones de evolución de la dinámica de un sistema físico a la economía. Un libro clásico de este período, escrito por un compatriota de Tinbergen y profesor de la Universidad de Chicago, es Theil (1971). En los modelos construidos en esta etapa hay un predominio de la teoría sobre los datos, se basan en relaciones lineales, y, en el caso de sistemas de ecuaciones, contienen pocas variables por las limitaciones existentes, tanto de datos como de medios de cálculo.

El tercer período (1975-2000) se caracteriza por una flexibilizacion de las hipótesis para construir modelos unida a un crecimiento continuo y acelerado de los datos disponibles y de los métodos de cálculo y almacenamiento. La posibilidad de series temporales largas, de relacionar decenas de variables y los avances en computación estimulan mode-

los más flexibles y adaptativos. Por ejemplo, la estructura dinámica de retardos en la transmisión de los efectos entre variables comienza a determinarse a partir de los datos, y no a especificarse *a priori.* También, se incia la incorporación de la heterogeneidad en las variables: datos atípicos, cambios estructurales, parámetros que cambian con el tiempo, heterocedasticidad condicional, etc. El crecimiento de los métodos de cálculo permite explorar la no linealidad con métodos no paramétricos de suavizado. Las relaciones entre variables se hacen más flexibles y los datos se utilizan para descubrir si una variable explica la tendencia de otras (cointegración) o no. En lugar de modelos causales con muchas variables aparece la posibilidad de utilizar modelos factoriales dinámicos, donde la relación entre las variables se determina empíricamente a través de variables no observables, o latentes, que se estiman a partir de las dadas. Dos libros característicos de estos cambios son Box y Jenkins (1976), que introduce en series temporales la estimación de modelos no lineales ARIMA y la determinación empírica de la dinámica de las variables, y Engle y Granger (1991) que establece el concepto de cointegración relacionándolo con el equilibrio económico. Un manual que describe bien el enfoque dominante en este período es Greene (1993), y la situación de los métodos de predicción en Clements y Hendry (1998).

El cuarto período se inicia en este siglo, con la aparición del big data, y estamos asistiendo a su rápido desarrollo. El nombre de "big data" se crea en 1997 por dos investigadores de la NASA para poner de manifiesto cómo el gran aumento de datos lleva al límite a los sistemas informáticos existentes, y en 2001 se caracteriza por las tres V (Velocidad, Volumen y Variedad). En 2001 se desarrolla la web 2.0 con participación de los usuarios y aparecen las redes sociales, Wikipedia y los blogs. Desde entonces, el crecimiento de los datos disponibles es exponencial; véase por ejemplo el informe COTEC, 2017. Muchos trabajos han analizado los cambios en la metodología estadística y econométrica como consecuencia del big data; véase como ejemplo de distintos enfoques Bühlmann y Van De Geer (2011); Varian (2014); Fan, Han y Liu (2014); Efron y Hastie (2016); Donoho (2017); Athey (2017); Giannone, Lenza y Primiceri (2017); Blazquez y Domenech (2018); Galeano y Peña (2019) y Hsiao (2020). Podemos concluir que en este siglo se produce un cambio de paradigma en el análisis de datos y, en particular, en la predicción económica, que desarrollaremos en la sección siguiente.

3. EL ENFOQUE DE PREDICCIÓN CON BIG DATA

La abundancia de variables estructuradas, como tablas de datos, y no estructuradas, como textos, imágenes o vídeos, ofrece nuevas posibilidades para la predicción y conduce a un cambio de perspectiva. En lugar de, como en el pasado, estimar el mejor modelo para los datos observados ahora se construyen reglas de predicción flexibles y heterogéneas con capacidad demostrada de prever bien datos diferentes de los utilizados para estimarlas. La metodología para obtenerlas se basa en los tres principios siguientes: (1) Se construyen utilizando las relaciones empíricas entre variables detectadas en la muestra; (2) se seleccionan por su capacidad predictiva fuera de la muestra;

(3) el predictor final utilizado combina distintos modelos, procedimientos y tipos de datos. A continuación, desarrollamos estos tres principios.

3.1. Utilizar las relaciones empíricas entre variables

En los modelos econométricos tradicionales las variables que se incluyen se determinan a partir de la relación teórica esperada entre la variable respuesta, o endógena, y las explicativas o exógenas. Esto no excluye que también se pueda explorar el efecto de otras variables disponibles, cuyo efecto *a priori* sea menos claro. En los modelos dinámicos los retardos con los que actúan las variables, en su caso, se suelen obtener empíricamente, con los datos observados. Sin embargo, al aparecer la posibilidad de incluir muchas más variables, nuevos datos, como textos, imágenes o vídeos, y modelar relaciones a muy corto plazo o muy desagregadas, donde la teoría es inexistente o muy débil, es más operativo explorar empíricamente qué variables muestran capacidad predictiva.

Por ejemplo, los modelos factoriales dinámicos (DFM, por sus siglas en inglés), que son en la actualidad los más utilizados para la predicción de muchas series económicas o empresariales, o las redes neuronales y el deep learning, que se utilizan para la predicción en *machine learning,* generan reglas de predicción donde la dependencia entre las series se transmite por ciertas variables no observadas, llamadas variables latentes o factores, cuya composición se determina a partir de los datos. En otros modelos, como los árboles de decisión, la regla de predicción se obtiene buscando las particiones de los valores de las variables más útiles para la predicción, de forma totalmente empírica. Finalmente, si mezclamos información espacial, temporal y de imágenes, vídeos o textos, por ejemplo para la predicción de las ventas de un determinado producto en un supermercado, las relaciones entre las variables tradicionales y las que podemos construir con la nueva información (píxeles de imágenes en la tienda, comentarios en las redes sociales, etc) son desconocidas y solo pueden determinarse empíricamente.

Tradicionalmente, en estadística trabajar con reglas empíricas, construidas a partir de la muestra, se ha considerado poco aconsejable por dos razones. En primer lugar, existe un fuerte riesgo de encontrar pautas que aparecen por azar en la muestra, que no serán efectivas en otros datos. En segundo lugar, se valora el principio de parsimonia: incluir los parámetros necesarios para la predicción, pero no más. Respecto a la primera razón, podemos encontrar variables que parecen ser efectivas para prever aunque no tengan relación causal con la respuesta. Son las llamadas relaciones espurias, donde dos variables sin conexión causal varían conjuntamente, generalmente por otra variable que influye sobre ambas en el período estudiado. Por ejemplo, la relación entre el número anual de burros en España y el presupuesto en educación de cada año en los años de desarrollo en España, que se movían en sentidos opuestos por el crecimiento del país. En los libros de estadística y econometría abundan estos ejemplos de relaciones no de causa efecto general, sino de covariación en un período. Sin embargo, a

veces estas relaciones entre variables independientes, pero con relación empírica en el período estudiado debido al efecto de otras, puede utilizarse con éxito para la predicción. Con datos masivos, es frecuente encontrar relaciones insospechadas entre conjuntos de variables que efectivamente mejoran la predicción. Incluir estas variables tiene sentido si comprobamos su capacidad predictiva fuera de la muestra, que es el segundo principio que explicamos en la sección siguiente.

La justificación de la parsimonia es que cada parámetro puede mejorar el ajuste dentro de la muestra, pero aumentar el error de predicción fuera de ella. No conviene, en consecuencia, introducir parámetros que no ayuden a la predicción. Este es el principio que lleva a los criterios de selección de modelos como el de Akaike, que estima el error esperado fuera de la muestra, y a los procedimientos de validación cruzada, que comentaremos a continuación.

Este enfoque pragmático es adecuado para predecir en situaciones nuevas, que están cambiando en el tiempo y dónde no existe una teoría contrastada para guiarnos en la construcción de reglas de predicción. Además, puede descubrirnos aspectos desconocidos del fenómeno descrito por la variable, o variables, a prever, y ser un primer paso para generar conocimiento en este campo y construir nuevas teorías. Para ello, las relaciones encontradas deben someterse a un escrutinio cuidadoso, con técnicas de diseño de experimentos, para detectar las verdaderas relaciones causales.

3.2. Elegir la regla de predicción por su capacidad predictiva fuera de la muestra

Es bien conocido desde los trabajos pioneros de Akaike (recogidos en Akaike, 1998), y Stone (1974) en los años 70 que el error de predicción dentro de la muestra no es un buen criterio para elegir modelos, y, desde los años 70, se han ido introduciendo en estadística otros métodos para seleccionar el mejor modelo. Los dos enfoques más utilizados son: (1) utilizar un criterio de selección que estime el error de predicción esperado fuera de la muestra, o que penalice el número de parámetros utilizados; (2) aplicar validación cruzada, *(cross validation)* y calcular el error de predicción dividiendo la muestra en dos partes, estimando el modelo en una de ellas y calculando en la otra el error de predicción fuera de la muestra. La ventaja del segundo enfoque es que no es necesario realizar hipótesis sobre la generación de los datos para aplicarlo, como ocurre con los criterios de selección de modelos que se construyen siempre bajo ciertas hipótesis. El enfoque de validación cruzada es más general por su carácter no paramétrico y su falta de restricciones.

Para aplicar correctamente la validación cruzada es imprescindible que la muestra de validación no se utilice en absoluto para ninguna decisión relacionada con la construcción del modelo y solo para la validación del mismo. Hay distintas formas de realizar

la validación cruzada para datos independientes, por ejemplo dividir la muestra en K partes al azar, dejar una parte fuera para validarlo y estimar el modelo con las restantes K-1. El proceso se repite para cada una de las K partes y se hace el promedio de los resultados obtenidos. Este método se llama K-validación cruzada. En el caso particular de $K=N$ el modelo se estima con N-1 datos y se valida con N. Estos métodos no son adecuados para series temporales, porque al dividir al azar destruimos el orden temporal de las observaciones.

Para datos temporales la manera más habitual de dividir la muestra en dos partes es utilizar los primeros T_1 períodos para construir el modelo y los siguientes T_2 para validarlo, donde $T=T_1+T_2$. Supongamos como ejemplo, que se desea comparar modelos por su predicción un período hacia delante y sea $T_0 \geq T_1$ el origen de estas predicciones. Para $T_0=T_1$ con el modelo estimado en T_1 se predice el valor de la serie para $t=v+1$. Para $T_0>T_1$ el modelo puede o no reestimarse y las tres estrategias más habituales son:

- Utilizar en todas las predicciones, con cualquier origen, los parámetros estimados en T_1. Este método se denomina predicción con estimación fija.

- Actualizar los parámetros con el origen de la predicción estimándolos en una muestra de tamaño T_1 que finaliza en el origen de la predicción. Se descartan las primeras observaciones y el intervalo de estimación es $(T_0\text{-}T_1+1, T_0)$. Este método suele llamarse *rolling forecast* o rodar las predicciones en español.

- Actualizar los parámetros con el origen de la predicción incluyendo todos los datos disponibles desde el inicio hasta el origen de la predicción, tomando como intervalo de estimación $(1, T_0)$. Este método se conoce como estimación recursiva y el tamaño de la muestra de estimación se incrementa con el origen de la predicción.

La comparación de modelos puede hacerse con cualquiera de estos tres procedimientos. El primero tiene la ventaja de que solo estimamos una vez y tiene sentido si se elige T_1 mucho mayor que T_2. El segundo nos permite estudiar la estabilidad de los parámetros a lo largo del tiempo. El tercero utiliza toda la información disponible en cada momento, pero hace más difícil la comparación de parámetros, al estar estimados con distinto tamaño muestral. En general, es interesante comparar los modelos con todas las predicciones posibles en la muestra de validación, que permite hacer $T_2\text{-}h+1$ predicciones de horizonte h para $h \leq T_2$.

Otros procedimientos para dividir la muestra son posibles, aunque aumentan la carga computacional. Peña y Sánchez (2005) mostraron como hacer predicciones con la mayoría de los datos de la serie temporal haciendo un tipo especial de validación cruzada. Este campo, sin embargo, debe desarrollarse mucho más para encontrar procedimientos más robustos y eficaces de dividir la muestra y validar los modelos elegidos.

Evaluar al modelo por su capacidad predictiva sustituye al enfoque de seleccionar las variables con los contrastes clásicos de significación sobre los coeficientes de un modelo estimado. De hecho, estos contrastes se han utilizado con mucha frecuencia para tomar decisiones sin fundamento y de forma inapropiada, construyendo modelos con poca capacidad predictiva. Por ejemplo, si tenemos una muestra pequeña, una variable con un efecto importante para la predicción de la variable respuesta puede no detectarse como significativa y, con una muestra muy grande, una variable prácticamente irrelevante para la predicción puede aparecer como muy significativa. En particular, en regresión simple el estadístico del contraste t es significativo cuando el coeficiente de correlación entre la variable respuesta y la que contrastamos es mayor en valor absoluto que $2/\sqrt{n}$, donde n es el tamaño muestral (veáse, por ejemplo, Peña [2002, pág. 275]). Por tanto, en una muestra de 20 observaciones, un efecto tiene que explicar más del 20% de la variabilidad para ser significativo y en una muestra de 200.000 datos es suficiente que explique el 0,005 de la variabilidad para serlo, aunque este efecto sea irrelevante en la práctica.

Esta dependencia tan fuerte de las conclusiones del tamaño muestral implica que el procedimiento habitual de incluir variables en un modelo cuando su *p*-valor es menor que 0.05, o su estadístico *t* es mayor en valor absoluto que dos, no conduce a buenas reglas predictivas. Como el análisis del big data ha puesto de manifiesto, los contrastes de significación tienden a rechazar cualquier hipótesis si el tamaño de la muestra es suficientemente grande. Por esta, y otras causas, la utilización de *p-valores* y de contrastes de significación en el análisis de datos ha sido formalmente desaconsejada por The American Statistical Association (ASA) en un comunicado oficial (Wasserstein y Lazar, 2016). Esta asociación ha publicado un número extraordinario de *The American Statistician* en 2019 con más de 40 trabajos que, desde distintos puntos de vista, recomiendan basarse en la estimación y no en contrastes de significación para tomar decisiones científicas sobre las relaciones entre variables. Es la primera vez en la historia de ASA que se emite un comunicado tan rotundo sobre los peligros de utilizar un método que se incluye en todos los programas de cálculos estadísticos para construir modelos o contrastar hipótesis científicas.

3.3. Combinar muchos modelos y tipos de datos

El paradigma clásico de la estadística es encontrar el mejor modelo. El concepto de modelo óptimo esta bien definido en entornos simples, bajo fuertes hipótesis sobre el proceso generador de los datos, pero empieza a desdibujarse cuando admitimos incertidumbre sobre este proceso o prescindimos de un modelo generador único. Sin un modelo óptimo que reproduzca el proceso generador, resulta razonable considerar todos aquellos que expliquen bien los datos y combinarlos después para construir la predicción. En el enfoque bayesiano, que admite incertidumbre sobre los modelos, hay una forma simple de abordar este problema. Supongamos que tenemos unos datos $D=\{y, X\}$ y un conjunto de modelos M_i, $i=1,...,I$, con probabilidades *a posteriori* $P(M_i|D)$. La esperanza condicionada de una observación futura y_f, $E(y_f|D)$, que es la predicción que minimiza el error cuadrático de predicción, tiene la forma:

$$E(y_f \mid D) = \sum_{i=1}^{I} E(y_f \mid M_i, D) P(M_i \mid D) \qquad [1]$$

y es una combinación lineal de las predicciones realizadas con cada modelo, $E(y_f \mid M_i, D)$, ponderadas por sus probabilidades *a posteriori,* $P(M_i \mid D)$ En el enfoque clásico los modelos carecen de probabilidades pero un procedimiento habitual es combinarlos proporcionalmente a su capacidad predictiva. Como la probabilidad *a posteriori* $P(M_i \mid D)$ depende de la capacidad predictiva, en la práctica ambos enfoques son similares, aunque el bayesiano tiene una justificación más clara. Existe una amplia literatura en combinación de modelos para la predicción. Véase Draper (1995); Meade e Islam (1998); Min y Zellner (1993); Yuan y Yang (2005); Koop y Potter (2004); Raftery *et al.* (2005); Bishop (2006) y Heitz *et al.* (2009). Sin embargo, además de estos métodos de combinación, que se ha aplicado mucho con modelos estadísticos y econométricos, se han desarrollado otros más orientados a las reglas de predicción que describimos a continuación. Estos métodos se conocen como *ensamble methods,* o combinación de métodos, en la literatura de *machine learning.*

Boosting: *combinar muchos modelos simples*

Este método fue introducido con mucho éxito para problemas de clasificación pero se ha extendido a problemas de predicción. La idea es crear una regla de predicción compleja combinando modelos muy simples que se ponderan por su precisión. En esencia, funciona como sigue para construir una regla de predicción lineal para una variable en función de un conjunto amplio de otras variables (que incluyen retardos de todas ellas):

- Seleccionar un predictor simple, P_0, por ejemplo, hacer una regresión lineal con una variable o un árbol de decisión (CART) (que se introducen en la sección siguiente) con pocas ramas.
- Calcular los errores de predicción y el error medio del predictor.
- Para $i = 1,\ldots,B$, calcular los residuos o parte no explicada por el predictor actual y construir un nuevo predictor tomando los residuos como nuevos datos. Volver al paso dos e iterar hasta que el nuevo predictor reduzca el error de predicción hasta un límite fijado.

Construir el predictor final como:

$$P_f = w_0 P_0 + w_1 P_1 + \ldots + w_B P_B \qquad [2]$$

Freund, Schapire y Abe (1999) presentan una simple introducción a este método con ejemplos de su aplicación. Observemos que cuando tomamos como predictor simple la regresión con una variable, el modelo final obtenido es en general diferente del que

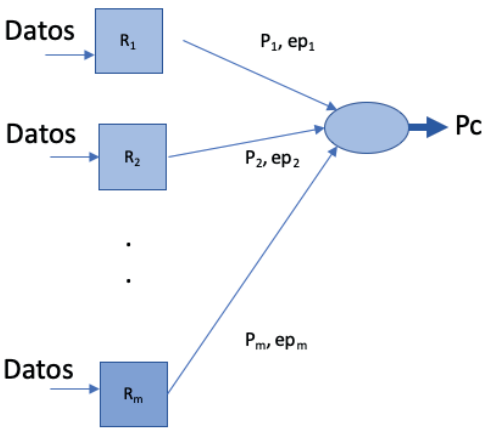
proporcionan los métodos clásicos de regresión paso a paso, ya que la regla final de predicción con boosting no corresponde a una estimación de un modelo de regresión múltiple por mínimos cuadrados. En estos métodos de selección de variables tradicionales vamos escogiendo nuevas variables por su capacidad explicativa de los residuos del último modelo estimado, pero también eliminamos las no significativas. El modelo final es simplemente un modelo de regresión múltiple estimado por mínimos cuadrados con las variables elegidas que son significativas. En boosting, el modelo final es una ponderación de los modelos simples, sin eliminar nada y combinando todos los modelos por su capacidad predictiva y, por tanto, es también un modelo lineal, pero donde los coeficientes de las variables que aparecen no se han estimado por mínimos cuadrados ni se han eliminado variables por contrastes de significación.

Bagging: *combinar el mismo tipo de modelo pero estimado con réplicas de los datos*

En este enfoque modificamos al azar los datos generados mediante muestras *bootstrap,* calculamos la predicción y los resultados obtenidos en cada muestra se promedian para obtener el predictor final. El nombre de *bagging, bootstrap aggregation,* se debe a Breiman (1996). La figura 1 ilustra el método. Se toman m conjuntos de datos por muestreo con reemplazamiento y se construye un predictor en cada conjunto, P_i, se calcula su error relativo promedio, ep_i, y las predicciones se promedian, o se combinan, teniendo en cuenta su error relativo, para obtener la predicción final, Pc. Si se utilizan las mismas variables se promedia, mientras que cuando son distintas queda la opción de ponderarlas.

FIGURA 1

COMBINACIÓN DE MODELOS MEDIANTE *BAGGING*



Fuente: Elaboración propia.

Bosques aleatorios *(random forests) combinar el mismo tipo de modelo pero estimado con réplicas de los datos y con variables diferentes*

Cuando se aplica la idea de perturbar la muestra mediante bagging en árboles de decisión la perturbación además de respecto a los datos suele hacerse también respecto a las variables. La idea es seleccionar al azar un conjunto de ellas en cada nodo para hacer la división, con lo que se crean un conjunto de árboles cuyas predicciones se promedian. También pueden combinarse teniendo en cuenta su precisión relativa.Estos son los llamados bosques aleatorios o *random forests.*

Combinar distintos tipos de datos

Con distintos tipos de datos podemos crear un modelo que los englobe a todos. Esto se ha hecho tradicionalmente en la predicción económica, por ejemplo incorporando datos mensuales en predicciones cuatrimestrales, pero, recientemente se han desarrollado enfoques para la predicción a muy corto plazo combinando datos de distinta frecuencia. Estos métodos se denominan de *nowcasting,* y han tomado el nombre de la predicción meteorológica *(Now and forecasting).* La literatura es ya extensa pero un trabajo pionero es Giannone, Reichlin y Small (2008). El método MIDAS *(mixed-data sampling),* combina datos temporales de distinta frecuencia; véase Kuzin, Marcellino y Schumacher (2011) y Meade e Islam (1998) para una descripción del mismo y ejemplos de su aplicación.

En otros casos mezclamos datos de series temporales y de sección cruzada, Galeano y Peña (2019), o textos, imágenes de video y audios con variables tradicionales. Un ejemplo reciente del uso de vídeos puede encontrarse en Sun *et al.* (2019).

4. MODELOS UTILIZADOS PARA LA PREDICCIÓN CON BIG DATA

A continuación, presentamos una breve introducción a los modelos estadísticos/econométricos y de machine learning más utilizados para hacer predicciones con grandes conjuntos de series temporales.

4.1. Modelos factoriales dinámicos

Los modelos factoriales dinámicos (DFM) son en la actualidad los más utilizados para la predicción de muchas series económicas o empresariales. Fueron introducidos en econometría por Geweke (1977) y Gary y Rothschild (1983) y en estadística por Engle y Watson (1981) y Peña y Box (1987). En estos modelos la dependencia entre las series es consecuencia de ciertas variables no observadas, llamadas variables latentes o factores, cuya composición se determina a partir de los datos. La estructura de un DFM de series

temporales implica una descomposición de los valores de cada serie en dos componentes: un componente común, que recoge el efecto en esa serie de los factores comunes, y otro específico o idiosincrático, que resume la dinámica propia de esa serie. Es decir, cada serie observada se descompone como:

$$x_{it} = \omega_{ij} f_t + e_{it} \qquad [3]$$

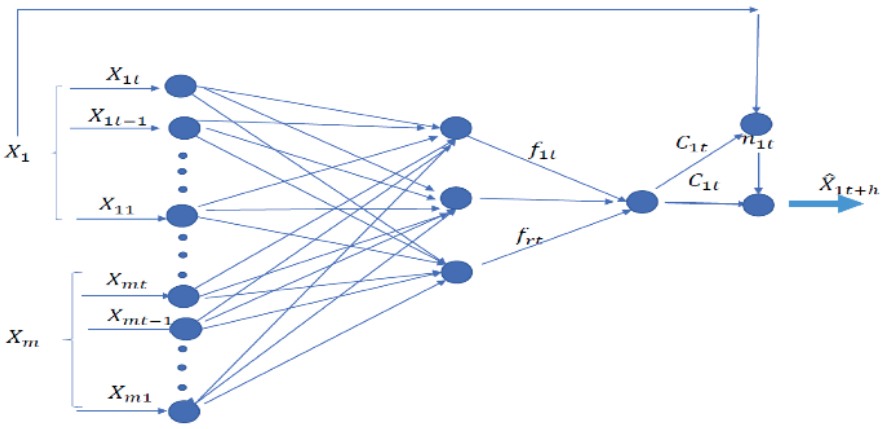
donde el valor de la serie i en el instante t es la suma de los efectos de los factores sobre esa serie en ese instante, más su término específico. Agrupando todas las series y todos los instantes temporales en una matriz de datos, X, que tendrá dimensiones T x m, los valores en esta matriz de datos se explican por el producto de la matriz de los valores de los factores, F, de dimensiones T x r, por la matriz de los efectos factoriales sobre cada una de las series, Ω, de dimensiones m x r, más la matrix de efectos específicos E, es decir:

$$X = F\Omega' + E \qquad [4]$$

La representación gráfica de un modelo factorial se presenta en las figuras 2 y 3. En la primera figura se trata de prever el valor de una variable $x_{i,t+h}$ utilizando sus valores pasados, $x_{i,t}$, $x_{i,t-1}$,... y también los valores de otras series $x_{1,t}$,...$x_{m,t}$ y sus retardos. Estas variables se combinan en r nudos, el número de factores, y la salida de cada nudo es una combinación lineal de las variables de entrada, con ciertos coeficientes ω, como se describe en la segunda figura para el caso $m=3$. Las salidas de estos nudos, que son los valores de los factores, f_t se combinan para formar la parte común de las series, C_t. El componente específico, n_t, se obtiene como una función lineal de los retardos de la variable a prever y ambos efectos se suman para dar lugar a la predicción. La figura 3 describe la opera-

FIGURA 2

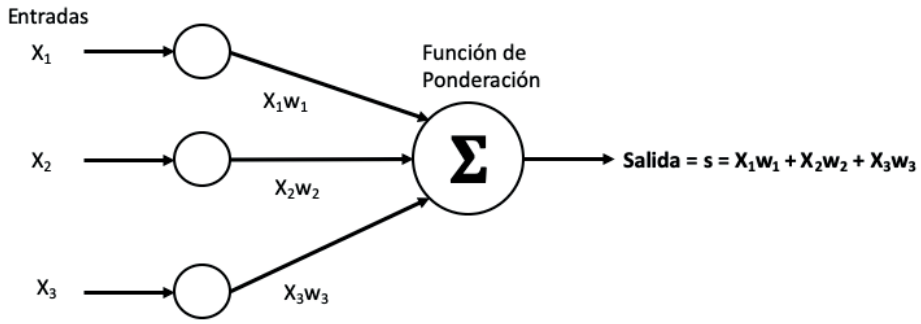
REPRESENTACIÓN GRÁFICA DE UN MODELO FACTORIAL DINÁMICO



Fuente: Elaboración propia.

FIGURA 3

DESCRIPCIÓN DE LAS OPERACIONES EN CADA NUDO DE UN MODELO FACTORIAL DINÁMICO



Fuente: Elaboración propia.

ción que se realiza en cada uno de los nudos: una combinación lineal de las variables de entrada ponderadas por ciertos pesos a estimar.

Para construir el DFM necesitamos determinar: (1) el número de factores necesarios para explicar los datos, r, y (2) los pesos en cada nudo. El número de factores se determina analizando los valores propios de las matrices de covarianzas de los datos. Peña y Tsay (2020) presentan una descripción detallada de los métodos existentes y Caro y Peña (2020) una propuesta reciente para determinar el número de factores y una comparación de diferentes enfoques.

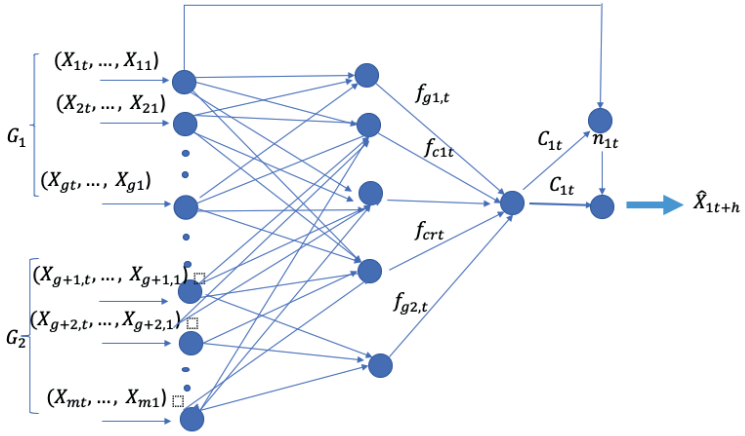
Los modelos factoriales dinámicos pueden incluir heterogeneidad como atípicos o estructura de grupos, véase Alonso, Galeano y Peña (2020). La figura 4 presenta un ejemplo de este tipo de modelo con dos grupos. Las series $1,...g$ pertenecen al primer grupo y las $g+1,...,m$ al segundo. Por ejemplo, corresponden a dos zonas geográficas distintas o a dos clases diferentes de variables. Tenemos ahora dos tipos de nodos o factores. Aquellos que reciben como entrada todas las variables y generan factores globales $f_{c1t},..., f_{crt}$ para todas las series, y aquellos que solo dependen de las observaciones de uno de los grupos y que generan factores específicos de grupo, que, en este caso, son uno para cada grupo, el f_{g1t} para el primero y el f_{g2t} para el segundo. Estos factores se combinan luego de la forma habitual para dar lugar a las predicciones incorporando la parte específica de la series. La ecuación general de este modelo es:

$$\mathbf{X} = \mathrm{F}_0\Omega'_0 + \sum_{s=1}^{S}\mathbf{F}_s\Omega_s^{'} + \mathbf{E}, \qquad [5]$$

donde el subíndice 0 indica el componente global y S el número de grupos.

FIGURA 4

MODELO FACTORIAL DINÁMICO CON DOS GRUPOS, R FACTORES COMUNES Y UNO ESPECÍFICO DE CADA GRUPO



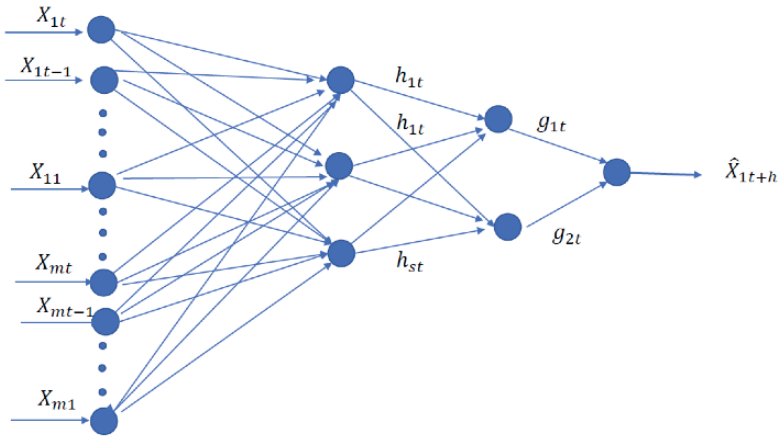
Fuente: Elaboración propia.

4.2. Redes neuronales y *deep learning*

Una forma alternativa de representar una relación cualquiera entre un grupo de variables es mediante redes neuronales. Este modelo considera un conjunto de variables de entrada y con ellas se forman combinaciones lineales, como en el DFM, que producen una respuesta no lineal. Las respuestas se combinan entre sí para formar nuevos factores que, de nuevo, actúan no linealmente. A diferencia de un DFM esto puede ocurrir

FIGURA 5

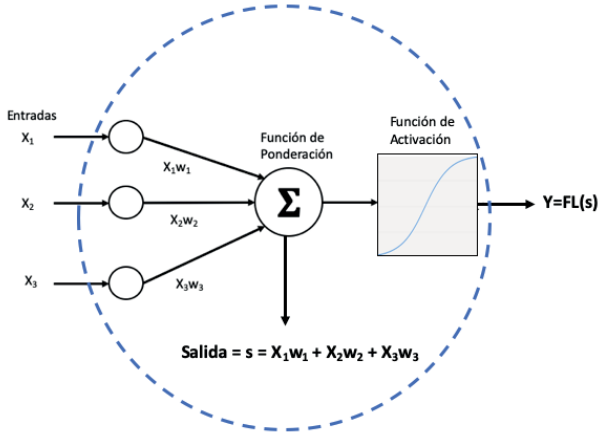
UNA RED NEURONAL CON DOS CAPAS INTERMEDIAS CON S Y 2 NUDOS



Fuente: Elaboración propia.

FIGURA 6

OPERACIONES EN CADA NUDO DE UNA RED NEURONAL



Fuente: Elaboración propia.

en distintas etapas o capas, hasta obtener la salida, que es la predicción de la variable de interés. El número de capas y de factores necesarios en cada capa, y su composición se determinan de forma empírica, de manera que la respuesta o predicción sea lo mejor posible.

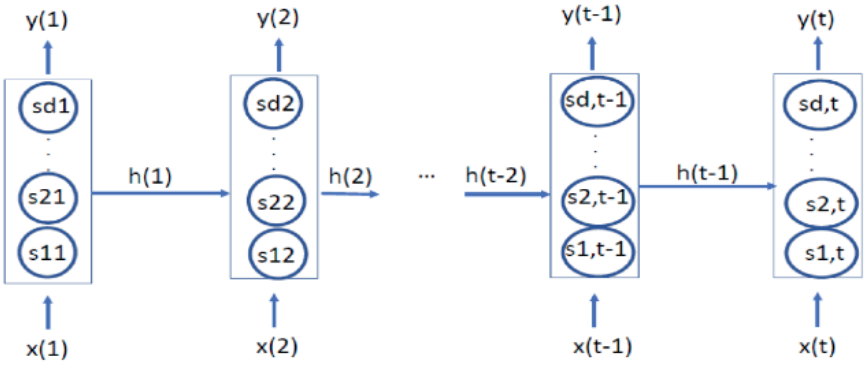
La figura 5 representa una red neuronal o perceptrón con dos capas, la primera con s nudos y la segunda con 2 nudos. En cada nudo se realizan las operaciones que se detallan en la figura 6: una combinación lineal de la entrada con ciertos pesos y una transformación no lineal, que suele ser la función logística. Observemos que una red neuronal (NN) con una capa y un nudo es equivalente al modelo logístico tradicional en estadística.

Las redes neuronales tradicionales no están pensadas para variables dinámicas y procesan todas las observaciones sin tener en cuenta su orden temporal. En los últimos años dentro del deep learning o aprendizaje profundo, se han desarrollado redes que procesan secuencialmente las observaciones. Un tipo de redes adaptadas a variables dinámicas son las *recurrent neural networks* (RNN), o redes neuronales recurrentes. La figura 7 presenta una de estas redes. Supongamos que queremos prever una variable $y(t)$ y disponemos de un conjunto de m variables $X(t)$. La variable $y(t)$ puede corresponder a los valores futuros de cualquiera de las variables explicativas $x_{i,t}$ con $i=1,\ldots,m$. Los datos se procesan secuencialmente. Inicialmente el vector de variables $X(1)$ entra como *input* en la red y se procesa para obtener la respuesta. En el período siguiente una parte de esa respuesta $h(1)$, la memoria del proceso, se introduce como *input* y se combina con el input $X(2)$ para generar la respuesta, o predicción y así sucesivamente. Los nudos en las capas ocultas de esta red tienen el mismo comportamiento que los descritos en

la figura 6. De esta forma, la predicción en t depende del input de los datos en t-1, pero también de las predicciones anteriores con memoria decreciente.

FIGURA 7

UNA RED NEURONAL RECURRENTE



Fuente: Elaboración propia.

4.3. ARBOLES DE DECISIÓN O *CART (CLASSIFICATION AND REGRESSION TREES)* Y RANDOM FOREST

Supongamos que tenemos N valores de una variable respuesta, y_t, y un conjunto de variables explicativas $X_t=(x_{1t},\ldots,x_{pt})$. Para construir un árbol de decisión o CART que nos permita hacer predicciones se procede como sigue: seleccionamos la variable x_i que conduzca a la mejor partición dicotómica del tipo $x_{it} < c$, *o*, $x_{it} \geq c$. Se desea dividir los datos para obtener la mejor predicción de la variable respuesta calculando la media de las observaciones que cumplen el criterio escogido para la división. Es decir, si llamamos $\bar{y}_1$ a la media de valores de la respuesta cuando $x_{it} < c$ y $\bar{y}_2$ a la media de la respuesta cuando $x_{it} \geq c$ entonces, llamando:

$$S(x_i,c) = min[\sum_{t\in(x_{it}<c)} (y_t-\bar{y}_1)^2 + \sum_{t\in(x_{it}\geq c)} (y_t-\bar{y}_2)^2], \qquad [6]$$

buscamos la variable y el punto de corte de la partición, c, que nos produce la mayor reducción en error cuadrático medio (MSE, por sus siglas en inglés). A continuación, se aplica el mismo método de división en los dos grupos creados, o ramas que salen del nudo creado. Es decir, se consideran primero solamente las observaciones que verifican $x_{it} < c$ y se busca una nueva variable y un punto de corte para ella que produzca la mayor reducción en el MSE de predicción. Este proceso se repite con las que verifican $x_{it} \geq c$. De esta manera, se obtiene una nueva partición, nuevos nudos y nuevas ramas,

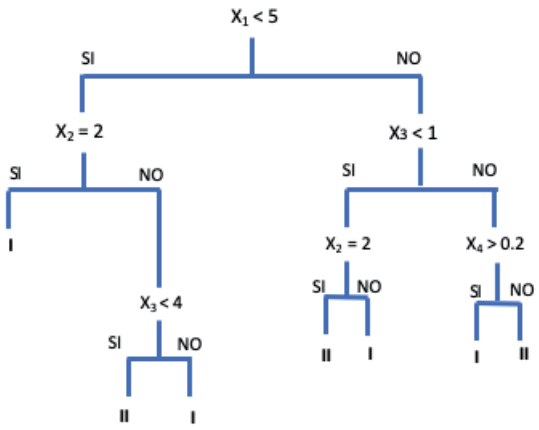
y el proceso se continúa mientras haya reducciones claras en el MSE. Si continuásemos hasta el final podríamos dividir la muestra hasta llegar a N grupos donde cada observación se prevería con su valor y el MSE sería cero. Obviamente, este resultado no es muy útil y generalmente el proceso finaliza cuando el número de observaciones que aparecen en los nudos finales es baja (menor que αN donde α es una cantidad pequeña, como 0,05).

Una vez obtenido este árbol máximo conviene ver si podándolo podemos obtener mejores resultados. Esto se hace eliminando los nudos donde la reducción en la suma de errores al cuadrado sea menor. Para equilibrar el número de nudos, que hacen el efecto de parámetros, con la reducción en la suma de errores al cuadrado se utiliza validación cruzada. Se comparan los modelos que parecen más adecuados en una muestra de validación donde los mejores modelos construidos se evalúan por su comportamiento fuera de la muestra.

Las reglas de predicción así construidas se denominan árboles de clasificación y regresión, (CART: classification and regression trees) y se aplican en ambos campos cuando tenemos variables cualitativas y cuantitativas. Son especialmente útiles para variables cualitativas o variables cuantitativas que afectan a la respuesta de forma no lineal que puede aproximarse por zonas constantes entre intervalos. La figura 8 ilustra un ejemplo simple de un árbol de decisión (CART). En la figura se desea prever una variable Y continua dadas un conjunto de otras cuatro variables explicativas $X=(x_1,x_2,x_3,x_4)$, tres continuas y una, la variable x_2, cualitativa. La predicción obtenida es el promedio de los valores que se sitúan en cada una de las ramas. Por ejemplo, para prever la respuesta

FIGURA 8

UN ÁRBOL DE CLASIFICACIÓN MUY SIMPLE CON UNA VARIABLE RESPUESTA Y CUATRO EXPLICATIVAS



Fuente: Elaboración propia.

de una variable con $X=(x_1=3, x_2=1, x_3=5, x_4=5)$ primero nos iremos por la rama de la izquierda, ya que x_1 es menor a cinco, luego por la derecha, ya que x_2 no es 2, y finalmente de nuevo a la derecha, ya que x_4 es mayor que 4. La media de las observaciones Y que verifican estas condiciones, $Y=Y(x_1\leq 5;\ x_2\neq 2;\ x4\geq 4)$ nos dará la predicción.

Los bosques aleatorios, o random forests, se obtienen por la combinación de muchos árboles de decisión con distintas muestras y variables, como hemos explicado anteriormente.

4.4. Estimación con regularización

Con muchas variables la estimación obtenida contiene, con frecuencia, demasiados parámetros. En estos casos, conviene estimar penalizando el número de parámetros, o su tamaño, en la muestra de entrenamiento. Una penalización muy utilizada es la suma de los valores absolutos de los parámetros introducidos. Minimizamos:

$$SSE_\lambda(\beta)=\sum_{t=1}^{N}(y_t-\sum_{j=1}^{p-1}X_{t,j}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|. \qquad [7]$$

Esta penalización, $\sum_{=1}|\ \ |$, la norma L1 del vector de parámetros, corresponde a la estimación Lasso introducida por Tibshirani (1996). Otros tipos de regularización son posibles; véase por ejemplo Peña y Tsay (2020).

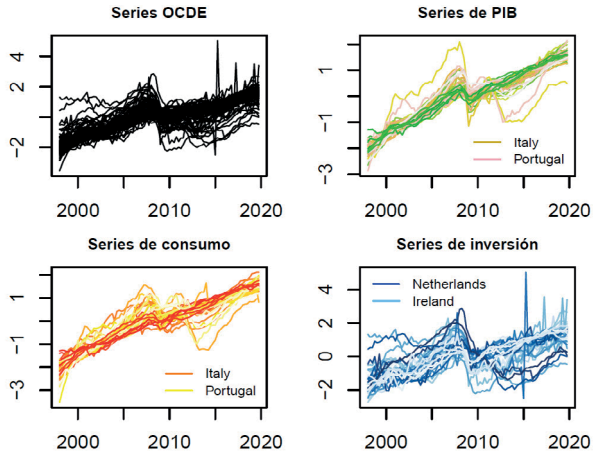
5. APLICACIÓN EMPÍRICA: PREDICCIÓN DE VARIABLES ASOCIADAS AL CICLO ECONÓMICO

El objetivo de este ejercicio es ilustrar el ajuste de un modelo factorial dinámico (DFM) y una red neuronal recurrente (RNN) para proporcionar predicciones a uno y tres períodos de un conjunto de tres variables macroeconómicas asociadas al ciclo económico de 35 países de la OCDE, que se indican en la tabla 1. Las variables son el PIB total, el gasto en consumo privado (CON) y la formación bruta de capital fijo (INV). Se tienen $N=105$ series trimestrales (tres series por cada país) y $T=88$ observaciones de tiempo, desde el primer trimestre de 1998 hasta el último trimestre de 2019. El conjunto de datos está disponible en OCDE Statistics (https://stats.oecd.org). Las series se presentan en la figura 9. Las series que presentan comportamientos distintos respecto al total de series son el PIB y el consumo de Italia y de Portugal. Las series de inversión de los Países Bajos y de Irlanda muestran gran variabilidad al final de la muestra.

Vamos a comparar las predicciones de los modelos univariantes de las series con las obtenidas por un modelo factorial y por una red neuronal recurrente. Con este ejemplo queremos ilustrar las dificultades con que se encuentra el analista al construir las reglas de predicción y no se pretende obtener el mejor modelo posible para prever estos datos.

FIGURA 9

SERIES TRIMESTRALES EN NIVELES DE PIB, CONSUMO E INVERSIÓN DESDE 1998(1T) HASTA 2019(4T)

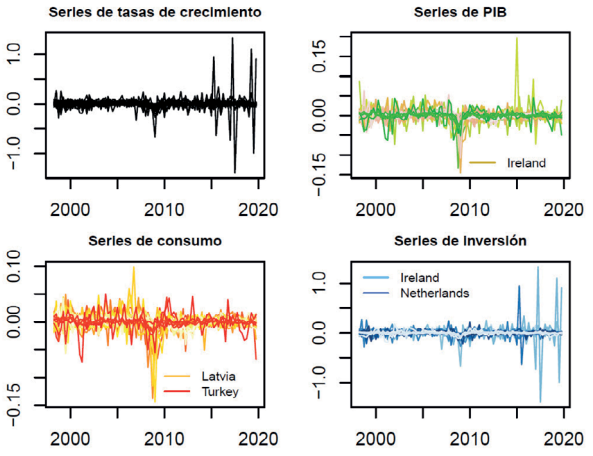


Fuente: Elaboración propia.

Las variables a prever son las tasas de crecimiento de cada serie por lo que, antes del análisis, se toman logaritmos en los datos y se diferencian para hacerlos estacionarios. La figura 10 presenta las series transformadas. Estos gráficos muestran el efecto negativo de la crisis financiera del 2008 en el desempeño económico de los países. Se observa que las series de PIB y consumo son más homogéneas, la escala de variación está en un intervalo del 15%, mientras que para las series de inversión la variación puede llegar

FIGURA 10

SERIES TRIMESTRALES EN TASAS DE CRECIMIENTO DE PIB, CONSUMO E INVERSIÓN DESDE 1998(1T) HASTA 2019(T4)



Fuente: Elaboración propia.

a ser del 100%, con grandes valores atípicos. Esta heterogeneidad va a condicionar los resultados de la predicción.

TABLA 1

PAÍSES DE LA OCDE INCLUIDOS EN LA MUESTRA DE DATOS REALES

Australia	Austria	Bélgica	Canadá	Chile
República Checa	Dinamarca	Estonia	Finlandia	Francia
Alemania	Hungría	Islandia	Irlanda	Israel
Italia	Japón	Corea	Letonia	Lituania
Luxemburgo	México	Los Países Bajos	Nueva Zelanda	Noruega
Polonia	Portugal	República Eslovaca	Eslovenia	España
Suecia	Suiza	Turquía	Reino Unido	Estados Unidos

El número de factores con los datos de la muestra de entrenamiento se estima por el contraste propuesto en Caro y Peña (2020) y esta misma metodología se aplica para estimar los *loadings* o coeficientes de los factores en las series, así como las series de los factores. Se encuentra un único factor común que explica el 78,20% de la variabilidad total (suma de las varianzas de todas las variables). A continuación, para cada serie, se obtienen los residuos al quitar la parte común, y se ajusta un modelo AR a estas series de componentes idiosincráticos. Para la estimación, cada vez que se incorpora una nueva observación se reestima el modelo factorial completo en la muestra ampliada. Los programas utilizados en *R* están en el paquete SLBDD, que se ha desarrollado para acompañar al libro Peña y Tsay (2020).

Para construir la red neural recurrente y hacer predicciones se utilizan las funciones del paquete rnn de R. Se elige una red formada por una capa oculta y dos nodos. El número de nodos se ha limitado a dos ya que al aumentarlo los resultados empeoraban. Para entrenar la red neuronal si tenemos información hasta el momento t y queremos hacer predicción un paso adelante de la serie $Y_{i,t+1}$, utilizamos como variables explicativas tres retardos de su propia serie, $Y_{i,t}$, $Y_{i,t-1}$, $Y_{i,t-2}$ y tres retardos de las otras variables explicativas, $y_{j,t}$, $y_{j,t-1}$, $y_{j,t-2}$ con $i \neq j$. Por ejemplo, si $i=1$, como tenemos 105 series en la muestra, entonces $j=2,\ldots,105$, y si queremos hacer predicciones a tres pasos adelante utilizamos como variables explicativas tres retardos de todas.

Así como la estimación del DFM es bastante directa, la construcción de la red neuronal es más compleja ya que hay que decidir los valores de varios parámetros como son la "tasa de aprendizaje" en el algoritmo de estimación o entrenamiento de la red y el número de "epochs", o número de veces que se recorre todo el conjunto de datos. Estos valores se han decidido por prueba y error y se han fijado para los resultados que presentamos con tasa de aprendizaje = 0,05 y número de epochs = 68. También, la

estimación de los parámetros no es determinista y al repetir la estimación los resultados pueden variar dependiendo del punto de inicio.

Las predicciones obtenidas para cada serie se comparan en la Tabla 2 con las de un modelo univariante ARIMA. En los tres casos, las predicciones se calculan de forma recursiva: se toma como muestra inicial de estimación o entrenamiento los datos de $t=1,2,...,71$. Con estos datos se decide la estructura del modelo factorial o de la red neuronal y se entrena el modelo para estimar los parámetros. A continuación, se hacen predicciones a uno y tres pasos desde los horizontes $t=71,...,86$ y $t=71....,84$, respectivamente, reestimando el modelo con todos los datos disponibles en el origen de la predicción. Para comparar los resultados se calcula la raíz del error cuadrático medio (RMSE, *root mean square error,* por sus siglas en inglés) y la desviación absoluta mediana (MAD, *median absolute deviation*) de los errores de predicción para cada una de las metodologías de predicción. Como la MAD utiliza medianas en lugar de medias no se ve afectada por valores extremos, con lo que ofrece una comparación más robusta que el RMSE.

Para las predicciones un período hacia delante, el DFM es el mejor modelo tanto por el criterio de RMSE como por MAD. La gran diferencia entre el RMSE del modelo de factores y el de la red neuronal se debe a que la RNN predice valores muy extremos, que no se tienen en cuenta al comparar con la MAD. Los resultados a uno y tres pasos son similares: el DFM es ligeramente mejor que el ARIMA en RMSE y ambos tienen mejor desempeño que la RNN.

TABLA 2

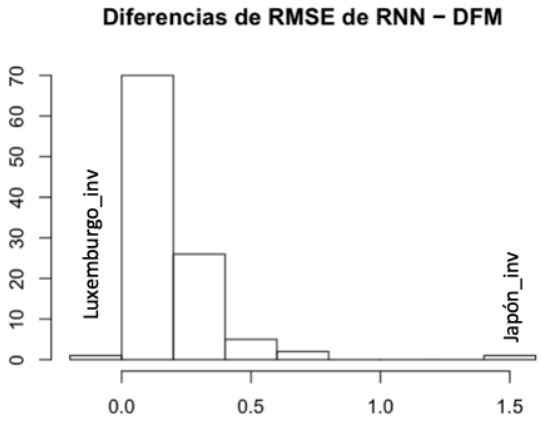
PRECISIÓN DE LAS PREDICCIONES CUANDO EL HORIZONTE DE PREDICCIÓN ES A UNO Y TRES PASOS, $h=1,3$

	DFM	RNN	ARIMA
h=1			
RMSE	0.0182	0.1761	0.0184
MAD	0.0157	0.0233	0.0162
h=3			
RMSE	0.0190	0.2100	0.0192
MAD	0.0140	0.0279	0.0135

En la figura 11 se presenta el histograma de las diferencias entre los RMSE de la red neuronal y los del DFM para las predicciones un período hacia delante. El DFM produce mejores predicciones en todas las series, exceptuando la predicción para la inversión en Luxemburgo, véase la figura 12, donde gana la RNN. La diferencia más abultada a favor del DFM se da para la inversión de Japón, véase la figura 13, donde la RNN presenta valores muy extremos en comparación con el DFM. Vemos como la heterogeneidad de las series de inversión provoca que sea en estas series donde se producen las mayores diferencias.

FIGURA 11

HISTOGRAMA DE LAS DIFERENCIAS ENTRE LOS RMSE DE LA RNN Y LOS DEL DFM

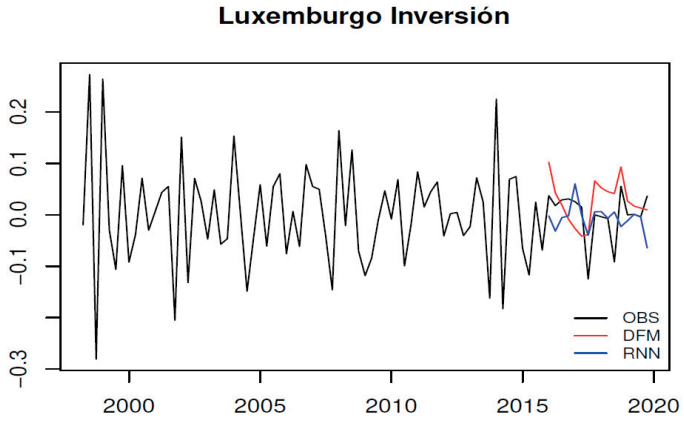


Fuente: Elaboración propia.

El histograma de las diferencias entre la MAD de los errores de predicción se representa en la figura 14. Según la MAD, la RNN predice mejor que el DFM en 9 de las 105 series. La inversión en Irlanda, figura 15, es la que presenta una mayor diferencia a favor de la RNN. La mayor diferencia a favor del DFM es para la serie de inversión en la República

FIGURA 12

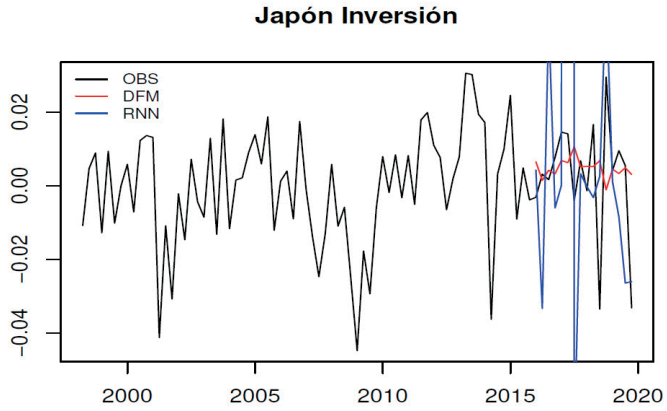
PREDICCIÓN DE LA TASA DE CRECIMIENTO DE LA INVERSIÓN DE LUXEMBURGO (NEGRO) USANDO UN MODELO DE FACTORES (AZUL) Y UNA RED NEURONAL RECURRENTE (ROJO)



Fuente: Elaboración propia.

FIGURA 13

PREDICCIÓN DE LA TASA DE CRECIMIENTO DE LA INVERSIÓN DE JAPÓN (NEGRO) USANDO UN MODELO DE FACTORES (AZUL) Y UNA RED NEURONAL RECURRENTE (ROJO)

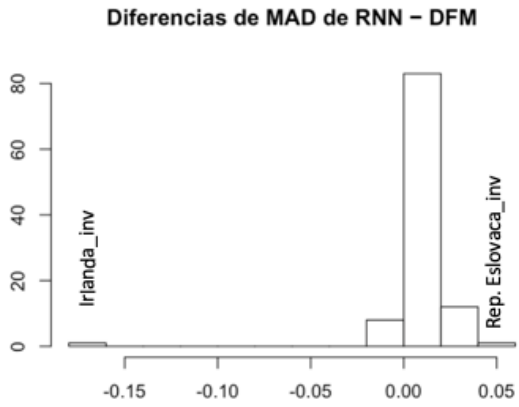


Fuente: Elaboración propia.

Eslovaca. Como puede verse en la figura 16, para esta serie la RNN predice un valor muy extremo al final de la muestra. Teniendo en cuenta ambos criterios, RMSE y MAD, la RNN solo supera al DFM en la predicción de la serie de Luxemburgo inversión.

FIGURA 14

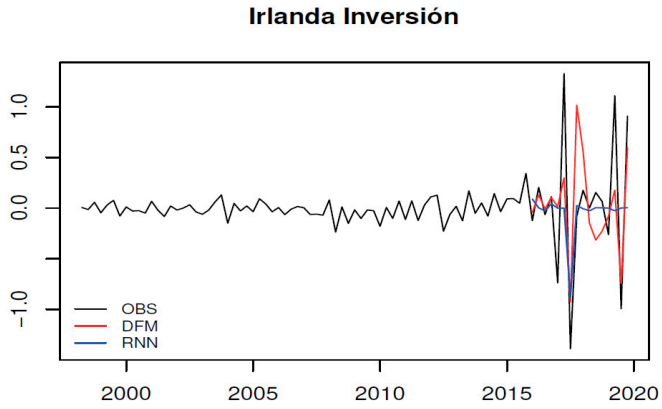
HISTOGRAMA DE LAS DIFERENCIAS ENTRE LA MAD DE LOS ERRORES DE LA RED NEURONAL RECURRENTE Y LA DEL MODELO DE FACTORES



Fuente: Elaboración propia.

FIGURA 15

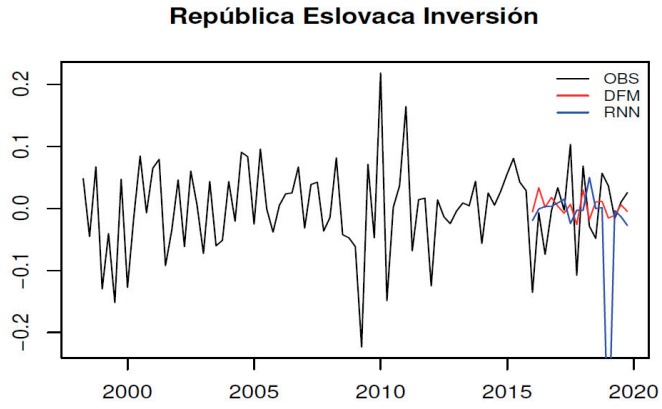
PREDICCIÓN DE LA TASA DE CRECIMIENTO DE LA INVERSIÓN DE IRLANDA (NEGRO) USANDO UN MODELO DE FACTORES (AZUL) Y UNA RED NEURONAL RECURRENTE (ROJO)



Fuente: Elaboración propia.

FIGURA 16

PREDICCIÓN DE LA TASA DE CRECIMIENTO DE LA INVERSIÓN DE LA REPÚBLICA ESLOVACA (NEGRO) USANDO UN MODELO DE FACTORES (AZUL) Y UNA RED NEURONAL RECURRENTE (ROJO)



Fuente: Elaboración propia.

6. CONCLUSIONES

Hemos mostrado cómo la abundancia actual de datos ha llevado a un nuevo paradigma en la construcción de predicciones económicas y hemos analizado algunas de sus consecuencias. Hace solo 20 años, a principios de este siglo, la situación de abundancia de datos que vivimos hoy era inimaginable. Este proceso de generación automática de

datos en muchos entornos va a acelerarse en los próximos veinte años: es esperable que el tratamiento de los datos recogidos para hacer predicciones automáticas vaya incorporándose poco a poco a todos los dispositivos que utilicemos, así como, a muchas de las actividades que realicemos. Los nuevos datos masivos irán refinando los métodos actuales para hacerlos más flexibles y adaptativos, lo que supondrá una reducción de la incertidumbre que modificará las estrategias de las organizaciones, de las empresas y de los individuos. Contaremos con predicciones frecuentes y fiables de nuestra situación económica, nuestra salud o nuestros estados de ánimo. Esto será posible por las grandes posibilidades de aprendizaje para datos muy desagregados que se obtienen de los datos masivos.

Sin embargo, las predicciones automáticas con métodos complejos de difícil interpretación entrañan riesgos, ya que pueden modificar las decisiones en direcciones equivocadas. Esto es más preocupante con reglas de predicción que, como ocurre con las redes neuronales, no permiten ver claramente las relaciones entre la respuesta y las variables involucradas. Con estas reglas si las predicciones son deficientes no es claro cómo actuar, ya que aumentar su complejidad puede llevarnos a modelar un ruido impredecible. Por otro lado, aunque los resultados sean buenos, no tenemos garantías de que continúen funcionando en el futuro.

En el ejemplo presentado con variables económicas agregadas, las NN no han supuesto ninguna mejora adicional sobre el modelo factorial dinámico lineal más simple. Este resultado es el esperado con series macroeconómicas, que tienen entre sí relaciones generalmente lineales y el modelo factorial se ajusta mejor a su comportamiento, mientras que las NN intentan explicar atípicos y series heterogéneas, como las de inversión en el ejemplo, estropeando el resutado promedio. Las ventajas de poder modelar la no linealidad que permiten las NN es especialmente útil con series muy desagregadas, observadas a intervalos cortos de tiempo, donde los efectos no lineales serán más pronunciados.

Referencias

Akaike, H. (1998). *Selected Papers of Hirotugu Akaike*. Springer.

Alonso, A. M., Galeano, P. y Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics,* Vol. 216(1), pp. 35-52.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science,* 355(6324), pp. 483–485.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Blazquez, D. y Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change,* 130, pp. 99–113.

Box, G. E. P. y Jenkins, G. M. 1976). *Time Series Analysis: Forecasting and Control.* Edición revisada. Holden-Day.

Breiman, L. (1996). Bagging predictors. *Machine learning,* 24(2), pp. 123–140.

Bühlmann, P. y Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications.* Springer Science & Business Media.

Caro, A. y Peña, D. (2020). A Test for the Number of Factors in Dynamic Factor Models. *Working Paper of the Statistics Department at Carlos III University of Madrid.*

Gary, CH. y Rothschild, M. (1983). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica,* 51(5), pp. 1281–1304.

Clements, M. y Hendry, D. (1998). *Forecasting Economic Time Series.* Cambridge University Press.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics,* 26(4), pp. 745–766.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological),* 57(1), pp. 45–70.

Efron, B. y Hastie, T. (2016). *Computer Age Statistical Inference.* Cambridge University Press.

Engle, R. y Granger, C. (1991). *Long-run Economic Relationships: Readings in Cointegration.* Oxford University Press.

Engle, R. y Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association,* 76(376), pp. 774–781.

Fan, J., Han, F. y Liu, H. (2014). Challenges of big data analysis. *National Science Review,* 1(2), pp. 293–314.

Freund, Y., Schapire, R. y Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence,* 14(5), pp. 771-780.

Galeano, P. y Peña, D. (2019). Data science, big data and statistics. *TEST,* 28(2), pp. 289–329.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent Variables in Socio-economic Models,* 33(6), pp. 583-–606.

Giannone, D., Lenza, M. y Primiceri, G. E. (2017). Economic predictions with big data: The illusion of sparsity. *CEPR Discussion Paper* No. DP12256.

Giannone, D., Reichlin, L. y Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics,* 55(4), pp. 665–676.

Greene, W. H. (1993). *Econometric Analysis.* Macmillan.

Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica,* 12, pp. iii–115.

Heitz, G., Gould, S., Saxena, A. y Koller, D. (2009). Cascaded classification models: Combining models for holistic scene understanding. *Advances in Neural Information Processing Systems,* pp. 641–648.

Hsiao, CH. (2020). An Econometrician's Perspective on Big Data. *Essays in Honor of Cheng Hsiao.* Emerald Publishing Limited.

Koop, G. y Potter, S. (2004). Forecasting in dynamic factor models using Bayesian model averaging. *The Econometrics Journal,* 7(2), pp. 550–565.

Kuzin, V., Marcellino, M. y Schumacher, CH. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting,* 27(2), pp. 529–542.

Meade, N. e Islam, T. (1998). Technological forecasting-Model selection, model stability, and combining models. *Management Science,* 44(8), pp. 1115–1130.

Min, C. K. y Zellner, A. (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics,* 56(1-2), pp. 89–118.

Peña, D. (2002). R*egresión y diseño de experimentos.* Madrid: Alianza.

Peña, D. y Box, G. E. P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association,* 82(399), pp. 836–843.

Peña, D. y Sánchez, I. (2005). Multifold predictive validation in ARMAX time series models. *Journal of the American Statistical Association,* 100(469), pp. 135–146.

Peña, D. y Tsay, R. S. (2020). *Statistical Learning for Big Dependent Data.* New York: Wiley NY.

Raftery, A. E., Gneiting, T., Balabdaoui, F. y Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review,* 133(5), pp. 1155–1174.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological),* 36(2), pp. 111–133.

Sun, CH., Shrivastava, A., Vondrick, C., Sukthankar, R., Murphy, K. y Schmid, C. (2019). Relational action forecasting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 273–283.

Theil, H. (1971). *Principles of Econometrics.* NY: Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 58(1), pp. 267–288.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives,* 28(2), pp. 3–28.

Wasserstein, R. L. y Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician,* 70(2), pp. 129-133.

Yuan, Z. y Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association,* 100(472), pp. 1202–1214.

CAPÍTULO II

Modelos de selección de carteras con muchos activos

Christian Brownlees
Jordi Llorens
Nuria Senar

Este capítulo trata el problema de selección de carteras de inversión con un gran número de activos financieros. En particular, se repasa la literatura en modelización de correlaciones condicionales dinámicas de elevadas dimensiones (*DCC,* por sus siglas en inglés). Consideramos diferentes tipos de especificaciones, en particular, la versión estándar del modelo DCC, el DCC con estructura de factores, y el DCC con regularización. Introducimos métodos de estimación específicamente diseñados para modelos de elevada dimensionalidad. Evaluamos su capacidad de predicción a través de una aplicación en selección de carteras de inversión con los constituyentes del índice S&P 500.

Palabras clave: volatilidad multivariante, correlaciones condicionales dinámicas, regularización de covarianzas, verosimilitud compuesta, selección de carteras.

1. INTRODUCCIÓN

El problema de selección de carteras es de suma importancia en el sector financiero. En la conocida teoría moderna de carteras –*MPT,* por sus siglas en inglés– desarrollada por Markowitz (1952), se demuestra que el problema de inversión únicamente depende de la media y la matriz de covarianzas de un vector de retornos de activos. Consecuentemente, el problema de selección de carteras en la MPT se traduce en un problema de estimación de la media y la matriz de covarianzas de los retornos. La evidencia empírica nos muestra que la matriz de covarianzas es un elemento clave en selección de carteras y su estimación ha despertado un gran interés tanto en la academia como en la industria. La estrategia más natural es construir la cartera mediante la matriz de covarianzas muestral, pero esto tiene algunas limitaciones. Por un lado –en la dimensión temporal– las volatilidades y correlaciones de los activos financieros son *dinámicas,* esto es, que varían a lo largo del tiempo. Por otro lado –en la sección cruzada– los métodos de estimación tradicionales de las covarianzas en contextos de *elevada dimensionalidad, i.e.* cuando el número de activos financieros es de un orden similar al número de observaciones en muestra– son imprecisos. Por ejemplo, véase Friedman, Hastie y Tibshirani (2008) o Pourahmadi (2013).

En este trabajo presentamos metodologías para lidiar con ambos tipos de retos, concretamente la familia de modelos GARCH-DCC. En las últimas dos décadas, la metodología GARCH-DCC introducida por Engle (2002) se ha establecido dentro de la literatura como una de las referencias clave tanto por su flexibilidad como su viabilidad de implementación. En síntesis, la estrategia consiste en modelizar por separado las varianzas y las correlaciones condicionales. Las varianzas condicionales de cada activo se modelizan individualmente vía GARCH (Bollerslev, 1986) mientras que la matriz de correlaciones condicionales se modeliza conjuntamente mediante el modelo de correlaciones condicionales dinámicas (*DCC,* por sus siglas en inglés). La modelización vía GARCH ha sido estudiada en profundidad y tiene ya una larga tradición en la literatura (Bollerslev, Engle y Nelson, 1994). En cambio, el modelo DCC fue desarrollado posteriormente y a día de hoy sigue estando en la frontera de la investigación en econometría financiera.

Este capítulo introduce el modelo DCC de Engle (2002) en su versión estándar y consideramos algunas de sus extensiones más importantes: el DCC con estructura de factores *(Factor DCC)* y el DCC con regularización lineal y no lineal *(Shrinkage DCC).* Una de las grandes ventajas de esta familia de modelos es que su estimación es escalable a dimensiones relevantes para aplicaciones prácticas.

En el modelo DCC estándar el proceso de correlaciones condicionales dinámicas viene determinado por los retornos estandarizados, es decir, los datos divididos por su volatilidad. El modelo se basa en una ecuación recursiva que garantiza que las matrices de correlaciones dinámicas son definidas positivas.

El modelo Factor DCC se basa en descomponer la matriz de covarianzas dinámica en un factor común y un componente idiosincrásico. En la literatura hay una larga tradición de trabajos que abordan la cuestión de la modelización de activos financieros mediante modelos de factores. Estos modelos se basan en el supuesto de que los precios y las volatilidades de diferentes activos vienen determinados por un pequeño número de factores, lo que determina su comovimiento. Los modelos de factores se pueden dividir en dos tipologías: factores latentes y factores observables. Algunos ejemplos que emplean factores latentes son Diebold y Nerlove (1986), Harvey, Ruiz y Sentana (1992), o el GARCH factorial ortogonal de Alexander y Chibumba (1996). Han (2006) y Aguilar (2009) combinan modelos multivariantes con factores latentes y volatilidad estocástica. Un ejemplo de modelo de factores latentes más reciente es Hallin *et al.* (2019), que emplea componentes principales dinámicos siguiendo la estrategia de Forni *et al.* (2015). Por el lado de los factores observables tenemos el reciente trabajo de De Nard, Ledoit y Wolf (2020), donde se utiliza el modelo DCC con estructura de factores. El modelo de factores que presentamos en detalle más adelante toma como referencia este último trabajo.

El modelo Shrinkage DCC combina la literatura en estimación regularizada de grandes matrices de covarianzas con la estimación dinámica de matrices de covarianzas. Es conocido, como mínimo desde Ledoit y Wolf (2004b), que cuando el número de variables es elevado con respecto al número de observaciones la matriz de covarianzas muestral suele tener un error de estimación preocupante. Esto ha motivado una amplia literatura que propone el uso de la regularización para aumentar la precisión en la estimación de la matriz de covarianzas (Bickel y Levina, 2008; Fan, Liao y Mincheva, 2013). Véase Pourahmadi (2013) para un repaso de la literatura relacionada. En una serie de trabajos, Olivier Ledoit y Michael Wolf han propuesto diferentes estimadores regularizados que han demostrado mejorar los resultados en aplicaciones financieras como la selección de carteras. Recientemente, Engle, Ledoit y Wolf (2019) incorporan esta metodología de regularización al modelo DCC. En este trabajo se ofrece un breve resumen de la versión del modelo con las técnicas de regularización lineal y no lineal.

Ilustramos los beneficios de las metodologías aquí introducidas a través de un ejercicio de selección de carteras de inversión con los constituyentes del índice S&P 500. El ejercicio en cuestión está estrechamente relacionado con los trabajos de Hautsch, Kyj y Malec (2015), Hautsch y Voigt (2019), Engle, Ledoit y Wolf (2019) y De Nard, Ledoit y Wolf (2020). En particular, evaluamos la precisión de los diversos estimadores de manera indirecta (Patton y Sheppard, 2009) mediante la volatilidad de la cartera (dinámica) de mínima varianza y mediante la ratio de información de la cartera de Markowitz con señal de *momentum.* Los resultados demuestran que las metodologías propuestas permiten construir carteras de inversión con rentabilidades ajustadas al riesgo superiores a los puntos de referencia. En particular, el valor diferencial que aportan los modelos Factor y Shrinkage DCC es significativo tanto desde el punto de vista económico como estadístico.

El modelo DCC pertenece a la familia de modelos generalizados de heteroscedasticidad condicional multivariante (MGARCH, por sus siglas en inglés), que ofrecen la posibilidad de estimar y predecir matrices de covarianzas dinámicas de un gran número de activos y que se han utilizado con éxito en aplicaciones de selección de carteras de inversión (Engle, Ledoit y Wolf, 2019), de gestión de riesgos (Ferreira, 2005) y de medición del riesgo sistémico (Brownlees y Engle, 2017). Fuera del sector financiero, estos modelos han sido empleados para modelizar sistemas de variables macroeconómicas como, entre otros, la interacción inflación-crecimiento (Conrad, Karanasos y Zeng, 2010) o la relación entre la evolución del petróleo y el oro (Bampinas y Panagiotidis, 2015). En Alessi, Barigozzi y Capasso (2009) se aplica, con éxito considerable, una combinación de modelo de factores dinámicos y MGARCH en series de inflación y en retornos de activos financieros. Para un repaso general de la literatura MGARCH, véanse, por ejemplo, Bauwens, Laurent, y Rombouts (2006), Engle (2009) Silvennoinen y Teräsvirta (2009) y de Almeida, Hotta, y Ruiz (2018).

Otra importante familia de modelos que ha suscitado un gran interés en la comunidad académica son los modelos de volatilidad estocástica, véanse Shephard (1996) y Broto y Ruiz (2004) para una revisión bibliográfica más detallada para modelos univariados, y Asai, McAleer y Yu (2006) para el caso multivariante. En resumen, la diferencia clave entre la familia de modelos MGARCH y la volatilidad estocástica multivariante reside en que los últimos modelizan la secuencia de matrices de covarianzas como un proceso estocástico, mientras que en los primeros el proceso es determinístico. En términos generales, se puede argumentar que los modelos de volatilidad estocástica ofrecen una mayor flexibilidad de modelización, pero pagan el precio de una mayor dificultad para llevar a cabo su estimación.

El resto del capítulo se estructura de la siguiente forma: la sección segunda sienta las bases metodológicas del capítulo; la subsección 2.1. introduce la notación de la versión canónica del modelo DCC y en la 2.2. tratamos su estimación; la subsección 2.3. explica cómo realizar predicciones con este modelo; la subsección 2.4. introduce el modelo DCC con estructura de factores y en la 2.5. consideramos el modelo DCC con regularización. La sección tercera presenta la metodología empleada para selección de carteras, en particular la cartera de mínima varianza (3.1.) y la cartera de Markowitz con señal de *momentum* (3.2.). La sección cuarta se dedica a la aplicación empírica; la sección 5 concluye el capítulo.

2. METODOLOGÍA

Esta sección fija la notación, así como las bases metodológicas sobre las que se asienta el capítulo. A lo largo de todo el capítulo, y salvo algunas excepciones donde no cabe ambigüedad, los escalares como x se denotan con minúsculas, los vectores como $\boldsymbol{x}$ en negrita minúscula, y las matrices como X con letra mayúscula. Un individuo de la sec-

ción cruzada (normalmente interpretado como el i-ésimo activo financiero) se denota con el subíndice $i=1,...,N$, y para la dimensión temporal se utiliza el subíndice $t=1,...,T$.

2.1. Modelo de correlaciones condicionales dinámicas (DCC)

Denotamos como $\{\mathbf{r}_t\}_{t=1}^{T}$ a la serie de retornos de N activos con media cero[1]. Se asume que el proceso generativo de los retornos viene dado por:

$$\mathbf{r}_t = \Sigma_t^{1/2}\mathbf{z}_t, \qquad \mathbf{z}_t \overset{i.i.d.}{\sim} \mathcal{D}(\mathbf{0}, I)$$

donde $\Sigma_t \equiv Var(\mathbf{r}_t \mid \mathcal{F}_{t-1})$ es definida positiva, $\mathcal{F}_{t-1}$ es el conjunto de información disponible en el período t-1, y $\mathcal{D}$ es una distribución multivariante estandarizada de la familia localización-escala (*e.g.,* Normal, t-Student, etc.). El objetivo principal es encontrar una buena predicción de Σ_t, es decir, de la matriz dinámica de covarianzas de los retornos. Dicha matriz puede ser reparametrizada como sigue:

$$\Sigma_t = D_t R_t D_t\,, \qquad [1]$$

donde D_t es la matriz (diagonal) de volatilidades dinámicas, con su i-ésimo elemento dado por $d_{i,t} \equiv \sqrt{Var(r_{i,t} \mid \mathcal{F}_{t-1})}$, y siendo $R_t \equiv Var(\boldsymbol{\varepsilon}_t \mid \mathcal{F}_{t-1}) = \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' \mid \mathcal{F}_{t-1}]$ la matriz de correlaciones dinámicas, donde $\boldsymbol{\varepsilon}_t = D_t^{-1}\mathbf{r}_t$. Esta reparametrización permite la desagregación del problema en dos partes: por un lado, la estimación de las volatilidades y, por el otro, la de la matriz de correlaciones. Asumimos que la volatilidad de cada activo sigue un proceso estacionario GARCH (Bollerslev, 1986), tal y como se recoge en la ecuación [2], donde a_i y b_i son escalares positivos tales que $a_i + b_i < 1$ y $d_i \equiv \sqrt{Var(r_{i,t})}$ es la volatilidad incondicional del i-ésimo activo:

$$d_{i,t}^2 = (1 - a_i - b_i)d_i^2 + a_i r_{i,t-1}^2 + b_i d_{i,t-1}^2\,. \qquad [2]$$

La dinámica de las correlaciones depende de los retornos estandarizados $\boldsymbol{\varepsilon}_t$ de acuerdo con la siguiente ecuación recursiva:

$$Q_t = (1 - \alpha - \beta)C + \alpha \boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}' + \beta Q_{t-1}\,, \qquad [3]$$

donde C es una matriz $N \times N$ positiva definida y α y β son escalares positivos tales que $\alpha + \beta < 1$. Nótese que la dinámica de Q_t es muy similar a la del modelo VECH escalar (Engle, 2009).

Bajo estos supuestos, $\{Q_t\}_{t=1}^{T}$ es una secuencia estacionaria de matrices definidas positivas. Sin embargo, su diagonal no es necesariamente unitaria y sus elementos no están

[1] Entendemos aquí $\mathbf{r}_t$ como el vector de retornos. Al estar considerando retornos de acciones, el supuesto de media zero en frecuencias diarias es razonable. Sin embargo, queda entendido que también podríamos interpretar $\mathbf{r}_t$ como los residuos de un modelo VARMA sin pérdida de generalidad en la exposición.

necesariamente acotados en el intervalo [-1,1], lo que significa que Q_t no puede ser considerada como una matriz de correlaciones propiamente dicha. Por su similitud con la matriz R_t, a Q_t la denominamos matriz de pseudo-correlaciones. Es sencillo reescalar la misma para obtener la matriz de correlaciones condicional:

$$R_t = \text{diag}(Q_t)^{-1/2}\, Q_t\, \text{diag}(Q_t)^{-1/2}.$$

Aunque reescalar parece una opción bastante natural, no es obvio que R_t sea la matriz de correlaciones más próxima a Q_t. De otra manera, R_t no es una proyección de Q_t hacia el espacio de matrices de correlaciones, y existen otras estrategias propuestas en la literatura para llevar a cabo esta normalización (Tse y Tsui, 2002; Brownlees y Llorens, 2020). Es posible considerar especificaciones más generales con un mayor número de retardos tanto en $\boldsymbol{\varepsilon}_t$ como en Q_t, pero empíricamente se ha documentado que añadir más retardos no suele mejorar la capacidad de predicción de una manera significativa. Para abreviar, utilizamos la notación $\mathbf{r}_t \sim \text{DCC}(1,1)$ para expresar que los retornos siguen el modelo descrito por las ecuaciones [1] a [3].

2.2. Estimación en el modelo DCC

2.2.1. Focalización en covarianzas y estimación por (cuasi) máxima verosimilitud

Desde el punto de vista de la estimación, se han de obtener, primero, las volatilidades para, en un segundo paso, estimar las correlaciones, pues estas dependen de ε_t –que a su vez depende de D_t–. Es posible estimar α, β y C conjuntamente por máxima verosimilitud, pero esto supone resolver un problema de optimización en $\binom{N}{2}+N+2$ variables, lo cual resulta poco atractivo a medida que el número de activos crece. En su lugar, se recurre a la técnica de focalización en la covarianza *(covariance targeting)* que consiste en reemplazar C por un estimador f de la matriz de covarianzas de ε_t que no dependa de α y β, es decir,

$$\hat{C} = f\left(\{\boldsymbol{\varepsilon}_t\}_{t=1}^{T}\right). \qquad [4]$$

Por ejemplo, en el modelo DCC canónico, $f\left(\{\boldsymbol{\varepsilon}_t\}_{t=1}^{T}\right) = \frac{1}{T}\Sigma_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'$. La técnica de focalización tiene una justificación clara en los modelos estacionarios GARCH multivariantes (como, por ejemplo, el VECH escalar), puesto que, en ese caso, la matriz interceptor es equivalente al segundo momento de los datos. Desde el punto de vista teórico, la técnica de focalización en el modelo DCC es inconsistente puesto que en general $C \neq \mathbb{E}\left[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right]$, algo que fue señalado en primer lugar por Aielli (2013), quien a su vez propuso una versión corregida de ese modelo, abreviada como cDCC. Sin embargo, la inconsistencia

es menor y en la práctica se sigue empleando la versión canónica del modelo DCC con focalización en covarianzas, lo cual es, naturalmente, mucho más conveniente desde la perspectiva computacional.

Los parámetros α y β se estiman maximizando el logaritmo de la cuasi-verosimilitud Gaussiana, que viene dada por:

$$\sum_{t=1}^{T} l_t(\theta) = \sum_{i=1}^{N} \mathcal{L}_{i,T}^{G}\left(\theta_i^{G}\right) + \mathcal{L}_{T}^{DCC}(\alpha,\beta),$$

donde $\mathcal{L}_{i,T}^{G}\left(\theta_i^{G}\right)$ es la verosimilitud logarítmica del GARCH del i-ésimo activo, $\theta_i^{G} := (d_i, a_i, b_i)'$, y $\theta = \left(\alpha,\beta,\theta_1^{G'},...,\theta_N^{G'}\right)'$ es el vector que reúne todos los parámetros del modelo. Nótese que la dependencia en la matriz C se omite ya que C se estima como en la ecuación [4]. También es posible asumir que las innovaciones siguen otras distribuciones como, por ejemplo, la *t-Student* multivariante (Pesaran y Pesaran, 2007), cuyos extremos son notablemente más gruesos que los de la distribución Gaussiana.

2.2.2. Verosimilitud compuesta

La evaluación de la verosimilitud logarítmica de la familia de modelos DCC requiere el cálculo del determinante de la matriz de correlaciones dinámica R_t, así como de su inversa R_t^{-1}. La complejidad de las computaciones mencionadas es aproximadamente de (TN), un cálculo progresivamente arduo a medida que la dimensión de la serie de retornos crece. Como el estimador de máxima verosimilitud se obtiene mediante métodos numéricos, la función de verosimilitud logarítmica ha de evaluarse múltiples veces, por lo tanto, la relevancia práctica de la cuestión es más que considerable.

El método de *verosimilitud compuesta* fue introducido por Pakel, Engle, Shephard y Sheppard (2017). Este se basa en la aproximación de la función de verosimilitud logarítmica conjunta empleando verosimilitudes marginales bivariantes. Definimos la verosimilitud logarítmica marginal del j-ésimo par de activos como:

$$l_{jT}(\alpha,\beta) = -\frac{1}{2}\sum_{t=1}^{T}\left(\log|R_{jt}| + \boldsymbol{\varepsilon}_{jt}' R_{jt}^{-1}\boldsymbol{\varepsilon}_{jt}\right), \qquad [5]$$

donde se entiende que $\boldsymbol{\varepsilon}_{jt} = (r_{j1t}/d_{j1t},\ r_{j2t}/d_{j2t})'$ y que R_{jt} es su correspondiente matriz de correlaciones condicionales. De tal manera, es posible aproximar la verosimilitud logarítmica tomando el promedio de todos los pares, como se representa en la ecuación [6] en la que $L = \binom{N}{2}$ es el número de todas las posibles parejas de N activos. A esta aproximación se la denomina verosimilitud compuesta,

$$\mathcal{L}_{LT}^{DCC}(\alpha,\beta) = \frac{1}{L}\sum_{j=1}^{L} l_{jT}(\alpha,\beta). \qquad [6]$$

La complejidad del problema queda, por tanto, reducida a $\mathcal{O}(TN^2)$. El proceso para una distribución *t*-Student es análogo al descrito en los párrafos previos. Cabe recalcar que naturalmente la verosimilitud logarítmica de cada pareja j depende de una matriz C_j constante 2 x 2 análoga a la matriz C de la ecuación [3], que estimamos vía focalización en covarianzas como en [4].

La versión más común de la verosimilitud compuesta utiliza solamente parejas *contiguas,* esto es, $X_{1t}=(r_{1t},r_{2t})',\ldots,X_{N-1t}=(r_{N-1t},r_{Nt})'$, simplificándose el problema en cuestión en $\mathcal{O}(TN)$. Pakel, Engle, Shephard y Sheppard (2017) demuestran que maximizar la verosimilitud logarítmica compuesta resulta en estimaciones consistentes (e incluso eficientes) de los parámetros α y β.

2.3. Predicción en el modelo DCC

La predicción a k períodos vista desde la fecha h de construcción de una cartera basada en un modelo DCC se puede realizar en tres pasos, dados por las ecuaciones [7], [8] y [9] que explicamos a continuación.

La predicción a k períodos vista desde la fecha h de las varianzas condicionales del proceso GARCH(1,1) para el i-ésimo activo viene dada por:

$$\mathbb{E}[d_{i,h+k}^2 \mid \mathcal{F}_h]=\begin{cases}(1-a_i-b_i)d_i^2+a_i r_{i,h}^2+b_i d_{i,h}^2 & k=1\\ d_i^2+(a_i+b_i)^{k-1}\left(\mathbb{E}[d_{i,h+1}^2 \mid \mathcal{F}_h]-d_i^2\right) & k>1\end{cases}. \qquad [7]$$

Para simplificar la notación recogemos las predicciones de cada activo a k períodos vista en la matriz diagonal $\hat{D}_{h+k}$ de dimensiones N x N.

Para la predicción de las correlaciones condicionales a k períodos vista desde h adoptamos la estrategia de Engle y Shephard (2001) donde se usa la aproximación $\mathbb{E}\left[R_{h+k} \mid \mathcal{F}_h\right]=\mathbb{E}\left[\boldsymbol{\varepsilon}_{h+k}\boldsymbol{\varepsilon}_{h+k}' \mid \mathcal{F}_h\right]\approx\mathbb{E}\left[Q_{h+k} \mid \mathcal{F}_h\right]$. De esta manera, la predicción a k períodos vista puede escribirse como:

$$\hat{R}_{h+k}\equiv\mathbb{E}\left[R_{h+k} \mid \mathcal{F}_h\right]=\begin{cases}(1-\alpha-\beta)C+\alpha\boldsymbol{\varepsilon}_h\boldsymbol{\varepsilon}_h'+\beta Q_h & k=1\\ C+(\alpha+\beta)^{k-1}\left(\mathbb{E}\left[R_{h+1} \mid \mathcal{F}_h\right]-C\right) & K>1\end{cases} \qquad [8]$$

Combinando ecuaciones [7] y [8], la predicción a k períodos vista de la matriz de covarianzas condicionales queda determinada por la siguiente ecuación:

$$\mathbb{E}\left[R_{h+k} \mid \mathcal{F}_h\right]=\mathbb{E}\left[\boldsymbol{\varepsilon}_{h+k}\boldsymbol{\varepsilon}_{h+k}' \mid \mathcal{F}_h\right]\approx\mathbb{E}\left[Q_{h+k} \mid \mathcal{F}_h\right]. \qquad [9]$$

En la selección de carteras de inversión, es común construir la matriz de covarianzas dinámicas en la fecha h en base a una única predicción de la matriz de covarianzas

para no realizar ninguna transacción hasta la próxima fecha de inversión $h + \mathcal{K}$, donde $\mathcal{K}$ es un número entero positivo. De esta manera, uno evita incurrir en los costes de transacción provenientes de una rotación excesiva de la cartera. Una estrategia común –véanse, por ejemplo, Engle, Ledoit y Wolf (2019) y De Nard, Ledoit y Wolf (2020) – es calcular el promedio de las predicciones a k períodos vista, esto es,

$$\hat{\Sigma}_h \equiv \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \hat{\Sigma}_{h+k} \ .$$

2.4. DCC con estructura de factores

Hay una larga tradición de modelos factoriales dinámicos en macroeconomía y finanzas. En macroeconomía, los modelos factoriales dinámicos se usan frecuentemente para ejercicios de predicción con *big data* (Forni *et al.,* 2000; Stock y Watson, 2002b; Peña y Poncela, 2004). En finanzas, se suele emplear modelos de factores para explicar el comportamiento de los retornos de las acciones. En particular, la teoría del arbitraje de Ross (1976) implica que los retornos de los activos se explican por un pequeño número de factores.

En la literatura financiera se han considerado tanto modelos de factores latentes como observables. En este trabajo nos centramos en estos últimos. Un ejemplo tradicional es el modelo CAPM donde el único factor es la prima de riesgo de mercado $r_{m,t} - r_{f,t}$, donde $r_{m,t}$ es el retorno del mercado y $r_{f,t}$ el retorno libre de riesgo. Para la serie $\{r_{m,t}\}$ se suele emplear la media ponderada de los retornos de todos los activos del universo en consideración (por ejemplo, todas las compañías disponibles en la base de datos CRSP cuyas acciones se comercian en NYSE, AMEX o NASDAQ) y para $\{r_{f,t}\}$ se suele usar el retorno de algún activo que se considere seguro como, por ejemplo, los *Treasury bills* a un mes en EE. UU. o el *Bund* alemán en Europa. Naturalmente, se pueden considerar más factores como el "Small Minus Big" o el "High Minus Low", pero en la aplicación empírica que consideramos en este trabajo éstos no aportan un gran valor añadido.

Es posible incorporar la estructura de factores al modelo DCC visto anteriormente. En este caso, el vector de retornos $\mathbf{r}_t$ viene dado por la siguiente ecuación:

$$\mathbf{r}_t = B\boldsymbol{f}_t + \boldsymbol{u}_t, \ \boldsymbol{u}_t \sim \text{DCC}(1,1)$$

donde $\boldsymbol{f}_t$ es un vector que recoge los K factores y B es una matriz N x K.

Es fácil ver que la matriz de covarianza de los retornos Σ_t se basa en la covarianza de los factores Σ_f y la covarianza de los residuos $\Sigma_{u,t}$ como sigue a continuación:

$$\Sigma_t = B\Sigma_f B' + \Sigma_{u,t}. \qquad [10]$$

Por lo tanto, para implementar el modelo DCC con estructura de factores es suficiente estimar primero B mediante N regresiones lineales por mínimos cuadrados ordinarios de las cuales se extraen los residuos $\boldsymbol{u}_t$. Luego, uno puede estimar $\Sigma_{u,t}$ mediante la metodología DCC estándar. La matriz Σ_f se puede estimar naturalmente vía la covarianza muestral de la serie de factores.

Nótese que también se podría haber considerado un modelo dinámico para los factores, y en ese caso $\Sigma_{f,t}$ se puede estimar fácilmente también mediante la metodología DCC. También es fácil lidiar con el caso en el que B es dinámica. Por ejemplo, en el caso $K = 1$, $B = Cov(\boldsymbol{r}_t, f_t)/Var(f_t)$, donde $Cov(\boldsymbol{r}_t, f_t)$ es un vector de N dimensiones con las correspondientes covarianzas. Si se cree que B es dinámica, entonces B_t se puede construir estimando $Cov(\boldsymbol{r}_t, f_t|\mathcal{F}_{t-1})$ y $Var(f_t|\mathcal{F}_{t-1})$ mediante modelos GARCH. Sin embargo, ninguna de estas generalizaciones parece aportar un gran valor diferencial en nuestra aplicación práctica (De Nard, Ledoit y Wolf, 2020), de manera que para ahorrar espacio únicamente se reportan los resultados del modelo descrito en [10].

2.5. DCC con regularización

La cuestión que aborda esta sección está estrechamente relacionada con el problema de predicción de alta dimensionalidad, también denominado como predicción con big data. Cuando la estructura de los datos es *ancha* –es decir, más variables que observaciones– solemos encontrarnos con problemas de imprecisión y sobreajuste (*overfitting*). En estos contextos, las metodologías de estimación con regularización han demostrado ser extremadamente efectivas en aprendizaje automático supervisado – clasificación y regresión. La razón detrás de esto es que hay un compromiso entre *sesgo* y *varianza* para cada metodología estadística que utilizamos. Dicho de otra forma, un método demasiado simple es generalmente incapaz de capturar patrones interesantes en nuestro conjunto de datos, pero por otro lado es más robusto a ligeras perturbaciones en los datos: su sesgo es elevado y su *varianza* es reducida; por otro lado, un método más complejo puede capturar prácticamente cualquier patrón y si es demasiado complejo también encontrará patrones donde en realidad únicamente hay ruido: en este caso, su sesgo es reducido pero su varianza es elevada, ya que el resultado de la estimación puede variar considerablemente dada una pequeña perturbación en los datos. Para un tratamiento más riguroso de la predicción con big data y los problemas de elevada dimensionalidad, véanse Hastie, Tibshirani y Friedman (2008) y Wainwright (2019).

Volviendo al contexto de la metodología DCC, la técnica de focalización en covarianzas requiere un estimador adecuado de la matriz C basado únicamente en los residuos estandarizados $\boldsymbol{\varepsilon}_t$ (ecuación [4]). La covarianza muestral es un estimador bastante defectuoso cuando la dimensionalidad del problema es elevada. Esto se debe a que no es posible estimar con exactitud $\mathcal{O}(N^2)$ parámetros con $\mathcal{O}(N^2)$ observaciones. En términos técnicos, cuando la *ratio de concentración* N/T aumenta, la teoría asintótica estándar

no describe adecuadamente el comportamiento real de la matriz de covarianzas. Esto ha motivado una amplia literatura que ha propuesto el uso de estimadores regularizados que funcionan bien cuando la ratio de concentración es elevada. En esta sección introducimos una clase de estimadores regularizados para aumentar la precisión en la estimación con respecto a la covarianza muestral.

El modelo DCC con regularización se define igual que su versión estándar con focalización en covarianzas salvo por el hecho de que la estimación de la matriz C se realiza mediante un estimador regularizado.

2.5.1. Regularización lineal

El método de regularización lineal de Ledoit y Wolf (2004b) cuenta con una fórmula clara, fácil de estimar e interpretar. De hecho, ésta es la combinación lineal convexa asintóticamente óptima de la matriz de covarianza muestral y de la matriz identidad.

Supongamos que $C = \mathbb{E}\left[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^{'}\right]$ –aunque esto no sea del todo correcto en el modelo DCC. El método de regularización lineal consiste en encontrar la combinación lineal:

$$C^{*} = \rho_1 \mu I + (1-\rho_1) S, \qquad [11]$$

entre la matriz escalar μI y la matriz de covarianzas muestral $S = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^{'}$ que minimiza la pérdida cuadrática esperada. El coeficiente ρ_1 es la denominada intensidad de regularización (*shrinkage intensity*) y μI el objetivo de regularización (*shrinkage target*). Nótese que, a su vez, la intensidad de regularización en este caso coincide con la ratio de mejora con respecto a la covarianza muestral S en términos de pérdida cuadrática esperada.

Es posible demostrar que este estimador es consistente bajo teoría asintótica general, es decir, cuando dejamos que N y T tiendan a infinito simultáneamente de manera que $N/T \to c < \infty$.

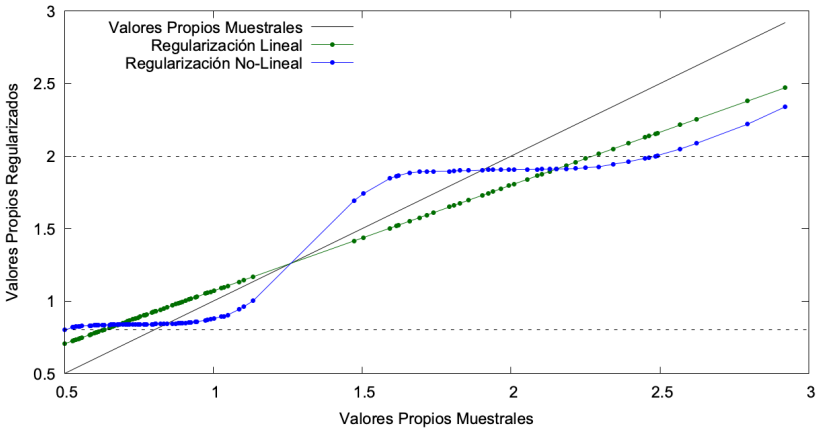
2.5.2. Regularización no lineal

La regularización lineal tiene la limitación de que modifica todos los valores propios hacia un mismo objetivo y con la misma intensidad, por lo tanto no permite capturar adecuadamente el hecho de que la regularización es un fenómeno esencialmente local y no global. La figura 1 muestra como el uso de la atracción local proporciona una mejor estimación de los valores propios de la matriz de covarianzas poblacional. En este ejemplo, la matriz covarianzas poblacional con $N = 100$ tiene 62 valores propios iguales a 0,8, y los 38 restantes son iguales a 2 (líneas punteadas horizontales). El tamaño muestral es $T = 1260$. La regularización lineal se calcula como en la sección 2.5.1 y

la regularización no-lineal mediante el procedimiento que explicamos a continuación.

FIGURA 1

REGULARIZACIÓN LINEAL Y NO LINEAL



Notas: Efecto de Atracción Local. La matriz de covarianzas poblacional con $N = 100$ tiene 62 valores propios iguales a 0,8, y los 38 restantes son iguales a 2 (líneas punteadas horizontales). El tamaño muestral es $T = 1260$. La regularización lineal se calcula como en la sección 2.5.1, y la regularización no-lineal como se describe en la sección 2.5.2, esto es, mediante el uso de la transformada de Hilbert con kernel de Epanechnikov.

Fuente: Elaboración propia.

Ledoit y Wolf (2020) introducen una fórmula analítica para la regularización no lineal óptima de matrices de covarianzas de elevadas dimensiones. Dicha fórmula se obtiene observando que el problema está estrechamente relacionado con la estimación (no paramétrica) de la función de densidad de los valores propios de la matriz de covarianzas muestral y de su transformada de Hilbert. En primer lugar, se calcula la descomposición espectral de la matriz de covarianzas muestral S como se define en la subsección 2.5.1., esto es,

$$S = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' = \sum_{i=1}^{N}\lambda_i u_i u_i', \qquad [12]$$

donde $\lambda_1,\ldots,\lambda_N$ son los valores propios muestrales en orden descendiente sin pérdida de generalidad y $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N$ son los respectivos vectores propios. Se considera la clase de estimadores *invariantes bajo rotación,* un marco conceptual desarrollado por Stein (1986). Bajo la pérdida de mínima varianza de Engle y Colacito (2006) y la pérdida cuadrática, la Proposición 2.1. en Ledoit y Wolf (2020) establece que el estimador óptimo dentro de esa clase se obtiene reemplazando λ_i por $\hat{d}_i = \boldsymbol{u}_i' C \boldsymbol{u}_i$ en [12]. Naturalmente esta cuantía es irrealizable ya que depende de la verdadera matriz C.

Bajo los supuestos 3.1.-3.4. de Leoit y Wolf (2020), el estimador "oráculo"

$$\forall i = 1,...,N \quad d^o(\lambda_i) = \frac{\lambda_i}{\left[\pi \frac{N}{T} \lambda_i f(\lambda_i)\right]^2 + \left[1 - \frac{N}{T} - \pi \frac{N}{T} \lambda_i \mathcal{H}_f(\lambda_i)\right]^2} \qquad [13]$$

minimiza el límite (con probabilidad 1) de la pérdida de mínima varianza bajo teoría asintótica general. Este estimador sigue siendo irrealizable pero representa un gran progreso con respecto a $\hat{d}_i$, ya que éste depende de N (N+1)/2 parámetros (libres) de C, mientras que d^o (λ_i) depende solamente de N parámetros (los valores propios de la matriz C). Dados los supuestos mencionados, la función $f(\lambda_i) \equiv \frac{dF}{d\lambda_i}$ y su transformada de Hilbert ($\mathcal{H}_f$) existen, donde F es el límite (con probabilidad 1) de la función de distribución empírica de los valores propios de S.

En la transformada de Hilbert reside gran parte de la clave, ya que es esta transformada la que actúa como una *fuerza de atracción local,* devolviendo valores muy positivos cuando hay una gran proporción de otros valores propios muestrales en puntos ligeramente mayores que λ_i. Similarmente, la transformada devuelve valores muy negativos cuando hay una gran proporción de otros valores propios muestrales en puntos ligeramente menores que λ_i. Si inspeccionamos la ecuación [13], vemos que cuando hay una gran densidad de otros valores propios muestrales ligeramente por encima de λ_i, el valor de la transformada de Hilbert aumenta y el oráculo empuja a λ_i hacia arriba, y viceversa.

Un estimador realizable de la matriz de covarianzas se puede obtener como se representa en la ecuación [14], donde $\tilde{d}_i$ está definida por la función de regularización no lineal asintóticamente óptima recogida en la ecuación [15].

$$\hat{C} \equiv \sum_{i=1}^{N} \tilde{d}_i \mathbf{u}_i \mathbf{u}_i' \qquad [14]$$

$$\forall i = 1,...,N \quad \tilde{d}_i(\lambda_i) \equiv \frac{\lambda_i}{\left[\pi \frac{N}{T} \lambda_i \tilde{f}(\lambda_i)\right]^2 + \left[1 - \frac{N}{T} - \pi \frac{N}{T} \lambda_i \mathcal{H}_{\tilde{f}}(\lambda_i)\right]^2}, \qquad [15]$$

donde $\tilde{f}$ y $\mathcal{H}_{\tilde{f}}$ son estimadores no paramétricos con una amplitud de banda local que es proporcional a cada valor propio λ_i.

Cabe observar que este estimador depende exclusivamente de fórmulas analíticas y por lo tanto no requiere de ningún cálculo numérico para su implementación práctica, lo que lo hace especialmente atractivo en aplicaciones financieras con centenares o incluso miles de activos financieros.

3. SELECCIÓN DE CARTERAS

Tal y como mencionamos en la introducción, este trabajo utiliza la conocida teoría moderna de carteras –o MPT, por sus siglas en inglés– como marco matemático para construir una cartera de activos. También se lo conoce como análisis media-varianza, ya que el problema de inversión se formula como un problema de optimización donde el vector de N pesos de la cartera $\boldsymbol{w}$ se escoge con el objetivo de maximizar el retorno esperado (media) para un nivel de riesgo dado (varianza). Usando notación matemática, el objetivo del inversor es minimizar la expresión:

$$\boldsymbol{w}'\Sigma\boldsymbol{w} = \gamma\boldsymbol{\mu}'\boldsymbol{w}$$

donde $\boldsymbol{\mu}$ es el vector de retornos esperados, Σ es la matriz de varianzas y covarianzas de los retornos y $\gamma \in [0,+\infty)$ representa la tolerancia al riesgo del inversor. Si $\gamma = 0$ la tolerancia es cero y como resultado se obtiene la cartera de mínima varianza, mientras que a medida que γ tiende a infinito, se obtienen carteras con rentabilidad esperada y riesgo que tienden a infinito simultáneamente a lo largo de la frontera eficiente. Entre los dos extremos, se obtienen carteras que encuentran un compromiso entre retorno esperado y riesgo.

Para poder implementar la estrategia en la práctica, es necesario usar datos históricos de retornos financieros para estimar los parámetros $\boldsymbol{\mu}$ y Σ que son generalmente desconocidos. También es fácil acomodar el caso en que la media o varianza son dinámicas, *i.e.* $\boldsymbol{\mu}_h$ y Σ_h, donde h representa una cierta fecha de inversión. Repitiendo el algoritmo para las fechas $h = 1,...,H$, se obtiene la secuencia de pesos $\{\boldsymbol{w}_h\}_{h=1}^{H}$. Puesto que en frecuencias diarias no hay una gran predecibilidad en la media condicional $\boldsymbol{\mu}_h$, en este trabajo nos centramos en metodologías de predicción de la matriz de covarianzas condicional Σ_h, que además es un ejercicio más interesante desde el punto de vista estadístico. Una opción natural en este contexto es analizar la cartera de mínima varianza (caso $\gamma = 0$). Otra opción es utilizar un proxy para el vector de retornos esperados como por ejemplo un índice de *momentum*. Los índices de *momentum* se emplean con frecuencia en estrategias de *trading* que consisten en comprar activos que han experimentado retornos muy positivos durante los últimos meses (generalmente se excluye el mes inmediatamente anterior), y vender aquellos que por el contrario han tenido un mal rendimiento en el mismo periodo.

En las dos subsecciones que siguen a continuación explicamos ambas estrategias en mayor detalle.

3.1. Cartera de mínima varianza

Consideramos el problema de minimización de la varianza de una cartera con N activos sin restricciones a la venta en corto. El problema de inversión en la fecha h queda formulado como,

$$\min_{\boldsymbol{w}_h} \quad \boldsymbol{w}_h' \Sigma_h \boldsymbol{w}_h \qquad [16]$$

$$\textit{sujeto a} \quad \boldsymbol{w}_h' \mathbf{1} = 1, \qquad [17]$$

donde $\mathbf{1}$ denota el vector unitario de dimensión N. Este problema tiene la solución analítica:

$$\boldsymbol{w}_h = \frac{\Sigma_h^{-1}\mathbf{1}}{\mathbf{1}'\Sigma_h^{-1}\mathbf{1}}.$$

Naturalmente, la estrategia consiste en obtener el vector $\boldsymbol{w}_h$ de pesos para cada fecha de actualización h reemplazando Σ_h por su predicción, $\hat{\Sigma}_h$. De esta manera, el vector de pesos factible viene dado por,

$$\hat{\boldsymbol{w}}_h = \frac{\hat{\Sigma}_h^{-1}\mathbf{1}}{\mathbf{1}'\hat{\Sigma}_h^{-1}\mathbf{1}}.$$

Estimar la cartera de mínima varianza presenta la ventaja evidente de ser un problema "limpio" en el sentido de que no es necesario realizar una estimación del vector de retornos esperados. Además, diversos autores han señalado que las carteras de mínima varianza no sólo tienen propiedades deseables a la hora de reducir el riesgo sino también en términos del binomio rentabilidad-riesgo, cuantificado mediante la ratio de Información.

3.2. Cartera de Markowitz con señal de momentum

En esta subsección adoptamos un enfoque más completo en tanto que consideramos un problema media-varianza (caso $\gamma \neq 0$), en particular, la cartera con señal predictiva basada en *momentum*. Al igual que en la sub-sección anterior, la solución a este problema tiene una expresión analítica sencilla que depende de (la inversa de) la matriz de covarianzas dinámicas. Sin embargo, a diferencia de la cartera de mínima varianza, la solución también depende del vector de retornos esperados, que en este contexto se estima mediante el factor de momentum como se describe en Jegadeesh y Titman (1993). Concretamente, para cada fecha h y cada activo, el factor *momentum* se calcula como la media geométrica de los retornos del último año excluyendo el último mes (*i.e.,* los últimos 21 días). Los *momentums* de los N activos de nuestro universo se recogen en el vector $\hat{\boldsymbol{m}}_h$.

En ausencia de restricciones a la venta en corto, el problema de inversión en la fecha h se formula como:

$$\min_{\boldsymbol{w}_h} \quad \boldsymbol{w}_h' \Sigma_h \boldsymbol{w}_h \qquad [18]$$

$$\textit{sujeto a} \quad \boldsymbol{w}_h' \boldsymbol{m}_h = b_h, \qquad [19]$$

$$\boldsymbol{w}_h' \mathbf{1} = 1, \qquad [20]$$

donde b_h es el objetivo de retorno esperado, que calculamos mediante la media aritmética del *momentum* de los activos en el quintil superior – donde el orden se determina también por *momentum*. Como se ve claramente, la restricción [19] obliga a que la solución pase por dar un peso suficientemente elevado a los activos con mayor *momentum* para alcanzar una rentabilidad objetivo b_h. Escribiendo el Lagrangiano y resolviendo un sistema de ecuaciones lineal 2 x 2, vemos que la solución a este problema viene dada por:

$$\boldsymbol{w}_h = c_1 \cdot \Sigma_h^{-1} \boldsymbol{m}_h + c_2 \cdot \Sigma_h^{-1} \mathbf{1},$$

donde,

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \frac{1}{AC - B^2} \begin{bmatrix} A \cdot b_h - B \\ C - B \cdot b_h \end{bmatrix}$$

$$A = \mathbf{1}' \Sigma_h^{-1} \mathbf{1}$$

$$B = \mathbf{1}' \Sigma_h^{-1} \boldsymbol{m}_h, \ y$$

$$C = \boldsymbol{m}_h' \Sigma_h^{-1} \boldsymbol{m}_h.$$

Nuevamente, la estrategia consiste en reemplazar Σ_h Σ_h por $\hat{\Sigma}_h$ y $\boldsymbol{m}_h$ por $\hat{\boldsymbol{m}}_h$, con lo que se obtiene el vector de pesos (factible):

$$\hat{\boldsymbol{w}}_h = \hat{c}_1 \cdot \hat{\Sigma}_h^{-1} \hat{\boldsymbol{m}}_h + \hat{c}_2 \cdot \hat{\Sigma}_h^{-1} \mathbf{1}.$$

4. APLICACIÓN EMPÍRICA

En esta aplicación realizamos un ejercicio de selección de carteras de inversión. Se compara el rendimiento de algunas de las metodologías mencionadas anteriormente de manera similar a Engle, Ledoit y Wolf (2019) y De Nard, Ledoit y Wolf (2020). Los datos utilizados son los precios de cierre diarios ajustados por dividendos y *splits,* $\{P_{i,t}\}$, de las acciones que constituyen el índice S&P 500 en el período que va de 04-01-2000 a 30-06-2019 (fuente: AlphaVantage). En cada fecha de inversión h, seleccionamos los

100 activos con mayor capitalización bursátil (por lo tanto, N=100). Construimos los retornos para cada activo i como $r_{i,t} = 100 \cdot \log(P_{i,t}/P_{i,t-1})$.

Para la familia de modelos DCC que incorporan estructura de factores utilizamos los datos de retornos diarios del factor "Prima de Riesgo de Mercado" de Fama y French (2015) de la página web de Ken French durante el mismo período[2]. Por lo tanto, el número de factores K=1. Este factor se calcula como el retorno ponderado por valor de todas las empresas de la base de datos CRSP registradas en los Estados Unidos que cotizan en los mercados NYSE, AMEX y NASDAQ, neto del tipo de interés a un mes del *US Treasury bill.*

Para esta aplicación, adoptamos la convención de que un "mes" equivale a 21 días consecutivos de comercialización. El período de validación de las diferentes estrategias va de 07-01-2005 a 30-06-2019, lo que representa un total de 173 meses (es decir, $h = 1,...,173$). La cartera se actualiza a principios de cada mes y se mantiene constante durante todo el mes. En cada fecha de actualización h, realizamos una predicción de la matriz de covarianzas $\hat{\Sigma}_h$ como se explica en la subsección 2.3. basada en los datos de los últimos cinco años hasta la fecha correspondiente –lo que equivale a un total de 1260 observaciones–. Los pesos de la cartera $\boldsymbol{w}_h$ son una función de (la inversa de) $\hat{\Sigma}_h$. En particular, consideramos dos estrategias conocidas: la cartera de mínima varianza (3.1.) y la cartera de Markowitz con señal de *momentum* (3.2.). Consecuentemente, el retorno de la cartera c en el día t del mes h se calcula como,

$$r_{c,t} = \boldsymbol{w}_h' \boldsymbol{r}_t .$$

Para predecir dicha matriz de covarianzas, los modelos de la familia DCC "estándar" que consideramos son los siguientes: DCC-SC, que utiliza la covarianza muestral de los residuos estandarizados para estimar la matriz C en la ecuación [3], DCC-LS, que en lugar de la covarianza muestral emplea el estimador de regularización lineal (2.5.1.), y DCC-NLS, que emplea regularización no lineal (2.5.2.). De manera similar, denotamos los modelos DCC que incorporan estructura de factores con las siglas AFM-DCC-SC, AFM-DCC-LS y AFM-DCC-NLS.

Para ampliar el rango de la comparativa, consideramos la versión escalar del modelo VECH (sVECH) con focalización en covarianzas. La matriz de covarianzas dinámica para este modelo se define como:

$$\Sigma_t = (1-\gamma_1-\gamma_2)\Sigma + \gamma_1 \boldsymbol{r}_{t-1}\boldsymbol{r}_{t-1}' + \gamma_2 \Sigma_{t-1} , \qquad [21]$$

donde aquí la igualdad $\Sigma = \mathbb{E}[\boldsymbol{r}_t \boldsymbol{r}_t']$ es exacta siempre y cuando $\gamma_1+\gamma_2<1$. Similarmente, utilizamos los acrónimos sVECH-SC, sVECH-LS y sVECH-NLS para identificar la versión

[2] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

del modelo que utiliza la covarianza muestral, el estimador con regularización lineal y el estimador con regularización no lineal, respectivamente. Todos los modelos introducidos en esta sección se estiman mediante verosimilitud compuesta (2.2.2).

En las tablas 1 y 3 reportamos las siguientes medidas de desempeño para cada cartera en consideración. El primer grupo de métricas nos proporcionan el resumen numérico de los retornos de la cartera, y son 1) AV: Retorno medio de la cartera en el período de validación. Anualizamos multiplicando el resultado por 252. 2) SD: Desviación estándar (o volatilidad) de la cartera en el período de validación. Anualizamos multiplicando el resultado por $\sqrt{252}$. Para evitar ambigüedad, la calculamos como sigue:

$$SD = \sqrt{252} \times \sqrt{\frac{1}{173 \cdot 21} \sum_{t=1}^{173 \cdot 21} r_{c,t}^2},$$

y 3) IR: Ratio de información, *i.e.* AV / SD.

El segundo grupo de métricas nos informa sobre la evolución y distribución de los pesos a lo largo del ejercicio empírico. En particular, estas son 4) TO: Rotación (turnover) mensual de la cartera que calculamos como $\frac{1}{172}\sum_{h=1}^{172} \| \hat{\mathbf{w}}_{h+1} - \hat{\mathbf{w}}_h^* \|_1$, donde $\|\cdot\|_1$ es la norma L^1 y utilizamos $\mathbf{w}_h^*$ para denotar el vector de pesos a final del mes h - que puede diferir de $\hat{\mathbf{w}}_h$ pues el primero tiene en cuenta la evolución de precios de cada activo durante el mes. Este índice cuantifica si los pesos de la cartera experimentan cambios grandes a lo largo del ejercicio, lo que en la práctica se traduce en mayores costes de transacción; 5) PL: Proporción de apalancamiento (*leverage*), calculada como $\frac{1}{173 \times N}\sum_{h=1}^{173}\sum_{i=1}^{N} \mathbf{1}_{\{\hat{\mathbf{w}}_{i,h}<0\}}$. Esta medida cuantifica la proporción activos que hemos vendido en corto a lo largo del ejercicio, 6) Max: Máximo peso de la cartera $\max_{i,t} \hat{\mathbf{w}}_{i,t}$, 7) Min: Mínimo peso de la cartera $\min_{i,t} \hat{\mathbf{w}}_{i,t}$, y 8) HI: Índice de concentración Herfindahl corrigiendo por posibles valores negativos. Un índice elevado se traduce en una cartera donde la riqueza está invertida en un número relativamente pequeño de activos.

Es legítimo preguntarse entonces por qué no considerar datos mensuales. En primer lugar, con frecuencia mensual no disponemos de suficientes observaciones para estimar modelos GARCH multivariantes con suficientes garantías. Por otro lado, usar datos diarios parece ofrecer mejores resultados incluso cuando la frecuencia de inversión es mensual (véase, por ejemplo, Ledoit y Wolf, 2017).

Para evaluar los resultados del ejercicio de mínima varianza (subsección 3.1.), tomamos como referencia la cartera equiponderada también conocida en la literatura como 1/N (De Miguel, Garlappi y Uppal, 2009). De acuerdo con este trabajo, en carteras con un número moderado de activos (entre 3 y 50) y una ventana de estimación de 120 meses, es difícil batir la cartera equiponderada con carteras Markowitz en términos de ratio de

información. De las medidas de desempeño presentadas al inicio de esta sección, la más importante en este contexto es sin duda SD, puesto que esta cartera está especialmente diseñada con el objetivo de minimizar esta métrica. Naturalmente, también son de interés el retorno medio de la cartera y su ratio de información, pero en este contexto son de importancia secundaria.

TABLA 1.

CARTERA DE MÍNIMA VARIANZA: *N*=100
PERÍODO DE VALIDACIÓN DE 07-01-2005 HASTA 30-06-2019

Modelo	*AV*	*SD*	*IR*	*TO*	*PL*	*Max*	*Min*	*HI*
DCC-LS	12.060	11.848	1.018	0.154	0.460	0.372	-0.159	0.013
DCC-NLS	12.185	11.732	1.039	0.109	0.463	0.373	-0.130	0.013
DCC-SC	12.057	11.874	1.015	0.171	0.461	0.373	-0.164	0.013
AFM-DCC-LS	14.145	11.649	1.214	0.139	0.425	0.427	-0.187	0.012
AFM-DCC-NLS	14.620	11.648	1.255	0.686	0.425	0.448	-0.192	0.012
AFM-DCC-SC	14.094	11.803	1.194	0.566	0.427	0.462	-0.209	0.012
sVECH-LS	12.966	11.881	1.091	0.198	0.443	0.353	-0.242	0.012
sVECH-NLS	12.955	11.834	1.095	0.151	0.441	0.353	-0.237	0.012
sVECH-SC	12.961	11.934	1.086	0.871	0.446	0.353	-0.245	0.012
1/N	10.757	19.645	0.548	0.056	0.000	0.010	0.010	0.010

Nota: Medidas de desempeño para varios estimadores de la cartera de mínima varianza. Las columnas AV y SD se reportan en términos anualizados (en porcentaje), es decir, multiplicando por $\sqrt{252}$. AV denota el retorno medio; SD es su desviación estándar; IR es la ratio de información; TO mide la rotación de la cartera; PL es la proporción de apalancamiento; Max y Min son el peso máximo y mínimo, respectivamente; y HI es el índice Herfindahl corrigiendo por valores negativos.

Fuente: Elaboración propia.

La significación estadística de los resultados se basa en el test de la diferencia de varianzas propuesto por Ledoit y Wolf (2011), que es consistente en presencia de heterocedasticidad y autocorrelación. En la tabla 2 se reportan los resultados del citado test. Las diferencias se dan en términos anualizados para hacerlos equiparables a los de la tabla 1, pero para el cálculo de significación se han utilizado los retornos diarios de la cartera $\{r_{c,t}\}$ en el período de validación sin anualizar.

En este ejercicio, la cartera que obtiene una menor volatilidad es AFM-DCC-NLS, lo cual pone en valor el uso de modelos de factores y la estimación por regularización no lineal. Por un lado, la regularización aporta un valor diferencial que es significativo tanto desde el punto de vista económico como estadístico para modelos DCC con y sin estructura de factores. La regularización no lineal ofrece mejores resultados que la regularización lineal, pero esta diferencia desaparece cuando se incorpora la estructura de factores en la ecuación. Por otro lado, las carteras basadas en modelos de factores parecen aportar un valor diferencial que es significativo desde el punto de vista económico pero no desde el punto de vista estadístico.

TABLA 2.

TEST DE DIFERENCIA DE VARIANZAS (LEDOIT Y WOLF, 2011) PARA VARIOS ESTIMADORES DE LA CARTERA DE MÍNIMA VARIANZA (*N*=100)

Modelo	*DCC-LS*	*DCC-NLS*	*DCC-SC*	*AFM-DCC-LS*	*AFM-DCC-NLS*	*AFM-DCC-SC*	*sVECH-LS*	*sVECH-NLS*	*sVECH-SC*	*1/N*
DCC-LS	–	-0.12***	0.03***	-0.20	-0.20	-0.05	0.03	-0.01	0.09	7.80***
DCC-NLS	0.12***	–	0.14***	-0.08	-0.08	0.07	0.15	0.10	0.20	7.91***
DCC-SC	-0.03***	-0.14***	–	-0.23*	-0.23*	-0.07	0.01	-0.04	0.06	7.77***
AFM-DCC-LS	0.20	0.08	0.23*	–	-0.00	0.15***	0.23	0.18	0.28	8.00***
AFM-DCC-NLS	0.20	0.08	0.23*	0.00	–	0.15***	0.23	0.19	0.29	8.00***
AFM-DCC-SC	0.05	-0.07	0.07	-0.15***	-0.15***	–	0.08	0.03	0.13	7.84***
sVECH-LS	-0.03	-0.15	-0.01	-0.23	-0.23	-0.08	–	-0.05***	0.05***	7.76***
sVECH-NLS	0.01	-0.10	0.04	-0.18	-0.19	-0.03	0.05***	–	0.10***	7.81***
sVECH-SC	-0.09	-0.20	-0.06	-0.28	-0.29	-0.13	-0.05***	-0.10***	–	7.71***
1/N	-7.80***	-7.91***	-7.77***	-8.00***	-8.00***	-7.84***	-7.76***	-7.81***	-7.71***	–

Nota: En cada celda se calcula la diferencia de la desviación estándar de la cartera de la columna menos la desviación estándar de la cartera de la fila. Las cifras se reportan en términos anualizados (es decir, multiplicando por $\sqrt{252}$). Los símbolos *, **, y *** indican que los resultados son significativos al nivel 10, 5 y 1%, respectivamente.

Fuente: Elaboración propia.

Si comparamos la familia de modelos DCC con la alternativa del VECH escalar, vemos que los modelos DCC ofrecen resultados significativamente mejores desde el punto de vista económico pero no estadístico. Por un lado, que los modelos DCC tengan un mejor desempeño tiene sentido porque permiten capturar la potencial heterogeneidad en la dinámica de las volatilidades de cada activo – algo que no permite el modelo VECH escalar. Sin embargo, el error que resulta de estimar un número de parámetros que crece linealmente en N podría eclipsar los beneficios de modelizar esa heterogeneidad. Por lo tanto, no sería adecuado concluir que hay que descartar los modelos VECH.

Finalmente, los resultados confirman que, independientemente de qué especificación utilicemos, la familia de modelos GARCH multivariante ofrece unos resultados que son significativamente mucho mejores que utilizar una cartera equiponderada.

TABLA 3.

CARTERA DE MARKOWITZ CON SEÑAL DE *MOMENTUM*: *N*=100
PERÍODO DE VALIDACIÓN DE 07-01-2005 HASTA 30-06-2019

Modelo	*AV*	*SD*	*IR*	*TO*	*PL*	*Max*	*Min*	*HI*
DCC-LS	14.339	13.556	1.058	0.373	0.466	0.445	-0.240	0.012
DCC-NLS	14.801	13.458	1.100	0.161	0.467	0.412	-0.204	0.013
DCC-SC	14.311	13.579	1.054	0.315	0.467	0.457	-0.245	0.012
AFM-DCC-LS	15.579	13.386	1.164	0.195	0.435	0.380	-0.301	0.012
AFM-DCC-NLS	16.207	13.370	1.212	2.119	0.436	0.379	-0.305	0.012
AFM-DCC-SC	15.168	13.512	1.123	0.523	0.440	0.397	-0.317	0.012
sVECH-LS	14.205	13.646	1.041	0.194	0.451	0.340	-0.277	0.012
sVECH-NLS	14.321	13.600	1.053	0.182	0.449	0.340	-0.237	0.012
sVECH-SC	14.133	13.698	1.032	0.213	0.452	0.340	-0.245	0.012
EW-TQ	10.697	21.075	0.508	0.072	0.000	0.050	0.010	0.010

Nota: Medidas de desempeño para varios estimadores de la cartera de Markowitz con señal de momentum. Las columnas AV y SD se reportan en términos anualizados (en porcentaje), es decir, multiplicando por $\sqrt{252}$. AV denota el retorno medio; SD es su desviación estándar; IR es la ratio de información; TO mide la rotación de la cartera; PL es la proporción de apalancamiento; Max y Min son el peso máximo y mínimo, respectivamente; y HI es el índice Herfindahl corrigiendo por valores negativos.

Fuente: Elaboración propia.

Para el ejercicio de cartera Markowitz con señal de *momentum* (subsección 3.2), consideramos los mismos estimadores que para el ejercicio de mínima varianza, y como referencia esta vez utilizamos la cartera equiponderada de las acciones en el quintil superior por *momentum* (EW-TQ). En este caso, tiene mucho más sentido evaluar el desempeño de las diferentes estrategias primordialmente en base a la ratio de información IR, que cuantifica la relación entre rentabilidad y riesgo. En un contexto "ideal", minimizar la varianza para un objetivo fijo de rentabilidad b_h debería ser equivalente a maximizar la

TABLA 4.

TEST DE DIFERENCIA DE RATIO DE INFORMACIÓN (LEDOIT Y WOLF, 2011) PARA VARIOS ESTIMADORES DE LA CARTERA DE MARKOWITZ CON SEÑAL DE *MOMENTUM* (*N*=100)

Modelo	*DCC-LS*	*DCC-NLS*	*DCC-SC*	*AFM-DCC-LS*	*AFM-DCC-NLS*	*AFM-DCC-SC*	*sVECH-LS*	*sVECH-NLS*	*sVECH-SC*	*EW-TQ*
DCC-LS	–	-0.04**	0.00	-0.11	-0.15	-0.06	0.02	0.00	0.03	0.55**
DCC-NLS	0.04**	–	0.05**	-0.06	-0.11	-0.02	0.06	0.05	0.07	0.59**
DCC-SC	-0.00	-0.05**	–	-0.11	-0.16	-0.07	0.01	0.00	0.02	0.55**
AFM-DCC-LS	0.11	0.06	0.11	–	-0.05***	0.04*	0.12	0.11	0.13	0.66***
AFM-DCC-NLS	0.15	0.11	0.16	0.05***	–	0.09***	0.17	0.16	0.18	0.70***
AFM-DCC-SC	0.06	0.02	0.07	-0.04*	-0.09***	–	0.08	0.07	0.09	0.61***
sVECH-LS	-0.02	-0.06	-0.01	-0.12	-0.17	-0.08	–	-0.01*	0.01*	0.53**
sVECH-NLS	-0.00	-0.05	-0.00	-0.11	-0.16	-0.07	0.01*	–	0.02*	0.55**
sVECH-SC	-0.03	-0.07	-0.02	-0.13	-0.18	-0.09	-0.01*	-0.02*	–	0.52**
EW-TQ	-0.55**	-0.59**	-0.55**	-0.66***	-0.70***	-0.61***	-0.53**	-0.55**	-0.52**	–

Nota: En cada celda se calcula la diferencia de la ratio de información de la cartera de la fila menos la ratio de información de la cartera de la columna. Las cifras se multiplican por $\sqrt{252}$ para facilitar la visualización. Los símbolos *, **, y *** indican que los resultados son significativos al nivel 10, 5 y 1%, respectivamente.

Fuente: Elaboración propia.

ratio de información IR, pero en la práctica esto no es así debido al error de estimación en el vector $\boldsymbol{m}_h$. Por lo tanto, enfocarse en la métrica SD no es tan apropiado en este contexto.

De manera análoga al ejercicio anterior, la significación estadística de los resultados se basa en el test de la diferencia de la ratio de información (Ledoit y Wolf, 2008), que es consistente en presencia de heterocedasticidad y autocorrelación. En la tabla 4 se recogen los resultados del citado test. Con el objetivo de facilitar la lectura de los resultados a través de las tablas 3 y 4, se reporta la diferencia en la ratio de información anualizado (es decir, multiplicando por $\sqrt{252}$ porque en IR = AV / SD el numerador se anualiza multiplicando por 252 y el denominador se anualiza multiplicando por $\sqrt{252}$). Sin embargo, para el test se han utilizado los retornos diarios de la cartera $\{r_{c,t}\}$ en el período de validación sin anualizar.

En términos generales, la cartera de Markowitz con señal de *momentum* ofrece una rentabilidad superior a la cartera de mínima varianza a expensas de una mayor volatilidad –como cabía esperar–. En términos de la ratio de información, ambas carteras ofrecen resultados comparables.

El mejor modelo por ratio de información es AFM-DCC-NLS. Como en el caso de la cartera de mínima varianza, estos resultados ponen en valor el uso de modelos de factores aunque los resultados no son estadísticamente significativos. Por otro lado, la regularización proporciona resultados significativamente mejores (económica y estadísticamente), y en este caso la regularización no lineal es significativamente superior a la lineal. El modelo VECH escalar ofrece resultados comparables desde el punto de vista estadístico, y ligeramente inferiores desde el punto de vista económico. El mismo comentario sobre la heterogeneidad que se hizo en la anterior subsección es de aplicación aquí. Finalmente, concluimos que, con independencia de la estrategia que utilicemos, todos los modelos GARCH multivariantes aquí considerados ofrecen resultados significativamente mucho mejores que los de una cartera que da el mismo peso a los activos que se encuentran en el quintil superior por *momentum*.

5. CONCLUSIONES

En este capítulo se han cubierto algunos de los últimos desarrollos en la literatura de modelización de covarianzas dinámicas de elevada dimensionalidad. Se ha hecho especial énfasis en el uso de técnicas computacionalmente escalables (verosimilitud compuesta, focalización en covarianzas) y en cómo introducir la estructura de factores de una manera sencilla. Hemos presentado los métodos de regularización de los valores propios de la matriz de covarianzas muestral y cómo pueden aplicarse de manera natural en el modelo DCC a través del método de focalización en covarianzas. Esto es de especial relevancia en contextos de elevada dimensionalidad, es decir, cuando la sección cruzada es de un tamaño comparable (o incluso superior) al número de observaciones.

Finalmente, a través de un ejercicio empírico con datos de acciones del índice S&P 500 se ha demostrado cómo la implementación de estos métodos permite construir carteras de inversión con mejores rentabilidades ajustadas por riesgo.

Ni que decir tiene que éste no es el primer trabajo en aplicar estimadores de covarianzas dinámicas de elevada dimensionalidad al problema de selección de carteras– y muy probablemente no será el último. Los resultados de este trabajo siguen la línea de los hallados por Engle, Ledoit y Wolf (2019) y De Nard, Ledoit y Wolf (2020), concluyendo que los modelos DCC factoriales conducen a carteras con menor volatilidad.

Cabe destacar que en la práctica los inversores deben tener en cuenta los costes de transacción para decidir qué estrategia seguir, y por lo tanto hay que encontrar el balance óptimo entre estrategias que aporten buenas rentabilidades ajustadas al riesgo y carteras que aporten un poco menos pero conlleven una menor rotación y con ello menores costes de transacción (véase, entre otros, Hautsch y Voigt, 2019)). En el reciente trabajo de Moura, Santos y Ruiz (2020) se hace una comparativa más amplia en la que también se incluyen modelos de volatilidad estocástica multivariante que resultan en carteras con mayor ratio de información después de costes de transacción. Los resultados de este trabajo también aportan evidencia de que los modelos DCC son los que mejor minimizan la volatilidad de la cartera. Sin embargo, según este trabajo, modelizar la matriz de covarianzas como un proceso Wishart con regularización hacia una matriz diagonal conduce a carteras más estables con menor rotación y mayor ratio de información una vez tenidos en cuenta los costes de transacción.

Referencias

Aguilar, M. (2009). A latent factor model of multivariate conditional heteroscedasticity. *Journal of Financial Econometrics,* 7(4), pp. 481–503.

Aielli, G. P. (2013). Dynamic conditional correlation: On properties and estimation. *Journal of Business and Economic Statistics,* 31.

Alessi, L., Barigozzi, M. y Capasso, M. (2009). Estimation and forecasting in large datasets with conditionally heteroskedastic dynamic common factors. *ECB Working Paper,* 1115.

Alexander, C. y Chibumba, A. (1996). Multivariate orthogonal factor garch. University of Sussex *Discussion Papers in Mathematics.*

Asai, M., McAleer, M. y Yu, J. (2006). Multivariate stochastic volatility: A review. *Econometric Reviews,* 25(2-3), pp. 145–175.

Baillie, R. y Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics,* 52(1-2), pp. 91–113.

Bampinas, G. y Panagiotidis, T. (2015). On the relationship between oil and gold before and after financial crisis: Linear, nonlinear and time-varying causality testing. *Working Paper series from Rimini Centre for Economic Analysis.*

Bauwens, L., Laurent, S. y Rombouts, J. (2006). Multivariate garch models: a survey. *Journal of Applied Econometrics,* 21, pp. 79–109.

Bickel, P. J. y Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics,* 36, pp. 2577–2604.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics,* 31, pp. 307–327.

Bollerslev, T., Engle, R. F. y Nelson, D. B. (1994). ARCH models. En: R. F. Engle y D. McFadden, editors, *Handbook of Econometrics,* pages 2959–3038. Elsevier.

Bollerslev, T., Engle, R. y Wooldridge, J. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy,* 96(1), pp. 116–31.

Bollerslev, T. y Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews,* 11(2), pp. 143–172.

Broto, C. y Ruiz, E. (2004). Estimation methods for stochastic volatility models: a survey. *Journal of Economic Surveys,* 18(5), pp. 613–649.

Brownlees, C. y Engle, R. (2017). SRISK: A Conditional Capital Shortfall Measure of Systemic Risk. *Review of Financial Studies,* 30(1), pp. 48–79.

Brownlees, C. y Llorens, J. (2020). Projected dynamic conditional correlations. *Disponible en SSRN.*

Conrad, C., Karanasos, M. y Zeng, N. (2010). The link between macroeconomic performance and variability in the uk. *Economics Letters,* 106(3), pp. 154–157.

DeMiguel, V., Garlappi, L. y Uppal, R. (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies,* 22(5), pp. 1915–1953.

De Almeida, D., Hotta, L. y Ruiz, E. (2018). Mgarch models: Trade-off between feasibility and flexibility. *International Journal of Forecasting,* 34(1), pp. 45–63.

De Nard, G., Ledoit, O. y Wolf, M. (2020). Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics (forthcoming).*

Diebold, F. y Nerlove, M. (1986). The dynamics of exchange rate volatility: a multivariate latent factor arch model. *Board of Governors of the Federal Reserve System (U.S.),* 205.

Ding, Z. y Engle, R. (2001). Large scale conditional covariance matrix modeling, estimation and testing. *Academia Economic Papers.*

Engle, R. (2002). Dynamic conditional correlation. *Journal of Business & Economic Statistics,* 20(3), pp. 339–350.

— (2009). *Anticipating Correlations: A New Paradigm for Risk Management.* Princeton University Press.

Engle, R. y Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics,* 24(2), pp. 238–253.

Engle, R. y Shephard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate garch. *Working Paper National Bureau of Economic Research,* 8554.

Engle, R. F., Ledoit, O. y Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics,* 37(2).

Fan, J., Liao, Y. y Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. *Journal of the Royal Statistical Society, Series B,* 75, pp. 603–680.

Ferreira, M. A. (2005). Evaluating interest rate covariance models within a value-at-risk framework. *Journal of Financial Econometrics,* 3(1), pp. 126–168.

Forni, M., Hallin, M., Lippi, M. y Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification and Estimation. *The Review of Economics and Statistics,* 82, pp. 540–554.

Forni, M., Hallin, M., Lippi, M. y Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics,* 185(2), pp. 359–371.

Friedman, J., Hastie, T. y Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics,* 9(3), pp. 432–441.

Hafner, C. M. y Herwartz, H. (2008). Testing for causality in variance using multivariate garch models. *Annales d'Economie et de Statistique,* 89, pp. 215–241.

Hafner, C. M. y Preminger, A. (2009). Asymptotic theory for a factor garch model. *Econometric Theory,* 25(2), pp. 336–363.

Hallin, M., Hotta, L. K., Mazzeu, J. H. G., Trucios-Maza, C. C., Pereira, P. L. V. y Zevallos, M. (2019). Forecasting conditional covariance matrices in high-dimensional time series: a general dynamic factor approach. W*orking Papers ECARES,* 2019-14.

Han, Y. (2006). Asset allocation with a high dimensional latent factor stochastic volatility model. *The Review of Financial Studies,* 19(1), pp. 237–271.

Harvey, A., Ruiz, E. y Sentana, E. (1992). Unobserved component time series models with Arch disturbances. *Journal of Econometrics,* 52(1-2), pp. 129–157.

Hastie, T., Tibshirani, R. y Friedman, J. (2008). *The Elements of Statistical Learning.* Springer Series in Statistics. New York, NY: Springer New York.

Haugen, R. A. y Baker, N. L. (1991). The efficient market inefficiency of capitalization–weighted stock portfolios. *The Journal of Portfolio Management,* 17(3), pp. 35–40.

Hautsch, N., Kyj, L. M. y Malec, P. (2015). Do high-frequency data improve high-dimensional portfolio allocations? *Journal of Applied Econometrics,* 30(2), pp. 263–290.

Hautsch, N. y Voigt, S. (2019). Large-scale portfolio allocation under transaction costs and model uncertainty. *Journal of Econometrics,* 212(1), pp. 221–240. Big Data in Dynamic Predictive Econometric Modeling.

Jegadeesh, N. y Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance,* 48(1), pp. 65–91.

Ledoit, O. y Wolf, M. (2004a). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management,* 30(4), pp. 110–119.

— (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis,* 88(2), pp. 365 – 411.

— (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance,* 15(5), pp. 850–859.

— (2011). Robust performances hypothesis testing with the variance. *Wilmott,* 2011(55), pp. 86–89.

— (2017). Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks. *The Review of Financial Studies,* 30(12), pp. 4349–4388.

— (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics (forthcoming).*

Markowitz, H. (1952). Portfolio selection. *Journal of Finance,* 7(1), pp. 77–91.

Moura, G. V., Santos, A. A. y Ruiz, E. (2020). Comparing high-dimensional conditional covariance matrices: Implications for portfolio selection. *Journal of Banking & Finance,* 118, 105882.

Pakel, C., Engle, R. F., Shephard, N. y Sheppard, K. (2017). Fitting vast dimensional time-varying covariance models. *NYU Working Paper,* No. FIN-08-009.

Patton, A. J. y Sheppard, K. (2009). *Evaluating Volatility and Correlation Forecasts,* pp. 801–838. Berlin: Springer Berlin Heidelberg, Berlin.

Peña, D. y Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics,* 119(2), pp. 291–321.

Pesaran, B. y Pesaran, H. (2007). Modelling volatilities and conditional correlations in futures markets with a multivariate t distribution. *Technical report, IEPR Working Paper,* No. 07.19.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation.* John Wiley & Sons, Inc.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory,* 13(3), pp. 341 – 360.

Shephard, N. (1996). *Statistical aspects of ARCH and stochastic volatility,* pages 1–67. Chapman & Hall, London, (edited by d.r. cox, david v. hinkley and ole e. barndorff-neilsen) edition. Reprinted in the Survey of Applied and Industrial Mathematics, issue on Financial and insurance mathematics, 3, 764-826, Scientific Publisher TVP, Moscow, 1996 (in Russian).

Silvennoinen, A. y Teräsvirta, T. (2009). *Multivariate GARCH Models,* pp. 201–229. Springer Berlin Heidelberg, Berlin, Heidelberg.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics,* 34(1), p. 1373.

Stock, J. H. y Watson, M. W. (2002a). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association,* 97, pp. 1167–1179.

— (2002b). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics,* 220, pp. 147–162.

Tse, L. D. Y. K. y Tsui, A. K. C. (2002). Evaluating the hedging performance of the constant-correlation garch model. *Applied Financial Economics,* 12, pp. 791–798.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

CAPÍTULO III

Árboles de decisión en economía: una aplicación a la determinación del precio de la vivienda

Máximo Camacho*
Salvador Ramallo
Manuel Ruiz Marín

En este trabajo describimos los árboles de decisión y repasamos su utilización en el análisis de datos económicos en el contexto de big data. De manera simplificada, realizamos un análisis comparativo de los algoritmos más utilizados en la construcción de árboles de decisión: CART, *bagging, random forests* y *boosting.* A modo de ilustración de estos procedimientos, examinamos cuáles son los determinantes del precio de la vivienda en Madrid en una muestra de 20.386 viviendas del portal Idealista en 2018. El fuerte componente no lineal de la relación del precio con sus determinantes es clave para destacar las ventajas de los árboles de decisión frente a propuestas más convencionales.

Palabras clave: CART, big data, precio de la vivienda.

* Los autores agradecen los comentarios de Fernando López y la financiación de los proyectos PID2019-107192 GB-I00 (AEI/10.13039/501100011033) y del Grupo de Excelencia de la Región de Murcia (Fundación Séneca) 19884/GERM/15. Los datos y programas de R que replican los principales resultados se encuentran en la web de los autores. Los errores cometidos son nuestra responsabilidad.

1. INTRODUCCIÓN

No se sabe con seguridad quién acuñó el termino *big data* por primera vez, ni siquiera si fue dentro o fuera del ámbito académico, aunque parece, en cualquier caso, que el mérito debería ser compartido (Diebold, 2012). Actualmente, el término recoge la esencia de un fenómeno de alto impacto en la ciencia y, en particular, en la economía y la empresa. Aunque no resulta fácil definir big data, las características de este tipo de análisis se pueden resumir en tres magnitudes: volumen, variedad y velocidad. Big data hace referencia a un gran volumen de datos, en numerosas ocasiones de naturaleza variada, organizados de manera estructurada o no estructurada, a veces disponibles (casi) en tiempo real y con una alta frecuencia, facilitado por los recientes avances en la tecnología de la información. Debido a estas características, el análisis estadístico y la visualización de los resultados obtenidos a partir de big data supone, en numerosas ocasiones, un reto.

Con el fin de superar este reto, se han propuesto en la literatura reciente numerosas técnicas de partición recursiva de datos encaminadas al tratamiento de big data. Estas técnicas se clasifican en dos grandes grupos: algoritmos supervisados y algoritmos no supervisados. Los algoritmos de aprendizaje supervisados son aquellos en los que tanto las variables de entrada como de salida están etiquetadas y son conocidas. Ejemplos de este tipo de algoritmos son la regresión lineal, regresión logística, maquinas de vectores de soporte, redes neuronales artificiales, k-vecinos mas cercanos y árboles de regresión y clasificación.

Por otro lado, la principal característica de los algoritmos de aprendizaje no supervisados es que pueden aprender a identificar procesos y patrones complejos sin que se les proporcione un etiquetado ni de las variables de entrada ni a las de salida. Están, por tanto, más orientados al aprendizaje de máquinas en *clustering* y reducción de la dimensionalidad. Ejemplos de algoritmos no supervisados son el análisis clustering por k-medias, inicialmente propuesto por MacQueen (1967), y el análisis de componentes principales.

En este capítulo nos centraremos en algoritmos supervisados basados en árboles de regresión y clasificación (*CART,* por sus siglas en inglés), popularizados tras la aparición de los algoritmos desarrollados por Breiman *et al.* (1984). Aunque no existe una herramienta de análisis de big data que sea superior a las demás en todos los escenarios, el comportamiento de los CART es bastante competitivo, y en algunos casos superior a otras alternativas como la regresión lineal, regresión logística, maquinas de vector soporte o redes neuronales artificiales (Kurt, Ture, y Kurum, 2008; Choubin *et al.,* 2018; Yang *et al.*, 2017).

Los CART se han utilizado en diversas ramas de la ciencia como en ecología (De'ath y Fabricius, 2000) para describir los tipos de hábitat de coral a partir de variables ambien-

tales; en medicina (Austin *et al.,* 2012) para predecir la mortalidad en pacientes con enfermedades cardiovasculares; en epidemiología (Austin, 2013) para clasificar y predecir enfermedades; en transporte (Ghasri, Rashidi, y Waller, 2017) para seleccionar modelos de transporte alternativos; o en psicología (Kitsantas, Moore y Sly, 2007) para estudiar los determinantes del consumo de tabaco.

Debido a la ingente cantidad de aportaciones en diversas ramas de la literatura (Loh, 2011; James *et al.,* 2013), en este trabajo introduciremos los CART poniendo especial atención a sus aplicaciones para la economía y la empresa. Aunque dedicaremos una sección completa a describir las aplicaciones más significativas de modelos CART en estas disciplinas, algunos ejemplos son la predicción de recesiones (Ng, 2014, Döpke, Fritsche y Pierdzioch, 2017); el abandono de clientes (Xie *et al.,* 2009); la predicción del fracaso empresarial (Gepp, Kumar y Bhattacharya, 2010); la detección de fraude financiero (Liu *et al.,* 2015) y la predicción de crisis financieras (Ward, 2017).

Los motivos por los que hemos seleccionado los CART para ilustrar el análisis de big data en economía y empresa son los mismos que están detrás del creciente éxito de esta metodología:

- Son una herramienta no paramétrica de regresión y clasificación de observaciones muy simple y flexible, permitiendo el uso de datos continuos y/o categóricos y bastante robusto a la presencia de datos atípicos. Por tanto, resultan una herramienta muy útil para capturar relaciones complejas entre las variables.

- Permiten operar con bases de datos muy grandes ya que funcionan con algoritmos muy eficientes desde el punto de vista computacional.

- Los resultados obtenidos con estas técnicas son muy fáciles de interpretar porque suelen ir acompañados de representaciones gráficas muy intuitivas que ayudan a discriminar cuáles son las variables explicativas más relevantes y cómo se relacionan con la variable dependiente.

En concreto, en este trabajo realizamos una introducción a los CART, centrándonos en datos de corte transversal aunque resultaría relativamente sencillo adaptarlos a datos de series temporales. Para ello, describiremos sus componentes y los métodos más habituales para su construcción y para la realización de predicciones económicas y empresariales. Acompañaremos la presentación de los CART con un ejemplo cuyo principal objetivo es esclarecer los determinantes del precio de la vivienda en Madrid en una muestra de 20.386 viviendas obtenidas a partir del portal Idealista en 2018. Como resultados más interesantes, obtenemos que el principal determinante del precio es la superficie de la vivienda, seguida del número de baños y habitaciones. La posición geográfica se encuentra también entre las variables explicativas más importantes. Como

ilustración de la capacidad de los CART para capturar relaciones complejas, mostraremos que la relación entre la posición geográfica y el precio de la vivienda es fuertemente no lineal. Por tanto, los métodos econométricos tradicionales serían incapaces de capturar la complejidad de esta relación.

El resto del artículo está estructurado como sigue. En la sección segunda, establecemos la notación y describimos los métodos de formación de un árbol de decisión. La sección tercera está destinada a desarrollar los métodos de *bagging, random forest* y *boosting* aplicados a controlar la inestabilidad de los árboles de decisión. En la cuarta sección repasamos algunos ejemplos de árboles de decisión aplicados al tratamiento de datos de economía y empresa. En la sección 5 concluimos.

2. ÁRBOLES DE REGRESIÓN Y CLASIFICACIÓN

Los árboles de decisión aparecen por primera vez en la literatura en el trabajo de Morgan y Sonquist (1963), aunque se popularizaron tras los algoritmos CART. Bajo estas siglas se encuentra un conjunto de algoritmos supervisados que tratan de modelizan una variable de respuesta a partir de un conjunto de p variables explicativas. Los árboles de decisión aplicados a variables de respuesta que toman valores numéricos con un orden intrínseco se denominan árboles de regresión. Cuando la variable de respuesta es cualitativa se aplican árboles de clasificación.

Un CART se entrena mediante un conjunto de datos que se usa para crear particiones recursivas, normalmente ortogonales, del espacio de definición del conjunto de variables explicativas. De esta manera las particiones son un conjunto de prismas dimensionales. Esto se consigue optimizando localmente en cada paso una función de pérdida cuya especificación depende de si el objetivo es la regresión o la clasificación.

En esta sección detallaremos la construcción del árbol de regresión y clasificación, definiremos la notación básica y los diferentes modelos de predicción. Aunque existen varios métodos para construir árboles de decisión, principalmente *CHAID (Chi-Square Automatic Interaction Detector), QUEST (Quick Unbiased Efficient Statistical Tree)* y CART, nosotros nos centraremos en los CART. Una comparativa de los algoritmos de construcción de CART más populares, como ID3, CART, C4.5 y See5/C5.0, se puede encontrar, por ejemplo, en Venkata y Kiruthika (2015)[1].

2.1. Construcción del árbol de regresión y clasificación

Con el fin de establecer la notación necesaria, consideremos una variable de respuesta Y que tratamos de modelizar a partir de un conjunto de p variables explicativas, $X = (X_1,...,X_p) \in \mathbb{R}^p$ El algoritmo CART se basa en encontrar las particiones del espacio

[1] Básicamente, los algoritmos se diferencian en las estrategias utilizadas para realizar las particiones.

de definición de X en un conjunto de M regiones terminales disjuntas $\{R_1,R_2,...,R_M\}$, que minimizan una determinada función de pérdida.

Para facilitar la descripción del algoritmo, consideremos un CART de particiones binarias con M nodos terminales, cada uno de los cuales $m \in \{1,...,M\}$ representa una rama del árbol que se caracteriza por las m^* escisiones que se producen en los nodos internos[2]. El espacio de parámetros que define cada rama lo denotaremos por $\theta_m^{m^*} = \left(v_{m,1}^{d_1}, s_{m,1}, ..., v_{m,m^*}^{d_{m^*}}, s_{m,m^*}\right)$, donde $X_{v_{m,i}}$ es la variable que determina la escisión i-ésima, con $i \in \{1,...,m^*\}$, del nodo terminal m, con $v_{m,i} \in \{1,...,p\}$. En el contexto de particiones ortogonales, para cada punto de escisión $s_{m,i}$, diremos que $d_i = 0$ si $X_{v_{m,i}} \leq s_{m,i}$ mientras que $d_i = 1$ si $X_{v_{m,i}} > s_{m,i}$. Por tanto, cada escisión i de una rama m dará lugar a dos subregiones, R_{mi}^0 cuando $d_i = 0$ y R_{mi}^1, cuando $d_i = 1$.

Como no es posible evaluar todas las particiones posibles, el objetivo de CART es proponer un mecanismo de partición recursiva en la que en cada paso se divide en dos subregiones una de las regiones creadas en el paso anterior. El criterio para realizar las particiones es encontrar aquella que minimiza una función de pérdida en el espacio de definición de las variables explicativas $\mathcal{X}$ y que dará lugar a M regiones finales disjuntas. La recursividad del algoritmo implica que el camino seguido por las particiones determinadas por la m-ésima rama es una sucesión de particiones θ_m^i tales que,

$$\theta_m^i = \theta_m^{i-1} \bigcup \left(v_{m,i}^{d_i}, s_{m,i}\right). \qquad [1]$$

con $i = 1,...,m^*$.

Supongamos que disponemos de una muestra de T observaciones, $\{(Y_1,\mathcal{X}_1),...,(Y_T,\mathcal{X}_T)\}_{t=1}^T$, donde Y_t es el valor de la variable dependiente o de respuesta y $\mathcal{X}_t$ es el vector de valores de las p variables explicativas para la observación t. Para cada escisión i que se produce en cada rama m definimos una función de predicción $f\left(\mathcal{X},\theta_m^i\right)$. Generalmente se trata de una medida relacionada con el valor que toma la variable dependiente condicionada a las regiones que se producen en la escisión. En árboles de regresión suele ser una medida de posición central y en árboles de clasificación la región se suele asociar al evento más probable. A la predicción para la subregión $R_{mi}^{d_i}$ de la rama m que se produce en la escisión i la llamaremos $\hat{Y}\left(R_{mi}^{d_i}\right)$.

Asociado a la función de predicción, en cada escisión de cada rama se define la función de pérdida,

$$L\left(Y, f\left(\mathcal{X},\theta_m^i\right)\right). \qquad [2]$$

[2] En vez de particiones binarias se podrían realizar particiones múltiples. Sin embargo, en la práctica no suelen funcionar bien porque particionan el espacio de definición demasiado rápido, dejando pocas observaciones para las nuevas particiones.

En árboles de regresión, la función de pérdida asociada a una escisión suele ser la suma del error cuadrático. Denotemos por $T_{mi}^{d_i}$ el número de observaciones que hay en la región $R_{mi}^{d_i}$. Si usamos la media como medida de posición central, podemos determinar la función de pérdida de cada escisión como:

$$L\left(Y, f\left(\mathcal{X}, \theta_m^i\right)\right) = \Sigma_{\mathcal{X}_t \in R_{mi}^0}\left(Y_t - \overline{Y}_{R_{mi}^0}\right)^2 + \Sigma_{\mathcal{X}_t \in R_{mi}^1}\left(Y_t - \overline{Y}_{R_{mi}^1}\right)^2 \quad [3]$$

donde, $\overline{Y}_{R_{mi}}^{d_i} = \frac{1}{T_{mi}^{d_i}} \Sigma_{\mathcal{X}_t \in R_{mi}^{d_i}} Y_t$ es la media de Y en la región $R_{mi}^{d_i}$ con $d_i \in .\ \{0,1\}$

En árboles de clasificación, la función de pérdida asociada a una escisión suele ser una medida del grado de impureza de las regiones resultantes. Supongamos que hay K clases distintas. Sea $I(\bullet)$ la función indicador que vale 1 si se cumple la condición dentro del paréntesis y 0 en caso contrario. Entre las diversas alternativas, vamos a elegir el índice de Gini para medir la impuridad o grado de impureza de la región m-ésima como:

$$IG\left(R_{mi}^{d_i}\right) = \Sigma_{k=1}^{K} P_k\left(R_{mi}^{d_i}\right)\left(1 - P_k\left(R_{mi}^{d_i}\right)\right), \quad [4]$$

donde,

$$P_k\left(R_{mi}^{d_i}\right) = \frac{1}{T_{mi}^{d_i}} \Sigma_{\mathcal{X}_t \in R_{mi}^{d_i}} I\left(Y_t = k\right),$$

es la proporción de observaciones de la región $R_{mi}^{d_i}$ que se clasifican como clase k de las K posibles. La función de pérdida de cada escisión se determina como la media ponderada de la impuridad de las regiones resultantes. Específicamente si $\mathcal{X} \in R_{mi}$

$$L\left(Y, f\left(\mathcal{X}, \theta_m^i\right)\right) = \frac{T_{mi}^0}{T_{mi}^0 + T_{mi}^1} IG\left(R_{mi}^0\right) + \frac{T_{mi}^1}{T_{mi}^0 + T_{mi}^1} IG\left(R_{mi}^1\right). \quad [5]$$

Si las regiones resultantes son muy puras, en el sentido de que contienen observaciones de una sola clase, el grado de impureza de esas regiones, medido por la expresión [4], será próximo a cero y, en ese caso, la función de pérdida también será próxima a cero. Cuanto mayor sea la mezcla de clases en las regiones, mayor será su grado de impureza y, por ende, mayor será el valor de la función de pérdida.

La descripción del algoritmo es muy sencilla. Al comienzo, todas las observaciones están en una única región que coincide con todo el espacio de definición de $\mathcal{X}$. Para realizar la primera escisión, seleccionaremos el indicador X_i y el punto de escisión s_i que den lugar a las dos regiones que minimicen la función de pérdida. Para cada una de esas dos regiones resultantes, volvemos a buscar el indicador y el punto de escisión que minimizan la función de pérdida, pero solo se realiza la escisión de la región que de lugar a una función de pérdida menor. El algoritmo se repite sucesivamente hasta que

se alcanza algún criterio que detiene al algoritmo. Algunas opciones habituales son que el número de observaciones que pertenecen a las regiones finales no sea inferior a un mínimo establecido por el usuario, que el árbol tenga un número máximo de regiones finales o que la reducción en la función de pérdida tenga que superar un mínimo para que tenga lugar una nueva escisión. En el apartado 2 de esta sección veremos una técnica más sofisticada.

Una vez que el algoritmo termina, la última etapa consiste en dar una predicción a las observaciones que se encuentran en cada una de las regiones terminales. En el caso de árboles de regresión, la predicción en la región m, será la media de las observaciones que pertenezcan a esa región, $\hat{Y}(R_m) = \overline{Y}_{R_m}$. Para los árboles de clasificación, la predicción en la región m será la clase k^* para la que se alcance el mayor valor de la proporción de observaciones de cada clase en esa región entre las $k=1,...,K$ clases posibles, $\hat{Y}(R_m) = P_{k^*}(R_m)$.

Para conocer la verdadera capacidad de predicción del árbol de decisión debemos examinar hasta qué punto es capaz de formar inferencia adecuada de las observaciones que no se han usado para generar el árbol. Como los árboles se encuentran dentro de lo que conocemos como métodos de aprendizaje supervisado basado en la experiencia pasada, resulta útil dividir la muestra en dos submuestras. La primera submuestra es la de entrenamiento o training, en la que se estima el CART. La segunda es la de evaluación o test, en la que se examina la habilidad del CART para formar inferencia de las observaciones que no se han usado para estimarlo[3].

Una vez que se ha estimado el árbol de decisión, una herramienta muy útil para interpretar el resultado obtenido es la importancia relativa propuesta por Breiman *et al.* (1983). La importancia relativa del indicador X_i entre los $(X_1,...,X_p)$ indicadores existentes viene determinada por el número de veces que ese indicador se ha utilizado para realizar las particiones ponderada por la reducción en la función de pérdida (error cuadrático o índice de Gini) que proporcionan las particiones en las que el indicador participa.

Supongamos que en un árbol se hacen un total de M^* subparticiones (una por cada nodo del árbol) y que el indicador que se usa para hacer la partición m-ésima es X_{v_m}, con $v_m \in \{1,...,p\}$ y $m \in \{1,...,M^*\}$. Si llamamos ∇L_m a la reducción de la función de pérdida que se produce en esa partición, la importancia relativa del indicador X_i en la construcción del árbol de decisión:

$$IR_i = \Sigma_{m=1}^{M^*} \nabla L_m I\left(X_{v_m} = X_i\right). \qquad [6]$$

[3] Sin pérdida de generalidad, describiremos las técnicas suponiendo que la submuestra de entrenamiento es la muestra total. La distinción entre submuestras de entrenamiento y evaluación se hará cuando se examine la habilidad predictiva de las distintas técnicas empleadas.

donde $i=1,...p$. Para facilitar la interpretación, el indicador se suele normalizar para que las importancias relativas de los p indicadores sumen 100[4].

Una segunda herramienta habitual en el contexto de árboles de decisión es el Gráfico de Dependencia Parcial, que mide el efecto de una variable explicativa sobre la variable dependiente. Sin pérdida de generalidad, supongamos que queremos medir el efecto *céteris páribus* de la primera variable explicativa X_1 sobre la predicción del árbol de decisión. El primer paso para calcular la dependencia parcial consiste en construir el árbol de decisión usando las técnicas que hemos descrito con anterioridad para la base de datos original. En el segundo paso, generamos un conjunto de posibles valores para X_1, que llamaremos $X_{1i} \in \{X_{11},...,X_{1N}\}$. Llamemos al conjunto formado por cada valor generado de X_1 y el resto de valores de las explicativas $\mathcal{X}(1i) = (X_{1i},X_2,...,X_p)$. Para cada uno de los elementos de este conjunto haremos una predicción con el árbol estimado y calcularemos la media:

$$\hat{Y}^*\left(X_{1i}\right)=\frac{1}{N}\Sigma^{N}_{i=1;\mathcal{X}(1i)\in R_k}\hat{Y}\left(R_k\right). \qquad [7]$$

El gráfico de $\left\{X_{1i},\hat{Y}^*\left(X_{1i}\right)\right\}$, para $i=1,...,N$, es el Gráfico de Dependencia Parcial de la variable explicativa X_1.

Se puede generalizar el Gráfico de Dependencia Parcial para cualquier subconjunto de variables explicativas incluidas en X. Aunque el efecto sobre la variable dependiente de combinaciones complejas de las variables explicativas va a ser difícil de interpretar, la opción que más se utiliza en la literatura es la que analiza el efecto de la combinación de dos variables explicativas, dando lugar a los gráficos que miden el efecto interacción. Sin pérdida de generalidad, supongamos que estamos interesados en medir el efecto conjunto de las dos primeras variables explicativas, X_1 y X_2, sobre la variable dependiente Y. Siguiendo el razonamiento de los Gráficos de Dependencia Parcial, el efecto parcial se aproxima generando pares de valores de las dos primeras variables explicativas, (X_{1i},X_{2i}), con $i=1,...,N$; para cada valor de i, predecimos el valor de la variable dependiente usando los valores reales del resto de variables explicativas y calculamos la media, $\hat{Y}^*_{\tau}\left(X_{1i},X_{2i}\right)$, siguiendo la lógica propuesta en la expresión [7]; por último, representamos gráficamente $\left\{X_{1i},X_{2i},\hat{Y}^*\left(X_{1i},X_{2i}\right)\right\}$, para $i=1,...,N$.

2.1.1. *Prunning*

La estimación del árbol de decisión con el algoritmo anterior sufre el riesgo de sobreajuste. Supongamos que el proceso generador de datos es $Y = f(\mathcal{X},\Theta_M) + u$. Como la estimación que realizamos con el árbol depende de sus regiones terminales y éstas de

[4] Otras veces se hace 100 la importancia relativa del indicador más importante y se relativizan las importancias relativas del resto de indicadores respecto al más importante.

los parámetros estimados $\hat{\Theta}_M$, podemos llamar al estimador resultante del árbol de decisión $\hat{f} = f\left(\mathcal{X}, \hat{\Theta}_M\right)$. La descomposición que viene determinada por el Error Cuadrático Medio (ECM) para ese estimador implica:

$$ECM = \left(E\hat{f} - f\right) + var\left(\hat{f}\right) + var\left(u\right) \qquad [8]$$

donde $f = f\left(\mathcal{X}, \Theta_M\right)$. Si obviamos el último componente, referido a la varianza irreducible del error, la estimación $\hat{f}$ puede modificar el sesgo y la varianza del árbol de decisión. Cuanto más complejo sea el árbol de decisión, medido como el número de regiones finales M, menor será el sesgo pero mayor será la varianza y esto provoca falta de robustez en los resultados provocando fallos en las predicciones. La poda de un árbol de decisión consiste en encontrar un árbol con un número de regiones óptimo que equilibre la balanza entre el sesgo y la varianza.

Sea $f = f\left(\mathcal{X}, \Theta_M\right)$ la función de pérdida calculada a partir de las M regiones del árbol. Para un árbol de regresión, la pérdida estimada será:

$$\Sigma_{m=1}^{M} \Sigma_{\mathcal{X}_t \in R_m} \left(Y_t - \hat{Y}\left(R_m\right)\right)^2. \qquad [9]$$

Si definimos $Q(R_m)$ como la proporción de observaciones de la región m sobre el total de observaciones, para un árbol de clasificación la pérdida será:

$$\Sigma_{m=1}^{M} Q\left(R_m\right)\left(1 - \Sigma_{k=1}^{K} P_k\left(R_m\right)^2\right), \qquad [10]$$

que se interpreta como la suma ponderada de la impureza de las regiones terminales.

Dado un valor de α, la técnica consiste en comenzar a generar el árbol con M_1 regiones. Seguidamente, aumentamos las escisiones sucesivamente incrementando el valor de las regiones terminales hasta que lleguemos a un árbol donde no se puedan encontrar más escisiones sin que las M_N regiones resultantes tengan menos observaciones que un número fijado. Para el valor dado de α, calculamos el valor de la expresión:

$$PR_{M_i}\left(\alpha\right) = L\left(Y, f\left(\mathcal{X}, \Theta_{M_i}\right)\right) + \alpha M_i \qquad [11]$$

para cada valor de M_i. De la colección de valores $\{PR_{M_1}(\alpha), \ldots, \{PR_{M_N}(\alpha)\}$, elegiremos el árbol con un número de regiones M para el que [11] alcance el valor más pequeño. A medida que el árbol es más complejo, la función de pérdida que aparece en el primer sumando tiende a caer. Sin embargo, el incremento en el número de regiones hace que el segundo sumando crezca en proporción al valor de α. Si α es cero el tamaño del árbol será muy grande y crecerá la varianza, mientras que si α es demasiado alto el árbol

resultante podría ser demasiado simple como para garantizar una mínima bondad del ajuste a los datos.

Para determinar el tamaño óptimo de que controla el *trade-off* entre el ajuste a los datos y la complejidad del árbol resultante, se suelen utilizar técnicas de validación cruzada. El primer paso de estas técnicas consiste en dividir la muestra en J submuestras seleccionadas aleatoriamente. Usaremos todas las muestras menos una, que llamaremos submuestra j, para estimar los árboles y dejaremos la submuestra j para evaluarlos. En segundo lugar fijamos un valor de α pequeño, que llamaremos α_1, y para ese valor se estiman árboles para distintos valores de M_i que se evalúan en la submuestra j. Para cada submuestra se elige el número de regiones que minimiza [11] evaluado en la muestra j. Seguidamente para cada una de las j - 1 submuestras se calcula la media $\overline{PR}(\alpha_1)$. Finalmente, se repite esta operación incrementando en cada paso el valor de α hasta un valor máximo α_L. El valor de α elegido será el valor,

$$\alpha^* = \arg\min_{\alpha_i} \left\{ \overline{PR}(\alpha_1), \ldots, \overline{PR}(\alpha_L) \right\} \qquad [12]$$

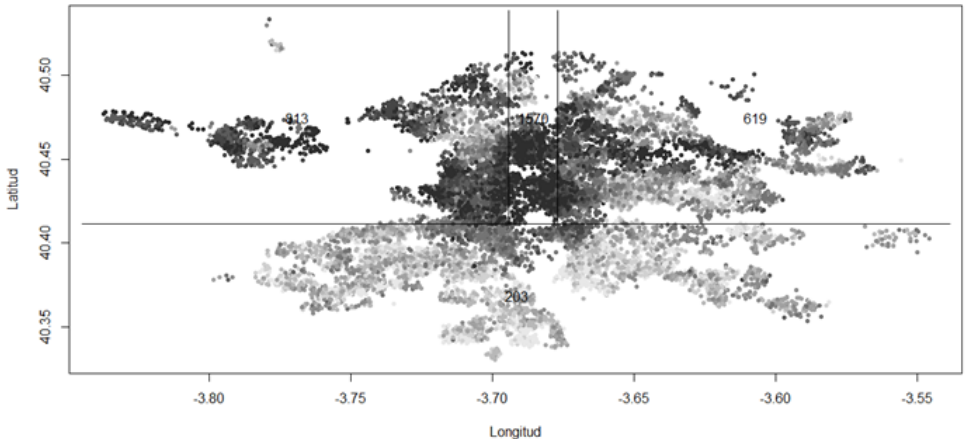
que minimiza la media.

2.2. Ejemplo de árbol de decisión

Para ilustrar el mecanismo de construcción de los árboles de decisión, vamos a usar una base de datos de precios de viviendas (en miles de euros) en el municipio de Madrid

FIGURA 1

PRECIO DE LA VIVIENDA Y PARTICIÓN



Fuente: Elaboración propia.

extraídos a partir de la base de datos del portal Idealista referido a las 20.385 viviendas que se anunciaron en el portal en enero de 2018[5]. La figura 1 muestra la localización en un mapa de los precios de la vivienda teniendo en cuenta su posición en términos de latitud y longitud de sus coordenadas geográficas. Para facilitar la inspección visual de los datos, cada vivienda aparece en el plano con un color más oscuro cuanto mayor es su precio. En la zona oeste no se venden viviendas en la zona central del distrito de Moncloa-Aravaca, donde se encuentra la Casa de Campo y en la zona norte donde se sitúan los montes del distrito Fuencarral-El Pardo.

Parece evidente que la localización geográfica tiene importancia en la determinación de los precios. En concreto, el mapa muestra que los precios más baratos se localizan en la latitud sur de Madrid donde están los distritos de la Latina, Carabanchel, Usera, Villaverde, Puente de Vallecas, Villa de Vallecas y Vicálvaro. Sin embargo, el precio de la vivienda se encarece para las viviendas que están en el centro, y en menor medida en el norte. Respecto a la zona centro-norte, los precios más altos están en la longitud central, donde se encuentran los distritos Centro, Retiro, Salamanca y Chamartín. En segundo lugar, aparecen las viviendas situadas en el oeste, en los distritos de Chamberí y algunas urbanizaciones de Fuencarral-El Pardo y de Moncloa-Aravaca. En tercer lugar, la zona situada más al este de la zona norte donde se encuentran Ciudad Lineal, Hortaleza y Barajas. Estas diferencias por longitud no están tan acentuadas en la zona sur.

El objetivo de un árbol de decisión donde las explicativas sean la latitud y la longitud consiste en particionar recursivamente este mapa geográfico dando lugar unas regiones para las que se minimice el error cuadrático entre el precio de cada vivienda y el precio medio de la región resultante de la partición a la que cada vivienda pertenezca. Con un fin únicamente ilustrativo, vamos a estimar un árbol de decisión con un máximo de tres particiones. Para visualizar el resultado, las cuatro regiones resultantes se superponen en el mapa que se muestra en la figura 1, junto con el precio medio de las viviendas localizadas en esas regiones. El precio estimado para una vivienda situada al sur es de unos 200.000 euros. Entre las viviendas del centro-norte, las que que se localizan en el este son las más baratas, con un precio esperado de unos 620.000 euros. Para las que se sitúan en el oeste estimamos un precio de unos 810.000 euros. Para las que se sitúan en el centro, estimamos un precio de un 1.600.000 euros aproximadamente.

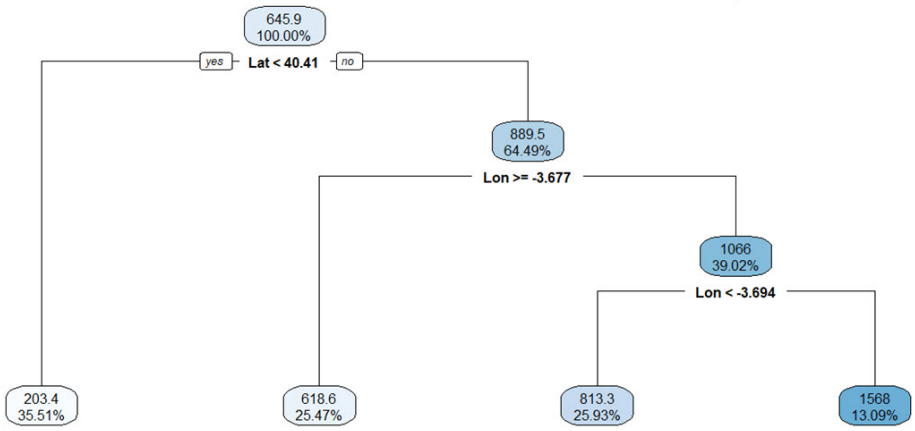
La figura 2 muestra el algoritmo recursivo que se ha utilizado para realizar las particiones. Al comienzo del algoritmo, todas las viviendas están situadas en una única región, para las que el precio medio es de 646.000 euros. El árbol comienza con la latitud y para un conjunto de puntos de escisión calcula la suma al cuadrado de la diferencia entre los precios de la vivienda y la media de las dos regiones resultantes a las que pertenecen las viviendas, tal y como aparece en la ecuación [10]. Acto seguido, realiza la

[5] Fan, Ong y Koh (2006) estiman un árbol de decisión mediante el algoritmo CART para analizar los determinantes del precio de la vivienda en Singapore. Aunque se fijan en características de la localización, como cercanía de colegios, no tienen en cuenta las coordenadas geográficas.

misma operación para longitud y un conjunto de puntos de escisión posibles para esa variable. Finalmente, elige la pareja de variable (latitud o longitud) y punto de escisión que minimizan [10]. En este caso, el mínimo lo ha encontrado para la variable latitud y con punto de escisión 40,41. Como se observa en figura 2, la región que se encuentra más al sur de esa latitud (pare izquierda de la figura) no se vuelve a particionar y se estima un precio medio para esas viviendas, que suponen un 35 % del total, de unos 200.000 euros.

FIGURA 2

ÁRBOL DE DECISIÓN



Fuente: Elaboración propia.

La región que queda al norte de la latitud 40,41 se vuelve a particionar de la misma manera que hemos descrito anteriormente, siendo ahora la longitud la variable elegida para realizar la partición y el punto de escisión que minimiza [10] de -3,677. La región que se encuentra al este este de esa longitud ya no se vuelve a particionar y el precio que se estima para las viviendas de esa región es de unos 620.000 euros. Sin embargo, la región que se encuentra al oeste de la longitud -3,677 se particiona de nuevo usando a la longitud como variable y el punto de escisión de -3,694. Las dos regiones son la que se encuentran al oeste de esa longitud para la que se estima un precio de la vivienda de unos 810.000 euros, y la región central (al este de -3,694 pero al oeste de -3,677) para la que se estima un precio de la vivienda de casi 1.600.000 euros.

Aunque didáctico, el ejemplo que hemos descrito hasta ahora es demasiado poco ambicioso. Como veremos más adelante, ni siquiera podemos asegurar que el efecto *ceteris paribus* de la latitud y la longitud sea el que hemos detectado en el ejercicio anterior porque el modelo es tan sencillo que puede esconder relaciones más complejas entre el precio de la vivienda y su posición geográfica cuando se añadan otras variables explicativas. El análisis de big data con árboles de decisión puede ser mucho más útil

con modelos más completos, como vamos a ilustrar en el análisis más completo del precio de la vivienda que proponemos.

Supongamos que queremos vender o comprar una vivienda en Madrid pero no sabemos cuál sería el precio adecuado al que deberíamos realizar la transacción. Las coordenadas geográficas, aunque potencialmente importantes, no parecen ser la única característica que tienen en cuenta compradores y vendedores a la hora de determinar los precios de sus transacciones. Generalmente, además de la posición geográfica, el mercado de la vivienda toma en cuenta una serie de características de la vivienda tales como superficie, planta, tipo (casa, piso, ático, estudio, chalet o dúplex), número de habitaciones y baños, si se encuentra en buen estado, si es exterior, o si la vivienda dispone de terraza, ascensor, aire acondicionado, garaje, piscina, trastero, armarios empotrados, o jardín.

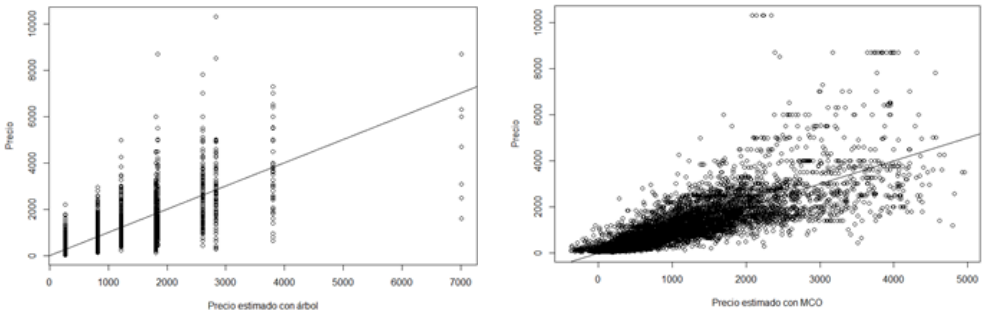
Otra ventaja de los árboles de decisión en el uso de grandes bases de datos es la sencillez con la que un usuario poco familiarizado con técnicas estadísticas sofisticadas puede utilizar el resultado del modelo para tomar decisiones económicas. Usando las técnicas de pruning que se han descrito anteriormente, el árbol que minimiza la expresión [11] es el que tiene nueve regiones. Un usuario sólo tiene que conocer el valor de las variables explicativas de la vivienda que quiere vender o comprar para que el árbol localice en qué región se encuentra la combinación de las explicativas y proponer como precio de la vivienda la media de las viviendas que se quieren comprar y vender en esa región. Por ejemplo, la vivienda de la última observación de la muestra se trata de un piso exterior de una habitación en segunda planta con ascensor y un baño, con unas coordenadas de longitud -3,702 y latitud 40,452, de 68 metros cuadrados, que no necesita reforma y no tiene jardín, trastero, piscina, aire acondicionado ni terraza. El precio propuesto por el árbol de decisión para esa vivienda es de 263.000 euros, algo más barato del precio en el que aparece en Idealista de 280.000 euros.

La predicción del árbol de decisión en cada una de las nueve regiones para todas las viviendas de la muestra, junto con el precio que se ofrecen en la plataforma aparece en el gráfico izquierdo de la figura 3. Para facilitar la interpretación, se ha incluido una línea que corta al eje y tiene pendiente unitaria. Esta figura permite apreciar cómo aumenta la heterogeneidad en el precio de las viviendas en las distintas regiones que se forman en el árbol de decisión a medida que aumenta el precio medio de la vivienda predicho para esas regiones. Como comparativa, en el gráfico derecho de la figura 3 aparece la estimación realizada por el modelo lineal estimado por mínimos cuadrados ordinarios. Como resumen de la bondad del ajuste de los modelos, el pseudo R^2, medido como uno menos la proporción del ECM sobre la varianza del precio alcanza un valor de 0,73 para el árbol de decisión y de 0,72 para MCO.

Un problema importante de la construcción de los árboles de decisión que hemos descrito es el de la falta de estabilidad ante pequeños cambios en el proceso de construcción del árbol. Aunque vamos a analizar este problema con detalle en la próxima

FIGURA 3

PREDICCIÓN DEL PRECIO DE LA VIVIENDA

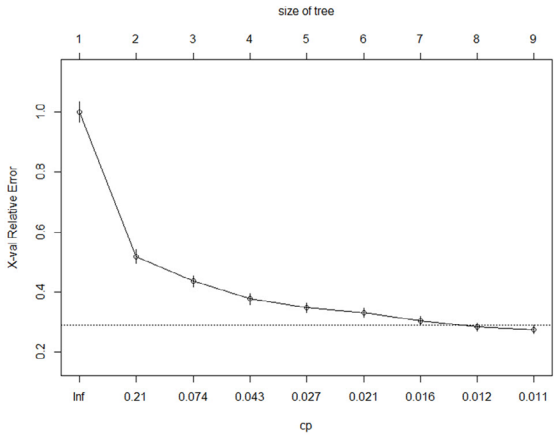


Fuente: Elaboración propia.

sección, podemos ilustrar uno de los efectos de la inestabilidad analizando el resultado de la validación cruzada que se ha utilizado para seleccionar un árbol con nueve regiones. En la figura 4 se muestra el valor de la expresión [11] para cada valor de α (escala inferior del gráfico) y del número de regiones que se eligen para cada α[6] Si elegimos el número de regiones como el valor que minimiza la expresión [11], el árbol debería tener nueve regiones.

FIGURA 4

VALIDACIÓN CRUZADA



Fuente: Elaboración propia.

[6] Para facilitar la interpretación, el valor de [11] se ha relativizado respecto del que tomaría el árbol sin particiones.

Sin embargo, Breiman *et al.* (1984) sugieren que se debería elegir no el árbol que minimice [11] sino el que tenga el menor número de regiones necesarias para las que se alcance el umbral de una desviación típica por encima del mínimo. Siguiendo este método alternativo, el árbol que cruza el umbral que aparece como una línea horizontal discontinua en el gráfico es el que tiene ocho regiones. Si flexibilizamos un poco el umbral, también serían defendibles árboles con un número de regiones comprendido entre tres y siete. Sin embargo, es evidente que modificar el número de regiones cambiaría drásticamente las predicciones del árbol, como se ilustra en la figura 3.

Además, el precio que se paga por la simpleza del proceso de construcción de árboles de decisión que hemos descrito hasta ahora es la inestabilidad asociada a que un error cometido en las primeras escisiones de una rama provoca un efecto dominó en las ramas inferiores que amplifica el efecto del error en el árbol resultante. La inestabilidad da lugar a árboles poco robustos puesto que pequeños cambios en los datos producen grandes cambios en la estimación del árbol de decisión, lo que reduce la capacidad de interpretación de los resultados y dificulta su uso para realizar predicciones de la variable dependiente.

3. INESTABILIDAD Y TRATAMIENTO

El control de la inestabilidad debe ir encaminado a proporcionar métodos de estimación de árboles de decisión menos condicionados a pequeños cambios en los datos manteniendo controlado el sesgo de los árboles resultantes. La racionalidad que hay detrás de estos procedimientos es la de construir el árbol de decisión a partir de la combinación de otros árboles más sencillos estimados con los datos disponibles. En este contexto, los métodos más utilizados son los que se denominan bagging y random forest.

La principal diferencia entre estos dos métodos se encuentra en la forma en la que se diseñan los árboles sencillos a partir de los que se construyen los árboles complejos. En bagging (Breiman, 1996), y su versión mejorada random forest (Breiman, 2001), la construcción de los árboles sencillos se realiza en paralelo a partir de boostraps de la muestra de datos original. Sin embargo, en boosting (Freund y Schapire, 1997), la construcción de los árboles resultantes es secuencial, acumulando la aportación de árboles sencillos de tal manera que cada nuevo árbol pone más peso en las observaciones que han sido mal estimadas por el acumulado anterior.

Usando la descomposición del ECM propuesta en la expresión [8], en bagging, se emplean árboles más complicados, con muy poco sesgo pero mucha varianza. Sin embargo, agregando muchos de estos árboles se consigue reducir la varianza sin apenas introducir sesgo. Por el contrario, en boosting, se emplean agregaciones de árboles muy sencillos (en ocasiones, con una sola escisión) y muy poca varianza pero mucho sesgo. Sin embargo, el sesgo cometido por la agregación anterior se va ajustando secuencialmente con la aportación de las nuevas particiones.

3.1. Bagging y random forest

La primera contribución en este tipo de algoritmos se debe a Breiman (1996), quien propuso la metodología de árboles de decisión bootstrap agregados, conocida como bagging. Este procedimiento funciona entrenando en paralelo un gran número, B, de árboles de decisión formando un "bosque", en lugar de ajustar un único árbol. Cada uno de estos árboles tiene poco sesgo y mucha varianza. En cada nueva predicción, todos los árboles que forman el "bosque" participan aportando su predicción. Como valor final, se toma la media de todas las predicciones (variables continuas) o la clase más frecuente (variables cualitativas).

En el caso particular de los árboles, la técnica de bagging ha demostrado incrementar en gran medida la precisión de las predicciones. El algoritmo se puede resumir en los siguientes pasos:

- Generar B conjuntos de entrenamiento mediante bootstrapping con reemplazamiento a partir de la muestra original.

- Entrenar un árbol con cada una de las B muestras del paso 1. Cada árbol se crea sin apenas restricciones y no se somete a pruning, por lo que tiene varianza alta pero poco sesgo.

- Para cada nueva observación, obtener la predicción de cada uno de los B árboles. El valor final de la predicción se obtiene como la media de las B predicciones en el caso de árboles de regresión y como la clase predicha más frecuente (moda) para árboles de clasificación.

En el proceso de bagging, el número de árboles creados no es un hiperparámetro crítico en cuanto a que, por mucho que se incremente el número, no se aumenta el riesgo de sobreajuste. A pesar de ello, desde un punto de vista computacional no conviene almacenar el resultado de un gran número de árboles si, a partir de un número determinado, añadir nuevos árboles no aporta reducciones significativas del ECM. Una opción es obtener el número de árboles que optimiza el ECM usando técnicas de validación cruzada, aunque esta opción suele ser computacionalmente muy intensiva.

Una manera directa de estimar el error de un modelo al que se aplica bagging sin necesidad de aplicar la validación cruzada es mediante el método *out-of-bag* (OOB), propuesto originariamente por Breiman (1996). En concreto, el método consiste en entrenar cada árbol generado en el bagging en aproximadamente de las observaciones, mientras que el restante, conocidas como OOB, se usa para evaluar el árbol ajustado. Esta técnica es similar al *leave-one-out,* dando lugar a aproximadamente $B/3$ predicciones. La predicción OOB consiste en promediar los valores de los $B/3$ que no se

han considerado en su entrenamiento en el caso de árboles de regresión, o en considerar la moda en el caso de árboles de clasificación.

Con un número suficientemente elevado de árboles (B) el error OOB puede llegar a ser equivalente al error de validación leave-one-out. Además, el método basado en OOB para estimar el error resulta conveniente cuando se aplica bagging en grandes conjuntos de datos, para los que aplicar la validación cruzada sería computacionalmente muy costoso.

En el caso particular de los árboles de decisión, la técnica de bagging ha demostrado incrementar la precisión de las predicciones realizadas con los árboles resultantes de los algoritmos CART. Sin embargo, con esta técnica se pierde en interpretabilidad (representación gráfica) y se hace más difícil identificar la importancia de cada variable en la estimación del modelo. Una manera de determinar las variables que tienen mayor importancia es identificar aquel predictor con menor varianza residual total, en caso de árboles de regresión o índice de Gini para árboles de clasificación, como resultado de las divisiones sobre cada predictor, promediado sobre todos los B árboles. Un predictor importante será el que consiga una reducción promedio mayor.

Tal y como hemos expuesto con anterioridad el proceso de bagging se basa en el hecho de que, promediando un conjunto de modelos con bajo sesgo, se consigue reducir la varianza y aumentar la precisión de la estimación, manteniendo el sesgo. Pero esta reducción de la varianza depende del grado de correlación de los árboles que constituyen el bagging. Si la correlación es alta, la reducción de varianza que se puede lograr es pequeña. Por ejemplo, cuando un predictor es muy importante o influyente casi todos los árboles generados por bagging usarán este predictor en la primera ramificación, de manera que los árboles generados acabarán siendo similares unos a otros, sus predicciones estarán altamente correlacionadas, y su promedio no producirá una reducción sustancial de la varianza con respecto a la de un solo árbol.

Para solucionar este problema, se han propuesto variaciones al procedimiento bagging que tienen como objetivo reducir la correlacion, sin necesidad de aumentar substancialmente el sesgo. La variación del bagging más conocida es el random forest, originariamente desarrollado en Breiman (2001). Como en el caso de bagging, en random forest se entrenan los árboles en una muestra bootstrap. Sin embargo, en vez de seleccionar el conjunto completo de p predictores para realizar las escisiones en los nodos del árbol, se selecciona un subconjunto aleatorio de $q < p$ de estos predictores. Este proceso se repite para cada uno de los B árboles individuales de manera que el predictor random forest, al igual que en el caso de Bagging, se obtiene promediando las B predicciones individuales para árboles de regresión o seleccionando la clase más frecuente para árboles de clasificación.

3.2. Boosting

Aplicados a árboles de decisión, boosting es una idea relativamente simple. Consiste en la agregación secuencial de árboles sencillos, conocidos en la literatura relacionada como *weak learners,* asignando en cada iteración un mayor peso a los datos con mayor error en la estimación realizada mediante la agregación de weak learners hasta la iteración anterior. El resultado final, conocido como *strong learner,* es una agregación de árboles como una suma ponderada para el caso de árboles de regresión o como una mayoría cualificada en el caso de árboles de clasificación. El control de la varianza está asegurado usando como weak learners a árboles muy sencillos en cada iteración, mientras que el sesgo se controla con el mecanismo de ajuste a los datos mal estimados hasta la iteración anterior.

La idea que hay detrás de las técnicas de boosting fue originariamente diseñada para métodos de clasificación binarios en el influyente trabajo de Schapire (1990). El autor demuestra que el strong learner puede clasificar mejor que un weak learner si añadimos secuencialmente dos weak learners adicionales aplicados a muestras filtradas para que se enfoquen en las observaciones más difíciles de clasificar por los weak learners anteriores[7]. Freund (1995) extiende este algoritmo a clasificaciones no necesariamente binarias.

Freund y Schapire (1996) contribuyeron a popularizar el boosting como algoritmo de clasificación al proponer unas reglas muy simples para clasificar a partir del algoritmo conocido como *Adaptive Boosting (Adaboost).* Cuando el weak learner es un árbol de clasificación, el input del algoritmo son las T observaciones, $\{(Y_1,\boldsymbol{\mathcal{X}}_T),...,(Y_1,\boldsymbol{\mathcal{X}}_T)\}$, donde $Y_t \in \{1,...,K\}$ hace referencia a la clase que toma la variable dependiente en t. Al comienzo del algoritmo, todas las observaciones tienen el mismo peso $D_1(t) = 1/T$ para todo t. Para simplificar la exposición, presentaremos la versión del algoritmo Discrete Adaboost que propone Ng (2014) para una clasificación binaria donde $Y_t \in \{-1,1\}$.

El algoritmo *Discrete Adaboost* consiste en iterar J veces los siguientes tres pasos:

1. Aplicamos el árbol de decisión a las observaciones iniciales ponderadas por los pesos, de tal forma que se minimice el error $\varepsilon_j = \Sigma_{t=1,\mathcal{X}_t \in R^t}^{T} w_j(t) Y_t \hat{Y}^j(R^t)$, donde $\hat{Y}^j(R^t)$ es la estimación en el paso j-ésimo en la región R_t con $\boldsymbol{\mathcal{X}}_t \in R_m^t$. El algoritmo solo continua si $\varepsilon_j > 0{,}5$.

2. Calculamos la ponderación de la clasificación de la iteración j,$\alpha_j = 0{,}5\log((1-\varepsilon_j)/\varepsilon_j)$.

3. Recalculamos los pesos para la siguiente iteración $w_{j+1}(t) = w_j(t) exp\left(-\alpha_j Y_t \hat{Y}^j(R^t)\right)$. Los pesos son normalizados para que sumen 1.

[7] En árboles de clasificación, es frecuente usar como *wear learner* los stumps, que suponen una única partición a partir de una única variable explicativa.

La clasificación final es $\hat{Y}^J\left(R^t\right)=sign\left(\widehat{SY}^J\left(R^t\right)\right)$, donde $\widehat{SY}^J\left(R^t\right)=\Sigma_{j=1}^J\alpha_j\hat{Y}^j\left(R^t\right)$, con $sign(v) =$ si $v >0$ y $sign(v) =$ -1 si $v < 0$[8]. Por tanto, podemos interpretar que la clasificación de una observación $\boldsymbol{x}_t$ se determinará según lo que decida la mayoría cualificada de clasificaciones en las iteraciones, con más peso aquellas iteraciones con menores errores.

La propuesta de Friedman *et al.* (2000) supone un avance significativo en los árboles de decisión estimados con boosting porque se establecen las bases estadísticas que permiten entender el buen ajuste de los algoritmos *Adaboost* a los datos y generalizar los algoritmos para incluir diversos escenarios estadísticos. Estos autores demuestran que los algoritmos se pueden interpretar como un problema de optimización de una función de pérdida y, usando funciones aditivas, se pueden resolver mediante algoritmos de Newton.

Por ejemplo, el algoritmo *Discrete Adaboost* se puede interpretar como la búsqueda del árbol de clasificación que minimiza la función de pérdida exponencial $E\left(exp\left(-Y_t\widehat{SY}^J\left(R^t\right)\right)\right)$. El clasificador que minimiza esta expresión es:

$$\widehat{SY}^J\left(R^t\right)=0.5\log\frac{p\left(Y_t=1\right)}{p\left(Y_t=-1\right)}, \qquad [13]$$

que coincide con la mitad del *odds ratio* de las probabilidades. Por similitud con modelos logísticos, el algoritmo Adaboost consiste en estimar el clasificador como una aproximación no paramétrica basada en modelos de regresión aditiva $\widehat{SY}^J\left(R^t\right)=\Sigma_{j=1}^J\alpha_j\hat{Y}^j\left(R^t\right)$. Friedman *et al.* (2000) muestran que el clasificador óptimo es el que minimiza el error ponderado del algoritmo Adaboost. Además, las ponderaciones que minimizan la función de pérdida exponencial y la ponderación óptima, coinciden con las de dicho algoritmo.

La versión del algoritmo para el caso de los árboles de regresión implica la minimización de la función de pérdida cuadrática $E(\left(Y-\Sigma_{j=1}^J\alpha_j\hat{Y}^j\left(R,\Theta_j\right)\right)^2)$, donde hemos hecho explícito que los árboles que se usan como weak learners en la iteración *j* dependen de los parámetros de escisión Θ_j. El problema de optimización consiste en encontrar el clasificador final con la secuencia $\{\alpha_j,\Theta_j\}_{j=1}^J$ que minimiza la función de pérdida.

Resulta útil ver el problema anterior como un algoritmo de optimización secuencial en el que, dado el acumulado de los weak learners hasta la iteración *j*-1, $S\hat{Y}^{j-1}$, en la iteración *j* tenemos que encontrar $\{\alpha_j,\Theta_j\}$ que minimizan:

[8] Se puede demostrar que si el error cometido por los weak learners es mayor que 0.5, el error de la clasificación final disminuye exponencialmente con .

$$E\left[\left(Y-\Sigma_{j=1}^{J-1}\alpha_i\hat{Y}^i\left(R,\Theta_j\right)-\alpha_j\hat{Y}^j\left(R,\Theta_j\right)\right)^2\right]. \qquad [14]$$

En muestras finitas, el objetivo en cada iteración es encontrar el árbol de decisión que minimiza la suma de los cuadrados de los errores que se han cometido hasta la iteración anterior[9]. Friedman *et al.* (2000) proponen un algoritmo de Newton secuencial para encontrar el óptimo de dicha función y sientan las bases del algoritmo *Gradient Boosting*.

El algoritmo Gradient Bossting, propuesto por Friedman (2001), usa el algoritmo de optimización numérica *Functional Gradient Descent* para encontrar iterativamente el mínimo de la función de pérdida. En cada iteración, $j \in \{1,...,J\}$ se busca la dirección y el tamaño de actualización encaminada a optimizar la reducción del valor de la función de pérdida. En concreto, en cada iteración, la dirección que más reduce la función de pérdida es la opuesta al gradiente de la función de pérdida. Por esa razón, buscaremos el weak learner que más se aproxima al negativo del gradiente. El tamaño de la actualización, que llamaremos c_j, será el que optimiza la caída de la función de pérdida en la dirección seleccionada.

Si definimos a la función de pérdida como $L\left(Y,\widehat{SY}(R)\right)$, el algoritmo Gradient Bossting es el resultado de la iteración de los siguientes pasos:

1. Inicializamos el algoritmo $\widehat{SY}^0 = \arg\min_{SY} \Sigma_{t=1}^{T} L\left(Y_t, SY\right)2$. Dado un número máximo de iteraciones Z, para cada $z = 1,2,...,Z$, buscamos la dirección óptima para minimizar la función de pérdida, que coincide con el negativo del gradiente de la función de pérdida.

   - Para cada $t = 1,2,...,T$ calcular

$$r_z^t = -\left[\frac{\partial L\left(Y_t,\widehat{SY}\left(R^t\right)\right)}{\partial \widehat{SY}\left(R^t\right)}\right]_{\widehat{SY}\left(R^t\right)=\widehat{SY}^{z-1}\left(R^t\right)} \qquad [15]$$

   A r_z^t se le conoce como pseudoresiduo.

   - Determinar el árbol de decisión que determina $\mathcal{X}$ que mejor se ajusta a r_z^t proporcionando A_i^t regiones terminales con $i = 1,2,...,K_z$.

[9] Para determinados casos, resulta útil establecer otras funciones de pérdida. Por ejemplo, la función de Huber es recomendable cuando para el tratamiento de bases de datos con atípicos.

- Para cada $i = 1,2,...,K_t$ calcular, $c_i^t = \arg\min_{SY}\left\{\Sigma_{x_t \in A_i^t} L\left(Y_t, \widehat{SY}^{z-1}\left(A_i^t\right)\right) + SY\right\}$.

- Actualizar $SY^z(R^t) = SY^{z-1}(R^t) + \Sigma_{i=1, x_t \in R^t \cap A_i^t} SY(A_i^t)$

3. Por último, el estimador es $\widehat{SY} = SY^{K_z}$.

Friedman (2001), incorpora un parámetro de ajuste que controla el tamaño de los saltos que se dan en el algoritmo en cada iteración, por lo que la actualización del strong learner se produce escalada por un parámetro μ:

$$\widehat{SY}^j\left(R^t\right) = \Sigma_{i=1}^{j-1}\widehat{SY}^i\left(R^t\right) + \mu\widehat{SY}^j\left(R^t\right), \qquad [16]$$

donde $0 < \mu \leq 1$. Cuanto más pequeño sea el valor del parámetro de ajuste mejor será la clasificación en el período de entrenamiento, pero mayor será el número de iteraciones necesarias para alcanzar el óptimo. Por otro lado, valores muy grandes pueden subestimar el número de árboles necesarios. Los resultados empíricos de Friedman (2001) sugieren usar valores $\mu \leq 0{,}1$.

Finalmente, Friedman (2002) incorpora un elemento adicional en el algoritmo *Stochastic Gradient Boosting:* el muestreo aleatorio de un porcentaje λ de observaciones que forman parte del periodo de entrenamiento. En concreto, en cada iteración del algoritmo, el nuevo árbol de decisión se ajusta empleando únicamente una fracción λ de datos del periodo de entrenamiento, extraída de forma aleatoria y sin reemplazo. Este procedimiento mejora el ajuste a los datos y agiliza la computación. En la práctica, para reducir tiempo de computación, se recomienda que λ sea menor cuanto mayor sea el número de variables explicativas disponibles.

Con el objetivo de que los árboles de clasificación se puedan usar para seleccionar de manera automática las variables explicativas más influyentes a la hora de formar los árboles de decisión, Friedman (2001) propone una medida de la importancia relativa en el algoritmo boosting. Siendo IR_p^j la importancia relativa de la variable p en el árbol que se estima en la iteración j, la importancia relativa de esa variable en el boosting será la media de su importancia relativa en los J árboles generados, $IR_p^{Bo} = \frac{1}{J}\Sigma_{i=1}^{j} IR_p^j$

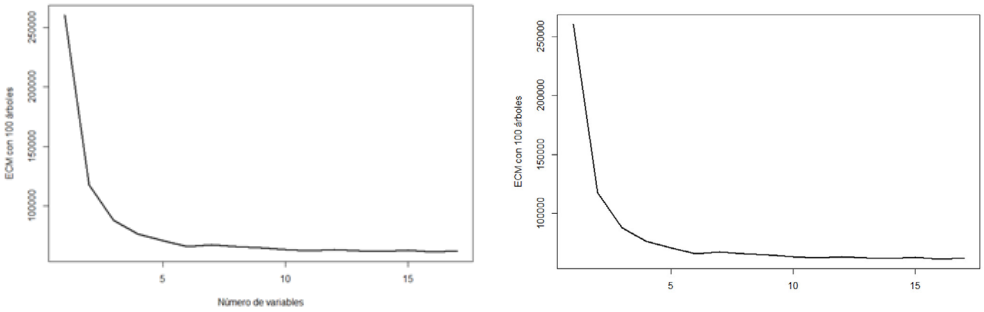
3.3. Ejemplo de bagging y boosting

En esta sección vamos a examinar cómo se ajustan las tres propuestas anteriores, bagging, random forest y boosting a los datos de los precios de la vivienda en Madrid. Si comenzamos con la estimación bagging, tenemos que elegir el número de árboles que se van a muestrear con bootstrap. El gráfico izquierdo de la figura 5 muestra el

ECM que se alcanza en función del número de árboles, con un máximo de 500 árboles. Como se puede observar, a partir de unos 100 árboles apenas se alcanzan reducciones significativas en el error. Por tanto, usaremos bagging con un número de boostraps de 100 árboles.

FIGURA 5

HIPERPARÁMETROS DE BAGGING Y RANDOM FOREST

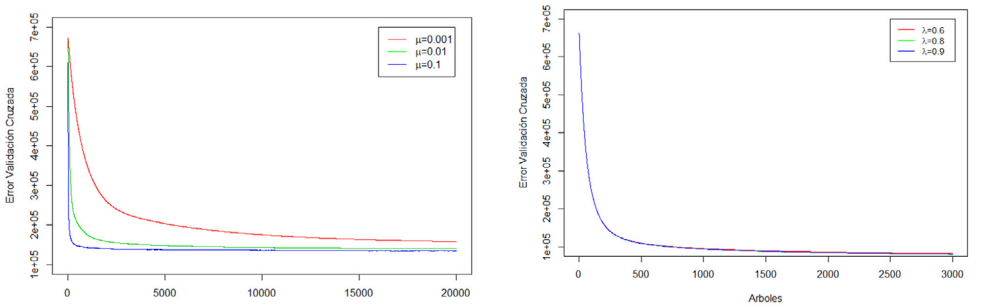


Fuente: Elaboración propia.

Para ilustrar las ventajas de random forest frente a bagging, el gráfico derecho de la figura 5 muestra cómo evoluciona el ECM con 100 árboles que se obtiene con un número de variables seleccionadas en cada muestreo que varía desde 1 a 17, que es el número total de variables explicativas de la muestra. Como podemos observar en la figura, a partir de 5 variables para cada muestreo, el error el árbol de regresión no es capaz de reducir el ECM de manera significativa.

FIGURA 6

HIPERPARÁMETROS DE BOOSTING



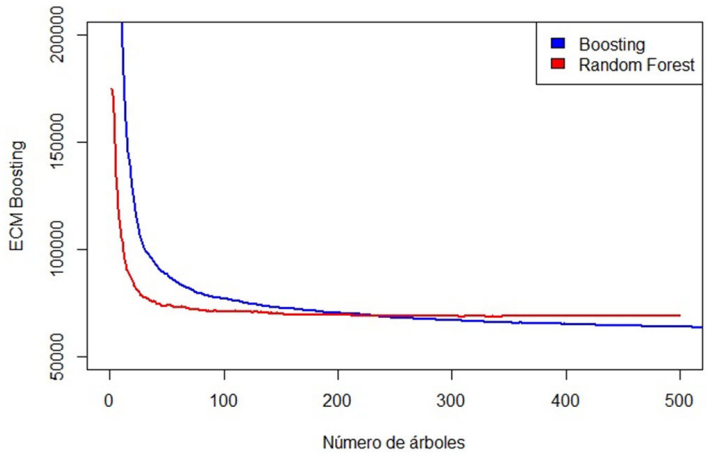
Fuente: Elaboración propia.

La figura 6 nos va a ayudar a optimizar los hiperparámetros que se usan para aplicar el algoritmo boosting. El gráfico de la izquierda muestra la evolución del ECM en función del número de árboles Z para distintos valores del parámetro de ajuste $\mu = 0{,}001, 0{,}01, 0{,}1$. En este gráfico podemos observar que el ECM se estabiliza para un número de árboles alrededor de 200 y que valores del parámetro de ajuste entre y no afectan mucho a la capacidad de ajuste a los datos del modelo. El gráfico de la derecha muestra que el ajuste del método Stochastic Gradient Boosting no depende mucho de la fracción de datos que usan para el entrenamiento, para valores de λ entre 0,6 y 0,9.

En la literatura de los árboles de decisión, la elección entre random forest o boosting ha resultado ser una cuestión empírica porque no hay un ganador absoluto entre ellos. Hamza y Larocque (2005) encuentran que random forest funcionaba mejor que boosting, aunque Gashler, Giraud-Carrier Martínez (2008) muestran que random forest no funciona peor cuantas más variables explicativas irrelevantes contenga la muestra de datos. En nuestro ejemplo, la figura 7 muestra que ninguna de estas dos opciones aparezca como una ganadora clara. En términos generales, El ECM de ambos algoritmos es muy parecido. El error que comete el bagging es un poco más reducido que el de boosting para un número de árboles inferior a 200. Sin embargo, a partir de 200 árboles, el algoritmo boosting se ajusta un poco más a los datos que el algoritmo random forest.

FIGURA 7

ECM DE RANDOM FOREST Y BOOSTING



Fuente: Elaboración propia.

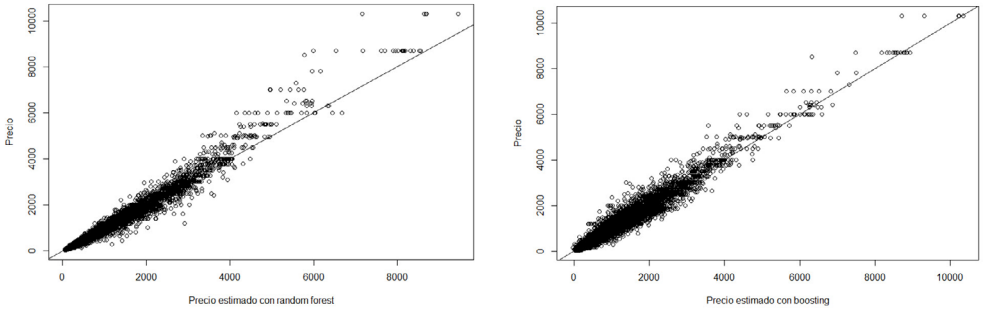
Una diferencia importante de los árboles de decisión estimados con random forest o con boosting respecto de los árboles estimados con CART es la formación de predicciones de la variable dependiente. La figura 8 muestra que la predicción obtenida con estos dos algoritmos no es tan particionada como la que se realiza con CART, como

mostraba el gráfico izquierdo de la figura 3. La forma de construcción del árbol con random forest y bagging, donde la inferencia de cada observación es una media de la inferencia realizada con cada uno de los árboles que forman parte de esos algoritmos hace que la inferencia final sea más parecida a la que estamos acostumbrados a obtener a partir de un modelo de estimación tradicional, como la predicción con MCO que aparece en el gráfico derecho de la figura 3.

Para comparar la bondad del ajuste de todos los modelos propuestos, hemos calculado el pseudo R^2 para estos dos algoritmos. En el caso de random forest el pseudo R^2 es de 0,97, mientras que en el caso de boosting el pseudo R^2 es de 0,96. Estos valores son claramente superiores a los que se alcanzaban con árboles estimados mediante CART (0,73) y con MCO (0,72). Sin embargo, debemos señalar que la habilidad de un modelo para ajustarse a los datos dentro de la muestra no necesariamente se tiene que corresponder con la capacidad predictiva del modelo.

FIGURA 8

PREDICCIÓN DEL PRECIO DE LA VIVIENDA CON RANDOM FOREST Y BOOSTING



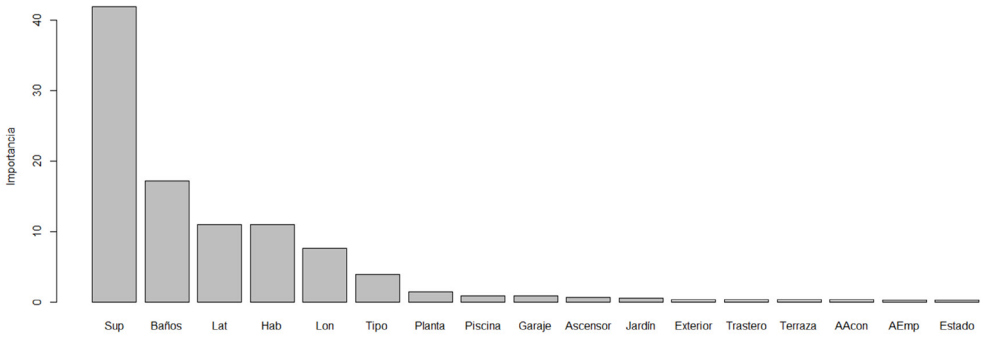
Fuente: Elaboración propia.

Para examinar la capacidad predictiva de los modelos, hemos dividido aleatoriamente la muestra en un dos grupos de datos. El grupo de entrenamiento supone el 75% de los datos y se ha usado para estimar los modelos. El grupo de evaluación lo forman el resto de datos y se usa para predecir con los modelos estimados en el grupo de entrenamiento. El *(out-of-sample)* pseudo R^2 que alcanzan los modelos MCO, CART, random forest y boosting es de 0,71, 0,72, 0,89 y 0,90. Como se esperaba, todos los modelos muestran un deterioro en la capacidad predictiva. Sin embargo, los modelos random forest y boosting siguen apareciendo como los modelos con mayor capacidad predictiva. Como los resultados que obtenemos con los algoritmos random forest y booting son muy parecidos, para simplificar el análisis nos centraremos en los resultados obtenidos con random forest.

La figura 9 muestra la importancia relativa de las variables que determinan el precio de la vivienda calculada a partir del método random forest[10]. La figura revela que la principal variable para determinar el precio de la vivienda es la superficie, seguida del número de baños. Posiblemente, estas dos variables estén correlacionadas y reflejen, junto con el número de habitaciones, la relación entre el precio y la dimensión de la vivienda. Las variables que determinan la posición de la vivienda también ocupan un papel relevante en la determinación del precio de la vivienda. La última variable importante es el tipo de vivienda, mientras que el resto de variables parecen tener una importancia relativa mucho menor.

FIGURA 8

IMPORTANCIA RELATIVA



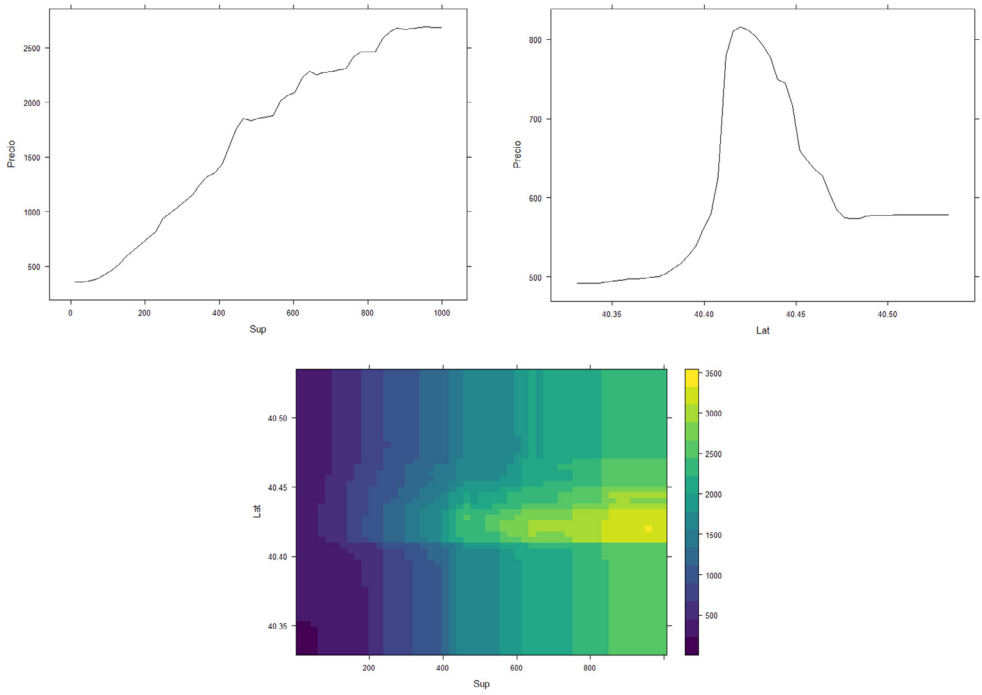
Fuente: Elaboración propia.

Una de las ventajas más significativas de los árboles de decisión estimados con random forest y boosting frente a modelos tradicionales como CART es su mayor capacidad de capturar relaciones no lineales entre las variables explicativas y la variable dependiente. La figura 10 muestra claramente que el efecto de la latitud en el precio de la vivienda es no lineal. Hasta una latitud de 40.41 el efecto sobre el precio es positivo, es decir, las viviendas del sur aumentan de precio conforme las viviendas se sitúan más al norte. Hasta una latitud de 40.43, donde se encuentran las viviendas del centro la latitud deja de tener importancia. Finalmente, para las viviendas situadas al norte de la latitud 40.45, la relación con el precio se invierte, siendo las viviendas más baratas conforme las viviendas se sitúan del centro hacia el norte. Para las viviendas muy en el sur o muy en el norte, la latitud no tiene efecto en el precio. La superficie, sin embargo, tiene una relación positiva y lineal con el precio de la vivienda. El gráfico inferior de la figura 10 muestra un efecto interacción no lineal entre esas dos variables.

[10] Cualitativamente, la importancia relativa que se obtiene con boosting es similar.

FIGURA 10

GRÁFICOS DE DEPENDENCIA PARCIAL: BOOSTING



Fuente: Elaboración propia.

En último lugar vamos a ilustrar la ventaja de los árboles de decisión sobre los métodos de estimación clásicos como MCO usando la figura 11. Si recordamos la figura 1, parece evidente que la relación entre la posición geográfica, medida a través de la latitud y la longitud de la vivienda, y el precio de la misma dista mucho de ser lineal. Sin embargo, si tratamos de estimar un modelo lineal entre el precio de la vivienda y sus características mediante MCO, tanto la latitud como la longitud aparecen con un coeficiente positivo. Eso explica que el efecto interacción que se muestra en el gráfico izquierdo de la figura 11, las viviendas con más valor aparezcan erróneamente en el noreste geográfico.

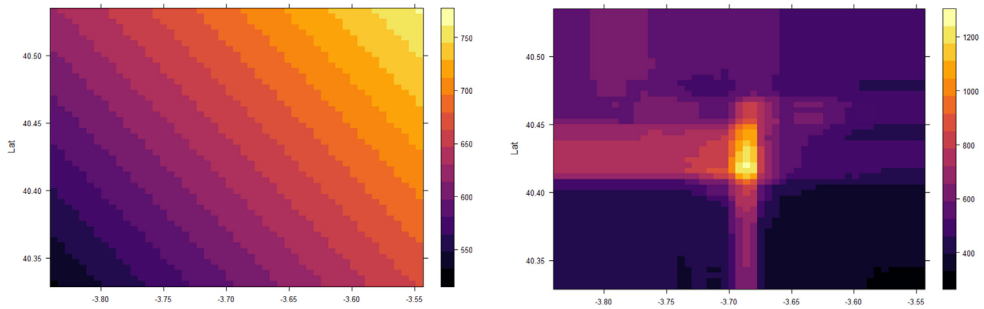
Sin embargo, el efecto interacción del gráfico derecho, que se ha obtenido a partir de un árbol de decisión estimado con el algoritmo random forest, muestra una relación claramente no lineal entre el precio de la vivienda y su posición geográfica. Las viviendas más caras se sitúan en el centro de Madrid. Hacia el sur, no aparece una relación evidente entre el precio de la vivienda y su posición geográfica excepto entre las que se encuentran en una longitud central, referida a las viviendas más caras. Para esas

viviendas, el precio cae cuando se sitúan más al sur. Las casas más caras se sitúan al norte de la latitud 40,41, especialmente las de latitud más central (entre -3,70 y -3,80). Al norte de la latitud 40,41, también ocurre que las viviendas son más baratas cuando nos alejamos del centro.

FIGURA 11

POSICIÓN GEOGRÁFICA Y PRECIO DE LA VIVIENDA

[MCO] [Arbol de decisión]



Fuente: Elaboración propia.

3.4. Extensiones

Una característica de los árboles de regresión, tal y como son normalmente implementados, es que las particiones del espacio de definición de las variables explicativas, se hacen de manera ortogonal a la variable de escisión, generando una partición formada por prismas rectangulares p-dimensionales. Esta forma de particionar está justificada por la reducción del espacio de búsqueda del algoritmo, lo que se traduce en ganancias computacionales. Sin embargo, las particiones ortogonales pueden fallar a la hora de encontrar particiones apropiadas en algunos casos, haciendo que el comportamiento del algoritmo sea mediocre.

Con el fin de paliar este problema, han aparecido en la literatura un conjunto de propuestas de partición alternativas. Una de las más populares está basada en la realización de particiones con fronteras oblicuas, como las que se proponen en los trabajos de Murthy, Kasif, y Salzberg (1994), Wickramarachchi *et al.* (2016) y Cantu-Paz y Kamath (2003). Además de la complejidad computacional, unos de los problemas de estas propuestas es que las realización de las particiones suele depender de diversos parámetros que deben ser elegidos por los usuarios, limitando la robustez de los resultados.

Otra opción es la que han propuesto Paez *et al.* (2019), quienes introdujeron una novedosa forma de generar particiones no ortogonales en el espacio de definición de las variables, produciendo un incremento en la complejidad en la construcción del árbol bastante modesto. En concreto, esta aproximación se basa en la aplicación de funcio-

nes base interactivas, lo que permite particionar con fronteras de cualquier forma funcional, manteniendo como caso particular las particiones oblicuas. La principal ventaja de esta técnica es su fácil implementación ya que permite seguir usando los mismos códigos que se usan tradicionalmente para particiones ortogonales.

La segunda extensión que vamos a considerar en esta sección es el tratamiento de datos perdidos *(missing data),* que resulta muy habitual en el análisis de datos económicos. Como se refleja en Tawala (2009), la forma en la que se ha resuelto este problema en el contexto de árboles de decisión es muy diversa. Una opción es eliminar de la base de datos las variables con datos perdidos o recortar la muestra para que todos los datos sean observados en todas las variables. Otra opción es imputar los datos que faltan con la media, la mediana o la moda de la variable explicativa correspondiente. También se han propuesto formas de imputación más complejas como las que se basan en algoritmos *Expectation Maximization (EM).*

Una de las alternativas a la imputación más habituales en algoritmos CART consiste en usar puntos de escisión *(splits)* subrogados. Supongamos que, en un determinado nodo, el árbol tiene que elegir entre la región izquierda y la derecha usando la variable de escisión X_p que no se observa en un elemento de la muestra ($X_{pt} = NA$). Para implementar esta medida, el primer paso consiste en estimar el árbol de decisión con el subconjunto completo de observaciones que excluye a t. En el segundo paso, se elige una variable (X_j) y un punto de escisión (s) que proporcionan la partición más parecida a la que se obtiene con X_p para el nodo determinado. Por último, para la observación t se realiza la partición usando la variable subrogada X_j.

4. OTRAS APLICACIONES ECONÓMICAS Y EMPRESARIALES

Aunque los primeros métodos de construcción de árboles de decisión aparecen en los años sesenta, la aplicación de estas técnicas a problemas económicos y empresariales es mucho más reciente. Sin ánimo de ser exhaustivos, en esta sección vamos a describir algunas de las aplicaciones más interesantes de estas técnicas para resolver problemas relacionados con economía y empresa.

Entre las aplicaciones de árboles de clasificación a la empresa, destacamos la que realizaron Qabbaah, Sammou y Vanhoof (2019) para mejorar la eficiencia de campañas de marketing realizadas a través de correo electrónico. Tirenni, Kaiser y Herrmann (2007) usan árboles de decisión para analizar la segmentación de los consumidores en empresas de aerolíneas. Tras una comparativa con otros modelos de clasificación, Gepp, Kumar, y Bhattacharya (2010) concluyen que los árboles de decisión son la mejor herramienta para predecir el fracaso empresarial. Xie *et al.* (2009) desarrollan una extensión de los árboles de decisión para anticipar el abandono de clientes de entidades financieras.

Relacionados con economía de la educación, Zeng *et al.* (2014) han aplicado técnicas de árboles de regresión para examinar los condicionantes de la elección de universidades en China. Mythili y Shanavas (2014) realizan una comparativa entre diversos algoritmos de clasificación para anticipar la mejora en los resultados de los estudiantes a partir de un conjunto de variables socio-económicas. Entre los métodos evaluados, destacan árboles estimados con random forest .

La aplicación de árboles de clasificación al análisis de ciclos económicos se inicia con el trabajo de Ng (2014), quien usa técnicas de boosting con stumps como weak learners para evaluar la capacidad de un elevado número de indicadores económicos para predecir las recesiones de EE. UU. con 3, 6, y 12 meses de adelanto. Döpke, Fritsche, y Pierdzioch, (2017) introducen como weak learners árboles más complejos en el algoritmo de boosting para examinar el efecto interacción de los indicadores económicos en la predicción de las recesiones en Alemania. Piger (2020) compara la habilidad de los árboles de decisión construidos con random forest y boosting junto con otras técnicas de machine learning y modelos *Markov-switching* para anticipar las recesiones en EE. UU. En sus resultados, Piger (2020) destaca la habilidad del algoritmo boosting sobre el resto de métodos evaluados. En relación con estos trabajos, Ward (2017) utiliza el algoritmo random forest para identificar episodios de crisis financieras en un conjunto amplio de países.

Rossi y Timmermann (2015) proponen un nuevo procedimiento para medir el riesgo en un modelo intertemporal de fijación de precios de los activos de capital (*ICAPM,* por sus siglas en inglés). Los árboles de regresión construidos con boosting se usan para realizar predicciones no lineales de las matrices de covarianza. En una comparativa entre cuatro metodologías diferentes, Liu *et al.* (2015) encuentran que los árboles de decisión construidos mediante random forest son la mejor opción para anticipar el fraude financiero, siendo la relación deuda-patrimonio (*debt-to-equity ratio*) la variable más significativa.

5. CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

El principal propósito de este trabajo es el de contribuir en la difusión de la metodología de árboles de decisión como una técnica de análisis de problemas relacionados con economía y empresa. Consideramos que esta herramienta es especialmente útil para resolver dos problemas a los que se enfrentan actualmente las técnicas de análisis tradicionales. El primer problema tiene que ver con el manejo de datos de gran volumen y variedad, conteniendo variables cuantitativas y cualitativas. El segundo problema es la detección de relaciones complejas entre las variables del modelo, difíciles de capturar con modelos lineales y/o paramétricos. Estamos convencidos que la versatilidad de los árboles de decisión puede ser especialmente útil en estos dos casos.

Para conseguir este objetivo, en el manuscrito hacemos una descripción autocontenida de la metodología de árboles de decisión, empezando por la construcción de

un árbol de decisión sencillo mediante el algoritmo CART. Seguidamente, mostramos las extensiones más habituales, entre las que destacan las que se basan en algoritmos random forest y boosting. Además, dedicamos una sección al tratamiento de extensiones menos habituales como las que producen particiones no ortogonales. Finalmente, dedicamos otra sección a mostrar numerosos ejemplos para ilustrar cómo los árboles de decisión han permitido esclarecer algunas relaciones entre variables relacionadas con la economía y la empresa.

Para facilitar el acercamiento de los economistas a estas técnicas, proponemos un ejemplo enfocado a examinar la determinación del precio de la vivienda en el municipio de Madrid. La base de datos, extraídos del portal de Idealista en 2018, contiene el precio de 20.385 viviendas, junto con un conjunto de potenciales variables explicativas. Entre ellas, las que destacan por su importancia final en la determinación del precio de la vivienda son la superficie, el número de baños y de habitaciones, la posición geográfica, y el tipo de vivienda. Usando el Pseudo-R^2, encontramos que la capacidad de los árboles de decisión para predecir el precio de la vivienda basados en los algoritmos random forest y boosting es similar, aunque muy superior a la árboles basados en CART y a la de mínimos cuadrados ordinarios. Además, mostramos cómo los árboles son capaces de capturar relaciones no lineales complejas entre, por ejemplo, la posición geográfica y el precio de la vivienda que las técnicas tradicionales serían incapaces de determinar.

Con el fin último de simplificar la adopción de estas técnicas en economía y empresa, hemos puesto a disposición de los lectores en la página web de los autores tanto la base de datos como un código en R que replica los principales resultados que se presentan en este trabajo. Confiamos en que este material sirva como primer paso de muchas aplicaciones en el contexto de la economía y la empresa en el futuro.

Referencias

Austin, P. C., Lee, D. S., Steyerberg, E. W. y Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical journal,* 54(5), pp. 657-673.

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D. y Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology,* 66(4), pp. 398-407.

Breiman, L., Friedman, J., Stone, C. J. y Olshen, R. A. (1984). *Classification and regression trees.* New York: Chapman and Hall.

Cantu-Paz, E. y Kamath, C. (2003). Inducing oblique decision trees with evolutionary agorithms. *Transactions on Evolutionary Computation,* 7(1), pp. 54-68.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F. y Klove, B. (2018). River suspended sediment modelling using the Cart model: A comparative study of machine learning techniques. *Science of the Total Environment*, 615, pp. 272-281.

De'ath, G. y Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology,* 81(11), pp. 3178-3192.

Diebold, F. X. (2012). On the origin(s) and development of the term Big Data. *Documento de Trabajo PIER,* 12-037

Döpke, J., Fritsche, U. y Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting,* 33(4), pp. 745-759.

Fan, G., Ong, S. y *Koh, H.* (2006). Determinants of house price: A decision tree approach. *Urban Studies,* 43(12), pp. 2301-2316.

Gashler, M., Giraud-Carrier, C. y Martínez, T. (2008). Decision tree ensemble: small heterogeneous is better than large homogeneous. *Proceesing of the 7th International Conference on Machine Learning and Applications,* pp. 900-905.

Ghasri, M., Rashidi, T. y Waller, S. (2017). Developing a disaggregate travel demand system of models using data mining techniques. *Transportation Research Part a-Policy and Practice,* 105, pp. 138–153.

Gepp, A., Kumar, K. y Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of forecasting,* 29(6), pp. 536-555.

Hamza, M. y Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation,* 75, pp. 629–643.

James, G., Witten, D., Hastie, T. J. y Tibshirani, R. J. (2013). *An introduction to statistical learning with applications in R.* New York: Springer-Verlag.

Kitsantas P., Moore T. y Sly D. (2007). Using classification trees to profile adolescent smoking behaviors. *Adictive Behaviors,* 32(1), pp. 9–23.

Kurt, I., Ture, M. y Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications,* 34 (1), pp. 366-74.

Liu, C., Chan, Y., Kazmi, S. y Fu, H. (2015). Financial fraud detection model: Based on random forest. *International Journal of Economics and Finance,* 7(7), pp. 178-188.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews – Data Mining and Knowledge Discovery,* 1(1), pp. 14-23.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability,* 1, pp. 281-297.

Morgan, J. y Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Assocciation,* 58, pp. 415–434.

Murthy, S., Kasif, S. y Salzberg, S. (1994). A System for Induction of Oblique Decision Trees. *Journal of Artificial Intelligence Research,* 2, pp. 1-32.

Mythili, M. y Shanavas, A. (2014). An Analysis of students' performance using classification algorithms. *IOSR Journal of Computer Engineering,* 16(1), pp. 63-69.

Ng, S. (2014). Boosting recessions. *Canadian Journal of Economics,* 47, pp. 1-34.

Páez, A., López, F., Ruiz, M. y Camacho, M. (2019). Inducing non-orthogonal and non-linear decision boundaries in decision trees via interactive basis functions. *Expert Systems with Applications,* 122, pp. 183-206.

Piger, J. (2020). Turning points and classification. En: P. Fuleky (ed.), *Macroeconomic forecasting in the era of big data: Theory and application.* Springer International Publishing.

Qabbaah, K., Sammour, G. y Vanhoof K. (2019). Decision tree analysis to improve e-mail marketing campaigns. International Journal of Information *Theories and Applications,* 26(1), pp. 3-36.

Rossi, A. y Timmermann, A. (2015). Modeling covariance risk in Merton's ICAPM. *Review of Financial Studies,* 28(5), pp. 1428–1461.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning,* 5, pp. 197–227.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data when using decision trees. *Applied Artificial Intelligence,* 23, pp. 373–405.

Tirenni, G., Kaiser, Ch. y Herrmann, A. (2007). Applying decision trees for value-based customer relations management: Predicting airline customers' future values. *Journal of Database Marketing and Customer Strategy Management,* 14, pp. 130–142.

Venkata, S. y Kiruthika, P. An overview of classification algorithm in data mining. International *Journal of Advanced Research in Computer and Communication Engineering,* 4(12), pp. 255-257.

Ward, F. (2017). Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors. *Journal of Applied Econometrics,* 32(2), pp. 359-378.

Wickramarachchi, D., Robertson, B., Reale, M., Price, C. y Brown, J. (2016). HHCART: An oblique decision tree. *Computational Statistics and Data Analysis,* 96, pp. 12-23.

Xie, Y., Li, X., Ngai, E. W. T. y Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications,* 36(3), pp. 5445-5449.

Yang, L. J., Liu, S., Tsoka, S. y Papageorgiou. L. G. (2017). A regression tree approach using mathematical programming. *Expert Systems with Applications,* 78, pp. 347-57.

Zeng, X., Yuan, S., Li, Y. y Zou, Q. (2014). Decision tree classification model for ppularity forecast of Chinese colleges. *Journal of Applied Mathematics,* Article ID 675806.

CAPÍTULO IV

Modelos predictivos del riesgo y aplicaciones a los seguros

Montserrat Guillen*
María Láinez
Ana M. Pérez-Marín
Eduardo Sánchez

El análisis del riesgo estudia los sucesos extraordinarios, qué los causa y cómo mitigar su probabilidad de ocurrencia. En seguros, básicamente importa la frecuencia y la severidad de los siniestros. Usando medidas como los cuantiles alejados de la mediana, la modelización predictiva permite detectar factores que afectan al riesgo. Tras la presentación de la regresión cuantílica como modelo básico y sus generalizaciones, se realiza una recopilación de casos de estudio en el ámbito asegurador, en situaciones de datos masivos y en particular en el análisis de datos telemáticos en seguros del automóvil.

Palabras clave: regresión cuantílica, extremos, seguros de automóvil, datos telemáticos, accidentes.

* Montserrat Guillen y Ana M. Pérez-Marín agradecen el apoyo de la Fundación BBVA en los proyectos de investigación en Big Data, del Ministerio de Ciencia e Innovación (proyecto número PID2019-105986GB-C21) y del programa ICREA Academia. Los autores agradecen las aportaciones de MAPFRE España y MAPFRE SA.

1. INTRODUCCIÓN

El análisis del riesgo tiene por objeto estudiar fenómenos extraordinarios, grandes accidentes y catástrofes que ocasionan cuantiosas pérdidas. Como campo científico, al igual que las ciencias actuariales, siempre ha quedado en tierra de nadie por su multidisciplinariedad; en economía, matemáticas, ingeniería, bioestadística y demografía. Sin embargo, en las últimas décadas, el impacto que el desarrollo tecnológico ha ejercido en el ámbito de los seguros y la gestión de riesgos ha impuesto al estadístico, o al científico de datos, como un perfil imprescindible en esta materia.

Resulta paradójico que el análisis de riesgos se acabe vinculando al big data cuando al hablar de riesgos solemos pensar en fenómenos catastróficos para los que no hay prácticamente ni información, ni antecedentes. Los grandes siniestros son infrecuentes, y se dan en circunstancias incontroladas que los hacen altamente peligrosos, donde además es difícil identificar una regularidad estadística, y con ello diseñar medidas para evitar su ocurrencia.

A pesar de la dificultad de recopilar información para el análisis de riesgos, existen métodos adaptados de por sí a la predicción de fenómenos de muy baja frecuencia. Por ejemplo, para el seguro de automóviles donde la probabilidad de que un conductor sufra un accidente durante un año no suele superar el 10 % en la mayoría de países desarrollados, o para el seguro de vida, donde se trabaja con probabilidades incluso mucho más reducidas, la metodología cuantitativa está bien establecida. Las bases de datos con miles de asegurados ya eran habituales en la segunda mitad del siglo XX, y las entidades aseguradoras han venido utilizando con total normalidad las técnicas de modelización que permiten analizar lo que en inglés se denominan los *rare events.*

Las nuevas tecnologías, los sensores y la internet de las cosas no han hecho más que ampliar las posibilidades del análisis de riesgos, para convertirlo en un auténtico paraíso de la información. Además, y añadido a los logros obtenidos, se ha generado una nueva demanda unánime en la sociedad: la necesidad de fomentar la prevención. La generación de datos para evitar catástrofes es un tema para el que los ciudadanos reclaman establecer niveles de vigilancia, incluso más exhaustivos de lo que sería considerado una invasión a su propia privacidad. Hay numerosos ejemplos de ello, desde la creación de avisos meteorológicos que informan a la población de la inminencia de fenómenos adversos tales como tormentas, huracanes o tsunamis, hasta la prevención de riesgos alimentarios, y lógicamente las pandemias. En consecuencia, el análisis de riesgos se ha transformado completamente en los últimos años, proporcionando respuesta no solo a aspectos meramente predictivos sino también preventivos. En definitiva, la disponibilidad de bases de datos de gran volumen ha multiplicado las posibilidades del análisis de riesgos y ha dado lugar a una nueva era en la elaboración de modelos predictivos, que sirven para anticipar patrones fuera de lo normal, y de modelos *prescriptivos,* cuya utilidad es crear sistemas de protección.

En este capítulo, se muestra cómo se puede implementar el análisis predictivo de los cuantiles para la modelización del riesgo, dejando atrás los modelos de regresión tradicionales focalizados en el análisis de la media. Seguidamente, se ofrecen algunas generalizaciones. También se incluyen varios ejemplos con resultados relativos al uso de datos telemétricos en el seguro de automóviles y, concretamente, en pólizas de pago por uso. Para terminar, se realiza una revisión de otras aplicaciones del big data en los seguros, y se concluye presentando algunas líneas emergentes en este ámbito.

2. PREDECIR FRECUENCIA Y SEVERIDAD ESPERADAS

El seguro es un mecanismo autoprotector en el que un colectivo solidario, formado por los asegurados, se hace cargo de compensar a sus miembros cuando alguno de ellos sufre un siniestro. El vínculo queda establecido a través de una entidad aseguradora, suscribiendo una póliza y realizando el pago de su correspondiente prima. En la mayoría de productos aseguradores, el principal escollo es anticipar cuál es la probabilidad de ocurrencia de un accidente, y si este se produce, cuál se prevé que sea su magnitud[1].

El planteamiento del problema como probabilidad de ocurrencia y seguidamente del coste económico del siniestro, es muy parecido al problema de impago de un crédito, donde por una parte, se modeliza la probabilidad de impago y, posteriormente, la cuantía esperada de la pérdida si el impago ya se ha producido. Sin embargo, el coste máximo total, o el parcial que queda por pagar, en una operación crediticia es una cuantía acotada. En la mayoría de siniestros en el sector asegurador, los límites no son tan claros *a priori,* ya que si bien los bancos, al conceder créditos con cuantía establecida, tienen un intervalo de oscilación de las pérdidas dentro de un margen, la inmensa mayoría de los productos aseguradores, incluso a pesar de que existan cláusulas sobre máximos de responsabilidad pactados contractualmente en las pólizas, pueden llegar a tener rangos de variación entre el coste del siniestro máximo y del siniestro medio que pueden calificarse de gigantescos o, a efectos prácticos, desconocidos. Con la excepción de la mayoría de los seguros de vida, en los que la indemnización ya queda fijada en la póliza y el siniestro solamente se produce una vez, el resto de seguros admiten que pueda producirse más de un accidente durante la vigencia de la póliza y, además como ya se ha mencionado, antes de que ocurran dichos accidentes existe una elevada incertidumbe sobre su potencial severidad. Los productos que tienen: severidad desconocida y posibilidad de reiteración de siniestro en un mismo periodo de cobertura son los más habituales, de hecho se conocen como los *seguros generales.* Tal es el caso por ejemplo, en un seguro de salud, donde se puede requerir asistencia en varias ocasiones, acudiendo a uno o más especialistas e incurriendo en gastos médicos difícilmente previsibles.

[1] Aquí no trataremos las consecuencias de los accidentes desde perspectivas ajenas a la compensación económica. La prevención de accidentes y gestión de riesgos pueden ir mucho más allá y considerar daños irreparables, como la pérdida de vidas humanas.

El cálculo de primas más básico consiste en multiplicar el número esperado de siniestros por su coste medio. Obtenida esta cantidad, que suele referirse a un periodo anual, se aplican ajustes de seguridad y recargos para gastos de administración y de adquisición, y se determina el precio final que pagará el tomador del seguro, lo que en términos técnicos se conoce como su *prima de transferencia del riesgo.* La suma de las primas de los asegurados de un mismo colectivo garantiza fondos suficientes para hacer frente de forma mancomunada y solidaria a todos los siniestros del colectivo. Debido a la gran responsabilidad asumida por las entidades aseguradoras en su compromiso de resarcir de las pérdidas a sus asegurados, el sector en su totalidad queda sujeto a una regulación férrea, a un nivel incluso más exigente que el aplicado en otras áreas de actividad del sector financiero. Todo ello implica un control de la solvencia de las entidades y, sobre todo, una garantía de corrección de los cálculos actuariales necesarios para proveer las primas. De ahí el papel fundamental del análisis de riesgos basado en los datos.

Como el principio del cálculo de precios se fundamenta en un modelo predictivo orientado a modelizar el valor esperado de una variable de conteo (la frecuencia de siniestralidad) y de una variable positiva (el coste o severidad), que generalmente se asume no acotada y como ya se ha comentado antes, es asimétrica a la derecha, los modelos lineales generalizados, árboles de clasificación, redes neuronales y *random forests,* entre otros, son los métodos de *machine learning* que se vienen utilizando con total normalidad en los departamentos actuariales de las entidades aseguradoras y que vinculan la siniestralidad a factores o características del objeto asegurado y de quien lo asegura. De ese modo, dichos modelos estadísticos predictivos sirven como base para establecer una prima suficiente, y distinta para cada tipología de cliente y cada contrato.

Uno de los grandes debates en el sector de los seguros actualmente surge a raíz del impulso que el big data ha ejercido en la personalización de las primas. La creciente disponibilidad de información permite que el número esperado y la cuantía esperada de los siniestros pueda ajustarse a un elevado número de características de riesgo particulares de quienes suscriben las pólizas, un número de factores a tener en cuenta que es muy superior al conjunto que se utilizaba décadas atrás. De ese modo, se ha visto incrementada la capacidad de diseñar sistemas de tarificación muy granulares que tienen en cuenta cada vez más información individual. La capacidad predictiva de los modelos y su adaptación a entornos con datos masivos choca entonces con el principio de mutualización. Y es en este punto donde emerge la inquietud de saber cuáles son los límites de la personalización de los precios, ya que si se pudiera llegar a predecir exactamente quién va a sufrir un accidente, y quién no, se acabaría estableciendo un precio para el primer grupo que sería igual al valor total de los accidentes que van a experimentar y un precio igual a cero para el segundo grupo, por lo que el propio concepto de la solidaridad en el seguro desaparecería. No existe un consenso sobre los límites de la ultra-segmentación de las primas, pero sí medidas que permiten detener un proceso

de individualización que conduzca a niveles de desigualdad excesiva de prima entre el colectivo de los asegurados[2].

Sin embargo, el big data abre una nueva perspectiva en el uso de los datos en los seguros y esa no es otra que la *prevención,* es decir, la predicción del riesgo anticipando la ocurrencia del siniestro y relegando el mero cálculo del precio a un segundo plano. A ello, ha contribuido muy notablemente la disponibilidad de información prácticamente en tiempo real.

2.1. Notación general y con datos telemáticos

Introducimos aquí la notación que va a utilizarse en el resto del capítulo. Se supone periodicidad anual en el contrato de seguro. Sea n el número de asegurados, sean N_i y S_{ij} respectivamente, el número de siniestros del asegurado i-èsimo, y la cuantía del j-ésimo siniestro del asegurado i, $i = 1,...,n, j = 1,...,N_i$ definida esta última únicamente si $N_i > 0$. Sean $X_{1i},...,X_{ki}$ las k características observables que suelen determinarse a partir de la formalización del contrato. Por ejemplo, la edad del asegurado, su antigüedad en la compañía y las características del objeto asegurado, como su superfície y localización en el caso de una vivienda, o marca, modelo, potencia y zona de conducción en un vehículo. Denominaremos $Z^*_{1i}(t),...,Z^*_{mi}(t)$, al conjunto de m variables telemáticas que están asociadas al objeto asegurado, es decir, que pueden medirse una vez ya está vigente el contrato e informan en tiempo real sobre su uso durante un periodo T y que se actualizan en intervalos de tiempo t. Denotaremos por $Z_{1i},...,Z_{si}$ a las s características anuales que resumen las observaciones telemáticas para cada asegurado i. Por ejemplo, en el caso del automóvil, con las tecnologías actuales puede medirse el total de kilómetros recorridos, número de trayectos realizados, la velocidad media de cada trayecto, frenazos, aceleraciones u otras medidas sin necesidad de tener localización exacta del vehículo. Cómo utilizar esta información telemática es uno de los objetivos de los modelos predictivos del riesgo en un entorno de datos masivos.

La información telemática permite conocer con detalle la exposición al riesgo, es decir, el intervalo de tiempo en el que realmente el asegurado puede tener un accidente que corresponde al momento en el que se encuentra conduciendo[3]. Hay casos en los que la exposición al riesgo es permanente, por ejemplo en los seguros de salud, pero en el seguro del automóvil a más kilómetros recorridos, mayor es la exposición y por lo tanto

---

[2] Hay factores que actualmente no pueden utilizarse para la determinación de precios. Por ejemplo, en la Unión Europea, como en un número creciente de países en el mundo, el principio de no discriminación impide que el sexo del asegurado pueda utilizarse como elemento diferencial en las tarifas, aunque sí puede servir internamente para analizar el riesgo que asume una entidad de seguros.

[3] Hay que tener en cuenta que los automóviles pueden sufrir percances aunque no están funcionando, por ejemplo estando aparcados pueden recibir un golpe de un tercero. Los siniestros de robo son un claro ejemplo también de exposición al riesgo con el vehículo parado. En el ejemplo concreto de recibir un golpe de un tercero, el siniestro estaría cubierto por la responsabilidad civil del culpable del golpe.

a igualdad de condiciones, quienes recorren más kilómetros tienen una probabilidad de sufrir accidentes sensiblemente superior a quienes recorren menos. El total de kilómetros recorridos en un año es una de las principales características telemáticas disponibles, pongamos Z_{1i}, que suele denotarse por D_i, y que puede utilizarse en el seguro de automóviles como una aproximación de la exposición al riesgo y, además, como elemento esencial en el pago por kilómetro.

2.2. Modelizar la frecuencia

El modelo de Poisson es el modelo básico para predecir el número esperado de siniestros y puede especificarse como:

$$E(N_i \mid X_{1i},...,X_{ki},Z_{1i},...,Z_{si}) = exp\,(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \gamma_1 Z_{1i} + ... + \gamma_s Z_{si}), \qquad [1]$$

siendo $\theta = (\beta_0,\beta_1,...,\beta_k,\gamma_1,...,\gamma_s)$ el vector de parámetros a estimar. Se suele usar la notación matricial $X_i^{a'}\theta$ para el predictor lineal, siendo $X_i^{a'} = \left(X_i^{'}, Z_i^{'}\right)$ el conjunto de regresores distinguiendo entre los que provienen del contrato $X_i^{'}$ y los que provienen de la telemetría $Z_i^{'}$. Se supone que N_i sigue una distribución de Poisson de parámetro $exp\left(X_i^{a'}\theta\right)$.

Cuando se utiliza una variable de exposición al riesgo, también denominada *offset*, D_i, para el i-ésimo individuo, el modelo se expresa como:

$$\begin{aligned} &lE(N_i \mid D_i, X_{1i},...,X_{ki},Z_{1i},...,Z_{si}) \\ &= D_i exp\left(\beta_0^o + \beta_1^o X_{1i} + ... + \beta_k^o X_{ki} + \gamma_1^o Z_{1i} + ... + \gamma_s^o Z_{si}\right) \\ &= exp\left(\log\left(D_i\right) + \beta_0^o + \beta_1^o X_{1i} + ... + \beta_k^o X_{ki} + \gamma_1^o Z_{1i} + ... + \gamma_s^o Z_{si}\right). \end{aligned} \qquad [2]$$

siendo $\theta^o = \left(\beta_0^o,\beta_1^o,...,\beta_k^o,\gamma_1^o,...,\gamma_s^o\right)$ el vector de parámetros a estimar.

Los parámetros se estiman por máxima verosimilitud y, como parte de los modelos lineales generalizados, se utilizan el conjunto de herramientas de inferencia de esta familia de modelos. De todos modos, como en muchos casos suele haber sobredispersión en los datos o un exceso de ceros, lo que se aconseja es usar algunas extensiones del modelo básico de Poisson como el modelo binomial negativo, que aquí omitimos.

2.3. Modelizar la cuantía

Para modelizar la cuantía de los siniestros se puede especificar un modelo para el coste, siendo cero si no ha habido ningún siniestro[4]. Para modelizar la cuantía, se puede utilizar un modelo Gamma donde la variable es estrictamente positiva. Así el modelo puede especificarse como:

[4] Si el asegurado ha sufrido más de un siniestro, se puede modelizar la media de los costes de los siniestros que ha sufrido cada asegurado.

$$E(S_i) = exp\ (\alpha_0 + \alpha_1 X_{1i} + ... + \alpha_l X_{ki} + \alpha_{l+1} Z_{1i} + ... + \alpha_{l+s} Z_{si}) \quad [3]$$

siendo S_i el coste medio de los siniestros del asegurado i, en caso de que este haya tenido algún siniestro, y 0 en caso contrario, $\alpha_0,..., \alpha_{l+s}$ el vector de parámetros a estimar en el modelo para las cuantías en las que se han incluido k características no-telemáticas del i-ésimo individuo y otras s de carácter telemático. Para simplificar, se han supuesto los mismo factores predictivos en los dos modelos: frecuencia y cuantía, pero el número y tipo de factores podría cambiar. El método de estimación es máxima verosimilitud. En el caso de la severidad, es difícil obtener resultados que permitan distinguir el impacto de los factores predictivos sobre la media, por lo que en la práctica no es extraño tomar solamente una constante y trabajar directamente con el importe medio de los siniestros.

3. PREDECIR CUANTILES

Dado que la frecuencia de siniestralidad es generalmente muy baja, la predicción del riesgo puede centrarse en algunos indicadores telemáticos, variables aleatorias continuas, que se sabe que están positivamente asociados a una mayor siniestralidad, por ejemplo, los excesos de velocidad o la conducción nocturna, entre otros, mediante la regresión cuantílica.

La regresión cuantílica es un modelo que especifica una relación entre los cuantiles de la variable respuesta R_i y un conjunto de covariables para el i-ésimo individuo. Si consideramos una especificación en la que distinguimos entre variables telemáticas y aquellas que no lo son, el modelo se expresa como:

$$Q_\tau(R_i \mid X_{1i},...,X_{ki},Z_{1i},...,Z_{si}) = \beta_0^\tau + \beta_1^\tau X_{1i} + ... + \beta_k^\tau X_{ki} + \gamma_1^\tau Z_{1i} + ... + \gamma_s^\tau Z_{si}, \quad [4]$$

siendo τ el nivel del cuantil a estimar (por ejemplo, 90 %, 95 %, 99 %), $Q_\tau(R_i \mid X_i^a) = X_i^a \theta^\tau$ el cuantil condicional a ajustar para la variable R_i, y $\theta^\tau = \left(\beta_0^\tau, \beta_1^\tau, ..., \beta_k^\tau, \gamma_1^\tau, ..., \gamma_s^\tau\right)$ el vector de parámetros a estimar. Se puede introducir una transformación en el predictor lineal de forma análoga a como se realiza en la media. Sin embargo, el procedimiento de estimación del modelo de regresión cuantílica generalizado es más complejo y no está disponible actualmente en los paquetes de estimación habituales.

4. OTROS MODELOS PREDICTIVOS DEL RIESGO

Recientemente, se han planteado nuevos modelos predictivos del riesgo, más allá de la regresión cuantílica, que amplían los modelos existentes en dos direcciones: a) para especificaciones no lineales del modelo de riesgo, como se acaba de mencionar en el párrafo anterior, y b) para otras medidas de riesgo como la esperanza condicionada de la cola. En este último caso, se establece un modelo conjunto en el que se estima simultáneamente un modelo para el cuantil y un modelo para la esperanza de los valores más

allá del cuantil. En el ámbito actuarial y de las ingenierías se supone que R_i es una variable aleatoria no negativa y, por lo tanto, el riesgo se encuentra en la parte derecha de la distribución, ya que se modelizan pérdidas y estas se suponen positivas. Sin embargo, en el ámbito financiero, cuando la variable respuesta son los rendimientos, que pueden ser positivos o negativos, entonces el riesgo se encuentra en los valores negativos, por lo que el interés reside en la cola izquierda de la distribución. Este cambio de signo, entre ambos entornos, ha propiciado un cierto distanciamiento entre las investigaciones y el uso de notaciones diferentes, que suelen confundir.

Los modelos conjuntos de medidas de riesgo surgen de la literatura financiera y, por lo tanto, en lugar de cuantiles se habla de valor en riesgo y en lugar de esperanza condicionada de la cola *(tail conditional expectation),* se utiliza el término *expected shortfall regression.* Para simplificar la notación, diremos que $Q_\tau(R_i \mid X_i^a)$ denota el cuantil condicional y que $E(R_i \mid R_i \geq Q_\tau(R_i \mid X_i^a), X_i^a) = CTE_\tau(R_i \mid X_i^a)$, siendo el modelo conjunto de regresión cuantílica y de regresión de esperanza condicional de la cola:

$$Q_\tau(R_i \mid X_i^a) = \beta_0^\tau + \beta_1^\tau X_{1i} + \ldots + \beta_k^\tau X_{ki} + \gamma_1^\tau Z_{1i} + \ldots + \gamma_s^\tau Z_{si}, \quad [5]$$

$$CTE_\tau(R_i \mid X_i^a) = \beta_0^{c\tau} + \beta_1^{c\tau} X_{1i} + \ldots + \beta_k^{c\tau} X_{ki} + \gamma_1^{c\tau} Z_{1i} + \ldots + \gamma_s^{c\tau} Z_{si}. \quad [6]$$

La estimación de los parámetros puede realizarse con un estimador de momentos o bien optimizando la función de pérdida correspondiente al modelo conjunto para ambas medidas de riesgo.

Los modelos anteriores permiten identificar factores que elevan el riesgo de ciertas variables respuesta, y, en particular, tienen interés las que están asociadas a una mayor accidentabilidad, como son el total de kilómetros recorridos y los excesos de velocidad en el caso del seguro de automóviles. En la siguiente sección se ilustran algunos casos de uso de los modelos anteriores, viendo el impacto de la inclusión de la información telemática.

5. EL SEGURO DE AUTOMÓVIL: PAGO POR KILÓMETRO

Una de las aplicaciones más reciente de la telemática es la utilización de sensores en los vehículos para monitorizar la conducción. No escapa a nadie que parte de este campo tiene como último objetivo lograr el transporte autónomo, en el que no sea necesaria la intervención de un humano al volante. Aunque se han logrado ciertos avances, y sobre todo en elementos de ayuda a la conducción como el control de velocidad, la distancia al vehículo precedente, el aparcamiento automático o el sensor de elementos alrededor del vehículo, no se vislumbra cuándo se podrá eliminar completamente el conductor. Sin embargo, la telemetría sí es una realidad y se está integrando de tal forma en los automóviles que el seguro no es ajeno a la disponibilidad de información que ello permite utilizar.

El pago por kilómetro es una de las formas de aseguramiento que más ha dado que hablar en los últimos años y se conoce bajo el acrónimo en ingles *PAYD, pay as you drive*[5]. Actualmente es posible comprar este tipo de producto en muchos países del mundo, incluida la mayoría de países de la Unión Europea. Dado que se aplica al seguro del automóvil, que es un seguro obligatorio, todo vehículo tiene una cuota mínima que corresponde a una prima básica, cuyo importe se incrementa de forma proporcional a la distancia recorrida por el vehículo.

Esta sección muestra un caso real de una cartera de vehículos en España que disponían de un dispositivo telemático de registro de datos de conducción instalado en los automóviles. Mostraremos tres aproximaciones al análisis del riesgo. En primer lugar, el tradicional, en el que la frecuencia de siniestralidad se explica por factores clásicos como la edad o la potencia del vehículo. Se muestra cómo la inclusión de indicadores telemáticos permite mejorar las predicciones incluso en una situación básica en la que no se utiliza la exposición al riesgo y el modelo para la frecuencia es el modelo de Poisson más simple. Al incluir la distancia total recorrida como *offset* en el modelo se aprecia la importancia del efecto de exposición al riesgo y se puede interpretar el resultado obtenido como una forma sencilla de obtener un coste por kilómetro adaptado a las características de riesgo. En un segundo paso, se presenta un modelo de regresión cuantílica y un modelo de regresión para la esperanza condicionada de la cola que permite detectar factores asociados a elevados valores de la exposición al riesgo. Este modelo permite detectar qué características influyen en exposiciones extremas al riesgo, percentil 90 %, lo que permite identificar segmentos de mayor peligrosidad que el resto. En un tercer caso, se muestra el comportamiento en curvas de referencia que comparan los kilómetros recorridos por encima de los límites de velocidad permitidos y los kilómetros recorridos para algunos asegurados que han sufrido un siniestro. Dicho análisis permite detectar pautas concretas que podrían asociarse a un incremento del riesgo.

5.1. Número de siniestros por kilómetro recorrido

En este apartado se muestra una aplicación empírica basada en una muestra de 11.937 asegurados que tienen una póliza PAYD en vigor durante todo el año 2018. Las variables de que se dispone se muestran en la tabla 1. Entre ellas se encuentran las variables clásicas utilizadas en tarificación (edad, antigüedad del carnet, antigüedad y potencia del vehículo y número de siniestros) así como variables telemáticas (kilometraje total recorrido durante el año, porcentaje de kilómetros recorridos en horario nocturno, por vías urbanas y por encima de los límites de velocidad).

En la tabla 2 se muestran los correspondientes estadísticos descriptivos. La edad media de los asegurados es 31,26 años, con 11,09 años de antigüedad media de carnet. La

[5] El PAYD también es conocido como *usage-based insurance (UBI)* cuando incorpora indicadores de uso del vehículo además de la distancia recorrida.

TABLA 1.

DEFINICIÓN DE LAS VARIABLES EN LOS DATOS SOBRE SEGUROS DE AUTOMÓVILES, 2018

Variable	*Descripción*
edad	Edad del asegurado a fecha 01.01.2018
carnet	Antigüedad del carnet de conducir (en años) a fecha 01.01.2018
antigveh	Antigüedad del vehículo (en años) a fecha 01.01.2018
potencia	Potencia del vehículo asegurado (en cc)
km_totales_mil	Distancia total (en miles de kilómetros) recorridos durante todo 2018
noctur	Porcentaje de kilómetros conducidos durante 2018 por la noche
velocidad	Porcentaje de kilómetros conducidos durante 2018 por encima del límite de velocidad
urban	Porcentaje de kilómetros conducidos durante 2018 por vías urbanas
nsin	Número de siniestros por responsabilidad civil con culpa durante 2018

Fuente: Elaboración propia.

antigüedad media del vehículo es de 10,85 años, y la potencia es de 98,33 en promedio. A lo largo de 2018 los asegurados de la muestra han conducido en promedio 8.661 kilómetros. La Figura 1 muestra el histograma del kilometraje total, que mide la exposición al riesgo, y que como puede apreciarse tiene una marcada asimetría positiva. Se tiene además que un 16.04 % de los kilómetros totales recorridos durante 2018 se han circulado por la noche, un 26.57 % en vías urbanas y un 0.31 % por encima del límite de velocidad permitida. En promedio han tenido 0.053 siniestros de responsabilidad civil con culpa durante 2018.

TABLA 2.

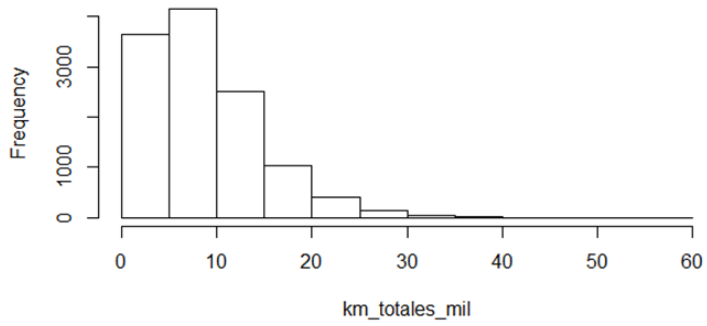
ESTADÍSTICOS DESCRIPTIVOS DE LAS VARIABLES EN LOS DATOS SOBRE SEGUROS DE AUTOMÓVILES, 2018

Variable	*Media*	*Desv. est.*	*Mínimo*	*Máximo*
edad	31,256	4,809	17,000	82,000
carnet	11,095	4,531	0,000	53,000
antigveh	10,859	4,751	0,085	46,118
potencia	98,334	27,554	23,000	418,000
km_totales_mil	8,662	5,983	0,001	56,639
noctur	16,037	16,862	0,000	100,000
velocidad	0,310	0,390	0,000	1,940
urban	26,572	16,497	0,000	100,000
nsin	0.053	0.235	0,000	2,000

Fuente: Elaboración propia.

FIGURA 1

HISTOGRAMA DEL KILOMETRAJE TOTAL DE 2018 EN MILES



Fuente: Elaboración propia.

En la tabla 3 se muestran los resultados de la estimación de los tres modelos de Poisson propuestos para el número de siniestros. El primero de ellos (modelo 1) incluye como variables explicativas únicamente las variables clásicas utilizadas en tarificación. Se ha excluido la edad, dado que está muy correlacionada con la antigüedad del carnet y esta última tiene un efecto más claro a la hora de explicar el número de siniestros. El segundo de ellos (modelo 2) incluye además de las variables clásicas, las variables telemáticas. Finalmente, el modelo 3 incluye las mismas variables explicativas que el modelo 2, pero además añade como offset el logaritmo del kilometraje anual en miles.

Respecto al modelo 1, se observa que la única variable con efecto significativo a la hora de explicar la siniestralidad es la antigüedad del carnet. En concreto, a más antigüedad, el número esperado de siniestros disminuye. Respecto al modelo 2, se observa (además del efecto significativo de la antigüedad del carnet) que los porcentajes de conducción nocturna y por vía urbana tienen un efecto significativo a la hora de explicar la siniestralidad. En concreto, conducir por la noche o por vía urbana se asocia a un mayor número de siniestros. Al añadir el logaritmo del kilometraje total (en miles) como offset, se producen algunos cambios en el efecto del resto de variables explicativas. La antigüedad del carnet deja de tener efecto significativo, en cambio sí lo tiene la antigüedad del vehículo, de manera que cuanto más antiguo es el vehículo mayor es el número esperado de siniestros. Por otra parte, respecto a las variables telemáticas, únicamente la conducción urbana tiene un efecto significativo a la hora de explicar la siniestralidad, de manera que, a más conducción urbana, mayor es el número de siniestros. Por lo que respecta a la bondad de ajuste de los modelos, se concluye que la inclusión de las variables telemáticas mejora el ajuste, dada la reducción que se produce en el valor del

AIC (criterio de información de Akaike) al pasar del modelo 1 al 2. Se observa además que la inclusión del logaritmo del kilometraje total como offset del modelo mejora el ajuste (el AIC vuelve a reducirse al pasar el modelo 2 al 3).

TABLA 3.

RESULTADOS DE LA ESTIMACIÓN DE LOS MODELOS DE REGRESIÓN DE POISSON PARA SINIESTROS CON CULPA SOBRE SEGUROS DE AUTOMÓVILES, MUESTRA DE 2018 (N=11,937)

Parámetro	*Modelo 1*	*Modelo 2*	*Modelo 3 (con* offset*)*
constante	-2,895***	-3,205***	-6,071***
sexo	0,110	0,079	0,024
carnet	-0,026**	-0,023*	-0,019
antigveh	-0,001	-0,003	0,024**
potencia	0,0020	0,002	0,002
noctur	–	0,007**	0,004
velocidad	–	0,071	-0,036
urban	–	0.006*	0,028***
ln(km_totales_mil)	–	–	1,000
AIC	5.019,715	5.011,130	4.996,762

Notas: *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05.
Fuente: Elaboración propia.

De los anteriores modelos deducimos que la inclusión de la información telemática mejora el ajuste del modelo.

5.2. Modelos de predicción de kilómetros e indicadores mediante cuantiles

En este apartado, ajustamos un modelo gamma para predecir la media del kilometraje anual en función de una serie de variables explicativas. Así mismo, ajustamos también un modelo de regresión cuantílica para los diferentes percentiles del kilometraje anual, en particular los percentiles 5 %, 25 %, 50 %, 75 % y 95 %. Los resultados se muestran en la tabla 5.

Por lo que respecta al modelo gamma que estima la media del kilometraje anual, todas las variables explicativas tienen efecto significativo. En concreto, los hombres realizan más kilómetros que las mujeres, por lo que están más expuestos al riesgo de sufrir un accidente. La potencia del vehículo, la conducción nocturna y por encima de los límites de velocidad se asocian también a un mayor kilometraje anual. Por otro lado, a medida que aumenta la conducción por vía urbana, así como la antigüedad del carnet de conducir y del vehículo, se reduce el kilometraje anual.

TABLA 4.

RESULTADOS DE LA ESTIMACIÓN DE LOS MODELOS DE REGRESIÓN GAMMA Y CUANTÍLICA A NIVELES 5 %, 25 %, 50 %, 75 % Y 95 % PARA EL TOTAL DE KILÓMETROS RECORRIDOS, MUESTRA DE 2018 (N=11,937)

		Percentil				
	Media	*5 %*	*25 %*	*50 %*	*75 %*	*95 %*
constante	9,717***	3205,329***	9380,727***	12977,480***	16678,573***	22657,529***
sexo	0,040***	56,727	169,423	233,063*	591,666***	1265,352***
carnet	-0,003*	-15,701*	-28,747*	-21,388	-19,840	-16,449
antigveh	-0,030***	-87,374***	-210,693***	-233,622***	-235,848***	-209,408***
potencia	0,001***	-0,104	1,555	3,977*	7,456**	15,803*
urban	-0,022***	-26,975***	-91,363***	-131,313***	-171,819***	-231,402***
noctur	0,005***	-5,213***	23,124***	39,614***	54,663***	66,698***
velocidad	0,094***	884,493***	971,182***	841,673***	733,010***	911,299**
AIC	234.724	238.928	236.640	236.947	240.428	250.092

Notas: *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05.

Fuente: Elaboración propia.

Respecto a los modelos de regresión cuantílica, se observa que el sexo únicamente resulta significativo para la estimación de percentiles elevados, y su efecto siempre es el mismo: ser hombre se asocia a valores más elevados del kilometraje o exposición al riesgo. Por otro lado, la antigüedad del carnet tiene efecto significativo y negativo para percentiles bajos, en concreto, a más antigüedad del carnet menores son los percentiles 5 % y 25 % del kilometraje total. La antigüedad del vehículo tiene efecto significativo y negativo para cualquier percentil, por tanto, se asocia a una reducción del kilometraje. La potencia del vehículo solo resulta tener efecto significativo para la estimación de percentiles elevados de la exposición al riesgo, en concreto a mayor potencia mayores son los percentiles 50 %, 75 % y 95 % del kilometraje. La conducción por vía urbana siempre se asocia con menores valores de los percentiles del kilometraje total, mientras que la conducción nocturna y por encima de los límites de velocidad se asocian con un mayor nivel de exposición al riesgo (excepto en el caso de la conducción nocturna para el percentil 5 %, que tiene efecto contrario).

El análisis realizado permite detectar qué características influyen en exposiciones extremas al riesgo (percentil 95 %). En concreto, se identifica a los hombres con vehículos nuevos y potentes, que circulan poco por vías urbanas, pero además con mayor proporción en horario nocturno y por encima del límite de velocidad, como el segmento más expuesto al riesgo y por lo tanto el de mayor peligrosidad.

5.3. Curvas de referencia, tiempo/kilómetros

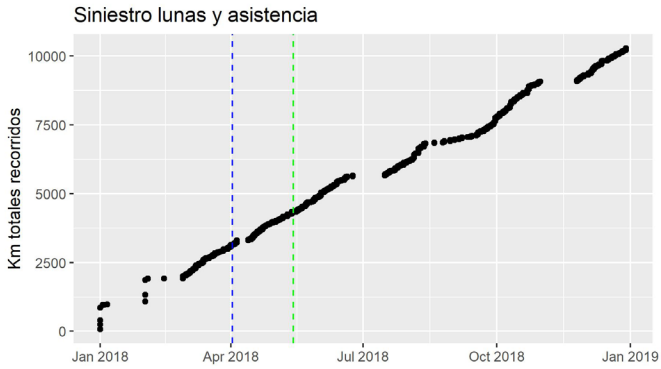
En este apartado se muestran algunas curvas de referencia para conductores de la muestra. Dichos diagramas permiten realizar un *tracking* de los conductores. En la figura 2

se muestra el total de kilómetros acumulados a lo largo del año 2018 y en la figura 3, se ve la acumulación de kilómetros recorridos con excesos de velocidad, cuyo patrón cambia notablemente en la segunda mitad del año.

El análisis longitudinal de los datos telemáticos permite establecer indicadores de riesgo sobre cambios inesperados de comportamientos y anomalías, pudiéndose utilizar como herramientas de prevención ante actitudes al volante que se asocian a mayor accidentabilidad.

FIGURA 2

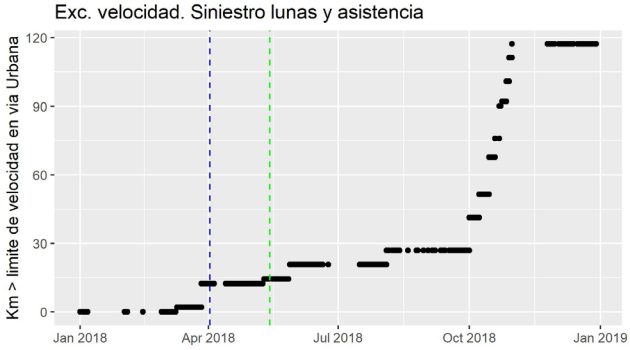
CONDUCTOR CON DOS SINIESTROS, TOTAL DE KILÓMETROS RECORRIDOS ACUMULADOS A LO LARGO DE 2018. LOS PERIODOS DE NO UTILIZACIÓN DEL VEHÍCULO SE ENCUENTRAN EN BLANCO



Fuente: Elaboración propia.

FIGURA 3

CONDUCTOR CON DOS SINIESTROS, TOTAL DE KILÓMETROS RECORRIDOS ACUMULADOS CON EXCESO DE VELOCIDAD EN VÍA URBANA A LO LARGO DE 2018. LOS PERIODOS DE NO UTILIZACIÓN DEL VEHÍCULO SE ENCUENTRAN EN BLANCO



Fuente: Elaboración propia.

6. OTRAS APLICACIONES DEL BIG DATA EN LOS SEGUROS

Para finalizar, realizamos un breve resumen de aportaciones en el ámbito del big data y los seguros, iniciando nuestro recorrido en un trabajo de Bologa, Bologa y Florea (2013) donde el análisis se centra en la detección del fraude. Efectivamente, la detección de comportamientos sospechosos ha sido una de las mayores preocupaciones de los aseguradores desde el inicio del siglo XXI, por cuanto supone un incremento de los costes. La efectividad de los métodos para identificar relaciones en redes dió un enorme impulso a la disciplina porque se demostró capaz de detectar núcleos de generación de fraude. Por ejemplo, talleres que de acuerdo con los asegurados incrementaban los costes de reparación. En el trabajo mencionado, se daban casos relacionados con el uso y recobro de asistencia sanitaria que nunca se utilizó, los autores concluyeron que la mejor estrategia era la combinación de la investigación de siniestros tradicional y los modelos predictivos.

Algunas contribuciones como Zhang (2017) no mejoran especialmente la capacidad predictiva de los modelos pero sí la ganancia en tiempo con la paralelización de algoritmos en el tipo de bases de datos de siniestros en el que se tienen varios millones de pólizas. En otros casos se propone la combinación de métodos como Lin *et al.* (2017) para el análisis de la siniestralidad o para la proyección de los beneficios futuros de un cliente de una entidad de seguros (Fang, Jiang Song, 2016).

Porrini (2017) plantea los grandes retos de los datos masivos que manejan las aseguradores, desde la perspectiva del marco regulatorio europeo. En su trabajo expone los principales elementos de preocupación: privacidad de los asegurados (los datos pueden contener información sensible), discriminación (los factores de tarificación pueden favorecer a determinados tipos de ciudadanos y perjudicar a otros) e impedimentos a la competencia (un mal uso de los datos puede dar ventajas a una empresa frente a otra). Arumugam y Bhargavi (2019) abordan el uso de datos masivos en seguros con un sistema de cálculo de primas a nivel teórico, pero a diferencia de Zhang *et al.* (2018) que sí establecen una correlación entre un indicador de riesgo y la ocurrencia de accidentes, los primeros no llegan a mostrar la implementación.

La pregunta más inquietante la lanzan Barry y Charpentier (2020) al cuestionar si el big data va a cambiar el sector asegurador. Hasta el momento, parece que los viejos modelos persisten y tener más datos significa solamente un incremento de los indicadores de riesgo más que un cambio de enfoque que revolucione el seguro. Meyers y Hoyweghen (2020) relatan con detalle un reciente experimento realizado en Bélgica para hallar evidencias de la relación entre la mejora de la conducción y la disminución de la siniestralidad mediante el uso de un gran volumen información sobre los asegurados. El estudio relata algunas dificultades como la mala calidad de recogida de datos a través del Smartphone, la comunicación con los asegurados (qué indicadores habría que darles) y finalmente, cómo aconsejarles, cuándo y cómo. Dichos autores alientan a

superar las dificultades de tarificar en base al comportamiento, es decir a los patrones de conducción observados en los asegurados.

7. CONCLUSIONES

Nos encontramos en una nueva era tanto por lo que respecta al fundamento de los seguros como a la reorientación de su finalidad.

La utilización de elementos telemáticos permite un mayor conocimiento del riesgo y los modelos de predicción de riesgos son buena muestra de la capacidad que tal información proporciona en la mejora de la modelización.

Los modelos de predicción del riesgo tienen una orientación diferente a los tradicionales y permiten detectar factores asociados a indicadores de mayor peligrosidad. En este sentido, el empleo de la regresión cuantílica frente a la metodología de modelos lineales generalizados presenta algunas ventajas por cuanto que permite conocer las causas del comportamiento de la variable respuesta en escenarios extremos, como los de mayor siniestralidad. Esto puede ayudar a la adopción de medidas preventivas que reduzcan la siniestralidad.

Quedan por resolver numerosos aspectos de la modelización del riesgo, como la elección del nivel de tolerancia o la propia medida de riesgo, dado que es bien sabido que dicha elección es crucial para la posterior interpretación de resultados.

Para finalizar, debemos añadir que en el entorno de los datos masivos, preocupan tres aspectos: los algoritmos de aprendizaje y su aceleración, la estabilidad estructural de los resultados, es decir, cada cuanto tiempo es válido un modelo que se alimenta constantemente de datos y, finalmente, la depuración de la información o el uso de indicadores sintéticos por cuanto deben ser capaces de recoger la esencia de aquello que están midiendo.

Referencias

Arumugam, S. y Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data,* 6(1), pp. 86.

Barry, L. y Charpentier, A. (2020). Personalization as a promise: Can Big Data change the practice of insurance?. *Big Data & Society,* 7(1). https://doi.org/10.1177/2053951720935143

Bologa, A. R., Bologa, R. y Florea, A. (2013). Big data and specific analysis methods for insurance fraud detection. *Database Systems Journal,* 4(4), pp. 30-39.

Fang, K., Jiang, Y. y Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. *Computers & Industrial Engineering,* 101, pp. 554-564.

Lin, W., Wu, Z., Lin, L., Wen, A. y Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee Access,* 5, pp. 16568-16575.

Meyers, G. y Hoyweghen, I. V. (2020). 'Happy failures': Experimentation with behaviour-based personalisation in car insurance. *Big Data & Society,* 7(1). https://doi.org/10.1177/2053951720914650

Porrini, D. (2017). Regulating Big Data effects in the European insurance market. *Insurance Markets and Companies,* 8, pp. 6-15.

Zhang, Y. (2017). Bayesian analysis of big data in insurance predictive modeling using distributed computing. *ASTIN Bulletin,* 47(3), pp. 943-961.

Zhang, H., Xu, L., Cheng, X., Chen, W. y Zhao, X. (2017). Big data research on driving behavior model and auto insurance pricing factors based on UBI. En: *International Conference On Signal And Information Processing, Networking And Computers* (pp. 404-411). Singapore: Springer.

CAPÍTULO V

La web corporativa y la supervivencia empresarial

Desamparados Blázquez*
Josep Doménech
Ana Debón

En la era digital, los sitios web corporativos son un potente canal de comunicación para cualquier empresa. Mantener un sitio web en línea implica una serie de costes, de forma que las empresas solamente invertirán recursos a este fin si son activas y financieramente saludables. Si una empresa atraviesa dificultades financieras, es probable que esto se manifieste en su web como una falta de actualizaciones o dejando de funcionar. Este capítulo analiza en qué medida los cambios en los sitios web corporativos reflejan la supervivencia empresarial. Hemos monitorizado los cambios en los sitios web de un panel de empresas entre los años 2008 y 2014. Los resultados muestran que los cambios en los sitios web corporativos reflejan fielmente el estado de las empresas. Así, hemos aportado una nueva fuente de información sobre la demografía empresarial.

Palabras clave: demografía empresarial, sitios web corporativos.

* Este trabajo ha sido financiado parcialmente por el proyecto PID2019-107765RB-I00 de la Agencia Estatal de Investigación.

1. INTRODUCCIÓN

La demografía empresarial es un aspecto clave de la economía, siendo objeto de la atención de gobiernos y decisores políticos. Así, la mayoría de los institutos nacionales de estadística, como Eurostat o el Instituto Nacional de Estadística de España (INE), llevan a cabo encuestas con un alto nivel de detalle para monitorizar la población de empresas, incluyendo su fundación, supervivencia y cierre. El creciente interés en la demografía empresarial radica en su importante papel en el crecimiento económico, productividad y empleo (Eurostat y OECD, 2007).

En la era digital, el relevante papel de Internet en la economía y la sociedad, junto con el desarrollo de sistemas computacionales avanzados, están abriendo nuevas vías para monitorizar la evolución de las actividades económicas (Blazquez y Domenech, 2014; Vaughan, 2014) y, así también, de la demografía empresarial. Internet y la World Wide Web (WWW) se han convertido en herramientas básicas en la vida diaria tanto de individuos como de empresas, y su importancia está en crecimiento tanto en países desarrollados como en desarrollo. Por un lado, para los consumidores la WWW constituye una herramienta útil para encontrar información sobre productos o servicios, y para comprarlos *online* si esta opción está disponible. Por otro lado, para las empresas, la WWW se presenta como un canal poco costoso tanto para ofrecer información sobre sus productos, servicios y actividades, como para realizar transacciones con sus clientes con mayor flexibilidad y agilidad.

En este contexto, las empresas han desarrollado masivamente sus sitios web corporativos para tener una presencia oficial en los canales digitales. En el caso de España, en el que se enmarca este estudio, en el año 2015 el 75% de las empresas ya estaban presentes en la WWW (INE, 2016).

Los sitios web corporativos constituyen la representación más formal y oficial de las empresas en Internet. Generalmente, las empresas los utilizan para describir sus actividades principales, sus productos y/o servicios, y sus planes, como, por ejemplo, la asistencia a ferias comerciales o la expansión a otros países. Por tanto, los contenidos de los sitios web corporativos están necesariamente conectados con la actividad de las empresas. Esta relación ha sido estudiada recientemente desde distintas perspectivas. Por ejemplo, se ha comprobado que los contenidos de los sitios web corporativos están relacionados, esto es, reflejan, entre otras cosas, la adopción de nuevas tecnologías en las empresas (Arora *et al.,* 2013; Youtie *et al.,* 2012), la realización de actividades innovadoras (Arora *et al.,* 2016; Gök, Waterworth y Shapira, 2015), el crecimiento empresarial (Li *et al.,* 2018) y la orientación exportadora empresarial (Blazquez y Domenech, 2014 y 2018).

Dado que, en efecto, las actividades de las empresas emergen en sus sitios web corporativos, surge la incógnita sobre si la inactividad de las empresas también se manifiesta

en sus sitios web. Mantener un sitio web funcionando en la red implica un serie de costes, como, por ejemplo, los pagos relacionados con el registro de nombres de dominio o con el mantenimiento del servidor donde se encuentre alojado el sitio web. Además, los costes se incrementan si las empresas desean mantener los contenidos y tecnologías de sus sitios web actualizados. Así, se espera que las empresas activas actualicen regularmente sus sitios web para incluir la descripción de nuevos productos y/o servicios, renovar el diseño del sitio web e incluir nuevas funcionalidades, o informar sobre ofertas o promociones. Dado que mantener un sitio web actualizado requiere que las empresas movilicen recursos (financieros, humanos o de ambos tipos), sería plausible que únicamente las empresas activas y financieramente saludables inviertan recursos a dicho fin. De este modo, si una empresa pasa a ser inactiva, es probable que esta situación se manifieste a través de su sitio web mediante una falta de actualizaciones, o directamente dejando de funcionar.

Hasta el momento, la mayor parte de las investigaciones sobre la supervivencia empresarial se ha centrado en los factores que contribuyen a mantener a las empresas activas. Factores como la edad de la empresa, su tamaño, productividad o rentabilidad, han sido muy estudiados y se consideran determinantes para la supervivencia de las empresas. Sin embargo, a pesar del importante papel que juega la WWW en el actual contexto empresarial, no existen estudios que relacionen la WWW con la supervivencia empresarial. Los sitios web corporativos son una fuente actualizada de información empresarial dado que son públicamente accesibles y ofrecen datos empresariales con un alto nivel de detalle, que además suelen ser actualizados con regularidad. Por estos motivos, se han empleado anteriormente para analizar algunas actividades y estrategias empresariales. Así, nuestra intención con este estudio es contribuir en esta línea, demostrando cómo la inactividad empresarial se manifiesta en los sitios web corporativos, discutiendo las implicaciones en términos de monitorización y formulación de políticas que ello conlleva.

En concreto, este estudio analiza en qué medida los sitios web corporativos son capaces de reflejar la supervivencia empresarial. Si una empresa atraviesa por dificultades financieras o muere, es muy probable que su sitio web deje de funcionar, hecho que ocurriría poco antes o después de que la empresa deje de estar activa. Si existe esta relación, entonces el estado del sitio web (fuera de servicio, sin modificar o actualizado), que se puede obtener y monitorizar de forma poco costosa, podría ayudar a realizar un seguimiento regular de la supervivencia empresarial. Para evaluar nuestra propuesta, hemos monitorizado y analizado los cambios en los estados de las empresas (activas o inactivas) y en sus sitios web corporativos para un periodo de siete años. Los resultados muestran con claridad que el estado de los sitios web corporativos refleja el estado de las empresas.

El artículo se estructura en otras cuatro secciones que siguen a esta introducción. En la siguiente sección, se presenta una revisión de la bibliográfica sobre la supervivencia

empresarial y sobre la detección de actividades económicas mediante análisis web. La tercera sección describe los datos que se han empleado y la metodología empleada para realizar el análisis empírico. La cuarta sección describe los resultados obtenidos, incluyendo un amplio análisis sobre las estimaciones de los modelos empleados. En la última sección, se exponen las conclusiones del estudio.

2. ANTECEDENTES

En esta sección se revisa la literatura relevante sobre el análisis de la supervivencia de las empresas y la detección de actividades económicas en el ámbito web. En primer lugar, se realiza un análisis de las publicaciones relacionadas con la supervivencia empresarial y se describen las variables relativas a las empresas que los investigadores suelen tener más en cuenta. Después, se describe cómo se pueden detectar las actividades empresariales a través de la web y otros datos en línea. Esto da pie a investigar la posibilidad de que la inactividad de una empresa pueda también detectarse a través de su sitio web.

2.1. Análisis de supervivencia de empresas

Por su grado de implicación en el éxito empresarial, en la estabilidad económica y en el crecimiento, la supervivencia de las empresas es un tema de enorme interés para los investigadores. Sin embargo, no fue hasta la década de los 90, impulsada por la creciente globalización económica, cuando la comunidad académica comenzó a realizar estudios y análisis al respecto. Las empresas empezaban a hacer frente a nuevos desafíos en un entorno más complejo y turbulento, por lo que se hacía más necesario que nunca determinar las características o acciones que podrían ayudarlas a sobrevivir.

El importante trabajo realizado por Evans (1987), junto con otros posteriores como Audretsch (1991), Mata y Portugal (1994) y Geroski (1995), ayudaron a ampliar el campo de la supervivencia de las empresas, con especial atención en el análisis sistemático de los factores industriales y empresariales que afectaban a la supervivencia, así como en qué dirección lo hacían.

En lo referente a las variables estructurales de las empresas, se han llevado a cabo amplias investigaciones acerca del tamaño y la antigüedad de las mismas, pues se consideran hechos estilizados relacionados con su supervivencia (Geroski, 1995). En general, se ha observado que el tamaño de una empresa aumenta su probabilidad de supervivencia, sobre todo si es de reciente creación (Agarwal y Audretsch, 2001; Cefis y Marsili, 2005; Geroski, Mata y Portugal, 2010). Las empresas más grandes suelen disponer de más recursos financieros y humanos, así como de una estructura sólida una vez alcanzan un determinado nivel de producción. Estos factores puedencontribuir a reducir su riesgo de mortalidad.

La antigüedad de las empresas sigue un patrón similar, ya que se ha observado que incrementa su probabilidad de supervivencia (Audretsch, Houweling y Thurik, 2000; Manjón-Antolín y Arauzo-Carod, 2008). Las empresas de mayor antigüedad han tenido la posibilidad de adquirir experiencia en cuanto al funcionamiento del mercado y las estrategias que les son más rentables, lo que puede ayudarles a sobrevivir si se comparan con las de nueva creación. Es decir, por lo general, la experiencia tiene una incidencia positiva en la supervivencia de las empresas.

Otras variables estructurales cuya relación con la supervivencia de las empresas han estudiado de manera persistente los investigadores son la estructura de deuda de las empresas, su nivel de productividad y su nivel de rentabilidad (Audretsch, Houweling y Thurik, 2000; Delmar, McKelvie y Wennberg, 2013; Görg y Spaliara, 2014). Estas variables guardan una estrecha relación con el nivel de éxito, la estabilidad y la salud de las empresas, por lo que son potencialmente influyentes en su probabilidad de supervivencia.

La intensidad tecnológica del sector de actividad en el que operan las empresas también se ha tenido en cuenta en otros estudios sobre la supervivencia empresarial (Esteve-Pérez y Mañez-Castillejo, 2008; Giovannetti, Ricchiuti y Velucchi, 2011). Los primeros resultados señalaron que las empresas tenían más dificultades para sobrevivir en sectores altamente tecnológicos. Sin embargo, más tarde se encontró un patrón opuesto: el suministro de productos y servicios de alta tecnología exige que las empresas desarrollen aptitudes sofisticadas y se centren en la innovación y el conocimiento, factores que pueden contribuir positivamente a su supervivencia, sobre todo en el complejo marco económico actual.

Estudios más recientes siguen proporcionando información sobre cómo los factores estructurales clásicos de las empresas (es decir, el tamaño, la antigüedad o las ratios financieros) y los factores contextuales (como las crisis financieras, la ubicación o el ciclo de vida específico de la empresa) contribuyen a aumentar o disminuir la probabilidad de supervivencia de una empresa (Basile, Pittiglio y Reganati, 2017; Gémar, Moniche y Morales, 2016; Guariglia, Spaliara y Tsoukas, 2016). Los resultados de la mayoría de los trabajos publicados tienen como objetivo principal servir de guía en los procesos de toma de decisiones por parte de los directivos, los cuales pueden hacer uso de esta información para promover o implementar estrategias que puedan contribuir a la supervivencia de las empresas.

Sin embargo, en ninguno de los estudios se analiza la relación entre la supervivencia de las empresas y los sitios web corporativos, cuyo papel en las estrategias empresariales es básico en la sociedad digital actual y se espera que adquiera importancia en el futuro. Si bien los datos contables son útiles para predecir la quiebra de una empresa, no muestran con exactitud la situación operativa y financiera de esta (Astebro y Winter, 2012), por lo que complementarlos con los datos en línea puede ayudar a conocer mejor su estado de salud.

Por todo ello, es importante confirmar en qué medida la situación del sitio web de una empresa está relacionada con el estado de actividad de la misma, además de determinar si la información que se ofrece en él puede utilizarse para evaluar la supervivencia de una empresa. A continuación, se realiza un análisis de las publicaciones acerca de la idoneidad de la web para reflejar la actividad empresarial, que es lo que dio pie a estudiar en este trabajo si también refleja la inactividad de una empresa o, lo que es lo mismo, su muerte.

2.2. Captura de la actividad económica de las empresas con datos de la WWW

Cada minuto del día, miles de personas, empresas y organismos públicos generan, publican y comparten información a través de Internet. Estas actividades en línea dejan una huella digital que se puede rastrear y que, si se procesa y analiza adecuadamente, puede ayudar a describir su comportamiento económico y social.

La detección de patrones de comportamiento y consumo y de actividades económicas y comerciales a través de los datos en línea es un campo de investigación incipiente cuya importancia está empezando a crecer al mismo ritmo que se expande Internet por todo el mundo. Esta expansión generalizada del uso de la red afecta a la forma en la que las empresas llevan a cabo sus actividades y negocios, pues se están viendo obligadas a dar el salto a Internet, dado el actual contexto digital. Para ello, lo que están haciendo es empezar a implantar sitios web, que son una representación oficial de su imagen y que pueden, al mismo tiempo, utilizarse como canal comercial.

De hecho, los sitios web son fuentes importantes de datos en línea cuyo potencial para detectar y supervisar las actividades económicas ha permanecido inexplorado hasta hace poco tiempo. Los sitios web tienen una estructura compleja que difiere de uno a otro, lo que hace que el proceso de extracción, procesamiento y análisis de la información resulte difícil de normalizar y automatizar para permitir la explotación masiva de datos en comparación con las bases de datos tradicionales. No obstante, los sitios web presentan también muchas ventajas, como, por ejemplo, que son accesibles al público, que proporcionan información actualizada y que pueden analizarse en cualquier momento, algo que no permiten las bases de datos tradicionales. En concreto, los sitios web corporativos atraen más atención porque son cada vez más las empresas que los adoptan y utilizan para reflejar sus características, productos y estrategias previstas, por lo que se han convertido en valiosas fuentes de información empresarial. Por todo ello, se están desarrollando metodologías y tecnologías específicas para extraer y analizar los datos web (Munzert *et al.,* 2015).

Los primeros trabajos sobre la detección de información económica o empresarial en los sitios web corporativos se publicaron hace más de una década. Siguiendo un enfoque

no automatizado, Overbeeke y Snizek (2005) captó diferentes dimensiones de la cultura corporativa mediante el análisis del texto y las imágenes de un conjunto de sitios web corporativos, mientras que Meroño- Cerdan y Soto-Acosta (2007) observó que el contenido web externo guardaba relación con el rendimiento de la empresa.

Las estrategias de responsabilidad social y sostenibilidad de las empresas, así como sus niveles de adopción, también se han detectado con éxito en los contenidos de los sitios web corporativos (Gallego Álvarez, García Sánchez y Rodríguez Domínguez, 2008; Tagesson *et al.,* 2009; Tang, Gallagher y Bie, 2015). Esto se ha llevado a cabo, por ejemplo, detectando el número de apariciones de diferentes palabras clave relacionadas con productos ecológicos (Albino, Balice y Dangelico, 2009). Esta medida se ha ampliado y se ha utilizado con éxito en otros estudios enfocados a los sectores de las nuevas tecnologías. En su trabajo, Libaers, Hicks y Porter (2010) encontró seis tipos de modelos de negocio para comercializar las nuevas tecnologías mediante el análisis automático de la frecuencia con la que determinadas palabras clave estaban presentes en los sitios web corporativos de las empresas en estudio.

Siguiendo un enfoque automático, Youtie *et al.* (2012) y Arora *et al.* (2013) aplicaron técnicas de extracción de información de sitios web *(web scraping)* y de análisis de contenido en los sitios web corporativos, incluido el recuento de palabras clave, para hacer un seguimiento de las estrategias de adopción de tecnología de las empresas de los sectores tecnológicos emergentes. La innovación es otro de los temas importantes que se han detectado recientemente a través de técnicas de *web mining.* Gök, Waterworth y Shapira (2015) y Arora *et al.* (2016) detectaron con éxito actividades de innovación de las empresas mediante el análisis de los contenidos de los sitios web corporativos. Por su parte, Li *et al.* (2018) realizó un seguimiento del crecimiento de las ventas de las empresas en un contexto de triple hélice.

El primer intento de generalizar el análisis automático de los sitios web corporativos para descubrir información económica lo presentó Domenech *et al.* (2012). Este trabajo expone una arquitectura de sistema de minería de datos web que gestiona el proceso de rastreo y análisis de los sitios web corporativos, el cual se probó con éxito para encontrar indicadores del tamaño de las empresas basados en la web. Este sistema lo adaptaron Blazquez y Domenech (2018) para detectar la orientación exportadora por parte de las empresas mediante el análisis automático de sus sitios web corporativos, pues un análisis manual previo detectó que los sitios web reflejan potencialmente dicha actividad empresarial (Blazquez y Domenech, 2014).

Sobre la base de investigaciones anteriores, en las que se demostraba que en los sitios web corporativos queda reflejada información económica y las actividades comerciales de las empresas, en este trabajo se formula la hipótesis de que también es posible detectar la inactividad de las empresas mediante el análisis de los datos obtenidos de los sitios web corporativos.

3. DATOS Y METODOLOGÍA

Esta sección describe, en primer lugar, la estructura de los datos empleados en el estudio y el proceso que se ha seguido para obtenerlos. En segundo lugar, describe la metodología de análisis, que se basa fundamentalmente en modelos de regresión logística multiperiodo para detectar la habilidad del estado de los sitios web corporativos para detectar el estado de actividad de las empresas, y en un modelo de duración que ayuda a comprender en mayor profundidad de qué forma el estado de los sitios web corporativos se relaciona con la supervivencia empresarial.

3.1. Datos

La muestra de datos inicial incluía 780 empresas[1] localizadas en España y pertenecientes a una diversidad de sectores económicos, incluyendo manufacturas, servicios y otros (pertenecientes a los códigos 10-95 de la clasificación NACE2[2] Rev.2), siendo todas activas y con sitio web en el año 2008.

La muestra se obtuvo mediante un muestreo aleatorio simple de la base de datos SABI (Bureau van Dijk, 2010), siendo candidatas para formar parte de la muestra todas las empresas de la base de datos que cumplieran cuatro criterios: ser activas, estar localizadas en España, pertenecer a alguno de los sectores de actividad anteriormente mencionados, y tener sitio web corporativo; todo ello, referido al año 2008, momento de partida del estudio. El conjunto completo de datos consiste en un panel de datos económicos y datos web para estas empresas, entre los años 2008 y 2014. La información económica se obtuvo de los registros financieros de las empresas accediendo a una versión más reciente de SABI en enero del año 2016, y el último año para el que se disponía de registros económicos completos era el año 2014, siendo por tanto este el último año que entró en nuestro estudio.

La información de los sitios web corporativos se obtuvo accediendo a ellos mediante la herramienta *Wayback Machine* del Archivo de Internet (Kahle y Gilliat, 2016), que se trata de un repositorio público y de libre acceso que contiene capturas sobre más de 400 billones de sitios web. Esta herramienta captura y almacena sitios web a diario, permitiendo así que los usuarios puedan acceder a ellos y conocer cómo han evolucionado a lo largo del tiempo. Pese a su potencialidad, presenta algunas limitaciones que deben ser tenidas en cuenta: no puede capturar los sitios web que no permiten ser explorados por los rastreadores web mediante el protocolo de exclusión "robots.txt"; su capacidad para capturar contenido flash es limitada; no explora la WWW al completo, de forma que algunos sitios web no son capturados y por tanto, no puede estudiarse

[1] De las 780 empresas, el 92 % eran pequeñas y medianas (pymes), en línea con la estructura productiva de España (DGIPYME, 2017).

[2] Clasificación estadística de actividades económicas de la Comunidad Europea (Eurostat, 2008).

su evolución a lo largo del tiempo; y de entre los sitios web capturados, no todos se capturan de forma frecuente, en algunos casos siendo incluso esta frecuencia inferior a una vez al año. Estas limitaciones impidieron que pudiéramos seguir la evolución de algunos sitios web.

Por estos motivos, aquellas empresas cuyos sitios web no encontramos en *Wayback Machine,* fueron eliminadas de la muestra de datos inicial. Así, obtuvimos finalmente una muestra de 720 empresas para formar parte del estudio, de las cuales 674 sobrevivieron durante los siete años incluidos en el estudio, mientras que las restantes 46 dejaron de estar activas en algún momento. En los análisis que se muestran en las siguientes secciones, solamente se incluyen los años 2010 a 2014 para poder comparar la evolución de los sitios web con respecto al año anterior y para alinear el estado del sitio web con el momento temporal en el cual la información financiera se encuentra. Para tener en cuenta los distintos momentos del tiempo en los cuales los datos empresariales están disponibles, la información de los registros financieros de las empresas se retrasó dos periodos, esto es dos años, en los análisis empíricos. Esto es, es posible conocer el estado de un sitio web en el momento t, pero en ese momento, las cuentas financieras más recientes disponibles corresponden al momento $t-2$. Además, algunas capturas de sitios web no estaban disponibles en algún año t específico. Todo esto dio como resultado un panel no equilibrado con un total de 3.254 observaciones, de las cuales 3.152 corresponden a empresas que sobrevivieron hasta el año 2014, mientras que las restantes 102 observaciones corresponden a empresas que dejaron de estar activas en el periodo estudiado.

Para registrar los cambios en los sitios web corporativos, el procedimiento consistió en buscar en *Wayback Machine* la URL del sitio web de cada empresa y revisar su página de inicio para cada año estudiado. Los cambios observados se codificaron mediante la variable *Estado_web,* que puede tomar cinco valores distintos dependiendo del estado del sitio web o tipo de cambio experimentado cada año. Estos cinco niveles se definen como sigue:

- Código 1: el sitio web está fuera de servicio. Esto incluye sitios web que no funcionan (por ejemplo, en los que aparece el código de error "404 Sitio No Encontrado") o aquellos cuyo nombre de dominio ha expirado o está en venta.

- Código 2: el sitio web permanece inalterado. Esto incluye los casos en los cuales el sitio web permanece exactamente igual con respecto a su captura del año anterior.

- Código 3: el sitio web ha experimentado cambios menores. Estos cambios incluyen la eliminación o incorporación de secciones, opciones, imágenes o contenidos.

- Código 4: el sitio web ha experimentado cambios mayores. Estos cambios hacen referencia fundamentalmente a un nuevo diseño web, de modo que el sitio web sea completamente diferente a su versión del año anterior, lo cual puede implicar, por ejemplo, un cambio en la tecnología empleada para construir el sitio web.

- Código 5: el sitio web no ha sido capturado por *Wayback Machine.* Estos casos se procesaron como datos faltantes y se eliminaron de la muestra final, dado que no era posible determinar el estado del sitio web.

El conjunto de datos también incluía variables económicas que se han relacionado clásicamente con la supervivencia empresarial. Estas variables, junto al estado de las empresas (activas o inactivas), se obtuvieron de la base de datos SABI y se complementaron con datos del *Boletín Oficial del Registro Mercantil (BORME)* para tener en cuenta las fusiones y adquisiciones. Concretamente, se obtuvieron las siguientes variables:

- $Activa_{i,t}$: variable dicotómica que toma el valor 1 si la empresa i está activa en el año t y 0 en caso contrario[3].

- $Tamaño_{i,t}$: variable cuantitativa medida como el logaritmo de número de empleados de la empresa i en el año t. Se utiliza como representación del tamaño de la empresa.

- $Edad_{i,t}$: variable cuantitativa medida como el número de años desde que la empresa i fue fundada hasta el año t. Se utiliza como representación de la experiencia de la empresa.

- $Deuda_{i,t}$: variable cuantitativa medida como el porcentaje de deuda de la empresa i en el año t.

- $Productividad_{i,t}$: variable cuantitativa medida como el valor añadido por empleado (en millones de euros) de la empresa i en el año t.

- $Rentabilidad_{i,t}$: variable cuantitativa medida como la ratio de la rentabilidad económica de la empresa i en el año t. Esta ratio, denominado *ROA* por sus siglas en inglés *(Return On Assets),* se obtiene dividiendo el beneficio de explotación entre los activos totales.

- $Alta\_tecnología_{i,t}$: variable dicotómica que toma el valor 1 cuando la actividad económica de la empresa i en el año t se considera de alta o media intensidad tecnológica siguiendo la clasificación de Eurostat (Eurostat, 2014), y 0 en caso contrario.

[3] Hemos considerado como inactivas a las empresas que presentaban los siguientes estados: en extinción; en disolución; en liquidación; en un concurso de acreedores finalizado, esperando a una disolución o liquidación ya ordenada; y en un concurso de acreedores en marcha (si este es el estado más reciente y no se dispone de información adicional), excepto si la empresa ha sido objeto de fusión o adquisición (Eurostat y OECD, 2007).

3.2. Regresión logística multiperiodo

En esta primera aproximación, la supervivencia de las empresas se estudió mediante modelos de regresión logística para múltiples períodos. Estos modelos nos sirven para examinar cómo algunas variables independientes están relacionadas con una variable dependiente cuando los datos utilizados como entrada incluyen individuos observados a lo largo del tiempo, que es el caso del presente estudio, y que se han aplicado con éxito en anteriores estudios de supervivencia de empresas (Bridges y Guariglia, 2008; Jacobson y Schedvin, 2015).

La variable dependiente en este estudio es $Active_{i,t}$ que indica si la empresa está activa o no, así pues la regresión logística es adecuada para analizar su relación con las covariables por ser una variable dicotómica. Además, los modelos utilizados incluyen efectos fijos temporales para dar cuenta de la cambiante situación económica y política que afecta a la probabilidad base de estar activo cada año. Analíticamente, el modelo se representa como:

$$\theta_{i,t} = \ln\left(\frac{P\left(y_{i,t}=1\right)}{1-P\left(y_{i,t}=1\right)}\right) = \beta' X_{i,t} + \gamma_t \qquad [1]$$

donde $\theta_{i,t}$ es el logit, $P\left(y_{i,t}=1\right)$ es la probabilidad de ocurrencia de que la empresa este activa, valor '1' de la variable dependiente $y_{i,t}$, β' es el vector de coeficientes de regresión, $X_{i,t}$ es el vector de covariables para la empresa i en el año t, y γ_t son parámetros específicos de tiempo que reflejan los eventos no observables que afectan a todas las empresas cada año.

Este modelo se utiliza para evaluar primero la relación entre el WWW y si las empresas están o no activas, ya que estima la probabilidad de que una empresa esté activa dado el estado de su sitio web en una primera especificación, y el estado de este sitio web y una serie de variables económicas en una segunda especificación. Ambas especificaciones del modelo controlaron la coyuntura económica o el efecto del período al incluir variables ficticias para cada año considerado en el estudio. En consecuencia, el primer modelo se definió de la siguiente manera:

$$\theta_{i,t} = \ln\left(\frac{P\left(Activa_{i,t}=1\right)}{1-P\left(Activa_{i,t}=1\right)}\right) = \beta_0 + \alpha Estado\_web_{i,t} + \gamma_t \qquad [2]$$

donde $P(Activa_{i,t}=1)$ es la probabilidad de que la empresa i esté activa en el año t, y el logit, $\theta_{i,t}$ se explica en base a las variables $Estado\_web_{i,t}$ y los efectos fijos temporales, capturados por γ_t.

Se especificó un modelo extendido al incluir también las variables económicas de las empresas que pueden afectar la supervivencia de la empresa de acuerdo con la litera-

tura. Las variables que finalmente se seleccionaron fueron aquellas que variaron con un nivel de significación admisible ($p<0.05$) entre ambos grupos de empresas y que no estaban altamente correlacionadas. Esta segunda especificación se definió de la siguiente manera:

$$\theta_{i,t} = \ln\left(\frac{P\left(Activa_{i,t}=1\right)}{1-P\left(Activa_{i,t}=1\right)}\right) \quad [3]$$
$$= \beta_0 + \alpha Estado\_web_{i,t} + \beta' Z_{i,t-2} + \rho Alta\_tecnologia_{i,t-2} + \gamma_t$$

donde $P(Activa_{i,t}=1)$ es la probabilidad de que la empresa i esté activa en año t, y el logit, $\theta_{i,t}$ se explica en base a la variable de la web corporativa, $Estado\_web_{i,t}$, el vector de variables cuantitativas económicas $Z_{i,t-2}$ que incluye $Tamaño_{i,t-2}$, $Deuda_{i,t-2}$, $Productividad_{i,t-2}$ y $Rentabilidad_{i,t-2}$, la variable categórica económica $Alta\_tecnología_{i,t-2}$ y los efectos fijos temporales, capturados por γ_t.

Una vez confirmada la relación entre el WWW y el estado de las empresas con ambas regresiones, se aplica un modelo de duración para estimar probabilidad de sobrevivir de la empresa un período de tiempo o más dado el estado del sitio web corporativo.

3.3. Análisis de supervivencia

La relación del estado del sitio web de la empresa con su duración, esta última definida como el tiempo transcurrido (durante el período observado) hasta que una empresa aparece no activa, se analizó a través de modelos de supervivencia (también conocidos como modelos de duración (Lancaster, 1990)). Estos modelos son útiles para predecir eventos como fallos o muertes en un sujeto (por ejemplo, empresa, máquina, sistema, producto o paciente). Específicamente, el tiempo y otras variables predictivas se consideran para estimar el riesgo de fallo o muerte durante un período de tiempo particular.

En el análisis de supervivencia, la función de riesgo $h(t)$ es la que se usa para realizar regresiones. En este estudio, se estimó la función de riesgo a través de un modelo lineal generalizado de cloglog, que es el equivalente a la versión de tiempo discreto del modelo de riesgo proporcional de Cox (*Cox Proportional Hazard Model*) (Jenkins, 1995). Este modelo se ha aplicado con éxito en anteriores estudios de supervivencia de empresas para datos recopilados anualmente (Görg y Spaliara, 2014; Guariglia, Spaliara y Tsoukas, 2016; Tsoukas, 2011), que es nuestro caso. El modelo de riesgo proporcional supone que la tasa de riesgo depende solo del tiempo en riesgo, $h_0(t)$ (el riesgo de referencia) y de la matriz de variables explicativas, X. Esta es la rapidez a la que las empresas mueren en el año t, siempre que hayan sobrevivido el año anterior, $t-1$. Se expresa como:

$$h\left(t,X\right) = h_0\left(t\right)\exp\left(\beta' X\right) \quad [4]$$

En particular, la función de riesgo de tiempo discreto (con efectos temporales específicos) se especifica de la manera siguiente:

$$h\left(t,X\right)=1-\exp\left[-\exp\left(\beta' X+\gamma_t\right)\right] \quad [5]$$

donde β' es el vector de coeficientes de los regresores que describe como la función de riesgo varia en respuesta a la matriz de variables explicativas o covariables X, y γ_t que captura los efectos temporales específicos sobre el riesgo.

Para este estudio, este modelo de duración se especificó de la siguiente manera:

$$h(t,\ Estado\_web)\ =\ 1\ -\ \exp[-\ \exp(\beta_0 + \alpha Estado\_web + \gamma_t)], \quad [6]$$

donde *h(t,Estado_web)* es la ratio de riesgo (*hazard rate*); es decir, la ratio a la que las empresas se vuelven inactivas en el momento t siempre que estuvieran activas en el año $t - 1$, que se modela a través de la variable explicativa *Estado_web* y el efecto temporal específico, γ_t.

4. RESULTADOS

Esta sección muestra, en primer lugar, una serie de gráficas, estadísticas descriptivas y comparaciones entre grupos para dar una visión general de las características de los datos. En segundo lugar, se muestran y comparan dos modelos de regresión logística multiperiodo para evaluar hasta qué punto la variable *Estado_web* captura el estado de actividad de la empresa. Finalmente, se complementan estos resultados con un modelo de duración.

4.1. Estadística descriptiva y comparaciones entre grupos

Las estadísticas descriptivas del conjunto de datos se muestran en la tabla 1. Como se puede observar, el conjunto está predominante formado por empresas activas (96,9 % de la muestra), que pertenecen a sectores de baja intensidad tecnológica (81 %) y cuyo nivel de deuda es moderado (61 %). La tabla 1 deja patente que no hay una alta correlación entre ningún par de variables, por tanto no hay riesgo de redundancia en la información o multicolinealidad a la hora de estimar los modelos de regresión.

Tras comprobar gráficamente la existencia de esta relación, se ha evaluado también el periodo temporal en el cual la inactividad de la empresa se refleja en el sitio web. La figura 1 muestra el estado de los sitios web como función del año en el cual las empresas pasan a estar inactivas (t), abarcando desde dos años antes de este momento ($t - 2$) hasta dos años después ($t + 2$). Por un lado, se puede observar que la proporción de

TABLA 1.

ESTADÍSTICAS DESCRIPTIVAS Y MATRIZ DE CORRELACIONES

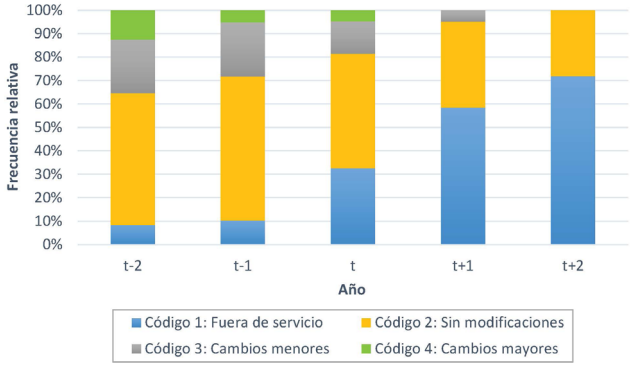
Variable	*Media*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1. Activa	0,969	0,174							
2. Estado_web	2,618	0,868	0,272						
3. Tamaño	3,855	1,241	0,046	0,149					
4. Edad	24,823	14,318	0,007	0,022	0,258				
5. Deuda	60,470	31,608	-0,127	-0,009	0,016	-0,140			
6. Productividad	0,270	4,071	0,007	0,021	-0,129	-0,042	0,046		
7. Rentabilidad	1,099	22,526	0,112	0,046	-0,005	0,005	-0,291	0,044	
8. Alta_tecnología	0,199	0,399	0,076	0,143	0,207	-0,062	-0,048	-0,021	-0,005

Notas: Procedimientos empleados: Coeficiente r de Pearson para parejas de variables cuantitativas; coeficiente biserial puntual para parejas de variable continua y binaria; coeficiente Phi para parejas de variables binarias; Eta para parejas de variable continua y categórica de más de dos niveles (Cohen *et al.,* 2002).

Fuente: Elaboración propia.

FIGURA 1

EVOLUCIÓN EN EL ESTADO DEL SITIO WEB DE LAS EMPRESAS QUE PASAN A SER INACTIVAS EN EL AÑO "t"



Fuente: Elaboración propia.

empresas con sitios web fuera de servicio se incrementa conforme se aproxima el año en el cual la empresa pasa a estar inactiva, pasando de representar un 8 % en el año $t-2$ a un 33 % en el año t. Esta proporción se incrementa hasta alcanzar el 59 % en el año $t+1$, y el 72 % en el año $t+2$. Por otro lado, la proporción de empresas cuyos sitios web han experimentado cambios mayores sigue la tendencia opuesta. Aproximadamente el 13 % de las empresas renovaron por completo sus sitios web dos años antes de pasar a estar inactivas, porcentaje que en el año t solamente supuso un 4,6 % y que a partir de dicho año, es de un 0 %. La evolución en la presencia de cambios menores es similar a la que acabamos de describir. En resumen, es destacable el hecho de que todos los

sitios web de las empresas inactivas están fuera de servicio o sin modificar dos después de que estas se hayan vuelto inactivas.

Así, este análisis revela que el estado de los sitios web corporativos refleja la supervivencia empresarial: cuando una empresa se vuelve inactiva, su sitio web permanece sin modificaciones o pasa a estar fuera de servicio en menos de dos años desde ese momento. Además, un pequeño porcentaje de sitios web pasan a estar fuera de servicio entre uno y dos años antes de la inactividad de la empresa, anticipando que la esta puede estar pasando por dificultades al haber dejado de prestar atención a su sitio web y de invertir recursos en mantenerlo.

Estos hallazgos de forma gráfica, se han confirmado también de forma estadística. La primera columna de la tabla 2 resume el comportamiento de los sitios web corporativos mostrando cómo se distribuye la variable *Estado_web* en la muestra. Se observa que la mayor parte de los sitios web han permanecido sin cambios (37,7 %) o han experimentado modificaciones menores (36,4 %) con respecto al año anterior. Solamente el 8,8 % de las observaciones presentaba un sitio web fuera de servicio, mientras el 17,1 % habían experimentado cambios importantes. Para ilustrar su asociación con otras variables, el valor numérico de la variable *Estado_web* también se ha incluido en la tabla 1.

Para evaluar si las distintas variables muestran un comportamiento distinto dependiendo del estado de la empresa (activa o inactiva), se han empleado técnicas estadísticas de diferencias entre grupos. Para las variables categóricas se ha empleado el test chi-cuadrado de Pearson, cuyos resultados se muestran en la tabla 2. Estos arrojan diferencias significativas en la intensidad tecnológica, estando las empresas activas más asociadas a

TABLA 2.

ESTADÍSTICAS DESCRIPTIVAS DE LAS VARIABLES CUALITATIVAS Y COMPARACIONES ENTRE GRUPOS

	Todas (N=3254) %	Activa *(0) (N=102)* %	Activa *(1) (N=3152)* %	*Chi-cuadrado (valor-p)*
Activa (0)	3,1			
Activa (1)	96,9			
Alta_tecnologia (0)	80,1	97,1	79,5	
Alta_tecnologia (1)	19,9	2,9	20,5	0,000
Estado_web(1)	8,8	50,0	7,5	
Estado_web(2)	37,7	40,2	37,6	
Estado_web(3)	36,4	7,8	37,3	
Estado_web(4)	17,1	2,0	17,6	0,000

Notas: Estado_web(1): Fuera de servicio; *Estado_web*(2): Sin modificaciones; *Estado_web*(3): Modificaciones menores; *Estado_web*(4): Modificaciones mayores.

Fuente: Elaboración propia.

sectores de alta y media intensidad tecnológica que las que se volvieron inactivas (20,5 % *vs.* 2,9 %). En cuanto al estado de los sitios web, las diferencias halladas entre las empresas activas e inactivas también son significativas. Para estas últimas la mayoría de los sitios web estaban fuera de servicio (50 %) o permanecían sin cambios (40,2 %), mientras que solamente en el restante 9,8 % de los sitios web se había realizado alguna modificación, menor o mayor.

En cambio, más de la mitad de los sitios web de las empresas activas presentaban cambios, especialmente cambios menores (37,3 %). Este era un comportamiento esperado dado que el diseño del sitio web forma parte de la imagen corporativa, que la mayor parte de empresas no renuevan con una periodicidad anual. Así, los cambios menores para mantener la información actualizada son los que las empresas activas realizan con mayor frecuencia. Por otro lado, los sitios web fuera de servicio son poco comunes entre las empresas activas (7,5 %). La presencia de sitios web sin modificaciones (37,6 %) es similar al caso de las empresas inactivas, de modo que este estado del sitio web no es tan indicativo del estado de la empresa como los casos en los que hay cambios en el sitio web.

En cuanto a las variables cuantitativas, la normalidad y la homogeneidad de la varianza se comprobó tanto gráficamente como numéricamente. Dado que ninguna de las variables cumplía con ambas asunciones, aplicamos el test no paramétrico de U Mann-Whitney, que se basa en la mediana (Anderson *et al.,* 2014). Estos resultados se muestran en la tabla 3. La mayoría de las variables económicas muestran valores estadísticamente diferentes para las empresas activas y las que pasaron a ser inactivas en algún momento del periodo observado. El logaritmo del número de trabajadores se muestra estadísticamente superior para las empresas activas (3,761 *vs.* 3,401), de forma que el tamaño de la empresa está relacionado hasta cierto punto con su permanencia.

En cuanto a la edad de la empresa, no se han encontrado diferencias significativas, de forma que aparentemente las empresas tienen la misma probabilidad de pasar a ser

TABLA 3.

ESTADÍSTICAS DESCRIPTIVAS DE LAS VARIABLES CUANTITATIVAS Y COMPARACIONES ENTRE GRUPOS

	Activa *(0)* *(N=102)*	Activa *(1)* *(N=3152)*	*U Mann-Whitney* *(valor-p)*
Tamaño	3,761	3,401	0,007
Edad	21,501	22,815	0,875
Deuda	88,990	60,810	0,000
Productividad	30,947	48,787	0,000
Rentabilidad	-3,775	2,000	0,000

Fuente: Elaboración propia.

inactivas independientemente de su edad. La deuda es mucho mayor para las empresas inactivas (88,99 % *vs.* 60,81 %), hecho indicativo del efecto negativo que tienen los altos niveles de deuda sobre la salud financiera de la empresa y, por tanto, en su continuidad. Además, según los resultados, las empresas activas están asociadas con niveles de productividad y rentabilidad significativamente mayores que las inactivas. Los altos niveles de productividad están conectados con un mejor funcionamiento general de la empresa, hecho que contribuiría a tener una mayor rentabilidad. Ambas variables muestran relación con un estado más saludable de las empresas, de forma que tal como se esperaba, las empresas financieramente más saludables mantienen su actividad más frecuentemente.

4.2. Modelos de regresión logística multiperiodo

En esta sección, profundizamos en el papel que juegan los sitios web corporativos sobre la probabilidad de que las empresas sean activas. Primero, hemos construido un modelo de regresión logística multiperiodo basado en la variable $Estado\_web$, como se especifica en la ecuación [2].

La tabla 4 muestra los resultados de la estimación de este modelo, incluyendo la estimación de los coeficientes de regresión (β), la razón de probabilidades (OR), los errores estandarizados (SE), valores z y valores p. La ratio OR, en este caso, es una medida de asociación entre los distintos estados posibles de los sitios web y el estado de las empresas, y se calcula como el exponente de los coeficientes de regresión. Por tanto, una ratio OR superior a 1 indica que la probabilidad de que una empresa sea activa crece en presencia de una variable independiente dada (en este caso, cada estado posible del sitio web). Si la ratio es inferior a 1, entonces dicha probabilidad decrece, mientras que si es igual a 1, entonces no hay asociación entre la variable independiente y la variable dependiente.

TABLA 4.

REGRESIÓN LOGÍSTICA MULTIPERIODO INCLUYENDO EL ESTADO DE LOS SITIO WEB. VARIABLE DEPENDIENTE: *ACTIVA*

Variables	β	*OR*	*SE*	*valor-z*	*valor-p*
(Constante)	3,483	32,557	0,598	5,827	0,000
Estado_web(2)	1,628	5,094	0,227	7,158	0,000
Estado_web(3)	3,423	30,661	0,390	8,786	0,000
Estado_web(4)	3,970	52,985	0,727	5,460	0,000
Observaciones	3.254				
Logaritmo de la verosimilitud	-349,349				

Notas: *Estado_web*(2): Sin modificaciones; *Estado_web*(3): Modificaciones menores; *Estado_web*(4): Modificaciones mayores. Se han incluido variables binarias para recoger el efecto temporal.

Fuente: Elaboración propia.

Para este modelo con datos web, los resultados muestran que los estados observados de los sitios web tienen un efecto estadísticamente significativo sobre la probabilidad de que una empresa sea activa. Conforme la actividad del sitio web aumenta, la probabilidad de que una empresa sea activa también lo hace. La estimación que corresponde al estado del sitio web "Sin modificaciones" (Código 2) es positiva, lo que indica que tener un sitio web que funciona, incluso aunque sus contenidos o apariencia no se hayan modificado en un año, incrementa la probabilidad de que una empresa sea activa con respecto a tener un sitio web fuera de servicio (Código 1, que se ha tomado como nivel base frente al que comparar). De hecho, la probabilidad de que una empresa con un sitio web sin modificar sea activa es 5 veces (o un 409,4 % superior) la probabilidad que tiene una empresa cuyo sitio web está fuera de servicio, como indica la ratio OR.

Actualizar los sitios web con menor (Código 3) o mayor (Código 4) alcance incrementa la probabilidad de que la empresa sea activa, como esperábamos. Además, el incremento que se produce es extraordinariamente alto en ambos casos. La probabilidad de que una empresa sea activa cuando ha modificado su web de forma moderada es 30 veces la probabilidad que tiene una empresa cuyo sitio web está fuera de servicio, y cuando la modificación es mayor, la probabilidad de ser activa se multiplica por 50. Estos resultados están en línea con lo que habíamos hipotetizado: las empresas saludables invierten más en mantener y actualizar sus sitios web. Así, cuanta más actividad se evidencia en sus sitios web, más probable es que sean activas. Hay que remarcar que esto no implica que actualizar los sitios web ayude a las empresas a mantenerse activas, pero sí que es un fuerte reflejo del estado de actividad de la empresa.

Tras evidenciar la relación entre el estado de los sitios web corporativos y el estado de las empresas, se ha hecho la estimación de la especificación ampliada que muestra la ecuación [3]. Esta incluye la variable sobre el estado del sitio web y las variables estructurales seleccionadas por su potencial relación con la supervivencia empresarial, así como por su variación significativa entre empresas activas e inactivas.

Como refleja la tabla 5, el efecto de cada estado de los sitios web sobre la probabilidad de que las empresas sean activas se mantiene positivo, elevado y estadísticamente significativo. En cuanto a las variables económicas, la única que muestra un efecto estadísticamente significativo es el nivel de deuda. Su coeficiente negativo indica que, a medida que el nivel de deuda de la empresa aumenta, su probabilidad de estar activa disminuye. Concretamente, esta probabilidad disminuye en 1, % por cada incremento porcentual en la deuda.

Aunque el resto de variables económicas mostraron diferencias a nivel univariante, no contribuyen a explicar el estado de la empresa a nivel multivariante. Por un lado, estas variables económicas está relacionadas con el estado de la empresa, pero solamente con un alcance limitado, dado que hay un gran número de factores, como las decisiones estratégicas de la empresa o situaciones concretas del mercado que pueden contri-

TABLA 5.

REGRESIÓN LOGÍSTICA MULTIPERIODO CON VARIABLES WEB Y ESTRUCTURALES. VARIABLE DEPENDIENTE: *ACTIVA*

Variables	*β*	*OR*	*SE*	*valor-z*	*valor-p*
(Constante)	3,917	50,249	0,974	4,023	0,000
Estado_web(2)	1,579	4,850	0,395	4,001	0,000
Estado_web(3)	2,490	12,061	0,522	4,768	0,000
Estado_web(4)	3,242	25,585	0,951	3,410	0,001
Tamaño	0,206	1,229	0,169	1,218	0,223
Deuda	-0,018	0,982	0,006	-3,053	0,002
Productividad	0,034	0,193	1,034	0,283	0,777
Rentabilidad	0,005	1,005	0,011	0,443	0,658
Alta_tecnología	10,811	49.563,01	960,4	0,019	0,985
Observaciones	3.034				
Logaritmo de la verosimilitud	-154,116				

Notas: Estado_web(1): Fuera de servicio; *Estado_web*(2): Sin modificaciones; *Estado_web*(3): Modificaciones menores; Estado_web(4): Modificaciones mayores. Se han incluido variables binarias para recoger el efecto temporal.

Fuente: Elaboración propia.

buir a que mueran empresas con todo tipo de características (pequeñas o grandes, más o menos productivas, de cualquier sector de actividad, etc.). Por otro lado, el estado del sitio web ha emergido como un claro indicador del estado de la empresa, así que las variables económicas no han podido complementar la información ofrecida por el sitio web.

Tras demostrar la relación existente entre el estado de los sitios web corporativos y el estado de las empresas, la siguiente sección va un paso más lejos para complementar los análisis y aportar una perspectiva distinta al estudio. A tal fin, se ha llevado a cabo un análisis de supervivencia, que se presenta a continuación.

4.3. Análisis de supervivencia

Esta sección describe el análisis de supervivencia que se ha realizado para modelizar el riesgo de muerte de una empresa en determinados momentos del tiempo dependiendo del estado del sitio web. Como los datos de este estudio tienen una periodicidad anual, hemos empleado un modelo de duración de tiempo discreto, como se especifica en la ecuación [6].

La tabla 6 muestra los resultados de la estimación para este modelo, incluyendo la estimación de los coeficientes de regresión (β), las ratios de riesgo (HR), los errores estandarizados (SE), valores z y valores p.

Las ratios de riesgo se calculan como el exponente de los coeficientes, y comparan la frecuencia con la que ocurre cierto evento en dos grupos distintos a lo largo del tiempo. En este caso, comparan la frecuencia con que se observan los distintos estados de los sitios web entre los grupos de empresas activas e inactivas. Por tanto, una ratio HR superior a 1 indica que el riesgo de muerte crece en presencia de un estado específico del sitio web. Si la ratio es inferior a 1, entonces dicho riesgo decrece, mientras que si es igual a 1, entonces no hay diferencia entre la supervivencia de los dos grupos que se han comparado.

TABLA 6.

MODELO DE DURACIÓN DE TIEMPO DISCRETO. VARIABLE DEPENDIENTE: *1- ACTIVA*

Variables	*β*	*HR*	*SE*	*valor-z*	*valor-p*
(Constante)	-4,942	0,007	1,029	-4,802	0,000
Estado_web(2)	-1,202	0,301	0,348	-3,454	0,001
Estado_web(3)	-2,480	0,084	0,484	-5,128	0,000
Estado_web(4)	-2,764	0,063	0,753	-3,668	0,000
Observaciones	3.194				
Logaritmo de la verosimilitud	-195,262				

Notas: *Estado_web*(2): Sin modificaciones; *Estado_web*(3): Modificaciones menores; *Estado_web*(4): Modificaciones mayores. Se han incluido variables binarias para recoger el efecto temporal.

Fuente: Elaboración propia.

Las estimaciones negativas y estadísticamente significativas de los coeficientes indica que las empresas cuyos sitios web no han sido modificados (Código 2), o que han realizado cambios menores (Código 3) o mayores (Código 4) con respecto al año anterior, tienen un riesgo de muerte significativamente menor con respecto a las empresas cuyos sitios web están fueran de servicio (Código 1, que corresponde al estado del sitio web que hemos tomado como base para comparar). Concretamente, la ratio de riesgo para el estado del sitio web "Sin modificaciones" (Código 2) indica que las empresas cuyos sitios web se han mantenido igual durante el último año tienen 0,301 veces el riesgo de muerte que tienen las empresas cuyos sitios web están fuera de servicio; esto es, su riesgo es un 69,9 % inferior. El riesgo de muerte para las empresas en cuyos sitios web se observan modificaciones menores (Código 3) es un 91,6 % inferior con respecto a las empresas cuyos sitios web están fuera de servicio, alcanzado este porcentaje un 93,7 % cuando los cambios observados son mayores (Código 4). Como podemos observar, el riesgo disminuye conforme la actividad del sitio web aumenta. Estos resultados son consistentes con los que arrojan las regresiones logísticas multiperiodo, y confirman la fuerte relación que existe entre el estado de los sitios web corporativos y la supervivencia empresarial.

5. CONCLUSIONES

La demografía empresarial es un importante campo de interés para los investigadores y los responsables políticos, ya que la creación y el fracaso de las empresas tienen un enorme impacto en la producción y el empleo en todas las economías. En el contexto actual, en el que las comunicaciones digitales y los contenidos de Internet reflejan el comportamiento mayoritario de la sociedad, surge un nuevo desafío: relacionar la demografía empresarial con la evolución de Internet.

En este trabajo se ha analizado y confirmado la conexión del estado de actividad de una empresa con el estado de actividad del sitio web corporativo. Ambas labores se han completado llevando a cabo un seguimiento de los sitios web corporativos y del estado de las empresas durante siete años, y analizando después su relación con regresiones logísticas y un modelo de supervivencia. La regresión logística estima que los cambios importantes que se introducen en el sitio web corporativo multiplican por más de 50 la probabilidad de que una empresa esté activa en comparación con un sitio web fuera de servicio. En términos de supervivencia, los cambios en el sitio web corporativo están relacionados con una reducción de más del 90 % del riesgo de mortalidad de la empresa. Que ambos métodos ofrezcan resultados similares implica que el sitio web corporativo es un indicador sólido y fiable del estado de actividad de una empresa.

Estos resultados abren nuevas posibilidades de supervisión de la demografía de las empresas. Los datos web captan el estado de una empresa, mientras que el acceso a los sitios web corporativos es abierto y barato. Esto significa que es posible crear indicadores en línea para pronosticar a corto plazo y controlar los índices de mortalidad de las empresas. A diferencia de los métodosde estadística oficial tradicional, basados en encuestas sobre una muestra de población y lentos en su procesamiento, la supervisión web es capaz de llegar rápidamente a toda la población de empresas con sitio web. La naturaleza de Internet hace que esto pueda llevarse a cabo en un periodo muy corto de tiempo, y permite que la información de las empresas se recupere y analice automáticamente. Esto, a su vez, facilita a los responsables políticos y demás usuarios de estadísticas oficiales la obtención de estimaciones a corto plazo de la demografía de las empresas, las cuales, con el tiempo, se convertirán en decisiones más fundamentadas.

Entre las limitaciones de este estudio, en primer lugar, cabe señalar que solo se analizó la página de inicio del sitio web; es decir, no se tuvieron en cuenta los cambios realizados en las secciones internas. En segundo lugar, la muestra únicamente incluye empresas con sede en España, por lo que cabe tener mucha cautela a la hora de generalizar los resultados a diferentes países. Por último, debemos destacar que describimos cómo se correlaciona el estado del sitio web con el estado de actividad de la empresa, sin análisis causal. Si bien esto resulta útil a efectos de supervisión, de nuestros resultados no se desprende que los directivos deban introducir cambios de forma continuada en los sitios web corporativos para incrementar la supervivencia de la empresa.

Referencias

Agarwal, R. y Audretsch, D. B. (2001). Does Entry Size Matter? The Impact of the Life Cycle and Technology on Firm Survival. *Journal of Industrial Economics,* 49, pp. 21-43.

Albino, V., Balice, A. y Dangelico, R. M. (2009). Environmental strategies and green product development: An overview on sustainability-driven companies. *Business Strategy and the Environment,* 18, pp. 83-96.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D. y Cochran, J. J. (2014). Statistics for Business & Economics. 12th. *Cengage Learning,* pág. 1120.

Arora, S. K., Li, Y., Youtie, J. y Shapira, P. (2016). Using the wayback machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology,* 67, pp. 1904-1915.

Arora, S. K., Youtie, J., Shapira, P., Gao, L. y Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics,* 95, pp. 1189-1207.

Astebro, T. y Winter, J. (2012). More than a Dummy: The Probability of Failure, Survival and Acquisition of Firms in Financial Distress. *European Management Review,* 9, pp. 1-17.

Audretsch, D. B. (1991). New-Firm Survival and the Technological Regime. *The Review of Economics and Statistics,* 73, pp. 441-450.

Audretsch, D. B., Houweling, P. y Thurik, A. R. (2000). Firm Survival in the Netherlands. *Review of Industrial Organization,* 16, pp. 1-11.

Basile, R., Pittiglio, R. y Reganati, F. (2017). Do Agglomeration Externalities Affect Firm Survival? *Regional Studies,* 51, pp. 548-562.

Blazquez, D. y Domenech, J. (2014). Inferring export orientation from corporate websites. *Applied Economics Letters,* 21, pp. 509-512.

— (2018). Web Data Mining for Monitoring Business Export Orientation. *Technological and Economic Development of Economy,* 24(2), pp. 406-428.

Bridges, S. y Guariglia, A. (2008). Financial constraints, global engagement, and firm survival in the United Kingdom: Evidence from micro data. *Scottish Journal of Political Economy,* 55, pp. 444-464.

Bureau van Dijk (2010). *SABI: Sistema de Análisis de Balances Ibéricos.* CD-ROM (Version 36.1).

Cefis, E. y Marsili, O. (2005). A matter of life and death: Innovation and firm survival. *Industrial and Corporate Change,* 14, pp. 1167-1192.

Cohen, J., Cohen, P., West, S. G. y Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* 3rd. Routledge, pág. 736.

DGIPYME (2017). *Estadísticas PYME: Evolución e indicadores.* Disponible en: http://www.ipyme.org/Publicaciones/Estadisticas-PYME-2016.pdf (Acceso 30 de marzo de 2017). Ministerio de Economía, Industria y Competitividad.

Delmar, F., McKelvie, A. y Wennberg, K. (2013). Untangling the relationships among growth, profitability and survival in new firms. *Technovation,* 33, pp. 276-291.

Domenech, J., Ossa, B. de la, Pont, A., Gil, J. A., Martinez, M. y Rubio, A. (2012). An Intelligent System for Retrieving Economic Information from Corporate Websites. *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).* Macau, China, pp. 573 -578.

Esteve-Pérez, S. y Mañez-Castillejo, J. A. (2008). The Resource-Based Theory of the Firm and Firm Survival. *Small Business Economics,* 30, pp. 231-249.

Eurostat (2008). *NACE Rev. 2 Statistical classification of economic activities in the European Communities.* EUROSTAT Methodologies and Working papers. Luxembourg: Office for Official Publications of the European Communities.

— (2014). *High-tech aggregation by NACE Rev. 2.* Eurostat indicators of High-tech industry and knowledge-intensive services. Luxembourg: Office for Official Publications of the European Communities.

Eurostat y OECD (2007). *Eurostat-OECD Manual on Business Demography Statistics.* Luxembourg: Office for Official Publications of the European Communities.

Evans, D. S. (1987). Tests of Alternative Theories of Firm Growth. *Journal of Political Economy,* 95, pp. 657-674.

Gallego Álvarez, I., María García Sánchez, I. y Rodríguez Domínguez, L. (2008). Voluntary and compulsory information disclosed online. The effect of industry concentration and other explanatory factors. *Online Information Review,* 32, pp. 596-622.

Gémar, G., Moniche, L. y Morales, A. J. (2016). Survival analysis of the Spanish hotel industry. *Tourism Management,* 54, pp. 428 -438.

Geroski, P. (1995). What do we know about entry? *International Journal of Industrial Organization* 13, pp. 421-440.

Geroski, P., Mata, J. y Portugal, P. (2010). Founding conditions and the survival of new firms. *Strategic Management Journal,* 31, pp. 510-529.

Giovannetti, G., Ricchiuti, G. y Velucchi, M. (2011). Size, innovation and internationalization: A survival analysis of Italian firms. *Applied Economics,* 43, pp. 1511-1520.

Gök, A., Waterworth, A. y Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics,* 102, pp. 653-671.

Görg, H. y Spaliara, M.-E. (2014). Financial Health, Exports and Firm Survival: Evidence from UK and French Firms. *Economica,* 81, pp. 419-444.

Guariglia, A., Spaliara, M.-E. y Tsoukas, S. (2016). To What Extent Does the Interest Burden Affect Firm Survival? Evidence from a Panel of UK Firms during the Recent Financial Crisis. *Oxford Bulletin of Economics and Statistics,* 78, pp. 576-594.

INE (2016). *Encuesta de uso de TIC y Comercio Electrónico en las empresas 2015-2016.* Disponible en: http://ine.es/dynt3/inebase/?path=/t09/e02/a2015-2016 (acceso 10 de octubre 2016).

Jacobson, T. y Schedvin, E. von (2015). Trade Credit and the Propagation of Corporate Failure: An Empirical Analysis. *Econometrica,* 83, pp. 1315-1371.

Jenkins, S. P. (1995). Easy Estimation Methods for Discrete-Time Duration Models. *Oxford Bulletin of Economics and Statistics,* 57, pp. 129-136.

Kahle, B. y Gilliat, B. (2016). *Wayback Machine.* Disponible en: http://archive.org/web/ (Acceso 27 de febrero de 2016).

Lancaster, T. (1990). *The econometric analysis of transition data.* Cambridge University Press.

Li, Y., Arora, S., Youtie, J. y Shapira, P. (2018). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation,* 76, pp. 3-14.

Libaers, D., Hicks, D. y Porter, A. L. (2010). A taxonomy of small firm technology commercialization. *Industrial and Corporate Change,* 25, pp. 371-405.

Manjón-Antolín, M. C. y Arauzo-Carod, J.-M. (2008). Firm survival: methods and evidence. *Empirica,* 35, pp. 1-24.

Mata, J. y Portugal, P. (1994). Life Duration of New Firms. *The Journal of Industrial Economics,* 42, pp. 227-245.

Meroño-Cerdan, A. L. y Soto-Acosta, P. (2007). External Web content and its influence on organizational performance. *European Journal of Information Systems,* 16, pp. 66 -80.

Munzert, S., Rubba, C., Meissner, P. y Nyhuis, D. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* Chichester, UK: John Wiley & Sons, Ltd.

Overbeeke, M. y Snizek, W. E. (2005). Web Sites and Corporate Culture: A Research Note. *Business & Society,* 44, pp. 346-356.

Tagesson, T., Blank, V., Broberg, P. y Collin, S.-O. (2009). What explains the extent and content of social and environmental disclosures on corporate websites: A study of social and environmental reporting in Swedish listed corporations. *Corporate Social Responsibility and Environmental Management,* 16, pp. 352-364.

Tang, L., Gallagher, C. C. y Bie, B. (2015). Corporate Social Responsibility Communication Through Corporate Websites: A Comparison of Leading Corporations in the United States and China. *International Journal of Business Communication,* 52, pp. 205-227.

TSOUKAS, S. (2011). Firm survival and financial development: Evidence from a panel of emerging Asian economies. *Journal of Banking & Finance,* 35, pp. 1736-1752.

VAUGHAN, L. (2014). Discovering business information from search engine query data. *Online Information Review,* 38, pp. 562-574.

YOUTIE, J., HICKS, D., SHAPIRA, P. y HORSLEY, T. (2012). Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. *Technology Analysis & Strategic Management,* 24, pp. 981-995.

CAPÍTULO VI

Predicciones financieras basadas en análisis de sentimiento de textos y minería de opiniones

Argimiro Arratia*

En este capítulo se describe la mecánica básica para construir un modelo de predicción que utiliza indicadores de sentimiento derivados de datos textuales. Enfocamos nuestro objetivo de predicciones en series de tiempo financieras y presentamos un conjunto de hechos empíricos que describen las propiedades estadísticas de los indicadores de sentimiento, con particular atención en aquellos indicadores extraídos de noticias sobre mercados financieros y cuya categorización de sentimientos se basa en diccionarios. El objetivo general es proporcionar pautas para los profesionales en el mundo de las finanzas para la adecuada construcción e interpretación de su propia información numérica dependiente del tiempo y que representa la percepción del público hacia las empresas, los precios de las acciones y los mercados financieros en general.

Palabras clave: análisis de sentimiento, emociones financieras, diccionario financiero, modelos de pronóstico.

* El autor agradece la ayuda recibida por Gustavo Avalos, Ariel Duarte y Martí Renedo en la elaboración de resultados presentados en la tabla 2 y la figura 4, y en la recopilación de software para el análisis de sentimiento. A Alejandra Cabaña por sus comentarios y correcciones de los aspectos estadísticos de este artículo.

1. INTRODUCCIÓN

Hoy en día existe una gran cantidad de compañías tecnológicas financieras *(fintech)* que ofrecen índices de sentimientos sobre los mercados, construidos a partir de opiniones publicadas en periódicos electrónicos, redes sociales, y otras fuentes de noticias en Internet. Además ahora existen incontables fuentes de grandes cantidades de datos, que están en continuo crecimiento, y que nos permiten desarrollar estrategias basadas en datos, en particular textos que recogen opiniones del público, en lugar de simplemente replicar y actuar sobre modelos econométricos clásicos.

Estos datos temporales que cuantifican las emociones del público general hacia compañías que cotizan en bolsa, personalidades y eventos del mundo de las finanzas, son utilizados por muchos inversores como información social en sus estrategias de compra y venta en los mercados financieros. La extensa investigación en finanzas conductuales ha mostrado evidencia del hecho que los inversores reaccionan a las noticias, y que, por lo general, muestran una mayor propensión a hacer un movimiento de inversión basado en malas noticias en lugar de buenas noticias (véase por ejemplo Chan, 2003; Brown y Cliff, 2004; Baker y Wurgler, 2007; Kumar y Lee, 2006; Tetlock, 2007). Esta actitud la atribuyen algunos autores a un rasgo de la psicología humana (Baumeister *et al.,* 2001; Rozin y Royzman, 2001), y otros a una estrategia específica de inversión en el mercado de valores (Engelberg, Reed y Ringgenberg, 2012).

El análisis de sentimiento de textos o minería de opinión trata con la categorización de las opiniones expresadas en un texto. En el contexto financiero el propósito es capturar las emociones de los inversionistas hacia el mercado financiero expresadas en redes sociales o medios informativos, y cuantificar estas emociones como variables numéricas que potencialmente serán predictoras del comportamiento de los mercados de valores.

Este objetivo es contrario al mantra de la teoría clásica del mercado, que sostiene que toda información se incorpora a los precios de las acciones tan pronto como se hace pública. Sin embargo, cada vez existen más indicios que apuntan a que la información se difunde gradualmente a través de la población de inversionistas y que esta difusión gradual afecta los precios de las acciones. Estudios recientes demuestran que existe alguna dependencia entre los precios del mercado y las historias publicadas por los medios.

Entre los primeros trabajos que reportan alguna relación, aunque débil, entre las noticias sobre empresas y el movimiento del precio, o la volatilidad, de las acciones de estas empresas tenemos los de Berry y Howe (1994) y de Mitchell y Mulherin (1994). Ambos trabajos, publicados en 1994, no tratan el sentimiento en la noticia, sino la relación del volumen de estas con la volatilidad o el volumen de negociación. Chan (2003) sí que analiza el tono de los titulares de noticias sobre compañías de diferentes capitalizaciones y observa un incremento en la variabilidad del precio de las acciones

de compañías de más baja capitalización después de la publicación de noticias negativas. El trabajo de Antweiler y Frank (2004) es pionero en el desarrollo de medidas de sentimiento de noticias mediante métodos de aprendizaje automático, para explicar el comportamiento del precio de las acciones. Estos investigadores emplean un algoritmo Naive Bayes para clasificar el sentimiento de los documentos a partir de un conjunto de entrenamiento construido manualmente, y obtienen de esta clasificación unas series de señales de compra y venta para el mercado. Concluyen que si bien estas señales pueden predecir la volatilidad del mercado, su capacidad predictiva de la variación del precio de las acciones es mínima.

En 2007, Tetlock publica un serio análisis de la técnica basada en léxicos, o "bolsa de palabras" *(bag-of-words),* para clasificar los sentimientos implícitos en las noticias del ámbito financiero (Tetlock, 2007). Esta técnica consiste en construir listas (bolsas) de palabras, donde cada una de estas listas se asocia con una categoría (p.ej. positivo o negativo). Mediante una clasificación de palabras basada en categorías extraídas del *Diccionario Psicosocial de Harvard,* Tetlock cuantifica el optimismo y el pesimismo contenido en la columna "Abreast of the Market" del *Wall Street Journal* y observa que los niveles altos de pesimismo reflejado por el conjunto de las noticias predicen caídas de los precios del mercado, a los que les sigue a continuación un movimiento inverso. Empleando técnicas similares basadas en diccionarios, Tetlock, Saar-Tsechansky y Macskassy (2008) descubren que la proporción de palabras de carga negativa en las noticias del *Dow Jones News* y *Wall Street Journal* predicen futuras variaciones en los beneficios de las compañías, lo cual explican como consecuencia del hecho de que el contenido linguístico en las noticias financieras captura los aspectos difíciles de cuantificar de los fundamentales de una compañía que son rápidamente incorporados en el precio de las acciones.

Una mejora a los métodos de análisis de sentimiento basados en diccionarios fue aportada posteriormente por Loughran y McDonald (2011), quienes demostraron que las listas de palabras desarrolladas para disciplinas sociales o humanísticas en general (como los diccionarios Harvard) clasifican erróneamente las palabras comunes en textos financieros. Como alternativa Loughran y McDonald desarrollan listas de palabras extraídas de los formularios anuales 10-K, con una carga sentimental positiva, negativa y otras cuatro categorías que reflejan más fielmente el tono en los textos financieros, y muestran que la proporción de palabras negativas en estos formularios 10-K afecta a la baja los rendimientos de las acciones.

Si bien no existe un consenso general sobre la posibilidad de predecir los movimientos en los precios de acciones en los mercados financieros mediante el análisis del sentimiento de las noticias (véase una extensa discusión al respecto en (Schoen *et al.,* 2013)), el hecho relevante es que esta área de investigación es actualmente una de las más populares en el desarrollo de modelos de pronóstico que hacen uso de grandes masas de datos, tanto a nivel académico como industrial.

2. LA MECÁNICA DEL ANÁLISIS DE SENTIMIENTO DE TEXTOS

Nuestra exposición se centra en el análisis de sentimientos de textos con referencia a un sujeto. Esto significa que nuestro objetivo es determinar si un documento, o una frase dentro de un documento, expresa emociones positivas, negativas o de otro tipo *hacia un sujeto específico.* Existen otras categorías de análisis de sentimiento textual, las cuales se describen en el libro de texto de Liu (2015). En aplicaciones financieras, nuestros sujetos de interés son empresas, mercados financieros, materias primas o cualquier otra entidad con valor económico. Esta información sentimental se utiliza para alimentar modelos de pronóstico de variables estocásticas que cuantifican el comportamiento del sujeto de interés; p.ej. precios, volatilidad, u otros indicadores estadísticos financieros.

Se propone a continuación un flujo de trabajo para construir modelos de pronóstico financiero basados en sentimientos extraídos de datos textuales. Este comprende las siguientes etapas:

- Creación y procesamiento de *corpus* textuales.
- Cálculo del sentimiento.
- Agregación de valores de sentimiento y construcción de indicadores.
- Modelización.

2.1. Gestión del *corpus* textual

La primera tarea fundamental es la recopilación de textos adecuados, y la aplicación de técnicas de minería de texto para limpiar y categorizar los términos dentro de cada documento. Los textos han de estar en formato electrónico y cada documento ha de tener un identificador único (por ejemplo, un nombre de archivo singular) y una marca de tiempo (*e.g.* fecha y/o hora de su publicación). Además, utilizando algún esquema de reconocimiento de entidades, se ha de identificar dentro de cada documento los sujetos de interés (*e.g.* el nombre o *ticker* de compañías cotizadas). De esta manera podremos agrupar los documentos por sujeto común, siendo posible que un mismo documento aparezca en dos grupos diferentes pertenecientes a dos sujetos diferentes.

Ejemplo 2.1. Los sujetos (por ejemplo, el nombre de una empresa o un indicador de su cotización-ticker) se pueden identificar mediante la concordancia de palabras clave o técnicas de reconocimiento de nombres de entidades[1]*. Alternativamente, algunos proveedores de noticias como Dow Jones Newswires incluyen etiquetas en sus archivos xml que indican la empresa de la que trata la noticia.*

[1] Consulte el software Stanford NER https://nlp.stanford.edu/software/CRF- NER.shtml

2.2. Cálculo del sentimiento

El análisis de sentimientos es básicamente un problema de clasificación de textos. Por lo tanto, se puede abordar este problema algorítmico de las dos formas posibles de clasificación mecánica: (1) aplicando un algoritmo de aprendizaje automático supervisado que se entrena con textos ya etiquetados como positivo o negativo (o cualquier otra emoción); o (2) utilizar un método de clasificación sin supervisión basado en el reconocimiento de algunos patrones sintácticos fijos, o palabras claves, que expresan un sentimiento específico (un diccionario o léxico de sentimiento). Esta última solución es utilizada con mayor frecuencia por investigadores y profesionales de las finanzas. Por lo tanto, en esta exposición priorizamos la descripción del método no supervisado basado en léxicos de sentimiento y únicamente daremos algunas indicaciones a la literatura sobre el enfoque de aprendizaje automático para la clasificación de sentimientos.

2.2.1. Método no supervisado basado en léxicos para la clasificación de sentimiento

El componente clave de este método de clasificación de texto es un conjunto de palabras, o patrones sintácticos, que denotan un sentimiento específico. Por ejemplo, la positividad se expresa con palabras como *bueno, admirable, mejor,* … y emoticonos como : -) o ; -], y otros símbolos similares que se utilizan a menudo en mensajes cortos como los de Twitter (Bifet y Frank, 2010; Go, Bhayani, y Huang, 2009). Estas *bolsas de palabras* (como comúnmente se les refiere en la jerga computista) conforman un *lexicón de sentimiento o diccionario.* Un ejemplo de diccionario de términos que caracterizan los sentimientos positivo y negativo, específico al ámbito financiero, es el recopilado por Loughran y McDonald (2011), una muestra del cual se presenta en la Tabla 1.

Dado un sentimiento S (por ejemplo, positivo, negativo, …), determinado por algún léxico $L(S)$, un algoritmo básico para asignar un valor numérico del sentimiento S a un documento, es contar el número de apariciones de términos de $L(S)$ en el documento. Este número da una medida de la fuerza del sentimiento S en el documento. Para comparar las fortalezas de dos sentimientos diferentes en un documento, es aconsejable relativizar estos números con el número total de términos en el documento. Hay muchas mejoras de esta función básica de valoración sentimental. Por ejemplo, si se considera una escala de valores predeterminados para los términos en el léxico (en lugar de que cada uno tenga el mismo valor de 1), o si se permiten valores negativos, además de positivos, para capturar la dirección del sentimiento, o se hacen otras consideraciones sobre el contexto donde, dependiendo de las palabras vecinas, los términos del léxico pueden cambiar sus valores o incluso pasar de un sentimiento a otro. Por ejemplo, *bueno* es positivo, pero *no es bueno* es negativo. Repasaremos algunas de estas variantes, pero para una exposición más detallada, consulte el libro de texto de Liu (2015) y sus referencias.

TABLA 1.

UNA MUESTRA DEL DICCIONARIO RECOPILADO POR LOUGHRAN Y MCDONALD (2011) A PARTIR DE REPORTES 10-K. CONSTA DE 2355 TÉRMINOS NEGATIVOS Y 354 POSITIVOS

Negativo	*Positivo*
Forbid	Collaborates
Adversity	Pleased
Intimidation	Rewarding
Barred	Stabilize
Distracted	Gaining
Diminishes	Profitably
Accusations	Achieve
Overstating	Effective
Declines	Inventive
Aftermaths	Revolutionize

Fuente: University of Notre Dame, Software Repository for Accounting and Finance (https://sraf.nd.edu/textual-analysis/resources/)

Formalicemos ahora un esquema general para definir una serie temporal de valoraciones de sentimiento para documentos con respecto a un objetivo específico (por ejemplo, una empresa o un instrumento financiero), basados en algún diccionario sentimental. Tenemos a mano $\lambda = 1, \ldots, \Lambda$ léxicos L_λ, donde cada uno define un sentimiento. Tenemos K posibles sujetos y recopilamos un flujo de documentos en diferentes instantes de tiempo $t = 1, \ldots, T$. Sea N_t el número total de documentos en el instante de tiempo t. Sea $D_{n,t,k}$ el n-ésimo documento, en el tiempo t y que menciona al k-ésimo sujeto, para $n = 1, \ldots, N_t$, $t = 1, \ldots T$ y $k = 1, \ldots K$.

Fijemos un lexicón L_λ y un sujeto G_k. Se define un valor numérico del sentimiento basado en el lexicón L_λ para el documento $D_{n,t,k}$, en tiempo t y referente al sujeto G_k mediante la siguiente ecuación:

$$S_{n,t}(\lambda, k) = \sum_{i=1}^{I_d} w_i S_{i,n,t}(\lambda, k) \qquad [1]$$

donde $S_{i,n,t}(\lambda, k)$ es el valor de sentimiento asignado al i-ésimo término singular que aparece en el documento y de acuerdo con el lexicón L_λ, siendo este valor cero si el término no está en el lexicón. I_d es el número total de términos en el documento $D_{n,t,k}$ y w_i es un peso, para cada término, que determina la forma en que se agregan las valoraciones de sentimiento en el documento.

Ejemplo 2.2. Si $S_{i,n,t} = 1$ (o 0 si es el caso que el i-ésimo término no está en el lexicón), para todo i, y $w_i = 1/I_d$ tenemos entonces la muy básica estimación de la densidad del

sentimiento en el texto utilizado, por ejemplo, en (Tetlock, Saar-Tsechansky y Macskassy, 2008; Loughran y McDonald, 2011) y muchos otros trabajos sobre análisis de sentimiento de textos, el cual da igual importancia a todos los términos en el lexicón. Un esquema de ponderación más refinado, que refleja diferentes niveles de relevancia de cada término con respecto al sujeto-objetivo, es considerar $w_i = \mathrm{dist}(i, k)^{-1}$*, donde* $\mathrm{dist}(i, k)$ *es una métrica de distancia entre el* i*-ésimo término y el* k*-ésimo sujeto, como en (Ding, Liu y Yu, 2008).*

El valor de sentimiento $S_{i,n,t}$ es un número real y puede descomponerse en factores $v_i \cdot s_i$, donde v_i es un número que representa un cambio de sentimiento (un *cambiador de valencia:* una palabra que cambia los sentimientos en la dirección opuesta) y s_i el valor de sentimiento *per se*.

Sobre los cambiadores de valencia

Definidos originalmente y analizado su efecto contrario sobre el sentimiento calculado en documentos de habla inglesa en (Polanyi y Zaenen, 2006), estas son palabras que pueden alterar el significado de una palabra polarizada, y que pertenecen a una de las cuatro categorías básicas siguientes: *negadores, amplificadores, de-amplificadores y conjunciones adversativas.* Un negador invierte el signo de una palabra polarizada, como en la frase: "esa empresa *no* es una *buena* inversión". Un amplificador intensifica la polaridad de una oración, como por ejemplo el adverbio *definitivamente* amplifica la negatividad en el ejemplo anterior: "esa empresa no es *definitivamente* una *buena* inversión". Los de-amplificadores (también conocidos como reductores), por otro lado, disminuyen la intensidad de una palabra polarizada (p. ej., "la empresa es *apenas* buena como inversión"). Una conjunción adversativa anula la polaridad de sentimiento de la cláusula precedente, p. ej., "Me gusta la empresa *pero* no vale la pena".

¿Debemos preocuparnos por los cambiadores de valencia? Si los cambiadores de valencia ocurren con frecuencia en nuestros conjuntos de textos, entonces no considerarlos en el cálculo de las valoraciones de sentimiento en la ecuación [1] generará una valoración inexacta del sentimiento del texto. Más aún en el caso de negadores y conjunciones adversativas que invierten o anulan la polaridad del sentimiento de la oración.

Para conjuntos de textos extraídos de redes sociales como Twitter o Facebook, se ha observado que la ocurrencia de cambios de valencia, en particular de negadores, es considerablemente alta (aproximadamente un 20% para varios trending topics[2]), por lo que en este contexto es importante tener en consideración este fenómeno gramatical.

En el ámbito financiero, hemos calculado la ocurrencia de los cambiadores de valencia (en inglés) en una muestra de 1,5 millones de documentos del conjunto de Dow Jones

[2] https://cran.r-project.org/web/packages/sentimentr/readme/README.html

Newswires. Los resultados de estos cálculos, que se pueden ver en la Tabla 2, muestran una baja incidencia de de-amplificadores y conjunciones adversativas (alrededor del 3%), pero los negadores aparecen en un número que puede merecer cierta atención.

Creación de léxicos de sentimiento

TABLA 2.
PORCENTAJE DE OCURRENCIA DE CAMBIADORES DE VALENCIA EN 1,5MM DOCS. DJN

Texto	Negadores	Amplificadores	De-amplificadores	Adversativas
DJN noticias	7,00	14,13	3,02	3,02

Un punto de partida para compilar un conjunto de palabras con una carga sentimental específica es utilizar un diccionario estructurado (preferiblemente en línea como WordNet[3]) que enumere sinónimos y antónimos para cada palabra. Luego comience con algunas palabras seleccionadas (palabras claves) que conlleven un sentimiento específico y continúe agregando algunos de los sinónimos al conjunto, y al conjunto que caracteriza al sentimiento contrario agregue los antónimos. Hay muchas formas inteligentes de hacer esta expansión desde palabras claves (o semillas) de sentimiento utilizando algún algoritmo de clasificación supervisado para encontrar más palabras que conlleven una emoción similar. Un ejemplo es el trabajo de Tsai y Wang (2014) donde tratan la expansión de un conjunto de palabras claves financieras utilizando el modelo continuo de bolsa de palabras aplicado a los informes financieros anuales obligatorios, conocidos como 10-K en EE. UU. Otro ingenioso esquema supervisado para construir léxicos basado en la teoría de redes se expone en (Rao y Ravichandran, 2009). Para una descripción más extensa sobre la creación de léxicos de sentimientos veáse (Liu, 2015, Cap. 7) y las referencias que allí se listan.

2.2.2. Método de aprendizaje automático supervisado para la clasificación de sentimientos

Otra forma de clasificar textos es mediante el uso de algoritmos de aprendizaje automático, que se fundamentan en un modelo previamente entrenado para generar predicciones. A diferencia del método basado en diccionarios, estos algoritmos no están programados para responder de cierta manera según las entradas recibidas, sino para extraer patrones de comportamiento de conjuntos de datos de entrenamiento pretiquetados. Los algoritmos internos que dan forma a la base de este proceso de aprendizaje tienen algunas componentes sólidamente fundamentadas en la Matemática y la Estadística, lejos de ser heurísticas arbitrarias. Algunos de los algoritmos de aprendizaje

[3] https://wordnet.princeton.edu/

automático más populares y robustos son Naïve Bayes, *Support Vector Machines* y *Deep Learning.* Las etapas para la clasificación de sentimientos en textos, utilizando modelos de aprendizaje automático son las siguientes:

Desarrollo y preprocesamiento de un corpus de textos. El proceso de aprendizaje parte de un corpus de textos clasificados manualmente, que después de la extracción de características, será utilizado por el algoritmo de aprendizaje automático para encontrar los parámetros que mejor se ajusten al modelo y evaluar la precisión en una etapa de prueba. Es por esto que la parte más importante de este proceso es el desarrollo de un buen conjunto de textos de entrenamiento. Este debe ser lo más grande posible y representativo del conjunto de datos que se analizarán. Una vez obtenido el corpus, se deben aplicar técnicas para reducir el ruido que generan las palabras sentimentales sin sentido, así como para aumentar la frecuencia de cada término mediante la derivación y lematización. Estas técnicas dependen del contexto al que se aplique. Esto significa que un modelo entrenado para clasificar textos de un campo determinado no podría aplicarse directamente a otro. Por tanto, es de suma importancia tener un corpus clasificado manualmente lo mejor posible.

Extracción de características. El enfoque general para extraer características consiste en transformar el texto preprocesado en una expresión matemática basada en la detección de la coocurrencia de palabras o frases. El texto transformado se divide en una serie de características, cada una de las cuales corresponde a un elemento del texto original de entrada.

Clasificación. Durante esta etapa, el modelo entrenado recibe un conjunto de características aún no vistas para estimar una nueva clase.

Para más detalles sobre todo este procedimiento de aprendizaje consúltese (Sebastiani, 2002; Liu, 2015).

Un ejemplo de algoritmo de aprendizaje automático para el análisis de sentimiento en textos es *Deep-MLSA* (Deriu *et al.,* 2016, 2017). Este modelo consta de un clasificador de red neuronal convolucional multicapa con tres estados que corresponden a sentimientos negativos, neutrales y positivos. Deep-MLSA se adapta muy bien a la corta longitud y el carácter informal de los tweets de las redes sociales, y ha resultado ser el mejor algoritmo de clasificación de polaridad de mensajes en la sección de " Sentiment Analysis in Twitter" de la competición *SemEval* (Nakov *et al.,* 2016).

2.3. Métodos para agregar valores de sentimiento y construir indicadores

Sean L_λ un lexicón que caracteriza algún sentimiento y G_k un sujeto-objetivo. Una vez calculados los valores del L_λ-sentimiento para cada documento que menciona el sujeto

G_k y siguiendo la rutina expuesta en la ecuación [1], se procede a agregar estos valores para cada instante de tiempo t disponible, y así obtener el L_λ-sentimiento sobre G_k y en el instante de tiempo t, que denotaremos por $S_t(\lambda, k)$, y formalmente describimos mediante la siguiente ecuación:

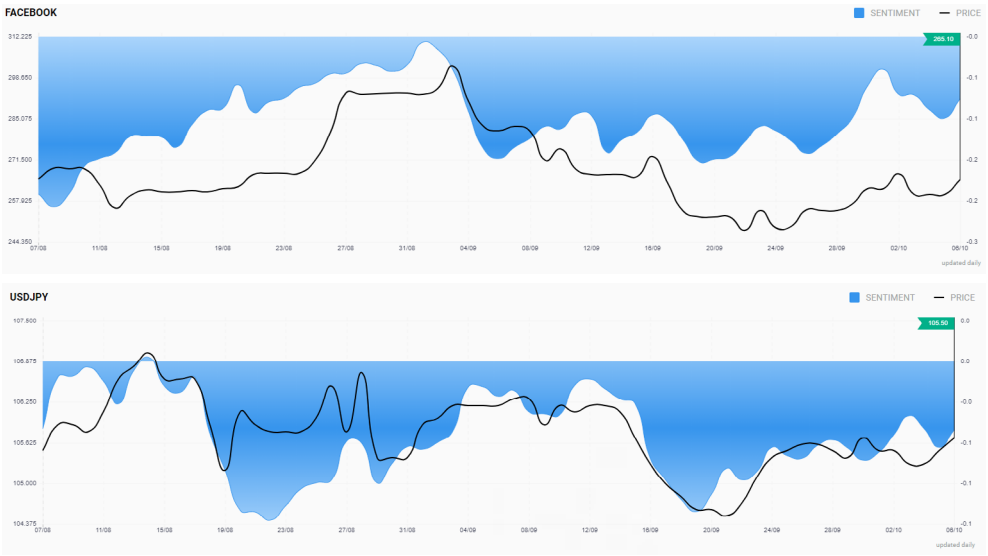
$$S_t(\lambda,k) = \sum_{n=1}^{N_t} \beta_n S_{n,t}(\lambda,k) \qquad [2]$$

Al igual que en la ecuación [1], los pesos β_n determinan la manera en que son agregados los valores de sentimiento para cada documento. Por ejemplo, si se considera $\beta_n = 1/\text{card}(D_{n,t,k})$, esta ponderación da más relevancia a los documentos de menor longitud. (card($D_{n,t,k}$) es la cardinalidad de $D_{n,t,k}$, esto es el número de términos que lo componen.)

Se obtiene de esta manera una serie temporal de valores de sentimiento, o un *indicador de sentimiento*, $\{S_t : t = 1, \ldots, T\}$, basado en el lexicón L_λ que define un sentimiento específico sobre el sujeto G_k. Variaciones de este indicador de sentimiento (basado en léxico L_λ) para G_k se pueden obtener aplicando cualquier filtro F a la serie S_t, y así obtener una nueva serie indicatriz $\{F(S_t) : t = 1, \ldots T\}$. Por ejemplo, si se aplica una media móvil se obtendrá una versión suavizada de la serie de sentimientos original. En la figura 1 se muestran ejemplos de activos financieros (*stock* de la compañía Facebook y la tasa de cambio USD/JPY) con indicador de sentimiento (en azul) sobrepuesto al

FIGURA 1

PRECIO DE FACEBOOK Y USD/JPY (EN NEGRO) SOBRE UN INDICADOR DE SENTIMIENTO COMPUESTO (EN AZUL)



Fuente: Cortesía de *Acuity Trading Inc.*

precio (negro). El indicador de sentimiento es una combinación de positivo y negativo, elaborado por la empresa *Acuity Trading Inc.*

2.4. Modelización

Existen básicamente dos enfoques para la modelización: utilizar los indicadores de sentimiento como información exógena que alimentamos a nuestros modelos de pronóstico, y se comprueba su relevancia para pronosticar movimientos de precios, retornos de precios u otras estadísticas del precio; o utilizarlos como criterios para seleccionar los sujetos (objetivos) de las noticias, que en nuestro caso de interés serán compañías cotizadas, y formar una cartera. Algunos ejemplos escogidos de la gran cantidad de investigaciones publicadas sobre el tema de la predicción y la gestión de carteras con análisis de sentimiento de textos son (Arias, Arratia y Xuriguera, 2013; Baker y Wurgler, 2007; Beckers, Kholodilin y Ulbricht, 2017; Heston y Sinha, 2017; Loughran y McDonald, 2011; Tetlock, 2007; Tetlock, Saar-Tsechansky y Macskassy, 2008; Uhl, Pedersen y Malitius, 2015). En la sección tercera damos un ejemplo (académico) de cómo utilizar indicadores de sentimiento para generar señales de inversión.

Para un tratamiento más extenso de los bloques de construcción para producir modelos de pronóstico basados en datos textuales descritos en esta sección consulte (Algaba *et al.,* 2020), y el tutorial para el paquete de R sentometric en (Ardia *et al.,* 2019).

3. NEGOCIACIONES BURSÁTILES CON SENTIMIENTO

Comencemos por recordar una regla de inversión muy popular entre los analistas técnicos (aquellos que se guían por patrones en las gráficas de precios para realizar inversiones [Achelis, 2001]). Esta es la estrategia de cruces de medias móviles (en notación $MM(p)$, donde p es el periodo de días sobre el que se calcula la media), que consiste en trazar dos medias móviles sobre el precio del activo, una de corto plazo $MM(c)$ y otra de mediano a largo plazo $MM(m)$, y la regla consiste en tomar posiciones largas (comprar) cuando la media móvil corta esté por encima de la larga $(MM(c) > MM(m))$, o tomar posiciones cortas (vender) en caso contrario.

Nuestra idea es aplicar esta regla de inversión, no al precio, sino a un indicador de sentimiento "bullish" (BULL) sobre el banco J.P. Morgan (NYSE:JPM), permitiendo posiciones en corto, y durante el periodo 1 de febrero 2018 a 15 de mayo 2020. Este indicador BULL se ha construido a partir de un diccionario de términos financieros con carga sentimental positiva. La media móvil corta que utilizamos es un MM (10) y la larga un MM (25), y contrastamos nuestra estrategia con la estrategia básica de no hacer nada, es decir "comprar-y-mantener" ("buy-and-hold"). La tabla 3 presenta los resultados de nuestra estrategia (MAx) y de "comprar-y-mantener" (BH) en términos de: retorno

acumulado (un 88% para nuestra estrategia contra un -14% para BH, debido en buena parte por tomar posiciones cortas en el momento de la crisis bursátil ocasionada por la COVID-19); retorno anualizado; cociente de Sharpe (anualizado); volatilidad (anualizada); caída porcentual máxima y de longitud máxima.

TABLA 1.

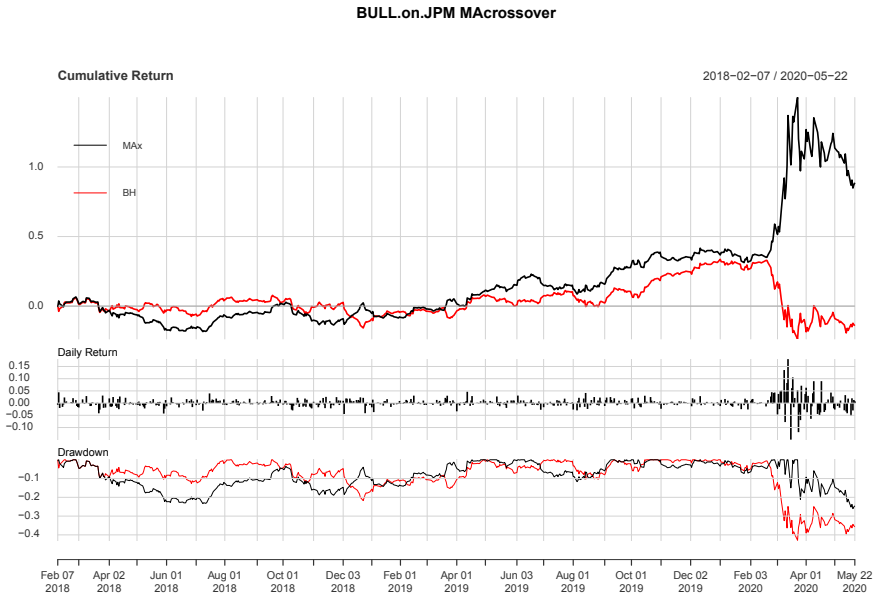
RESULTADOS DE TRADING CON SENTIMIENTO BULL SOBRE JPM CONTRA "COMPRAR-Y-MANTENER" (BH)

	MAx	*BH*
Retorno Acum.	0.8866467	-0.14067188
Retorno Anual	0.3246418	-0.06493845
Cociente de Sharpe (anual.)	0.9167684	-0.18304064
% veces ganancia	0.5106383	0.50531915
Volatilidad (anual.)	0.3541154	0.35477613
Caída % max.	-0.2603982	-0.42870098
Longitud caída max.	284	149

Fuente: Elaboración propia con el software R.

FIGURA 2

RETORNO ACUMULADO Y CAÍDA MÁXIMA DE LAS ESTRATEGIAS DE INVERSIÓN MAX (NEGRO) Y BH (ROJO)

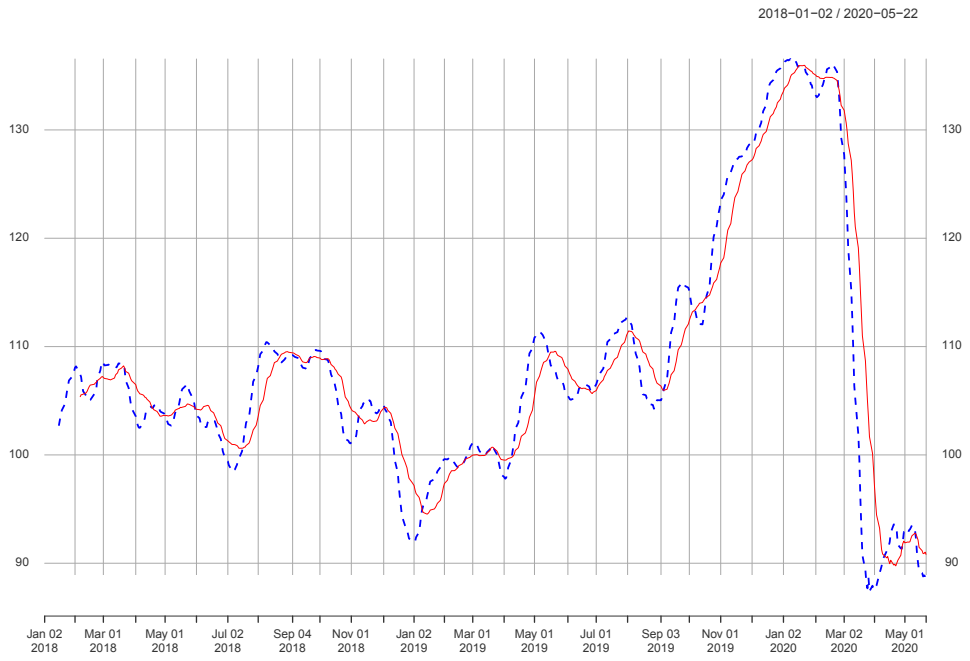


Fuente: Elaboración propia con el software R.

Es evidente que nuestra estrategia de inversión sentimental en JPM gana por goleada a una actitud pasiva. En la figura 2 se muestran la evolución del retorno acumulado y caídas de las estrategias MAx y BH. En la figura 3 se pueden observar los puntos de cruces de las medias móviles M M (10) y M M (25) para JPM en el periodo considerado de inversión.

FIGURA 3

CRUCES DE *M M* (10) (AZUL SEGMENTADA) Y *M M* (25) (ROJO CONTINUA) PARA JPM



Fuente: Elaboración propia con el software R.

4. PROPIEDADES EMPÍRICAS DE LOS INDICADORES DE SENTIMIENTO

En esta sección se presentan algunas de las propiedades empíricas más notables observadas sobre los indicadores de sentimiento público construidos en distintos contextos. Sirvan estas observaciones de orientación para una mejor comprensión de los datos, su potencial predictivo y utilidad en la construcción de modelos de pronóstico o sistemas de inversión.

Antes debemos advertir que, según un estudio realizado por Kumar y Lee (2006), estas propiedades empíricas de los indicadores de sentimiento parecen ser causados por y afectar principalmente a los inversores minoristas. Porque se acepta que los inversores institucionales son más racionales en la toma de decisiones de inversión basadas en

información pública (no por poseer una psique más estable, menos emocional, sino más bien y en gran parte debido a una mayor automatización de sus procesos de negociación y toma de decisiones, lo cual minimiza la intervención del factor humano en estos eventos). En consecuencia, es el inversor minorista quien se ve más afectado por el tono del sentimiento en las noticias financieras y más propenso a actuar en consecuencia, lo que hace que los precios de las acciones se desvíen de sus valores fundamentales. Por lo tanto, el análisis del sentimiento de los textos financieros y sus aplicaciones tiene más sentido en los mercados con una alta participación de inversores minoristas (principalmente de economías desarrolladas, como EE. UU. y Europa), en contraposición a los mercados emergentes. En estos mercados desarrollados, los inversores institucionales aún podrían explotar las desviaciones de los precios de las acciones de los valores fundamentales debido al comportamiento de los inversores minoristas impulsado por las noticias.

Listamos y comentamos a continuación aquellos hechos observables más comúnes para un amplio espectro de indicadores de sentimiento en el ámbito financiero.

4.1. Correlación entre el volumen de noticias y volatilidad del precio

Cuanto más noticias hay sobre una empresa cotizada, mayor será la volatilidad del precio de su acción en el mercado.

Esta dependencia entre el volumen de noticias y volatilidad de precios se ha observado para varias empresas cotizadas, en distintos mercados y utilizando diferentes fuentes de texto. Por ejemplo, esta relación se ha observado con mensajes extraídos de Twitter sobre empresas que aparecen en el S&P 500 por Arias, Arratia y Xuriguera (2013). Adicionalmente, Aouadi, Arouri y Teulon (2013), Dimpfl y Jank (2015), y Hamid y Heiden (2015) demuestran que para diferentes mercados de valores (Francia, S&P 500, Dow Jones) el volumen de búsquedas en Google (o "Google Trends"), tomado como indicador de atención sobre una compañía o tema económico, tiene poder predictivo de la volatilidad del mercado (en el sentido de una relación causa-efecto). Más recientemente Arratia y López (2020) demostraron esta relación causal del volumen de búsquedas en Google sobre bitcoins y el precio de las cripto-monedas de mayor liquidez.

4.2. Mayor volumen de noticias alrededor de las fechas de publicación de beneficios

El volumen de noticias sobre una empresa tiende a aumentar significativamente en los días cercanos a la publicación de beneficios obtenidos por la empresa (informes periódicos, usualmente trimestrales, de obligatoriedad por ley). Este hecho ha sido reportado y analizado por Tetlock, Saar-Tsechansky y Macskassy (2008) para las noticias que aparecen en el *Wall Street Journal* y el *Dow Jones Newswires* entre 1980 y 2004, para empresas listadas en el índice S&P 500. En su estudio los autores elaboraron un histograma

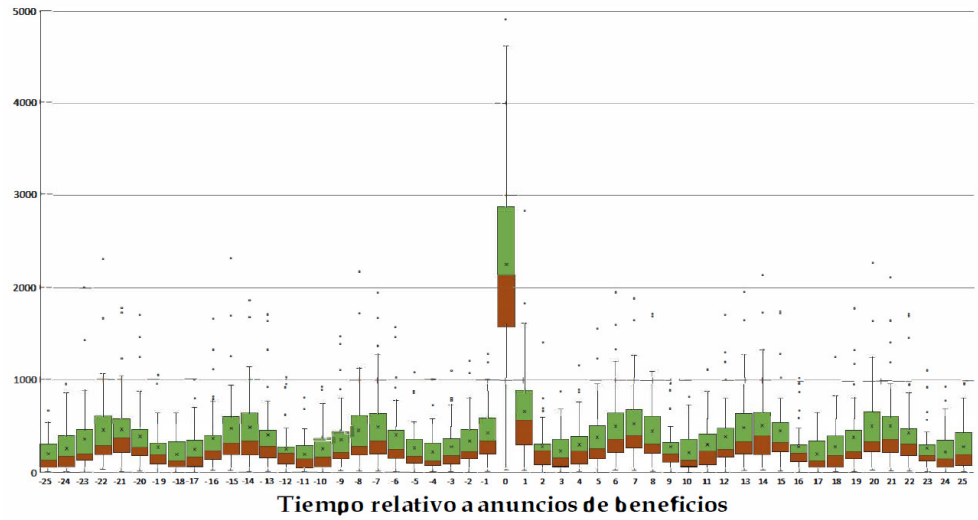
que describe la relación entre el número de noticias específicas sobre la empresa y el número de días desde (respectivamente, hasta) el último (respectivamente, próximo) anuncio de beneficios de la empresa, observando una mayor concentración en los registros cercanos a las fechas de beneficios, que se representan por el 0 del histograma. Este hecho lleva a sugerir a los autores una posible relación de dependencia estadística entre las noticias específicas sobre una empresa y los fundamentales de la empresa.

Hemos reproducido este análisis con datos actuales, para corroborar (y actualizar) las observaciones de Tetlock, Saar-Tsechansky y Macskassy (2008) (figura 4). En nuestra versión consideramos todas las noticias de empresas cotizadas en el S&P 500 que se publican en el *Dow Jones Newswires* desde 2012 a 2019. Sin embargo, consideramos que el efecto de las noticias en el comportamiento de los mercados financieros es de corto alcance y, por tanto reducimos el intervalo de análisis a 25 días antes y 25 días después de la publicación de beneficios.

El día de publicación de beneficios se representa en el 0. Para cada noticia correspondiente a una empresa, se calcula el número de días *hasta* el próximo reporte de beneficios, o el número de días *después* del último reporte de beneficios por esa empresa. Cada noticia contribuye solamente una vez al volumen de noticias en día $t \in [-25, 25]$, antes o después del reporte de beneficios (y no dos veces como en (Tetlock, Saar-Tsechansky y Macskassy, 2008)), y se realiza un *boxplot* de este volumen para cada período t de días.

FIGURA 4

VOLUMEN DE NOTICIAS PUBLICADAS EN EL *DOW JONES NEWSWIRES,* DESDE 2012 A 2019, ALREDEDOR DE DÍAS DE ANUNCIO DE BENEFICIOS PARA EMPRESAS COTIZADAS EN EL S&P 500



Fuente: Elaboración propia.

En todos los períodos, la distribución del número de noticias es muy asimétrica (todas las medias son más grandes que las medianas) y sus colas son pesadas por la derecha, excepto en el día que se hacen públicos los beneficios, donde parece más simétrica. En esta nueva gráfica podemos ver que el aumento más destacado en el volumen de noticias es el día exacto de publicación de beneficios, y el día inmediatamente posterior también tiene un aumento anormal con respecto al resto de la serie de volúmenes, lo que indica un aumento de noticias después de los hechos. Por lo demás se observa una estacionalidad semanal, seguramente debido a que el número de noticias depende del día de la semana, con menos volumen los fines de semana. El número de observaciones extremas cada día es pequeño: como máximo cinco empresas superan el límite estándar (1,5 veces el rango intercuartílico) para declarar el valor como un "valor atípico". Entonces, no podemos concluir de nuestra representación de la cobertura de los medios sobre los anuncios de beneficios que los sentimientos en las noticias pueden pronosticar indicadores fundamentales de una empresa (por ejemplo, precio-beneficios, precio-valor-libro, etc.) como se concluye en (Tetlock, Saar-Tsechansky y Macskassy, 2008), excepto quizás por las pocas empresas más populares en las noticias, los valores atípicos en nuestra gráfica. Sin embargo, especulamos que el sentimiento en las noticias después de las publicaciones de beneficios es el tipo de información útil para operar en corto, tal como ha sido considerado por Engelberg, Reed y Ringgenberg (2012).

4.3. Sentimientos de orientación negativa están más vinculados a movimientos en los mercados financieros que los sentimientos de orientación positiva

Este hecho se ha observado para varios mercados y empresas cotizadas en distintos sectores industriales. Algunas fuentes donde se han reportado ejemplos de este fenómeno son (Arias, Arratia y Xuriguera, 2013; Loughran y McDonald, 2011; Tetlock, Reed y Ringgenberg, 2008) y (Chan, 2003), aunque este último usa datos físicos de la prehistoria de internet. Allí se observa que esta mayor frecuencia en la dependencia entre negatividad y mercados es más evidente cuando se trata de empresas de media y pequeña capitalización.

4.4. Mayor efecto de los sentimientos sobre las compañías de capitalización media y pequeña

Este hecho se sugiere y es analizado por Chan (2003). Está relacionado con el hecho de que son los inversores minoristas quienes principalmente negocian en función del sentimiento de las noticias, y este tipo de inversores no mueven acciones de empresas de gran capitalización, un hecho que se argumenta extensamente en (Kumar y Lee, 2006).

5. HERRAMIENTAS PARA LA INFERENCIA ESTADÍSTICA

Para inferir algunas propiedades de los datos, útiles en la modelización de predictores, yendo más allá de la pura estadística descriptiva, es deseable tener a mano y poner en práctica un conjunto de contrastes estadísticos que desglosen las posibles relaciones entre los indicadores de sentimiento y las variables a predecir. En las próximas secciones pasamos revista a esos contrastes más utilizados en la inferencia de propiedades de las series temporales de nuestro interés.

Ejemplos de aplicación de estos contrastes estadísticos para determinar las posibles relaciones entre los indicadores de sentimiento de noticias y variables financieras se encuentran en los trabajos (Mendoza, 2018) y (Arratia y López, 2020).

5.1. Estacionaridad

En el contexto de variables económicas o sociales, se deberian intercambiar? normalmente observamos únicamente una realización del proceso estocástico subyacente que define estas variables. No es posible obtener muestras sucesivas o realizaciones independientes de dicho proceso. Para poder estimar las características "transversales" del proceso, como su media y su varianza, debemos asumir a partir de su evolución "longitudinal" que las propiedades transversales (distribución de las variables en cada instante en el tiempo) son estables en el tiempo. Esto conduce al concepto de estacionaridad.

Un proceso estocástico (serie temporal) es estacionario (o estrictamente estacionario) si las distribuciones marginales de todas las variables son idénticas y las distribuciones finito-dimensional de cualquier conjunto de variables dependen solo de la longitud de los retrasos que las separan. En particular, si la varianza es finita, la media y la varianza de todas las variables son las mismas. Más aún, la distribución conjunta de cualquier conjunto de variables es invariante por traslaciones (en el tiempo).

En general, estacionaridad estricta es una condición muy fuerte y difícil de verificar en la práctica, ya que se han de tener las distribuciones conjuntas para cualquier conjunto de variables dadas por el proceso. Nos conformamos entonces con una versión débil de estacionaridad, a saber, que sean invariantes por traslaciones temporales el primer y segundo momento de la distribución. Una serie temporal es débilmente estacionaria si EX_t es constante y $EX_{t+h}X_t$ solamente depende de h.

La estacionaridad de una serie temporal puede ocasionalmente ser evaluada mediante un contraste de Dickey y Fuller (1979). Este no es exactamente un contraste de la hipótesis nula de estacionaridad, sino más bien un contraste de la existencia de una raíz unitaria en un proceso autoregresivo. La hipótesis alternativa es que el proceso, o bien es estacionario, o que es estacionario en tendencia (*i.e.,* estacionario después de remo-

verle la tendencia). Es importante tener en cuenta que la aplicación del contraste de Dickey-Fuller asume implícitamente que el proceso observado se rige por un modelo autoregresivo. Asi que se debe al menos realizar previamente un contraste de autoregresión tipo ACF o PACF a los datos (cf. (Arratia, 2014, Ch. 2)).

5.2. Independencia

Antes de usar cualquier indicador como predictor, es importante determinar si existe alguna dependencia, en un sentido estadístico, entre el objetivo Y y el predictor X. De todos los contrastes estadísticos de dependencia existentes nuestro predilecto es el (relativamente nuevo) contraste basado en *correlaciones de distancias,* concebido por Székely *et al.* (2007).

Dadas dos variables aleatorias X e Y (posiblemente multivariadas), la correlación de distancias entre X e Y se calcula a partir de una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$, mediante el siguiente esquema:

- Calcular la distancia euclídea entre todos los pares de observaciones de cada vector de muestras, $X_i - X_j$ y $Y_i - Y_j$, $1 \leq i, j \leq n$, para así obtener 2 matrices $n \times n$ de distancias, una por cada vector.

- Centrar doblemente cada elemento de la matriz de distancia: a cada elemento restar la media de su fila y la media de su columna, y sumar la media de la matriz.

- Finalmente, calcular las covarianzas de las n^2 distancias centradas.

La correlación de distancias se obtiene al normalizar las covarianzas anteriores de tal forma que cuando $X = Y$ el resultado es 1. Se puede demostrar que cuando $n \to \infty$, las covarianzas de las distancias convergen a 0 si y solamente si los vectores X e Y son independientes.

A partir de esta somera descripción que hemos dado, deben ser obvios los puntos fuertes de la correlación de distancias, a saber: (1) caracteriza independencia; (2) se puede calcular para vectores numéricos, no únicamente para escalares; (3) al basarse en distancias, X e Y pueden tener dimensiones diferentes; (4) es invariante por rotaciones.

El contraste de independencia consiste en probar la hipótesis nula de cero correlación de distancias. Los p valores se obtienen mediante técnicas de *bootstrap*. El paquete energy de **R** (Rizzo y Szekely, 2018) incluye las funciones dcor y dcor.test para calcular la correlación de distancias y el correspondiente contraste de independencia.

5.3. Causalidad de Granger

Es también importante evaluar la posibilidad de causalidad (y no solo dependencia) de un proceso aleatorio X_t hacia otro proceso aleatorio Y_t. En nuestro caso X_t es un indicador temporal de sentimiento e Y_t la serie de retornos del precio de un activo, o cualquier otro estadístico del precio que deseamos predecir. La idea básica de causalidad se debe a Granger (1969), quien la formuló de la siguiente manera: X_t causa Y_t, si Y_t se puede predecir mejor utilizando la información del pasado de X_t junto con el pasado de Y_t, que con solamente el pasado de Y_t. Formalmente se ha de considerar un modelo bivariado de autoregresión sobre X_t y Y_t, con Y_t dependiente de los pasados de X_t y Y_t, junto con un modelo lineal autoregresivo solo para Y_t, y probar la hipótesis nula de "X_t no causa Y_t", lo cual significa probar que todos los coeficientes que acompañan las observaciones pasadas de X_t en el modelo bivariado autoregresivo son cero. Bajo la presunción de que los datos tienen una distribución normal, se puede evaluar esta hipótesis nula mediante un F-test. Este modelo vectorizado de autoregresión para probar la causalidad de Granger se debe a Toda y Yamamoto (1995), y tiene la ventaja de funcionar bien aún en presencia de series no estacionarias. Para mayor detalle y ejemplos de aplicaciones en series temporales financieras ver (Arratia, 2014, Ch. 3).

Existen otras propuestas más recientes para probar causalidad entre series temporales basados en métodos no paramétricos, métodos de kernel y teoría de la información, entre otros, que hacen frente a la no linealidad y no estacionariedad, pero sin tener en cuenta la presencia de información secundaria (causalidad condicional), ver p. ej. (Diks y Wolski, 2015; Marinazzo, Pellicoro, y Stramaglia, 2008; Wibral *et al.,* 2013). Para un contraste no paramétrico de causalidad condicional, ver (Arratia, Cabaña y Serès, 2016).

5.4. Selección de variables

El análisis de causalidad revela cualquier relación causa-efecto entre los indicadores de sentimiento y cualquiera de las funciones del precio de los activos financieros que tengamos como objetivo a predecir. El siguiente paso es analizar estos indicadores de sentimiento, individualmente o en conjunto, como variables independientes en un modelo de regresión para cualquiera de las variables financieras. Una razón fundamental para poner las variables juntas podría ser, al menos, lo que podrían tener en común semánticamente. Por ejemplo, juntar en un modelo de regresión todas las variables que expresan un sentimiento pesimista hacia el mercado *(bearish),* o todas aquellas que expresan un sentimiento optimista *(bullish)*.

No obstante, en un período de tiempo determinado, no todas las variables de uno de estos grupos pueden causar tanto a la variable objetivo como a algunas del grupo, y su adición en el modelo podría agregar ruido en lugar de información de valor. Por tanto, conviene tratar con un modelo de regresión que *discrimine la importancia de las variables.*

Aquí es donde entonces proponemos implementar una regresión LASSO con todas las variables bajo consideración que potencialmente expliquen (causen) la variable objetivo. El método LASSO, debido a Tibshirani (Tibshirani, 1996), optimiza el error cuadrático medio entre la variable objetivo y la combinación lineal de los regresores, sujeto a una penalización de tipo L_1 sobre los coeficientes de los regresores, lo que equivale a eliminar aquellos que son significativamente pequeños, eliminando así aquellas variables que aportan poco al modelo. LASSO no tiene en cuenta las posibles dependencias lineales entre los predictores, que pueden dar lugar a inestabilidades numéricas, por lo que recomendamos la verificación previa de las posibles correlaciones entre las variables a incluir en la regresión, descartando una de cada par de variables que estén altamente correlacionadas. Alternativamente, se puede intentar agregar una penalización de tipo L_2 en los coeficientes de los regresores, lo que conduce a una red elástica *(elastic net)*.

6. GUÍA DE SOFTWARE

Se dan aquí algunas sugerencias de software desarrollado en los lenguajes *R* y Python para el análisis de sentimiento de textos.

R

En la actualidad existen varias funciones integradas en diversos paquetes de *R* para analizar el sentimiento en un documento y construir indicadores de sentimiento. A continuación, hacemos una breve revisión de las herramientas a disposición en *R*, diseñadas exclusivamente para el análisis de sentimientos en textos. Esta lista no es de ninguna manera exhaustiva, ya que continuamente se publican nuevas actualizaciones debido al creciente interés en el campo. Además otras herramientas de análisis de sentimientos ya están incluidas implícitamente en paquetes de minería de textos más generales como *tm* (Meyer Hornik y Feinerer, 2008), *openNLP* (Hornik, 2019) y *qdap* (Rinker, 2020). De hecho, muchos de los paquetes actuales específicos para el análisis de sentimiento tienen una fuerte dependencia en los paquetes de minería de textos antes mencionados, como también en otros paquetes pertenecientes a la *CRAN Task View on Natural Language Processing*[4].

SentimentAnalysis (2019-03): diseñado para el análisis de sentimiento de textos basado en diccionarios. El paquete contiene varios diccionarios generales (p. ej. Harvard IV), o específicos del lenguaje financiero (como el diccionario recopilado por Lougran-McDonald), y permite la creación de diccionarios propios. Para esto último incorpora una rutina de regularización LASSO, como herramienta estadística para seleccionar términos relevantes basado en variables características de contexto (Feuerriegel y Proellochs, 2019).

[4] https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

RSentiment (2018-07): sirve para analizar el sentimiento de una oración en inglés y asignar una valoración correspondiente al nivel del sentimiento. Puede clasificar oraciones en las siguientes categorías de sentimientos: Positivo, Negativo, Muy Positivo, Muy Negativo, Neutral. Para un vector de oraciones, cuenta el número de oraciones en cada categoría de sentimiento. Al calcular la valoración, se tienen en cuenta los cambiadores de valencia (Bose, 2018).

sentimentr (2019-03): diseñado exclusivamente para calcular la polaridad en el sentimiento en un texto (Rinker, 2019).

sentometrics (2019-11): un sistema integrado para la construcción y agregación de series temporales de sentimiento textual. Contiene todas las funciones necesarias para implementar cada una de las etapas del proceso descrito en la sección 2 para crear modelos de predicción basados en el sentimiento extraído de las noticias (Ardia *et al.,* 2019).

quanteda (2019-11): diseñado para realizar análisis cuantitativo de datos textuales (Benoit y *et al.,* 2019).

syuzhet (2017): sirve para extraer sentimiento y graficar el árbol de derivación de sentimiento de textos (Jockers, 2017).

Python

Para los programadores de Python, también hay una gran cantidad de opciones para el análisis de sentimientos. De hecho, una búsqueda rápida de "Sentiment Analysis" en el índice de paquetes de Python (PyPI)[5] devuelve alrededor de 6000 registros. Aquí incluimos una lista reducida de los módulos más relevantes.

Vader: Valence Aware Dictionary for sEntiment Reasoning es un modelo basado en reglas (Hutto y Gilbert, 2014), principalmente entrenado en el análisis de textos sociales (por ejemplo, textos de medios sociales, reseñas de películas, etc.). Vader clasifica las oraciones en tres categorías: positivas, negativas y neutrales, con valores que representan las proporciones de las partes del texto que caen en cada categoría (la suma es 1 o cercana). También proporciona una valoración *compuesta* que se calcula sumando los valores normalizados de cada palabra en el léxico[6]. Una implementación de Vader se encuentra en la librería de uso general para el Procesamiento Natural del Lenguaje *nltk*.

TextBlob: para cualquier texto (en inglés) la librería TextBlob[7] provee de funciones para calcular el sentimiento en términos de polaridad, con valores continuos en [−1, 1],

[5] https://pypi.org/

[6] https://github.com/cjhutto/vaderSentiment\#about- the-scoring

[7] https://textblob.readthedocs.io/en/dev/quickstart.html\#sentiment-analysis

y subjetividad, con valores en [0, 1]. Para los valores de subjetividad, 0 significa muy objetivo y 1 es muy subjetivo.

Pattern: es un paquete multipropósitos para minería de la web, tareas de procesamiento natural de lenguaje (NLP, en inglés), aprendizaje automático y análisis de redes. El sentimiento se estima como combinación de polaridad y subjetividad, y estos se pueden obtener a nivel de documento o de palabras (Smedt y Daelemans, 2012).

pycorenlp: provee de una interface para la librería en Java de la Standford CoreNLP de la cual hereda varias funcionalidades[8]. La lista completa de las funciones en el coreNLP pueden verse en el sitio web[9].

Referencias

Achelis, S. B. (2001). *Technical Analysis from A to Z.* New York: McGraw Hill.

Algaba, A., Ardia, D., Bluteau, K., Borms, S., y Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys,* 34(3), pp. 512–547.

Antweiler, W. y Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance,* 59(3), pp. 1259–1294.

Aouadi, A., Arouri, M., y Teulon, F. (2013). Investor attention and stock market activity: Evidence from france. *Economic Modelling,* 35, pp. 674–681.

Ardia, D., Bluteau, K., Borms, S., y Boudt, K. (2019). *The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment.*

Arias, M., Arratia, A. y Xuriguera, R. (2013). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST),* 5(1), p. 8.

Arratia, A. (2014). *Computational Finance: An Introductory Course with R.* Atlantis Press.

Arratia, A., Cabaña, A. y Serès, A. (2016). Towards a sharp estimation of transfer entropy for identifying causality in financial time series. En: ECML-PKDD, *Proc. 1st Workshop MIDAS.* CEUR-WS. org.

Arratia, A. y López, A. (2020). Do google trends forecast bitcoins? stylized facts and statistical evidence. *Journal of Banking and Financial Technology*.

Baker, M. y Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives,* 21(2), pp. 129–152.

[8] https://pypi.org/project/pycorenlp/

[9] https://stanfordnlp.github.io/CoreNLP/other- languages.html

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., y Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology,* 5(4), pp. 323–370.

Beckers, B., Kholodilin, K. A., y Ulbricht, D. (2017). Reading between the lines: Using media to improve german inflation forecasts. *Technical report, DIW Berlin Discussion Paper.*

Benoit, K. y *et al*. (2019). *quanteda: Quantitative Analysis of Textual Data.* Version 1.5.2.

Berry, T. D. y Howe, K. M. (1994). Public information arrival. *The Journal of Finance,* 49(4), pp. 1331–1346.

Bifet, A. y Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. En: *International Conference on Discovery Science,* pages 1–15. Springer.

Bose, S. (2018). *Rsentiment: Analysis of Sentiment of English Sentences.* Version 2.2.2.

Brown, G. W. y Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance,* 11(1), pp. 1–27.

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics,* 70(2), pp. 223–260.

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., Luca, V. D., y Jaggi, M. (2016). Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. En: *Proceedings of the 10th international workshop on semantic evaluation,* pp. 1124–1128.

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., y Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. En: *Proceedings of the 26th international conference on world wide web,* pp. 1045–1052.

Dickey, D. A. y Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association,* 74(366a), pp. 427–431.

Diks, C. y Wolski, M. (2015). Nonlinear granger causality: Guidelines for multivariate analysis. *Journal of Applied Econometrics.*

Dimpfl, T. y Jank, S. (2015). Can internet search queries help to predict stock market volatility? *European Financial Management,* 22(2), pp. 171–192.

Ding, X., Liu, B., y Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. En *Proc. of the 2008 Int. Conf. on Web Search and Data Mining,* pp. 231–240. ACM.

Engelberg, J. E., Reed, A. V., y Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics,* 105(2), pp. 260–278.

Feuerriegel, S. y Proellochs, N. (2019). *SentimentAnalysis: Dictionary-Based Sentiment Analysis.* Version 1.3-3.

Go, A., Bhayani, R., y Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford,* 1(12).

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica,* 37, pp. 424–438.

Hamid, A. y Heiden, M. (2015). Forecasting volatility with empirical similarity and google trends. *Journal of Economic Behaviour and Organization,* 117, pp. 62–81.

Heston, S. L. y Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal,* 73(3), pp. 67–83.

Hornik, K. (2019). *openNLP: Apache OpenNLP Tools Interface. R Package Version 0.2.7.*

Hutto, C. J. y Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. En: *8th Int. AAAI Conf. on Weblogs and Social Media.*

Jockers, M. L. (2017). *Syuzhet: Extract Sentiment and Plot Arcs from Text.* Version 1.0.4.

Kumar, A. y Lee, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance,* 61(5), pp. 2451–2486.

Liu, B. (2015). *Sentiment Analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press.

Loughran, T. y McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance,* 66(1), pp. 35–65.

Marinazzo, D., Pellicoro, M., y Stramaglia, S. (2008). Kernel method for nonlinear granger causality. *Physical Review Letters,* 100(14), pp. 144103.

Mendoza, D. (2018). Indices de sentimiento en el ámbito financiero. Master's thesis, Dept. Matemáticas y Estadística, Universitat Autónoma de Barcelona.

Meyer, D., Hornik, K., y Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software,* 25(5), pp. 1–54.

Mitchell, M. L. y Mulherin, J. H. (1994). The impact of public information on the stock market. *The Journal of Finance,* 49(3), pp. 923–950.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., y Stoyanov, V. (2016). Semeval-2016 task 4: Sentiment analysis in twitter. En: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* pp. 1–18.

Polanyi, L. y Zaenen, A. (2006). Contextual valence shifters. En: *Computing Attitude and Affect in Text: Theory and Applications,* pp. 1–10. Springer.

Rao, D. y Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. En: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics,* pp. 675–682 . Association for Computational Linguistics.

Rinker, T. W. (2019). *sentimentr: Calculate Text Polarity Sentiment.* Version 2.7.1.

— (2020). *qdap: Quantitative Discourse Analysis.* Version 2.3.6. Buffalo, New York.

Rizzo, M. L. y Szekely, G. J. (2018). *energy: E-Statistics: Multivariate Inference via the Energy of Data.* R package version 1.7-4.

Rozin, P. y Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review,* 5(4), pp. 296–320.

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., y Gloor, P. (2013). The power of prediction with social media. *Internet Research.*

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR),* 34(1), pp. 1–47.

Smedt, T. D. y Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research,* 13(Jun), pp. 2063–2067.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., *et al.* (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics,* 35(6), pp. 2769–2794.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance,* 62, pp. 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., y Macskassy, S. (2008). More than words: Quantifying language to measure firm's fundamentals. *The Journal of Finance,* 63(3), pp. 1437–1467.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 58(1), pp. 267–288.

Toda, H. Y. y Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics,* 66(1-2), pp. 225–250.

Tsai, M.-F. y Wang, C.-J. (2014). Financial keyword expansion via continuous word vector representations. En: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 1453–1458.

Uhl, M. W., Pedersen, M., y Malitius, O. (2015). What's in the news? using news sentiment momentum for tactical asset allocation. *The Journal of Portfolio Management,* 41(2), pp. 100–112.

Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Linder, M., Lizier, J., y Vicente, R. (2013). Measuring information-transfer delays. *PLoS ONE,* 8(2), e55809.

CAPÍTULO VII

Desarrollos con *big data* para el análisis coyuntural en los bancos centrales

Corinna Ghirelli
Samuel Hurtado
Javier J. Pérez
Alberto Urtasun

Los bancos centrales utilizan intensivamente datos estructurados (micro y macro) para el desarrollo de sus funciones, entre las que destaca el seguimiento en tiempo real de la actividad económica. El desarrollo tecnológico está permitiendo integrar nuevas fuentes de datos masivos, más granulares y disponibles con mayor frecuencia, en muchos casos no estructuradas. Esto supone un desafío importante desde el punto de vista de la gestión, el almacenamiento, la seguridad y la confidencialidad. Este capítulo analiza las ventajas y los retos de estas nuevas fuentes, y describe algunos casos de éxito de su incorporación en el ámbito del análisis económico y la previsión.

Palabras clave: big data, datos masivos, predicción económica, análisis textual, incertidumbre económica, datos de prensa.

1. INTRODUCCIÓN[1]

El desarrollo de las nuevas tecnologías y las redes sociales ha abierto la posibilidad de utilizar nuevas fuentes de datos, que presentan características específicas en términos de volumen, nivel de detalle, frecuencia y (o falta de) estructura. En los últimos años se han desarrollado una gran cantidad de aplicaciones que explotan estas nuevas fuentes de datos en las áreas de economía y finanzas, particularmente en los bancos centrales (BC). En el área específica del análisis económico, las nuevas fuentes de datos presentan un potencial significativo para los BC, incluso teniendo en cuenta que estos ya venían haciendo un uso muy intensivo de datos estadísticos, tanto individuales (microdatos) como agregados (macroeconómicos) para realizar sus funciones.

En particular, estas nuevas fuentes están permitiendo a los BC, entre otras áreas de aplicación:

- La ampliación de la información utilizada para llevar a cabo las funciones de estabilidad financiera, supervisión bancaria y de sistemas de pagos (Broeders y Prenio, 2018; Fernández, 2019; Alonso y Carbó, 2020; Moreno Bernal y González Pedraz, 2020; Nyman, Kapadia, Tuckett *et al.,* 2018; Carlsen y Storgaard, 2010; Duarte, Rodrigues y Rua, 2017; Gil, Pérez, Sánchez y Urtasun, 2017).

- La aplicación de nuevas metodologías de análisis económico (véase, por ejemplo, Fernández-Villaverde, Hurtado y Nuño, 2019).

- Un análisis más profundo (datos más detallados) y un seguimiento más preciso (disponibilidad de datos casi en tiempo real) de la actividad económica (Kapetanios y Papailias, 2018; Thorsrud, 2018; Aprigliano, Ardizzi y Monteforte, 2017; Duarte, Rodrigues y Rua, 2017; Gotz y Knetsch, 2019; D'Amuri y Marcucci, 2017; Ferrara y Simoni 2019; Carriére-Swallow y Labbé, 2013) o alguna faceta de interés específico para los BC, como el seguimiento del crédito (Petropoulos *et al.,* 2019).

- Una mejor cuantificación de la confianza y la incertidumbre de los agentes, y sus expectativas de inflación o crecimiento (Baker, Bloom, y Davis, 2016; Ghirelli, Pérez y Urtasun, 2019 y 2020; Aguilar *et al.,* 2020).

- Una mejor valoración de la política económica y un aumento de la capacidad de simular medidas alternativas, debido principalmente a la disponibilidad de microdatos que pueden utilizarse para caracterizar de manera precisa la heterogeneidad de los agentes y, por tanto, llevar a cabo un análisis más profundo y

[1] Este capítulo descansa en gran medida en las ideas y resultados desarrollados en Ghirelli *et al.* (2019) y Ghirelli *et al.* (2020).

preciso de su comportamiento (Chetty, Friedman y Rockoff, 2014; Pew Research Center, 2012).

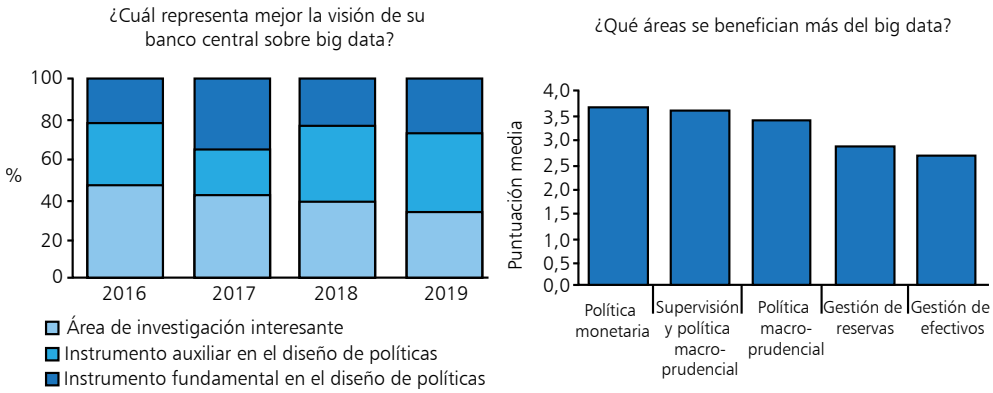
- La mejora de las estadísticas disponibles, como las relativas al turismo (Lacroix, 2019; Artola y Galán, 2010) o al mercado de la vivienda (Loberto, Luciani y Pangallo, 2018).

Según la encuesta anual del portal de información Central Banking[2], en 2019 más del 60% de los BC utilizaron fuentes de *big data* en sus operaciones, y dos tercios de ellos las utilizaron como instrumento principal o auxiliar en el proceso de diseño de sus políticas (véase la figura 1).

El resto de este capítulo está estructurado como sigue. En la sección segunda se proporciona una breve descripción de las principales ventajas y desafíos relacionados con la utilización de estas nuevas fuentes de datos por parte de los BC. En la tercera sección se describen algunos casos de éxito en que los BC han utilizado nuevas fuentes de datos para llevar a cabo sus funciones. En particular, nos centramos en la medición de la incertidumbre económica con artículos de prensa (apartado 3.1.), el uso de los informes regulares de un BC como herramienta de comunicación sobre el estado de la economía (apartado 3.2.), y algunas aplicaciones en el ámbito de la predicción macroeconómica (apartado 3.3.).

FIGURA 1

ALGUNOS RESULTADOS DE LA ENCUESTA DE CENTRAL BANKING (2019) SOBRE EL USO DE BIG DATA EN LOS BANCOS CENTRALES (ECONOMÍAS AVANZADAS Y EMERGENTES)



Fuente: Central Banking (2019) (https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central-banks-2019-survey-results).

[2] Disponible en https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central-banks-2019-survey-results. A la encuesta respondieron 58 bancos centrales, con una muestra importante tanto de países avanzados como de emergentes.

2. NUEVAS FUENTES DE DATOS: OPORTUNIDADES PARA LOS BANCOS CENTRALES[3]

Los BC utilizan de manera intensiva bases de datos estructurados para llevar a cabo sus funciones, en las áreas de supervisión bancaria, estabilidad financiera o política monetaria. Se utilizan datos de naturaleza micro y datos macro. Algunos ejemplos de micro datos son los balances de las empresas (véanse, por ejemplo, Menéndez y Mulino, 2018; Banco de España, 2018), la información relativa al volumen de crédito concedido por las instituciones financieras a personas y empresas, o los datos relacionados con las decisiones financieras de los agentes (véase, por ejemplo, Banco de España, 2017). En cambio, las principales fuentes de información en el ámbito de la macroeconomía suelen ser las cuentas nacionales o datos elaborados por los propios BC, así como un conjunto amplio de información publicada por otros organismos como, por ejemplo, datos de seguridad social (Ministerio de Seguridad Social), precios de las acciones (Bloomberg) o precios de la vivienda (portales inmobiliarios).

Gracias a los avances tecnológicos las fuentes de información se están ampliando de manera significativa, en particular en lo que se refiere a su granularidad y frecuencia. En muchos casos se puede obtener información casi en tiempo real sobre decisiones individuales realizadas por individuos o empresas, y la mayoría de las veces a frecuencias más altas que con fuentes de datos tradicionales. Por ejemplo, los datos de transacciones de tarjetas de crédito, que se pueden utilizar para aproximar las decisiones de consumo de los hogares, están potencialmente disponibles en tiempo real con un coste reducido, especialmente si se compara con el coste de realizar encuestas de hogares en todo el país.

La disponibilidad de grandes cantidades de información supone enfrentarse a retos importantes en cuanto a su gestión, a las necesidades y capacidades de almacenamiento, a los costes, seguridad y confidencialidad de la infraestructura necesaria, o a la calidad de los datos (sobre este último punto, véase Einav y Levin, 2014). Además, la gestión óptima de grandes conjuntos de datos estructurados y no estructurados requiere la integración de nuevos perfiles profesionales (científicos e ingenieros de datos) en los BC e implica la necesidad de su completa transformación digital. Asimismo, la diferente naturaleza de las nuevas fuentes de información requiere la asimilación y el desarrollo de técnicas para transformar y sintetizar los datos brutos en formatos que puedan incorporarse al análisis económico. Por ejemplo, las técnicas de análisis textual permiten procesar la información contenida en textos y convertirla en datos estructurados. Este ocurre, por ejemplo, con la información que se obtiene, entre otros, de noticias de prensa, redes sociales (Facebook y Twitter), portales de búsqueda web (por ejemplo, de vivienda o trabajo).

[3] La discusión en esta sección presenta las principales ideas de nuestro trabajo conjunto con Juan Peñalosa en Ghirelli *et al.* (2019).

Las nuevas fuentes de datos están ampliando la frontera de la estadística, en particular en el campo de las estadísticas no financieras. Este es el caso, por ejemplo, de las iniciativas para adquirir mejores medidas de precios en la economía utilizando técnicas de *web-scraping*[4] (Loberto, Luciani y Pangallo, 2018), o determinados elementos del comercio exterior, como la estimación de los movimientos turísticos mediante el seguimiento de redes móviles u otras fuentes (véase Hardy *et al.,* 2017; Lacroix, 2019; Artola y Galán, 2010). Los países en desarrollo, que se enfrentan a mayores dificultades para establecer infraestructuras estadísticas sólidas, están comenzando a utilizar las nuevas fuentes de datos, incluso para realizar estimaciones de algunos agregados de las cuentas nacionales (véase Hammer, Kostroch y Quirós, 2017). Asimismo, la ampliación de las fronteras de la estadística también está afectando al "monopolio público" sobre la información, dado que gran parte de la nueva información son propiedad de empresas privadas, y mucha de la información es incluso de libre acceso a través de la web.

3. ALGUNAS APLICACIONES EN EL ÁMBITO DE LOS BANCOS CENTRALES

3.1. Uso de la prensa para medir la incertidumbre

Las aplicaciones que involucran el análisis de texto han adquirido una importancia especial en el área del análisis económico. Con estas técnicas, se puede obtener la información relevante de un texto y luego sintetizarla y codificarla en forma de indicadores cuantitativos. Primero, el texto se prepara (preprocesamiento), eliminando los elementos del texto que no son relevantes (artículos, palabras no relevantes, números) y cortando las palabras a su raíz (es decir, quitando las terminaciones). Segundo, la información relevante se sintetiza principalmente mediante el cálculo de la frecuencia de palabras o grupos de palabras. Intuitivamente, la frecuencia relativa de grupos de palabras relacionados con un determinado tema permite evaluar la importancia relativa de este tema en el texto. Los datos de texto son una nueva fuente de información valiosa, ya que reflejan los principales acontecimientos actuales que afectan a las decisiones de los agentes económicos y están disponibles en tiempo real.

Una rama reciente de la literatura se centra en construir indicadores de incertidumbre económica a partir de artículos de prensa. El trabajo más influyente y seminal de esta literatura es Baker *et al.* (2016). Estos autores construyeron un índice de incertidumbre acerca de las políticas económicas para Estados Unidos (*Economic Policy Uncertainty Index, EPU*), basado en el volumen de artículos de periódicos que contienen palabras relacionadas con los conceptos de incertidumbre, economía y política. A raíz de este artículo, muchos investigadores han utilizado en sus análisis indicadores de incertidumbre basados en textos, proporcionando evidencia empírica de los efectos negativos de aumentos de la incertidumbre sobre la actividad económica como, por ejemplo,

[4] El web-scraping es un proceso en el que se navega automáticamente en sitios web para extraer contenido y datos de esas páginas.

Meinen y Roehe (2017) para Alemania, Francia, Italia y España;Fontaine, Didier y Razafindravaosolonirina (2017) para China, o Colombo (2013) y Azqueta-Gavaldon *et al.* (2020) para la eurozona.

El resto de esta sección presenta dos estudios que, siguiendo estas metodologías, desarrollan indicadores de incertidumbre acerca de las políticas económicas para: (1) la economía española y (2) para los principales países de América Latina (Argentina, Brasil, Chile, Colombia, México, Perú y Venezuela). Estos indicadores se han construido a partir de la prensa española y se utilizan actualmente en las tareas de seguimiento y previsión económica del Banco de España.

3.1.1. Incertidumbre acerca de las políticas económicas en España

Si bien Baker, Bloom y Davis (2016) también elaboraron un índice EPU para España basado en los dos principales periódicos generalistas españoles (*El País* y *El Mundo*), Ghirelli, Pérez y Urtasun (2019) desarrollan un nuevo índice de incertidumbre acerca de las políticas económicas para España, que se basa en la metodología de Baker, Bloom y Davis (2016) pero amplía la cobertura de prensa de dos a siete periódicos[5], extiende su cobertura temporal (a partir de 1997 en lugar de 2001), y enriquece las palabras clave utilizadas en las expresiones de búsqueda.

El indicador aumenta o disminuye en momentos relacionados con eventos asociados con un aumento o una disminución de la incertidumbre económica *a priori,* como los ataques terroristas del 11 de septiembre de 2001 en Estados Unidos, el colapso de Lehman Brothers en septiembre de 2008, la solicitud de ayuda financiera de Grecia en abril de 2010, la solicitud de ayuda financiera para reestructurar el sector bancario y los bancos de ahorro en España en junio de 2012, el referéndum del *brexit* en junio de 2016, o los episodios de tensión política en la región española de Cataluña en octubre de 2017. Ghirelli, Pérez y Urtasun (2019) muestran la existencia de una relación dinámica significativa entre este indicador y las principales variables macroeconómicas, de manera que incrementos inesperados en el indicador de incertidumbre acerca de las políticas económicas tienen efectos macroeconómicos adversos. Específicamente, un aumento de la incertidumbre reduce significativamente el PIB, el consumo y a la inversión. Este resultado está en línea con los resultados de la literatura empírica acerca de la incertidumbre económica.

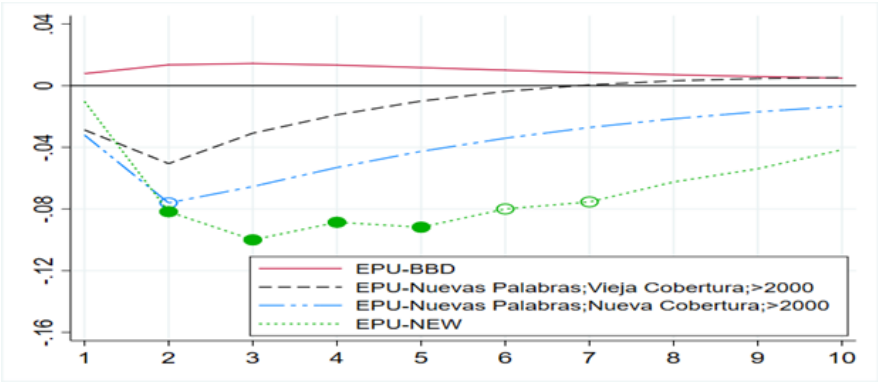
Además, en ese artículo se muestra la importancia de enriquecer el conjunto de palabras clave utilizado en las expresiones de búsqueda y ampliar la cobertura de prensa y temporal en la construcción del índice. Estas conclusiones se muestran en la figura 2,

[5] El nuevo índice se basa en los cuatro periódicos de carácter generalista más leídos en España y los tres principales periódicos de negocios: *El País, El Mundo, El Economista, Cinco Días, Expansión, ABC, La Vanguardia.*

que compara las respuestas macroeconómicas a perturbaciones de incertidumbre, usando versiones alternativas del índice EPU, en que se varían una por una las dimensiones mencionadas anteriormente (riqueza de las palabras, cobertura temporal y cobertura de prensa), pasando del índice EPU construido por Baker, Bloom, y Davis (2016) al nuevo índice construido por Ghirelli, Pérez y Urtasun (2019). Todas estas dimensiones son importantes ya que todas contribuyen a obtener el signo negativo esperado en las respuestas. Sin embargo, ampliar la cobertura temporal es clave para mejorar la precisión de las estimaciones y obtener resultados significativos, así como la cobertura de prensa.

FIGURA 2

RESPUESTA DEL PIB DE ESPAÑA ANTE DISTINTAS PERTURBACIONES DE INCERTIDUMBRE CONSTRUIDAS CON VERSIONES ALTERNATIVAS DEL ÍNDICE EPU



Nota: La figura muestra la función impulso-respuesta de la tasa de crecimiento del PIB español hasta 10 trimestres después de un shock positivo de una desviación estándar en el índice EPU para España. El eje horizontal representa los trimestres desde el choque. El eje vertical mide la tasa de crecimiento del PIB español (en puntos porcentuales). Los círculos completos (vacíos) indican significatividad estadística al 5% (10%); la línea continua indica la ausencia de significatividad estadística. EPU-BBD: índice EPU para España proporcionado por Baker, Bloom y Davis (2016). EPU-NEW: índice EPU para España elaborado por Ghirelli, Pérez y Urtasun (2019). Los modelos de autorregresión vectorial (VAR) incluyen el índice EPU, el spread entre los bonos alemanes y los bonos españoles, la tasa de crecimiento del PIB y la tasa de crecimiento del índice de precios al consumidor (IPC); el EPU global se incluye como variable exógena.

Fuentes: Banco de España y Policy-Uncertainty-web.

3.1.2. Incertidumbre acerca de la política económica en América Latina

La literatura también demuestra que la incertidumbre económica en un país puede tener efectos indirectos en otros países, así como ramificaciones globales. En este marco, Ghirelli, Pérez y Urtasun (2020) desarrollan índices de incertidumbre acerca de las políticas económicas para los principales países de América Latina (AL): Argentina,

Brasil, Chile, Colombia, México, Perú y Venezuela. El objetivo es doble. Primero, medir la incertidumbre acerca de las políticas económicas en los países de AL, y obtener una narrativa de los "choques de incertidumbre" y sus efectos potenciales sobre la actividad económica en estos países. Segundo, explorar en qué medida esos choques en AL tienen el potencial de extenderse a España. Este último país representa un caso relevante para estudiar este tipo de contagio internacional, dados sus importantes vínculos económicos con la región latinoamericana.

Los indicadores de incertidumbre se construyen siguiendo la misma metodología descrita en la sección anterior, es decir, contando los artículos de los siete diarios españoles más importantes que contienen palabras relacionadas con los conceptos de economía, política e incertidumbre. Sin embargo, además, se modifican las búsquedas para que se adapten a cada uno de los países latinoamericanos de interés[6]. Estos indicadores se basan en la prensa española y, por tanto, reflejan puramente la variación en la incertidumbre en los países de AL que es relevante para la economía española. El supuesto es que la prensa española refleja con precisión la situación política, social y económica de la región de AL, dados los estrechos vínculos económicos y culturales existentes, incluido el idioma común para la mayoría de países. Por esta razón, se puede afirmar que los índices basados en la prensa española proporcionan medidas relevantes acerca de la incertidumbre política en esos países. Esta metodología también es coherente con una rama de la literatura que utiliza la prensa internacional para calcular indicadores de texto para un amplio conjunto de países (véanse, por ejemplo, Ahir, Bloom y Furceri, 2019, y Mueller y Rauh, 2018).

En el estudio de referencia se muestra como una mayor incertidumbre en América Latina afecta a las empresas españolas más expuestas a la región. En particular, los resultados empíricos muestran que una perturbación positiva inesperada en el índice EPU agregado para AL, genera una caída significativa en la tasa de variación de la cotización de dichas empresas. El resultado se mantiene cuando se realizan ejercicios individuales para todos los países de AL considerados en el trabajo. Y lo confirman las pruebas de placebo, que consideran empresas españolas que cotizan en la bolsa española pero que no tienen intereses económicos en la región latinoamericana. En un segundo conjunto de resultados, Ghirelli, Pérez y Urtasun (2020), muestran como los *shocks* en los índices EPU de AL afectan a las siguientes variables macroeconómicas españolas: las exportaciones, la inversión extranjera directa de España a América Latina y el PIB español.

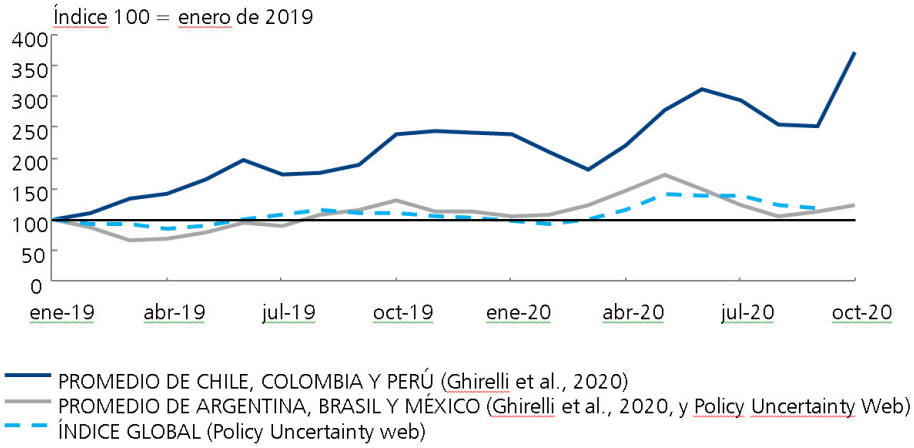
Además de los resultados discutidos, de un carácter más estructural, la disponibilidad de estos índices en tiempo real permite su uso para el seguimiento de la incertidumbre

[6] En particular, (1) se requiere que cada artículo contenga el nombre del país de interés de AL; (2) entre el conjunto de palabras clave relacionadas con la política, se incluye también el nombre del banco central y el nombre del lugar de trabajo del gobierno del país de interés. Para obtener más detalles, véase Ghirelli, Pérez y Urtasun, (2020).

sobre las políticas económicas en los países de referencia de América Latina, lo que, dados los resultados apuntados en el párrafo anterior, resulta relevante para la estimación de corto plazo *(nowcasting)* de las principales macromagnitudes españolas, pero también de estos países. En la figura 3 se muestra la evolución reciente de estos índices, desde enero de 2019 hasta octubre de 2020, agrupando los países analizados y normalizando el valor de los distintos índices a 100 en enero de 2019. Como puede observarse, la crisis asociada a la COVID-19 causó un aumento generalizado de la incertidumbre en todos los países de AL y a nivel global. No obstante, los países en los que también se vienen registrando tensiones sociales e institucionales presentaron aumentos mucho más marcados, como se refleja con claridad en el agregado de Chile, Perú y Colombia, al compararlo con el de Argentina, Brasil y México.

FIGURA 3

EVOLUCIÓN RECIENTE DE LA INCERTIDUMBRE SOBRE LAS POLÍTICAS ECONÓMICAS EN ALGUNAS ECONOMÍAS DE AMÉRICA LATINA



Nota: Medias móviles de tres meses.

Fuentes: Banco de España y Policy-uncertainty-web.

3.2. La narrativa sobre la economía como previsión sombra: un análisis con informes trimestrales del Banco de España

Otra de las técnicas de minería de texto consiste en el uso de métodos de diccionario para el análisis de sentimiento. Un diccionario es una lista de palabras asociadas con sentimientos positivos y negativos. Estas listas se pueden construir de varias formas, desde técnicas puramente manuales hasta técnicas de aprendizaje automático. El análisis de sentimiento a su vez se basa en búsquedas en bases de datos de texto y requiere que el investigador tenga acceso a los textos. En su versión más simple, las búsquedas permiten calcular la frecuencia de términos positivos y negativos en un texto. El índice

de sentimiento se define como la diferencia (con algunos pesos) entre las dos frecuencias, es decir, un texto tiene un sentimiento positivo (negativo) cuando la frecuencia de términos positivos es mayor (menor) que la de los términos negativos.

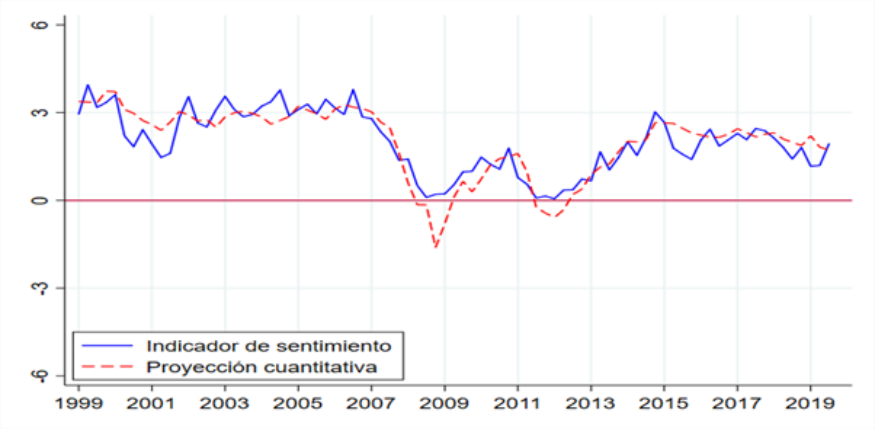
En este arpartado, se proporciona un ejemplo de análisis de sentimiento para mostrar la utilidad de los datos de texto (Diaz-Sobrino *et al.,* 2020). Para ello, se utiliza la técnica más básica de análisis de sentimiento, es decir, el simple conteo de palabras contenidas en un diccionario creado por los autores de este capítulo. El ejercicio se basa en el *Boletín Económico Trimestral de la economía española* del Banco de España, publicado *online* desde el primer trimestre de 1999. El objetivo es construir un indicador (a partir del primer trimestre 1999) que refleje el sentimiento de los informes de perspectivas económicas del Banco de España. El análisis muestra que el indicador refleja correctamente las previsiones cuantitativas del PIB del Banco de España. Esto significa que la narrativa cualitativa de los informes contiene información similar a la transmitida por las previsiones cuantitativas.

Para realizar el análisis, primero se crea un diccionario en español de términos positivos y negativos que se usan típicamente en el lenguaje económico para describir la economía. Por ejemplo, palabras como "crecimiento" o "aumento" entre términos positivos, y "disminución" o "reducción" entre términos negativos. Para disminuir los casos de signos incorrectos, se ignoran estos términos cuando aparecen alrededor (dentro de nueve palabras antes o después) de palabras que tienen un significado opuesto, como "desempleo" o "déficit". Por último, se asigna un peso de +1 y -1 a las cuentas de términos positivos y negativos en cada texto, se suma las cuentas ponderadas de los términos del diccionario y se divide el número resultante por la longitud del texto. A continuación, se compara el índice con las previsiones de crecimiento del PIB realizadas cada trimestre por el Banco de España, que en la mayor parte de la muestra considerada fueron registradas internamente pero no publicadas *online*.

Los resultados muestran una relación dinámica significativa entre ambas series: el indicador de sentimiento sigue fielmente el ciclo español, así como la evolución de las previsiones cuantitativas. Además, la comparación muestra que los informes son informativos no solo en el horizonte de previsión a corto plazo, sino también en el horizonte a uno a dos años. El indicador de sentimiento muestra la correlación más alta con las previsiones a dos años. La figura 4 muestra el indicador textual (línea azul continua) frente a la previsión de crecimiento del PIB realizada por el Banco de España para el horizonte a dos años (línea roja discontinua). Esta evidencia sugiere que la narrativa contenida en los informes del Banco de España refleja de manera muy fiel las previsiones cuantitativas de crecimiento del PIB obtenida por esta misma institución. Esto significa que un lector "sofisticado" podría haber inferido las previsiones de crecimiento del PIB del Banco de España a partir de sus informes.

FIGURA 4

PREVISIONES CUANTITATIVAS E INDICADOR DE SENTIMIENTO DE LOS INFORMES DEL BANCO DE ESPAÑA



Nota: La figura muestra el indicador de sentimiento (línea azul continua) frente a las previsiones cuantitativas del Banco de España (línea roja discontinua).

Fuente: Banco de España.

3.3. Previsiones con nuevas fuentes de datos

Normalmente, los ejercicios de previsión de los bancos centrales se llevan a cabo combinando indicadores cualitativos *(soft)* con indicadores cuantitativos *(hard),* que representan el conjunto de información publicado por las instituciones de estadística. La principal limitación de los datos cuantitativos es que se publican con cierto retraso y con baja frecuencia (por ejemplo, trimestralmente). En cambio, los indicadores cualitativos proporcionan información cualitativa (por lo tanto, de menor calidad que los datos cuantitativos) acerca de la coyuntura económica con una frecuencia más alta que los datos cuantitativos, como por ejemplo las encuestas de confianza de empresas y consumidores. La utilidad de los indicadores cualitativos es máxima al inicio del trimestre, cuando falta información macroeconómica, y disminuye tan pronto como se publican indicadores *hard* (Ferrara y Simoni, 2019).

Los indicadores de texto son un nuevo tipo de indicadores cualitativos. En comparación con los tradicionales, basados en encuestas, los indicadores textuales muestran las siguientes características: (1) suponen un menor coste, ya que no se basan en encuestas mensuales sino en suscripciones a servicios de prensa; (2) proporcionan más flexibilidad, ya que se pueden seleccionar las palabras clave en función de las necesidades específicas y obtener la serie temporal completa (mirando a los textos pasados), mientras que en una encuesta, se debería incluir una nueva pregunta y en consecuencia la serie temporal empezaría a partir de ese momento.

El resto de esta sección presenta tres aplicaciones en que se muestra como las nuevas fuentes de datos pueden mejorar los ejercicios de previsión económica. La primera se basa en el análisis de sentimiento. La segunda aplicación muestra cómo el aprendizaje automático puede mejorar la precisión de las técnicas de previsión disponibles. Finalmente, la tercera aplicación valora la importancia relativa de indicadores basados en nuevas fuentes de datos, como Google Trends y transacciones de tarjetas de crédito.

3.3.1. Un método "supervisado"

Este ejercicio se centra en construir un indicador textual para mejorar el seguimiento de la actividad económica (para detalles, véase Aguilar *et al.,* 2020). Para ello, se utiliza un procedimiento similar a lo que se ha descrito anteriormente para la elaboración del indicador de incertidumbre acerca de las políticas económicas, es decir, descansa en el número de artículos en la prensa española que contienen palabras clave específicas, y se usan los mismos periódicos para ello. Además, en este caso, se construye un diccionario de palabras positivas y negativas que se suelen utilizar cuando se describe la evolución de la tasa de crecimiento del PIB, la variable objetivo de interés, para identificar correctamente el tono de los artículos de prensa y, en particular, hasta qué punto se está tratando de repuntes o desaceleraciones de la economía. En concreto, el indicador se construye en la siguiente manera:

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones positivas. Se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recuperacion*" o una de las siguientes palabras (aceler*, crec*, increment*, recuper*, aument*, expansi*, mejora*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de 5 palabras entre sí. Se ignora "crecimiento económico" dado que tiene un tono neutro (se usa para describir indiferentemente un crecimiento negativo o positivo).

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones negativas: se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recesión*" o "crisis" o una de las siguientes palabras (descen*, ralentiz*, redu*, disminu*, contraccion*, decrec*, desaceler*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de cinco palabras entre sí.

Sucesivamente, para cada periódico, se calcula la diferencia entre el total de artículos relacionados con las recuperaciones y las recesiones, y si divide el valor resultante por

el número total de artículos económicos publicados en el mismo periódico en cada mes. Esta proporción, primero, se estandariza, a continuación, se computa una serie agregada tomando la media entre las series de los distintos periódicos y, finalmente, se le quita la media para que tenga media 0. El panel derecho de la figura 5 muestra el indicador textual (línea azul continua) contra la tasa de crecimiento del PIB (línea roja y discontinua). Aguilar *et al.* (2020) muestran que el indicador textual tiene poder predictivo para la previsión de la tasa de crecimiento del PIB español, a través de un ejercicio de previsión de PIB (datos del PIB no revisados) a corto plazo en (pseudo) tiempo real.

Una de las principales ventajas de los indicadores basados en periódicos es que se pueden actualizar en tiempo real y son de alta frecuencia. Estas ventajas han sido extremadamente valiosas desde el brote de la COVID-19, cuando los indicadores de confianza tradicionales basados en encuestas no han proporcionado señales correctas sobre la actividad económica[7]. Como ejemplo, el panel derecho de la figura 5 se muestra el indicador textual con una frecuencia semanal alrededor del confinamiento (14 de marzo de 2020). Su evolución refleja correctamente la drástica reducción de la actividad económica española en esa época.

3.3.2. Un método "no supervisado"[8]

Este ejercicio se compone de dos partes: en una primera parte se utiliza el método *Latent Dirichlet Allocation (LDA)* (Blei, Ng y Jordan, 2003) para extraer un conjunto de artículos de prensa indicadores cuantitativos que representen la importancia de determinados temas a lo largo del tiempo. La segunda parte utiliza los datos resultantes del método LDA para mejorar las previsiones del PIB español a través de un modelo de aprendizaje automático.

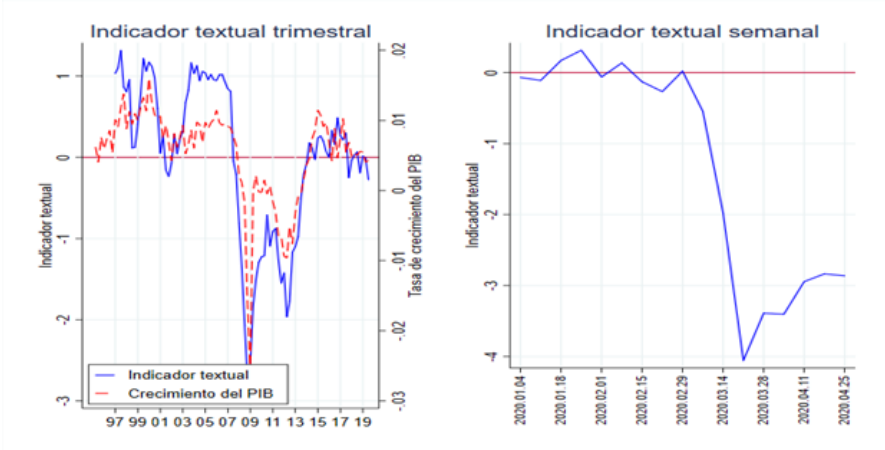
El LDA es un método de aprendizaje no supervisado, lo que significa que la definición de los temas no la decide el investigador, sino que es el resultado de ejecutar el modelo sobre los datos. El primer paso del proceso es construir un *corpus* con datos de texto. En este caso, se trata de una base de datos de más de 780.000 observaciones que contienen todas las noticias publicadas por *El Mundo* (uno de los principales periódicos españoles) entre 1997 y 2018, extraídas del repositorio de prensa española Dow Jones. A continuación, se tiene que procesar la base de datos de manera que se obtenga una versión del *corpus* que excluya puntuación, números o caracteres especiales, en que todo el texto esté en minúscula y que excluya las palabras más comunes como artículos y conjunciones. También es útil reducir las palabras a su raíz básica (la parte de la palabra que captura su significado central) eliminando algunas variaciones de las palabras como, por ejemplo, los tiempos verbales.

[7] En Aguilar *et al.* (2020), se compara este indicador textual con el indicador de sentimiento económico (ESI) de la Comisión Europea y se muestra que, para España, el primero mejora significativamente la previsión a corto plazo del PIB en comparación con el ESI.

[8] Véase Ghirelli *et al.* (2020).

FIGURA 5

UN INDICADOR TEXTUAL PARA ANTICIPAR LA EVOLUCIÓN DEL PIB DE ESPAÑA



Nota: La figura de la derecha muestra el indicador textual trimestral de la economía (línea azul y sólida) frente a la tasa de crecimiento del PIB español (línea roja y discontinua) hasta junio de 2019. La figura de la izquierda muestra el indicador textual semanal de enero a marzo de 2020.

Fuentes: Banco de España e Instituto Nacional de Estadistica (INE).

El segundo paso es representar el conjunto de textos con un modelo "bolsa-de-palabras" *(bag-of-words)*: en términos prácticos, una tabla con una fila para cada texto y una columna para cada posible palabra. Cada celda contiene números que indican cuántas veces aparece cada palabra en cada texto (notar que por construcción la matriz contiene muchos ceros porque la mayoría de las palabras de un diccionario extenso no aparecen en la mayoría de los textos).

A continuación, el algoritmo LDA procesa esta representación "bolsa-de-palabras" para tratar de identificar 128 temas diferentes de que se trata en la base de datos[9] y asignar a cada texto la probabilidad de que pertenezca a cada uno de los esos temas. Intuitivamente, el algoritmo analiza los textos y determina qué palabras tienden a aparecer juntas y cuáles no, asignándolas de manera óptima a diferentes temas para minimizar la distancia entre textos asignados a un tema determinado y maximizar la distancia entre textos asignados a diferentes temas. El resultado final es una base de datos que contiene, para cada trimestre de 1997 a 2018, el porcentaje de artículos que tratan cada uno de los 128 temas identificados por el modelo de aprendizaje no supervisado. Además, cada artículo se procesa también con un diccionario de términos positivos y negativos, y los resultados se agregan en indicadores trimestrales que

[9] En los modelos LDA, el investigador debe elegir el número de temas que quiere extraer. En general, la cantidad de temas se elige minimizando medidas de la bondad del modelo LDA.

representen en qué medida los artículos relacionados con cada tema tienen un tono positivo o negativo.

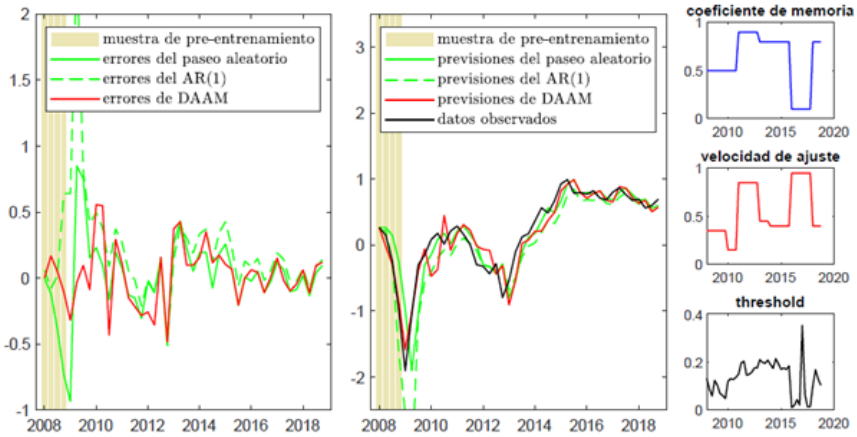
En la segunda parte del ejercicio, se recurre a un modelo de aprendizaje automático para explotar los datos resultantes del método LDA y obtener previsiones del PIB español. El aprendizaje automático se refiere a una gama muy amplia de métodos y algoritmos utilizados en diferentes campos, como la visión artificial. En el contexto de la economía, se utilizan máquinas de vectores de soporte, bosques aleatorios y redes neuronales para analizar microdatos sobre millones de consumidores o empresas y encontrar correlaciones, patrones de comportamiento e incluso relaciones causales. Para la previsión de series temporales, se utilizan técnicas *ensemble* (como *boosting* y *bagging*) para construir modelos de previsión sólidos combinando de manera óptima una gran cantidad de modelos más débiles (para más detalles sobre estas técnicas, véase Barrow y Crone, 2016).

Para este ejercicio se utiliza uno de estos modelos ensemble para hacer la previsión del PIB: un modelo de agregación doblemente adaptativo que utiliza los resultados del método LDA (denominado DAAM-LDA, de su acrónimo en inglés). Una ventaja de este modelo es que puede adaptarse a cambios en las relaciones entre los datos a lo largo del tiempo. Los ingredientes del modelo son un conjunto de 128 modelos de series temporales simples y de bajo desempeño; en particular, cada modelo es la regresión del crecimiento trimestral del PIB español sobre su valor retrasado y el peso de cada tema en cada trimestre, y su tono prevalente, representado por un indicador de positividad y negatividad. La estimación se hace en tiempo real, de manera que los modelos se estiman cada trimestre y se guarda la previsión del primer trimestre fuera de la muestra. La mayoría de los estos modelos muestran un desempeño débil fuera de la muestra por dos razones: (1) el peso de cada tema en la prensa y su positividad o negatividad son indicadores con una relación señal/ruido relativamente baja; (2) la mayoría de los temas identificados por el método LDA no están realmente relacionados con la economía. Efectivamente, solo cuatro de los 128 modelos muestran un mejor desempeño que un simple paseo aleatorio. La ventaja de los métodos ensemble es que están diseñados específicamente para construir modelos sólidos a partir de un conjunto de modelos débiles. Además, el investigador no tiene que decidir qué temas son útiles y cuáles no: en cada periodo el modelo descarta automáticamente cualquier tema que no proporciona buenas previsiones en los períodos anteriores. En este caso, se ha decidido combinar las previsiones a través de una función umbral a un parámetro que varía a lo largo del tiempo[10]. Es decir, el modelo "ensemble" analiza el desempeño reciente de cada modelo de partida para decidir si utilizarlo en el cómputo del promedio del siguiente trimestre o no (pero todos los modelos que se consideran en el promedio tienen igual peso).

[10] Una manera más sofisticada seria construir una función de ponderación no lineal que trasforme el desempeño de cada modelo en el trimestre actual en su ponderación óptima para el trimestre sucesivo. En este caso, la función de ponderación óptima es muy parecida a una función umbral a un parámetro, lo que explica que se haya elegido esta simplificación en este ejercicio.

FIGURA 6

RESULTADOS DEL EJERCICIO DE PREVISIÓN EN TIEMPO REAL DEL CRECIMIENTO DEL PIB ESPAÑOL TRIMESTRAL



Nota: El DAAM-LDA es el modelo de agregación doblemente adaptativo con datos LDA presentado en esta sección.

Fuente: Banco de España.

La figura 6 resume los resultados de este experimento y también muestra el umbral que se utiliza en cada momento del tiempo, así como el parámetro de memoria y la velocidad de ajuste del umbral óptimo cada año. El modelo DAAM-LDA proporciona mejores previsiones respecto a un modelo de paseo aleatorio, incluso si solo cuatro de los 128 modelos débiles que utiliza como ingredientes realmente lo hacen (véase tabla 1). Si nos centramos en los últimos cuatro años de la muestra (2015-2018), los resultados se pueden comparar también con los más recientes modelos de previsión del PIB a corto plazo actualmente en uso en el Banco de España (la previsión oficial del Banco de España y el modelo Spain-STING, véase Camacho y Perez-Quiros, 2011). Para este período de muestra, el modelo DAAM-LDA funciona mejor que el paseo aleatorio, el

TABLA 1

PREVISIONES DEL PIB ESPAÑOL: RAÍZ CUADRADA DEL ERROR CUADRÁTICO MEDIO EN EJERCICIOS FUERA DE MUESTRA EN TIEMPO REAL

	RW	AR (1)	BdE	DAAM-LDA	Spain-Sting
2008-2018	0,29	0,476	0,082	0,24	---
2015-2018	0,11	0,155	0,76	0,097	0,121

Nota: La tabla muestra la raíz cuadrada del error cuadrático medio fuera de muestra para las previsiones de crecimiento del PIB trimestral español obtenidas por los siguientes modelos (es orden): la caminata aleatoria, el modelo AR (1) simple, la previsión oficial del Banco de España, el modelo de agregación doblemente adaptativo con datos LDA y el modelo Spain-STING.

Fuente: Banco de España.

modelo AR (1) simple y el modelo Spain-STING. No obstante, las previsiones oficiales del Banco de España muestran un desempeño superior en comparación con los métodos estadísticos considerados en esta sección.

3.3.3. Previsión del consumo privado con Google-trends, tarjetas de crédito e indicadores de incertidumbre

El ejercicio presentado en esta sección resume el trabajo de Gil *et al.* (2018). El objetivo del trabajo es averiguar si las nuevas fuentes de información pueden ayudar a predecir el consumo privado de los hogares. Normalmente, los datos oficiales para la medición de las decisiones de gasto de los hogares privados son los datos de las cuentas nacionales que están disponibles con una frecuencia trimestral (datos hard). Por esta razón, los datos cuantitativos se suelen combinar con indicadores cualitativos, de naturaleza más cualitativa, pero con frecuencia más elevada (véase la discusión a principio del apartado 3.3. de este capítulo). El objetivo de este ejercicio es testar el poder predictivo de nuevas fuentes de datos juntos con los datos más tradicionales, tanto hard como soft.

En particular, se consideran las siguientes fuentes de datos mensuales: (1) datos de cajeros automáticos (ATM), que comprenden a la retirada de efectivo en terminales de cajeros automáticos y pagos en puntos de venta (POS) con tarjetas de débito y crédito; (2) indicadores Google-trends, que proporcionan indicadores del comportamiento de consumo basados en patrones de búsqueda en Internet proporcionados por Google; (3) medidas de incertidumbre económica y de política[11].

Estos indicadores se combinan con otros tradicionales cuantitativos (afiliados a la seguridad social; Índice de comercio minorista; Índice de actividad en los servicios) y cualitativos (índice PMI de servicios; Índice de confianza del consumidor) en un modelo de múltiples frecuencias en el que los indicadores entran en frecuencia mensual, y la variable objetivo, el consumo privado de la contabilidad nacional, lo hace en trimestral. Para calcular el desempeño de cada grupo de indicadores, se consideran distintos modelos, que se diferencian en el conjunto de indicadores incluidos en cada uno. Los modelos estimados incluyen indicadores de cada grupo por separado, varios grupos al mismo tiempo y diferentes combinaciones de modelos individuales. Como referencia, se considera un modelo de paseo aleatorio en el cual se incluye en los trimestres futuros la última tasa de crecimiento trimestral observada para el consumo privado. Se evalúa el desempeño de cada modelo para la previsión en el horizonte a corto plazo (trimestre actual, *nowcast*), y también de a 1 a 4 trimestres.

Los principales resultados se muestran en la tabla 1, y pueden resumirse como sigue. En primer lugar, entre los modelos que utilizan únicamente indicadores de cada grupo,

[11] Medido alternativamente por el índice de volatilidad bursátil IBEX y el índice EPU basado en texto proporcionado por Baker, Bloom y Davis (2016) para España.

los que utilizan indicadores cuantitativos y tarjetas de pago ofrecen un mejor desempeño que los demás en las previsiones a corto plazo y, algo menos, en las previsiones a 1 y a 4 trimestres (véase Panel A del cuadro). Los errores cuadráticos medios relativos (RMSE) son en casi todos los casos menores que uno, aunque solo en algunos casos las previsiones de los modelos son estadísticamente diferentes de las obtenidas por el modelo de paseo aleatorio trimestral. En general, los otros modelos no mejoran sistemáticamente la previsión de un paseo aleatorio trimestral. Las dos principales excepciones son el modelo con indicadores cualitativos para los horizontes de previsión a corto plazo y los basados en Google Trends para las previsiones a más largo plazo. Estos resultados son coherentes con el supuesto que los indicadores de Google Trends proporcionan información actual sobre los proyectos de compras futuras.

En segundo lugar, el Panel B del cuadro muestra los resultados de la estimación de modelos que incluyen indicadores cuantitativos agregando, a la vez, indicadores que pertenecen a los otros grupos (indicadores cualitativos, tarjetas de pago, indicadores de incertidumbre, indicadores de Google Trends). En general, la precisión de las previsiones a corto plazo no mejora cuando se incluyen más indicadores, excepto para los indicadores soft. No obstante, las previsiones más a largo plazo mejoran significativamente cuando se expande el modelo de referencia, especialmente cuando se añaden los indicadores de incertidumbre y los de Google trends para las previsiones a 4 trimestres.

En tercer lugar, se destaca que la combinación (promedio) de modelos con grupos individuales de indicadores mejora el desempeño de la previsión en todos los casos y a todos los horizontes (ver Panel C del cuadro). En particular, la combinación de las previsiones de modelos que incluyen indicadores cuantitativos con aquellos con tarjetas de pago ofrece, en general, el mejor desempeño de previsión para todos los horizontes. Al mismo tiempo, agregar las previsiones obtenidas con indicadores cualitativos parece añadir valor en las previsiones a corto plazo. Además, combinar un amplio conjunto de modelos proporciona un RMSE menor respecto al obtenido por el paseo aleatorio trimestral en la previsión a cuatro trimestres.

En conclusión, Gil, Pérez, Sánchez y Urtasun (2018) muestran que aunque los indicadores tradicionales proporcionen una buena previsión del consumo privado en tiempo real, las nuevas fuentes de datos añaden valor, sobre todo aquellas relacionadas con tarjetas de crédito, pero también, en menor medida, los indicadores de Google Trends y los indicadores de incertidumbre, cuando se combinan con otras fuentes.

TABLA 1

ESTADÍSTICOS RMSE (RAÍZ CUADRADA DEL ERROR CUADRÁTICO MEDIO) RELATIVOS: RATIO DE CADA MODELO RESPECTO A UN PASEO ALEATORIO TRIMESTRAL [A]

Panel A: Modelos con indicadores de cada grupo:									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
Indicadores cuantitativos (``hard'') [b]	0.84	0.75 *	0.79	0.75 **	0.81	0.80	0.98	0.97	1.00
Indicadores cualitativos (``soft'') [c]	1.01	0.85	0.85	1.11	1.05	1.05	1.09	1.10	1.19 *
Tarjetas de pago (cuantía, cu) [d]	0.79	0.82	0.88	0.65 ***	0.84	0.69 **	0.74 **	0.84	0.83
Tarjetas de pago (numeros) [d]	1.05	1.15	1.13	0.90	1.10	0.98	0.75 **	0.81	0.79
Indicadores de incertidumbre [e]	1.06	0.97	0.99	1.00	1.05	1.06	0.94	1.00	1.02
Google: agregado de todos los indicadores	1.04	1.06	1.06	0.85	1.03	1.03	0.71 **	0.79	0.79
Google: bienes duraderos (retrasado)	1.04	0.97	0.98	0.96	1.04	1.04	0.85 *	0.93	0.93
Panel B: Modelos con indicadores de grupos diferentes:									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
cuantitativos \& cualitativos	0.69 **	0.78	0.77	0.67 ***	0.76 *	0.72 *	0.79 *	0.82 *	0.80 *
cuantitativos \& Tarjetas de pago (cu) [d]	0.90	0.82	0.91	0.67 ***	0.79	0.78	0.86	0.89	0.91
cuantitativos \& Incertidumbre	0.88	0.86	0.75	0.74 **	0.91	0.93	0.69 **	0.76	0.76
cuantitativos \& Google (agregado)	0.85	0.76	0.77	0.81 *	0.94	0.89	0.77 **	0.81 *	0.82
cuantitativos \& Google (duraderos)	0.91	0.95	0.87	0.69 **	0.83	0.88	0.72 **	0.76 *	0.77 *
Panel C: Combinación de modelos:									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
Todos los modelos [f]	0.66 **	0.71 **	0.69 **	0.65 ***	0.77 *	0.65 **	0.73 **	0.78 *	0.78 *
Hard \& Tarjetas de pago (cu) [d]	0.62 **	0.69 **	0.71 **	0.53 ***	0.69 **	0.52 ***	0.79 *	0.86	0.84
Hard \& Tarjetas de pago (cu) [d] \& Soft	0.65 **	0.67 **	0.67 **	0.65 ***	0.74 **	0.59 ***	0.83 *	0.89	0.92
Hard \& Soft	0.65 **	0.66 **	0.66 **	0.77 **	0.75 **	0.69 **	0.91	0.94	1.02
Hard \& Google (duraderos)	0.77 **	0.78 **	0.76 **	0.74 ***	0.83	0.78 *	0.85	0.91	0.90

Notas: Los asteriscos denotan los resultados del test de Diebold-Mariano, cuya hipótesis nula es que dos métodos de previsión proporcionan resultados de igual precisión. Se utiliza una función de pérdida cuadrática. El número en cada celda representa el diferencial de pérdida del método mencionado en su línea horizontal en comparación con la alternativa de paseo aleatorio trimestral. * (**) [***] Indica el rechazo de la hipótesis nula al nivel del 10% (5%) [1%] de significatividad. [a] Errores de previsión calculados como la diferencia con la primera publicación de los datos de consumo privado. Las previsiones se generan de forma recursiva durante la ventana móvil de 2008T1 (m1) a 2017T4 (m3). [b] Datos del registro de afiliados a la Seguridad Social; Índice de comercio minorista; Índice de actividad del sector servicios. [c] PMI de servicios; índice de confianza del consumidor. [d] Serie agregada de tarjetas de crédito vía POS y cajeros automáticos. [e] Volatilidad del mercado de valores (IBEX); índice de incertidumbre de la política económica (EPU). [f] Combinación de los resultados de 30 modelos, incluyendo modelos en los que se incluyen los indicadores de cada bloque por separado, modelos que incluyen el bloque cuantitativo y cada uno de los otros bloques, y versiones de todos los modelos mencionados que incluyan además valores retrasados de las variables.

Referencias

Aguilar P., Ghirelli, C., Pacce, M. y Urtasun, A. (2020). Can news help to measure economic sentiment? An application in Covid-19 times. *Documento de Trabajo,* No. 2027. Banco de España.

Ahir, H., Bloom, N. y Furceri, D. (2019). The World Uncertainty Index. *Working Paper,* 19–027. Standford Institute for Economic Policy Research.

Alonso, A. y Carbó, J. M. (2020). Machine Learning in Credit Risk: Measuring the Dilemma Between Prediction and Supervisory Cost. *Documento de Trabajo*, No. 2032. Banco de España.

Aprigliano V., Ardizzi, G. y Monteforte, L. (2017). Using the payment system data to forecast the Italian GDP. *Working paper,* No. 1098. Bank of Italy.

Artola C. y Galán, E. (2012). Tracking the future on the web: construction of leading indicators using internet searches. *Documento de Trabajo,* No. 1203. Banco de España.

Azqueta-Gavaldon A., Hirschbühl, D. Onorante, L. y Saiz, L. (2020). Sources of economic policy uncertainty in the euro area: an unsupervised machine learning approach. *Working Paper,* No. 2359. European Central Bank.

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

Banco de España (2017). *Survey of Household Finances, 2014: methods, results and changes since 2011.* Artículo Analítico No. 1/2017. Banco de España.

— (2018). Central Balance Sheet Data Office. Annual results of non-financial corporations 2017. Banco de España.

Barrow, D. K. y Crone, S. (2016). A comparison of AdaBoost algorithms for time series forecast combination. *International Journal of Forecasting,* 32(4), pp. 1103-1119.

Bhattarai, S., Chatterjee, A. y Park, W. Y. (2019). Global spillover effects of US uncertainty. *Journal of Monetary Economics*. https://doi.org/10.1016/j.jmoneco.2019.05.008

Biljanovska, N., Grigoli, F. y Hengge, M. (2017). Fear Thy Neighbor: Spillovers from Economic Policy Uncertainty. *Working Paper* No., 17/240. International Monetary Fund.

Blei, D. M., Ng, A. Y. y Jordan, M. I. (January 2003). Lafferty, John (ed.). Latent Dirichlet Allocation. *Journal of Machine Learning Research.*

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives,* 28(2), pp. 153-176.

Bodas, D., García, J., Murillo, J. Pacce, M., Rodrigo, T., Ruiz, P., Ulloa, C. Romero, J. y Valero, H. (2018). Measuring Retail Trade Using Card Transactional Data. *Working Paper,* No. 18/03. BBVA Research.

Broeders, D. y Prenio, J. (2018). Innovative Technology in Financial Supervision (Suptech) - The Experience of Early Users. Financial Stability Institute Insights on Policy Implementation. *Working paper,* No. 9. Bank for International Settlements, July.

Camacho, M. y Perez-Quirós, G. (2011). Spain-Sting: Spain Short-Term Indicator of Growth. *The Manchester School,* 79, pp. 594–616.

Carlsen, M. y Storgaard, P. E. (2010). Dankort payments as a timely indicator of retail sales in Denmark. *Working paper,* No. 66. Bank of Denmark.

Carriére-Swallow Y. y Labbé, F. (2013). Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting,* 32, pp. 289-298.

Central banking (2019). *Big data in central banks: 2019 survey results.* https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central- banks-2019-survey-results

Colombo, V. (2013). Economic policy uncertainty in the US: Does it matter for the euro area? *Economics Letters,* 121(1), pp. 39-42.

Chetty, R., Friedman, J. y Rockoff, J. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *The American Economic Review,* 104(9), pp. 2633-2679.

D'Amuri F. y Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting,* 33(4), pp. 801-816.

Diaz-Sobrino, N., Ghirelli, C., Hurtado, S., Pérez, J. J. y Urtasun, A. (2020), The narrative about the economy as a shadow forecast: an analysis using Bank of Spain quarterly reports. *Documento de Trabajo* (próxima aparición). Banco de España.

Duarte, C., Rodrigues, P. M. y Rua, A. (2017). A mixed frequency approach to the forecasting of private consumption with ATM/POS data. *International Journal of Forecasting,* 33(1), pp. 61-75.

Einav, L. y Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy,* 14, pp. 1-24.

Farné, M. y Vouldis, A. T. (2018) A Methodology for Automatised Outlier Detection in High-Dimensional Datasets: An Application to Euro Area Banks' Supervisory Data. *Working Paper,* No. 2171. European Central Bank.

Fernández, A. (2019). Artificial intelligence in financial services. Analytical Articles. *Boletín Económico,* No. 2/2019. Banco de España.

Fernández-Villaverde, J., Hurtado, S. y Nuño, G. (2019). Financial Frictions and the Wealth Distribution. *Working Paper,* No. 26302. National Bureau of Economic Research. September.

Ferrara, L. y Simoni, A. (2019). When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage. *Working paper,* No. 2019-04. Center for Research in Economics and Statistics.

Fontaine, I., Didier, L. y Razafindravaosolonirina, J. (2017). Foreign policy uncertainty shocks and US macroeconomic activity: Evidence from China. *Economics Letters,* 155, pp. 121-125.

Ghirelli, C., Hurtado, S., Pérez, J. J. y Urtasun, A. (2020). New data sources for central banks. En: S. Consoli, D. Reforgiato Recupero, y M. Saisana, *Data Science for Economics and Finance: Methodologies and Applications.* Springer (próxima aparición).

Ghirelli, C., Peñalosa, J., Pérez, J. J. y Urtasun, A. (2019). Some implications of new data sources for economic analysis and official statistics. *Boletín Económico,* No. 2/2019. Banco de España. Mayo.

Ghirelli, C., Pérez, J. J. y Urtasun, A. (2019). A new economic policy uncertainty index for Spain. *Economics Letters,* 182, pp. 64-67.

— (2020). Economic Policy Uncertainty in Latin America. *Documento de Trabajo,* No. 2024. Banco de España.

Gil, M., Pérez, J. J., Sánchez, A. J. y Urtasun, A. (2018). Nowcasting Private Consumption: Traditional Indicators, Uncertainty Measures, Credit Cards and Some Internet Data. *Documento de Trabajo,* No. 1842. Banco de España.

Gotz, T. B. y Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting,* 35, pp. 45-66.

Hammer, C. L., Kostroch, D. C. y Quirós, G. (2017). Big Data: Potential, Challenges and Statistical Implications. *IMF Staff Discussion Note,* 17/06. International Monetary Fund.

Hardy, A., Hyslop, S., Booth, K. B., Aryal, J., Gretzel, U. y Eccleston, R. (2017). Tracking tourists' travel with smartphone-based GPS technology: a methodological discussion. *Information Technology & Tourism,* 17, pp. 255-274.

Kapetanios, G. y Papailias, F. (2018). Big Data & Macroeconomic Nowcasting: Methodological Review. *ESCoE Discussion Paper,* 2018-12. Economic Statistics Centre of Excellence.

Lacroix R. (2019). The Bank of France datalake. En: Bank for Inter- national Settlements (ed.), The use of big data analytics and artificial intelligence in central banking, *IFC Bulletins,* vol. 50. Bank for International Settlements.

Loberto, M., Luciani, A. y Pangallo, M. (2018). The potential of big housing data: an application to the Italian real-estate market. *Working paper,* No. 1171. Bank of Italy.

Menéndez, A. y Mulino, M. (2018). Results of non-financial corporations in the first half of 2018. *Boletín Económico,* No. 3/2018. Banco de España.

Meinen, P. y Roehe, O. (2017). On measuring uncertainty and its impact on investment: Cross-country evidence from the euro area. *European Economic Review,* 92, pp. 161-179.

Moreno Bernal, A. y González Pedraz, C. (2020). Análisis de sentimiento del informe de estabilidad financiera. *Documento de Trabajo,* No.2011. Banco de España.

Mueller, H. y Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review,* 112(2), pp. 358-375.

Nyman R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P. y Smith, R. (2018). News and narratives in financial systems: exploiting big data for systemic risk assessment. *Staff Working Paper,* No. 704. Bank of England.

Pew Research Center (2012). Assessing the Representativeness of Public Opinion Surveys. Mimeo. https://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

Petropoulos, A., Siakoulis, V., Stavroulakis, E. y Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. En: Bank for International Settlements (ed.), The use of big data analytics and artificial intelligence in central banking, *IFC Bulletins,* vol. 50. Bank for International Settlements.

Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of business cycles, *Journal of Business & Economic Statistics,* pp.1-17.

Trung, N. B. (2019). The spillover effect of the US uncertainty on emerging economies: a panel VAR approach. *Applied Economics Letters*, 26(3), pp. 210-216.

Apéndice

EXPLICACIONES ADICIONALES SOBRE EL CONTENIDO DE LAS FIGURAS

Figura 2. Respuesta del PIB de España ante distintas perturbaciones de incertidumbre construidas con versiones alternativas del índice EPU

La figura muestra la función impulso-respuesta de la tasa de crecimiento del PIB español hasta diez trimestres después de un *shock* positivo de una desviación estándar en el índice EPU para España, y se ha elaborado a partir del modelo ilustrado en Ghirelli, Pérez y Urtasun (2019). Se trata de un modelo VAR estimado por MCO que incluye las siguientes variables endógenas: el indicador EPU (en niveles), el diferencial de la deuda soberana española a 10 años sobre el bund alemán, el PIB real (en frecuencia trimestral y en tasas de crecimiento) [a su vez, el consumo real agregado de los hogares o la inversión real de bienes de capital], la tasa de inflación (tasas de crecimiento trimestrales del IPC) y una variable que controla por la incertidumbre global representada por el índice EPU global de Baker, Bloom y Davis (2016), que se asume exógena. En este modelo se incluyen un numero óptimo de retardos que se identifican a partir del criterio de la información de Schwarz. Para la identificación de los impulsos-respuestas de una variable sobre las otras variables endógenas se utiliza la descomposición de Cholesky de la forma reducida de la matriz de varianza y covarianza.

En la figura 2 se investiga el papel relativo de (1) la riqueza de palabras clave, (2) la cobertura de prensa y (3) cobertura temporal en la conducción de los resultados. La figura compara las respuestas del PIB al impulso de las versiones alternativas del EPU en las que se modifica una de las dimensiones mencionadas anteriormente a la vez, pasando de la versión EPU-BBD (el EPU construido como en Baker, Bloom y Davis, 2016) al nuevo índice EPU, propuesto en Ghirelli, Pérez y Urtasun (2019). La línea *"EPU-Nuevas palabras, vieja Cobertura; >2000"* se refiere a un índice que se construye usando las palabras clave de Ghirelli, Pérez y Urtasun (2019), pero manteniendo la cobertura de prensa y de tiempo con las del índice EPU-BBD. Por tanto, comparar esta respuesta con la de EPU-BBD permite comprobar la relevancia de enriquecer las palabras clave en las búsquedas. Del mismo modo, para la línea *"EPU-Nuevas palabras, nueva Cobertura; >2000"* se utilizan las palabras clave y la cobertura de prensa de Ghirelli, Pérez y Urtasun (2019), pero se mantiene la cobertura de tiempo como en el índice EPU-BBD. Comparando las respuestas de los choques en *"EPU-Nuevas palabras, vieja Cobertura; >2000"* y *"EPU-Nuevas palabras, nueva Cobertura; >2000",* se puede apreciar la contribución de ampliar la cobertura de prensa al construir el índice. Finalmente, comparando los resultados de *"EPU-Nuevas palabras, nueva Cobertura; >2000"* con los obtenidos con EPU-NEW se puede comprobar la importancia de incrementar el periodo temporal. De acuerdo con la figura, todas las dimensiones (i) - (iii) son importantes, ya

que todas contribuyen a obtener el signo negativo esperado en las respuestas del PIB. Además, la definición de la cobertura del periodo es clave para mejorar la precisión de las estimaciones y obtener resultados significativos.

Figura 4. Previsiones cuantitativas e indicador de sentimiento de los informes del Banco de España

La figura 4 muestra el indicador de sentimiento del *Boletín Económico Trimestral* del Banco de España frente a las previsiones cuantitativas elaboradas por el Banco de España, y se ha elaborado en Ghirelli, Hurtado, Pérez y Urtasun (2020). Para ello, se considera el *Boletín Económico Trimestral de la economía española* publicado online por el Banco de España desde el primer trimestre de 1999. En concreto, se considera el apartado inicial que contiene los principales mensajes. Con esta información se construye un indicador de sentimiento del primer trimestre de 1999.

El índice de sentimiento se construye de la siguiente manera. Primero, se crea un diccionario de términos positivos y negativos en castellano. Para construir la lista de palabras, se lee una muestra de los informes considerados y se identifican los términos que se usan más frecuentemente para describir la situación macroeconómica (considerando adjetivos, adverbios, verbos y sustantivos). Se seleccionan 47 términos (raíces de palabras), eliminando las terminaciones de las palabras. En segundo lugar, se asigna una puntuación igual a +1 (-1) a las palabras que expresan un sentimiento positivo (negativo). En tercer lugar, se cuenta cuántas veces aparece cada palabra del diccionario en cada texto y se pondera cada resultado con su puntuación. En quinto lugar, se suman todas las apariciones ponderadas de cada texto y se divide el número resultante por la longitud total del texto.

Figura 5. Un indicador textual para anticipar la evolución del PIB de España

La imagen de la derecha del figura 5 muestra el indicador textual de la economía con frecuencia trimestral frente a la tasa de crecimiento del PIB español hasta junio de 2019. La figura de la izquierda muestra el mismo indicador textual con frecuencia semanal de enero a marzo de 2020. El indicador textual de la economía se ilustra en Aguilar *et al.* (2020).

Dicho indicador textual es un indicador de sentimiento basado en el análisis de texto de artículos publicados en periódicos españoles desde 1997. Se computa utilizando la base de datos de Factiva y se tienen en cuenta siete periódicos nacionales *(ABC, El País, El Mundo, La Vanguardia, Expansión, Cinco Días, El Economista).* Básicamente, este indicador captura el tono económico de los artículos de noticias publicados en la prensa

española, reflejando el equilibrio entre el número de noticias que contienen palabras clave relacionadas con repuntes y recesiones en el ciclo económico español. Para su construcción se siguen los siguientes pasos.

En primer lugar, con frecuencia mensual, se hacen tres tipos de búsquedas en cada uno de los periódicos mencionados:

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones positivas (pos): se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recuperacion*" o una de las siguientes palabras (aceler*, crec*, increment*, recuper*, aument*, expansi*, mejora*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de cinco palabras entre sí. Se ignora "crecimiento económico" dado que tiene un tono neutro (se usa para describir indiferentemente un crecimiento negativo o positivo).

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones negativas (neg): se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recesión*" o "crisis" o una de las siguientes palabras (descen*, ralentiz*, redu*, disminu*, contraccion*, decrec*, desaceler*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de cinco palabras entre sí.

- Se cuenta el número de artículos que hablan sobre noticias económicas o financieras (total).

En segundo lugar, para cada periódico, se calcula la siguiente proporción: (pos-neg)/total. A esta proporción se la estandariza utilizando el periodo comprendido entre enero 1997 y febrero 2020. De esta manera la volatilidad de las distintas series es comparable entre los distintos periódicos. Por último, se computa una serie agregada tomando la media entre las series de los distintos periódicos y se le quita la media del periodo comprendido entre enero 1997 y febrero 2020.

CAPÍTULO VIII

Predicción de series temporales basada en *Machine Learning:* aplicaciones económicas y financieras

Lorenzo Pascual
Esther Ruiz*

En este capítulo se realiza una breve revisión de aplicaciones empíricas de procedimientos de predicción basados en Machine Learning (ML) en el contexto de series temporales económicas y financieras. Dada la naturaleza reciente y cambiante de dichos procedimientos, esta revisión es limitada e incompleta. Nuestra contribución es la descripción de los ámbitos de aplicación donde ML ha sido utilizado con éxito.

Palabras clave: árboles de decisión, big data, LASSO, redes neuronales, regularización.

* Esther Ruiz agradece la ayuda financiera de la Agencia Estatal de Investigación (PID2019-108079GB-C21/AIE/10.13039/501100011033).

> "So the question is, what is the real world? Is it simple or complex? Machine learning shows that there are examples of complex worlds. We should approach complex worlds from a completely different position than simple worlds. For example, in a complex world one should give up explain-ability (the main goal in classical science) to gain better predictability."
>
> *Vapnik* (2006)

1. INTRODUCCIÓN

La predicción es un instrumento fundamental para la toma de decisiones en un amplio abanico de aplicaciones entre las que naturalmente están áreas del ámbito económico como, por ejemplo, organización de empresas, macroeconomía, microeconomía y finanzas. La importancia de la predicción explica el gran interés académico y práctico por la metodología cuyo objetivo final es obtener procedimientos para predecir fenómenos futuros que sean lo más precisos posible.

Los procedimientos "tradicionales" de predicción se basan en modelos estadísticos relativamente sencillos que tratan de aproximar de la mejor forma posible el supuesto proceso generador de la variable que se quiere predecir. Estos modelos pueden ser lineales o no lineales, univariantes o multivariantes, paramétricos o no paramétricos, estacionarios o no estacionarios, pero, en cualquier caso, la estimación de sus parámetros puede realizarse utilizando datos con dimensiones relativamente pequeñas o medianas. Sin embargo, en las últimas décadas, dos fenómenos han propiciado la aparición de nuevos procedimientos de predicción basados en lo que se conoce como *Machine Learning*[1] *(ML)* supervisado[2]. En primer lugar, la capacidad de ejecución en tiempo real de los ordenadores y, por lo tanto, su velocidad de procesamiento, han crecido de forma espectacular durante las últimas décadas, disminuyendo drásticamente el coste de computación. En segundo lugar, la arquitectura de sistemas permite el almacenamiento y tratamiento de cantidades ingentes de datos, lo que se conoce como *big data* (BD), que pueden clasificarse en datos estructurados, obtenidos de forma habitual mediante representaciones numéricas, y datos no estructurados, como textos, vídeos o imágenes; ver, por ejemplo, Garboden (2019) para una descripción de datos BD en el contexto de la predicción macroeconómica. La principal diferencia entre los procedimientos de predicción tradicionales y los basados en ML es que estos últimos utilizan

[1] Queremos pedir disculpas por la extensa utilización de términos en el idioma inglés a lo largo de este capítulo. La razón principal es que los conceptos asociados a dichos términos son conocidos cuando se utiliza el inglés, pero se crearían confusiones si utilizáramos traducciones literales al español que no son habituales en esta literatura. Nuestras disculpas por adelantado a la Real Academia Española.

[2] Supervisado se refiere a que la variable a predecir sirve para guiar el proceso de estimación. Aunque los procedimientos no supervisados quedan fuera de esta revisión, hay aplicaciones muy interesantes donde dichos procedimientos se han implementado para la predicción económica; ver, por ejemplo, Aromi (2020) que construye un indicador de incertidumbre basado en el contenido publicado en el *Wall Street Journal*, para predecir el ciclo económico en USA.

modelos estadísticos y algoritmos flexibles que permiten extraer información de datos de alta dimensión pero que, en general, no tratan de representar el mecanismo que ha generado los datos (únicamente se predice, no se trata de explicar). Sin embargo, es importante señalar que no existe una frontera bien definida entre los llamados procedimientos tradicionales de predicción y los procedimientos ML; ver, por ejemplo, las discusiones en Barker (2020), Januschowski *et al.* (2020) e Israel, Kelly y Moskowitz (en prensa) para excelentes descripciones de las diferencias entre procedimientos "tradicionales" y ML.

Además de en predicción, ML supervisado tiene importantes aplicaciones en problemas de clasificación. En algunos casos, no existe una clara frontera entre predicción y clasificación. Por ejemplo, cuando se clasifica a un cliente en función de si va a devolver o no un crédito, se podría considerar como una predicción de una variable cualitativa binaria. Sin embargo, en este capítulo, nos centraremos en la predicción de variables cuantitativas continuas. Existen también procedimientos no supervisados de ML que se utilizan principalmente para reducir la dimensión de los datos y para encontrar grupos homogéneos. Estos procedimientos de procesamiento de datos suelen utilizarse como paso previo a la implementación de procedimientos supervisados para la predicción. Los procedimientos más populares dentro de ML no supervisado son componentes principales, análisis factorial y análisis de conglomerados.

Los algoritmos de ML son computacionalmente intensivos y están diseñados para identificar de forma sistemática patrones y relaciones en los datos que sirvan de base para realizar la predicción. La mayoría de los procedimientos ML para la predicción se han desarrollado en el ámbito de variables independientes e idénticamente distribuidas (iid); véase, por ejemplo, Varian (2014) para regresiones supervisadas o Biau y Scornet (2016) para árboles de decisión. Sin embargo, en este capítulo nos centraremos en procedimientos de predicción para datos dinámicos en los que los propios retardos de la variable a predecir pueden aparecer entre los predictores. Nuestro objetivo es revisar brevemente las principales aplicaciones empíricas de procedimientos de predicción basados en ML en el contexto de series temporales económicas y financieras. Dada la naturaleza reciente y cambiante de dichos procedimientos, esta revisión es necesariamente limitada e incompleta. Por ejemplo, para limitar la discusión, no hemos incluido en esta revisión aplicaciones de procedimientos bayesianos para el tratamiento de BD; ver, por ejemplo, las descripciones en Varian (2014) y Scott y Varian (2014).

El resto del capítulo se organiza como sigue. La sección segunda describe brevemente los principales procedimientos de predicción con ML en el contexto de series temporales. La sección tercera describe aplicaciones empíricas de ML en el contexto de la predicción económica y financiera. Finalmente, la sección cuarta concluye.

2. PROCEDIMIENTOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE SERIES TEMPORALES

2.1. El problema de predicción

El objetivo es obtener predicciones de y_{T+h}, donde h es el horizonte de predicción, basadas en observaciones $\{y_1,...,y_T\}$, obtenidas normalmente en espacios equidistantes del tiempo. La predicción es univariante cuando y_{T+h} es un escalar, es decir, el objetivo es la predicción del valor futuro de una única variable, mientras que es multivariante cuando y_{T+h} es un vector. Para realizar la predicción de y_{T+h}, se puede disponer además de observaciones de un conjunto de predictores, $x_{i,t}$, i=1,...,N, t=1,...T. Al hablar de gran dimensión de los predictores, podemos encontrarnos con dos situaciones diferentes. En primer lugar, están los datos "anchos" en los que el número de predictores, N, es muy grande relativo al número de observaciones, T. Los datos pueden ser "largos" cuando hay muchas observaciones en relación al número de predictores; ver Diebold *et al.* (2019) que también definen los datos "densos" cuando la frecuencia de observación es muy elevada. El problema de predicción es obtener predicciones h-periodos-hacia-adelante mediante el siguiente procedimiento directo[3]:

$$\hat{y}_{T+h|T} = g_h\left(y_T,..,y_1,x_{1,T},...,x_{1,1},...,x_{N,T},...,x_{N,1}\right), \qquad [1]$$

dónde $g_h(\cdot)$ es una función, posiblemente no lineal y desconocida, de los datos disponibles en el momento T; ver, por ejemplo, Gu, Kelly y Xiu (en prensa b), Huber y Stuckenschmit (en prensa), Varian (2014) y Caro y Peña (2021) para una detallada descripción del problema de predicción de series temporales basado en procedimientos ML. Para simplificar la descripción posterior, supondremos que las predicciones se realizan un-periodo-hacia-adelante, es decir, $h=1$, y suprimiremos el subíndice de la función $g_h(\cdot)$. Los procedimientos ML para la predicción tratan de obtener predicciones de y_{T+1} lo más precisas posible cuando el modelo verdadero, es decir, la función $g_h(\cdot)$, es desconocido y/o muy complejo. Es por ello que estos procedimientos están especialmente diseñados para realizar predicciones cuando las bases de datos son de gran dimensión y, por lo tanto, es esperable que pueda haber no-linealidades y relaciones complejas difíciles de especificar *a priori.*

Los procedimientos de predicción ML requieren, en general, seleccionar una arquitectura (por ejemplo, en los procedimientos de *Artificial Neural Networks (ANN),* hay que seleccionar el número de nodos y capas y, en los procedimientos *Support Vector Machine (SVM),* hay que seleccionar las funciones kernel) y/o algunos parámetros, denominados hiperparámetros (por ejemplo, en los procedimientos de regularización hay que selec-

[3] Coulombe *et al*. (2020) señalan que el procedimiento directo es el más habitual en ML frente al procedimiento iterativo que se utiliza frecuentemente cuando las predicciones se basan en modelos "tradicionales" de series temporales.

cionar previamente los parámetros de regularización), que reducen la complejidad del modelo y que deben ser seleccionados con anterioridad a la propia estimación de los parámetros del modelo correspondiente. El procedimiento de selección de los hiperparámetros, estimación de los parámetros y evaluación de las predicciones se basa en dividir la muestra en tres submuestras: la submuestra de entrenamiento, en la que, dados los hiperparámetros, se estiman los parámetros; la submuestra de validación, en la que se eligen los hiperparámetros mediante la minimización de una determinada función de pérdida; y la submuestra de contraste, en la que se evalúan las predicciones. La figura 1 resume el procedimiento de estimación, validación y predicción de los procedimientos ML. Una vez seleccionados los hiperparámetros, los parámetros del modelo vuelven a estimarse utilizando conjuntamente las submuestras de entrenamiento y de validación. La submuestra de contraste es la que habitualmente se conoce como periodo "fuera-de-muestra". En una gran mayoría de aplicaciones de predicción con ML, los hiperparámetros son seleccionados mediante validación cruzada y las predicciones son evaluadas utilizando el error de predicción cuadrático medio (EPCM); ver, por ejemplo, Barrow y Crone (2016b) sobre la validación cruzada y Coulombe *et al.* (2020) para una comparación con otras medidas de evaluación. Es importante señalar que, en los procedimientos de predicción ML, tanto la muestra de entrenamiento como la de prueba deben ser suficientemente grandes como para permitir encontrar patrones en los datos.

FIGURA 1

PROCEDIMIENTOS DE PREDICCIÓN ML

Muestra de entrenamiento (estimación): **Dados** los hiperparámetros, se estima el modelo.	Muestra de validación (selección del modelo: **Selección de hiperparámetros** que minimicen la función de pérdida de las predicciones obtenidas con el modelo estimado en la muestra de entrenamiento.	Muestra de contraste (predicción): Evaluación de las predicciones fuera-de-muestra: EPCM.
Hiperparámetros: Regularización: Constantes Arboles: Nº de ramas Redes Neuronales: Función de activación	Función de pérdida: Error de predicción cuadrático medio (EPCM) Procedimiento: Validación cruzada	

Fuente: Elaboración propia.

Los procedimientos ML de predicción pueden clasificarse en dos grandes grupos dependiendo de que traten de reducir la dimensión de los predictores en [1] o de que traten de representar funciones $g(\cdot)$ complejas. Al primer grupo pertenecen los modelos de regresión con regularización y los procedimientos basados en SVM que se utilizan habitualmente para datos "anchos". El segundo grupo está compuesto por los populares

modelos basados en árboles de decisión y por las redes neuronales o *deep learning.* A continuación se describen brevemente los procedimientos más populares dentro de cada uno de estos dos grandes grupos.

2.2. Reducción de la dimensión: regresión con regularización y Support Vector Machine

Cuando el número de predictores en un modelo de regresión es muy elevado, algunos de ellos pueden ser redundantes o no tener suficiente información como para ser incluidos en el modelo predictivo. En estos casos, antes de estimar, es necesario regularizar, es decir, forzar a los parámetros de algunos predictores a que tomen valores relativamente pequeños y/o seleccionar algunos predictores, para poder recuperar grados de libertad y mejorar la capacidad predictiva del modelo. Cuando la función $g(\cdot)$ es lineal, el modelo predictivo es el siguiente modelo de regresión

$$y_{T+1} = \beta' z_T + \varepsilon_{T+1}, \qquad [2]$$

donde $Z_T=(y_T,...,y_{T-p},x_{1,T},...,x_{1,T-p},...,x_{N,T},...,x_{N,T-p})$ es el vector de predictores, con p siendo el número de retardos en el modelo; ver Eklund y Kapetanios (2008) para una descripción de los procedimientos para grandes bases de datos basados en el modelo [2]. Los $(N+1)$ x p parámetros desconocidos, β, son estimados introduciendo una penalización convexa para forzar los parámetros en el vector β hacia cero en lo que se conoce como regresión con regularización. Entre los procedimientos de regresión con regularización más populares están la regresión *ridge* que restringe los parámetros hacia zero, reduciendo el sobreajuste, pero no selecciona predictores y, por lo tanto, no se genera un modelo más interpretable. Alternativamente, en el estimador conocido como *Least Absolute Shrinkage and Selection Operator (LASSO),* se produce tanto la reducción de los parámetros como la selección de variables, mejorando tanto la precisión de las predicciones como la interpretabilidad del modelo al seleccionar solamente un subconjunto de los predictores disponibles. Cuando entre los predictores aparecen tanto un número alto de variables como sus retardos, Simon *et al.* (2013) introducen el *sparse group* LASSO (sg-LASSO). El sg-LASSO reduce la dimensión tanto entre variables como entre sus retardos. Babii, Ghysels y Striankas (2020) consideran sg-LASSO cuando las variables en la regresión en [2] están observadas con distintas frecuencias y analizan las correspondientes propiedades del estimador y predictor. Otro procedimiento de regresión con regularización muy popular es *elastic net,* en el que se introduce una segunda penalización.

Finalmente, dentro de los modelos que tratan de estimar [2] está SVM, un procedimiento de minería de datos originalmente desarrollado por Vapnik en los Laboratorios Bell en 1995 para clasificación. Cuando se aplica a regresión, SVM se conoce como *Support Vector Regression (SVR)*; ver, por ejemplo, Smola y Scholkopf (2004) y Awad

y Khanna (2015) para descripciones detalladas de SVR no solo aplicadas a regresiones lineales si no también no-lineales.

2.3. Funciones de predicción desconocidas y/o complejas

En esta subsección, describimos procedimientos ML diseñados para obtener predicciones cuando la función $g(\cdot)$ en [1] es compleja y/o desconocida. Entre los procedimientos más populares están los árboles de decisión y los tres procedimientos propuestos para mejorar su comportamiento predictivo: 1) *Bagging*; 2) *Random Forest*; y 3) *Boosting*,[4] y las redes neuronales.

2.3.1. Árboles de decisión

La idea básica de un árbol de decisión es segmentar el rango de valores de los predictores en un número finito de subregiones de tal forma que, dentro de cada subregión, los predictores son más homogeneos y se pueden utilizar procedimientos muy simples para la predicción como, por ejemplo, la media muestral de todas las observaciones en cada subregión. En los árboles de decisión, las particiones en regiones son binarias para todos los predictores y sus retardos, es decir, de cada nodo, solo pueden salir dos ramas. Los umbrales se determinan sucesivamente para minimizar el EPCM. Finalmente, la predicción de y_{T+1} se obtiene utilizando el valor de los predictores a través del árbol que nos va dirigiendo mediante decisiones binarias a una de las subregiones (nodos terminales del árbol).

En la construcción de un árbol de decisión, se seleccionan los mejores predictores y retardos para minimizar el EPCM. Sin embargo, es frecuente que, en este proceso, se obtengan árboles con muchas ramas que pueden generar sobreajuste y, consecuentemente, predicciones no óptimas. Por esta razón, en la práctica, se realiza una "poda" del árbol introduciendo en la función objetivo una penalización sobre el número de nodos terminales.

Bagging: el procedimiento Bagging (*Bootstrap Aggregation*) está diseñado para evitar el gran EPCM que se suele observar en las predicciones obtenidas mediante árboles de decisión debido a que el número de ramas que se utiliza finalmente para la predicción es relativamente pequeño. Bagging se basa en generar B muestras bootstrap de los predictores y construir el árbol de decisión correspondiente en cada una de las réplicas *bootstrap*. Es muy importante tener en cuenta que, en el caso de datos temporales, el remuestreo para obtener réplicas bootstrap de la muestra de entrenamiento ha de realizarse con especial cuidado para no destruir la dependencia temporal existente en los datos. Existen en la literatura diversas técnicas Bootstrap que tienen en consideración esta característica como, por ejemplo, el *wild bootstrap* o el *block bootstrap*; ver, por

[4] Es importante señalar que Boosting puede utilizarse también en contextos diferentes de los áboles de decisión. Sirve para mejorar cualquier regla de predicción simple.

ejemplo, Pan y Politis (2016). La predicción final es la media de todas las predicciones obtenidas en cada una de las réplicas *bootstrap*.

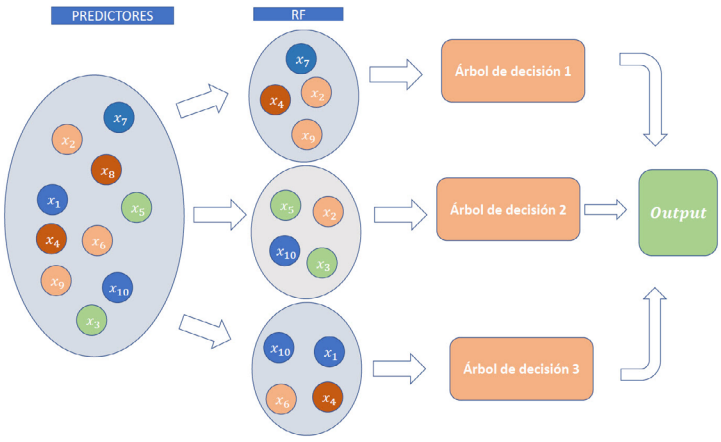
Bülmann y Yu (2002) han propuesto también el *Subbagging* basado en *subsampling* en lugar de Bootstrap y demuestran que aunque es computacionalmente más sencillo, tiene el mismo nivel de precisión que Bagging.

Random Forest: Random Forest (RF), propuesto originalmente por Breiman (2001), es un procedimiento diseñado para reducir la correlación entre árboles de decisión cuando estos son generados mediante Bagging, por el hecho de que los mismos predictores aparecen en todas las réplicas *bootstrap*. RF también se basa en construir árboles de decisión a partir de réplicas *bootstrap* de la muestra de entrenamiento. Sin embargo, mientras que Bagging utiliza en cada árbol todos los regresores existentes (p), RF selecciona de forma aleatoria, solamente $m < p$ de los p predictores, generalmente $m = \sqrt{p}$; ver la figura 2. Nótese que m vuelve a ser otro hiperparámetro que debe seleccionarse mediante validación cruzada en la muestra de entrenamiento. Junto con m, es necesario también determinar otros hiperparámetros como el número de réplicas Bootstrap o el número de árboles de decisión y el número mínimo de observaciones en los nodos terminales (poda). Tsay y Chen (2019) señalan que es necesario un estudio detallado de las propiedades de RF en el caso de que existan predictores fuertes dado que estos deben ser incluidos en cualquier modelo para obtener un buen ajuste.

Boosting: Mediante el procedimiento Boosting se generan árboles de forma secuencial a partir de un árbol inicial relativamente sencillo. En cada paso, se mejora el ajuste del

FIGURA 2

PROCEDIMIENTO DE SELECCIÓN DE ÁRBOLES MEDIANTE RANDOM FOREST



Fuentes: Elaboración propia.

árbol anterior mediante un nuevo árbol construido para los residuos en lugar de para la variable a predecir directamente. Con los ajustes a los residuos se mejora el árbol, mejorando áreas específicas donde el ajuste del árbol anterior no funciona correctamente. Al final del proceso, la predicción es una media ponderada de las predicciones de todos los árboles intermedios, dando más peso a los árboles últimos, que son los que están mejor calibrados. Nuevamente, el número de árboles, que es un hiperparámetro muy sensible que puede producir sobrejuste en el modelo, y la ratio de aprendizaje del Boosting, deben calibrarse mediante validación cruzada.

Existen variantes populares del Boosting, como el *Gradient Boosting,* que transcribe el método boosting como un algoritmo de optimización donde se minimiza una función de pérdidas adecuada y usa el método del gradiente en lugar de los residuos resultantes del ajuste previo para encontrar la siguiente mejora del árbol anterior. Hay que tener cierta precaución porque el Gradient Boosting tiende rápidamente al sobreajuste.

2.3.2. Redes neuronales artificiales

Las redes neuronales artificiales han ganado una gran popularidad en el área de predicción (a pesar del escepticismo sobre que las funciones de pérdida asociadas a estas redes son mayoritariamente no-convexas) debido a su excepcional comportamiento (ayudadas por procedimientos de gradiente descendente). Las ANN son procedimientos basados en datos con muy pocos supuestos previos sobre los modelos generadores de dichos datos. Las redes neuronales son modelos matemáticos computacionalmente intensivos que intentan imitar la forma en que los humanos aprenden cuando reciben información. Estos modelos se consideran "cajas negras" en términos de interpretabilidad con respecto a la relación entre las distintas variables. Sin embargo, su estructura permite capturar relaciones complejas entre los predictores y la variable a predecir sin que sea necesario especificar la forma correcta de dicha relación ya que la propia red neuronal tratará de identificar dichas relaciones a partir de los datos disponibles para su entrenamiento.

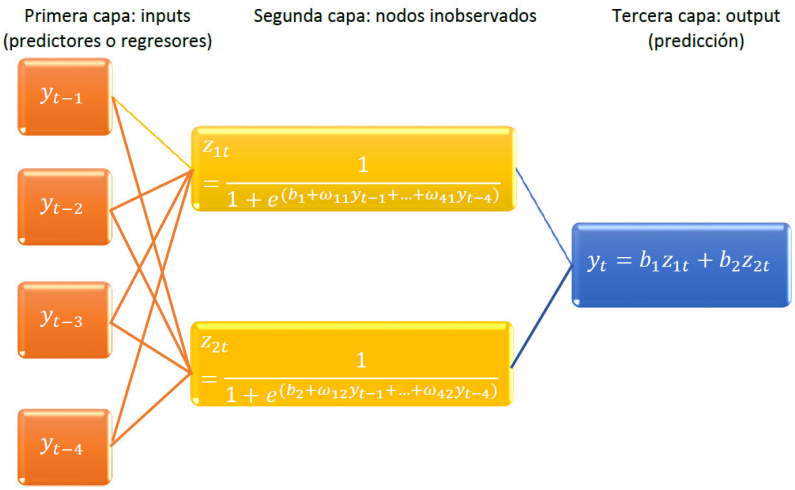
La primera aplicación de predicción utilizando ANN aparece en la tesis de Hu (1964) para la predicción meteorológica; ver Zhang, Patuwo y Hu (1998) para una revisión de la literatura sobre ANN a finales del siglo XX. Investigaciones recientes han validado los buenos resultados predictivos de las ANN mostrando que, si las redes son suficientemente grandes, casi todos los mínimos locales son muy similares al mínimo global; ver Choromanska *et al.* (2015). Una descripción detallada de las redes neuronales para la predicción de series temporales aparece en Palit y Popovic (2005).

Existen numerosos tipos de redes neuronales, siendo una de las más populares la conocida como "*multilayer feedforward networks*", que está definida por: (1) una capa de entrada, formada por varios nodos (o neuronas) que se limitan a recibir los valores de los

predictores; (2) una capa de salida relacionada con la variable que queremos predecir; y (3) entre ambas, una serie de capas ocultas cada una con un número de nodos que reciben información únicamente de los nodos de la capa inmediatamente anterior, y de cuyos nodos sale la información que servirá de *input* a cada uno de los nodos de la capa siguiente. Todos los nodos de una capa están conectados con todos y cada uno de los nodos de la capa siguiente, y los valores que los conectan y relacionan se conocen como "pesos". Finalmente, en cada nodo es necesario definir una función de activación que transforma convenientemente la información que llega de cada uno de los nodos de la capa anterior. Entre las funciones de activación más populares están la lineal, la exponencial, y la logística o sigmoide[5]. En la figura 3 se representa una red neuronal autorregresiva con 4 retardos y una única capa con dos nodos.

FIGURA 3

RED NEURONAL AUTORREGRESIVA CON 4 RETARDOS Y UNA ÚNICA CAPA CON DOS NODOS



Fuente: Elaboración propia.

Para determinar todos los hiperparámetros que aparecen en la red se utiliza un procedimiento de estimación iterativo que utiliza los errores recursivamente para actualizar los pesos. En concreto, el error que se comete en el nodo final o de salida se distribuye hacia atrás a través de todos los nodos ocultos que llegan a dicho nodo final, de tal manera que a cada nodo se le asigna una parte de dicho error ajustando iterativamente

[5] Cuando se utiliza la función de activación logística, las redes funcionan mejor cuando, tanto los predictores como la variable a predecir, están en la escala [0,1]. También se mejora considerablemente el resultado de una red si aquellos predictores que presenten una asimetría importante se transforman mediante la transformación logarítmica antes de convertirlos a la escala [0,1]. También es importante señalar que las redes no tienen un mecanismo de selección automática de los predictores que tienen que ser definidos y tratados previamente.

los pesos de cada nodo. Este procedimientos de optimización secuencial se conoce como "*back propagation of error*", y el hiperparámetro que controla el cambio o la actualización de los pesos de una iteración a la siguiente se conoce como ratio de aprendizaje, que suele frecuentemente fijarse en 0,1. El entrenamiento efectivo de una ANN sigue siendo una tarea difícil dado que el diseño de arquitecturas eficientes es más un arte que una metodología bien estructurada; ver De Stefani *et al.* (2019).

Nótese que una red neuronal sin capas ocultas, es decir, con una capa de entrada con tantos nodos como predictores, y una capa de salida con un único nodo y una función de activación lineal, es equivalente al modelo de regresión lineal. Si la función de activación es la logística, la red se corresponde con el modelo de regresión logística. Sin embargo, los estimadores de los parámetros (pesos) obtenidos a partir de la red neuronal pueden diferir de los obtenidos en los modelos clásicos, porque el procedimiento de estimación difiere de mínimos cuadrados ordinarios (MCO) para el modelo de regresión lineal y de máxima verosimilitud (MV) para el modelo de regresión logística.

Finalmente, debemos mencionar las redes neuronales recurrentes (RNN), que son especialmente interesantes en la predicción de series temporales; ver Lipton, Berkowitz y Elkan (2015) para una revisión crítica. Este tipo de redes fueron concebidas en la década de los 80, pero no se popularizaron en aquel momento por la dificultad de entrenarlas debido a sus elevados requerimientos computacionales. Las redes descritas anteriormente tienen una función de activación que solo actúa en una dirección, hacia adelante, desde una capa hacia la siguiente, es decir, redes que no recuerdan valores previos; ver, por ejemplo, la figura 3. Alternativamente, una RNN incluye conexiones que apuntan hacia atrás, es decir, permiten una especie de retroalimentación entre las neuronas dentro de una misma capa. De esta forma, cada neurona recibe dos entradas, la entrada correspondiente a la capa anterior y a su vez la salida del instante anterior dentro de la misma capa. Esto implica que cada neurona recurrente tiene dos conjuntos de hiperparámetros, uno que se aplica a la entrada de datos que recibe de la capa anterior y otro conjunto aplicado a la entrada de datos correspondiente al vector salida del instante anterior. De esta forma la estructura de estas redes permite una conexión recurrente en el tiempo, y se podría decir que una neurona recurrente tiene en cierta forma memoria.

En general, es importante notar que las redes necesitan un gran número de observaciones para su aprendizaje y las predicciones fuera del rango de valores de las variables con las que han aprendido pueden tener malas propiedades. Además, la convergencia del proceso de estimación a un mínimo global suele ser complicada, por lo que el proceso de búsqueda debe inicializarse en distintos puntos aleatorios. Las redes neuronales pueden tener problemas relacionados con sobreajuste y un gran esfuerzo computacional; ver, por ejemplo, Aras y Kocakoc (2016). Para solucionar estos problemas se han propuesto procedimientos como el *Extreme Learning Machine (ELM)*; ver, por ejemplo, Wang *et al.* (2018) para una descripción.

3. ALGUNAS APLICACIONES EN ECONOMÍA Y FINANZAS

Existe un creciente interés por la utilización de ML en la predicción de variables relacionadas con el área económica, tanto a nivel macro, micro o finanzas. En esta sección revisaremos brevemente algunas de estas aplicaciones para ilustrar el potencial de ML en la predicción económica; ver también Chakraborty y Joseph (2017) para una excelente revisión de la literatura sobre predicción ML en el contexto económico con atención especial a tres casos de interés para bancos centrales. Además, revisaremos comparaciones empíricas de los procedimientos ML utilizados para la predicción de variables económicas con el objetivo de determinar las circunstancias en las que dichos procedimientos son ventajosos frente a procedimientos de predicción "tradicionales".

3.1. Predicción macroeconómica

En el contexto de la predicción macroeconómica, los procedimientos ML descritos anteriormente, se han utilizado fundamentalmente en cuatro áreas de aplicación. En primer lugar, se han utilizado en el contexto de grandes conjuntos de predictores potenciales para reducir su dimensión. La segunda área de aplicación de procedimientos ML es para la predicción de magnitudes macroeconómicas cuando los predictores son no estructurados, como, por ejemplo, textos de noticias o consultas en Google. Los procedimientos ML también se han utilizado para la predicción en caso de relaciones y dependencias temporales no-lineales como las que se suelen observar en épocas de crisis. Finalmente, algunos autores han propuesto utilizar ML para la combinación de predicciones. A continuación, describiremos brevemente algunas aplicaciones empíricas realizadas dentro de cada una de estas áreas.

Reducción de dimensión

En el contexto de la predicción macroeconómica, los procedimientos de regularización se han utilizado frecuentemente para reducir el número de predictores; ver Diebold *et al.* (2020) quienes describen varios trabajos aplicados en los que la regresión con regularización se ha utilizado para la descripción y análisis de variables económicas.

Una de las áreas más populares de aplicación de procedimientos de regularización en el contexto de la predicción macroeconómica es en el contexto de modelos de factores dinámicos. Fan, Ke y Wang (2020) proponen una estrategia consistente basada en extraer factores comunes de los predictores y trabajar con las variables decorreladas. Jokubaitis, Celov y Leipus (en prensa) combinan LASSO y componentes principales para mejorar la predicción de los componentes del producto nacional bruto (PNB) en Estados Unidos (EE. UU.) y la Unión Europea (UE). Concluyen que los procedimientos de regularización mejoran a los procedimientos tradicionales de extracción de factores comunes basados en componentes principales, identificando conjuntos razonables de predictores. Sin embargo, Kim y Swanson (2018) concluyen que los métodos de regularización no mejoran con respecto a los modelos de factores, por lo que sugieren

que, en la práctica es razonable, reducir primero la dimensionalidad mediante la extracción de factores para luego utilizar las regresiones aumentadas con factores estimadas mediante ML; véase, por ejemplo, Bai y Ng (2008,2009), Schumacher (2010) y Umbach (2020) que también proponen utilizar procedimientos de regularización en el contexto de regresiones predictivas aumentadas con factores para la selección de los predictores. Panagiotelis *et al.* (2019) también concluyen, en el contexto de la predicción de magnitudes macroeconómicas en Australia que es difícil mejorar las predicciones de modelos sencillos de extracción de factores, con entre 20 y 40 predictores.

Los procedimientos de regularización también han sido implementados con éxito en el caso de modelos de predicción en los que los parámetros pueden cambiar en el tiempo; ver, por ejemplo, Kapetanios y Zikes (2018). Finalmente, Hillebrand, Lukas y Wei (en prensa) han propuesto una combinación de procedimientos de regularización junto con Bagging para reducir el EPCM en presencia de predictores débiles, definidos como aquellos predictores para los que el incremento de la varianza debido a la estimación de los parámetros correspondientes es mayor que el cuadrado del sesgo que se introduce si no se incorporaran dichos predictores en el modelos predictivo.

Regresores no estructurados

Las predicciones de magnitudes macroeconómicas se basan cada vez más en predictores no-estándar que frecuentemente requieren la utilización de procedimientos de ML; ver, por ejemplo, Clements y Fritsche (en prensa) quienes describen varias aplicaciones en las que se utiliza la información en textos para la predicción. Existe, por ejemplo, una literatura creciente en la que el ciclo económico se relaciona con consultas en Google, noticias y con otras variables no-estándar. Aprigliano, Ardizzi y Monforte (2019) predicen el PNB en Italia en función de los volúmenes y cantidades de diferentes medios de pago como, cheques, transferencias, tarjetas de crédito, etc. Como hemos comentado anteriormente, Babii, Ghysels y Striankas (2020) predicen el crecimiento del PNB en USA en función de datos basados en textos de noticias, mediante sg-LASSO. También Kalamara *et al.* (2020) utilizan datos basados en textos de noticias para predecir el crecimiento del PNB en Reino Unido. Concluyen que las ganancias en predicción son más pronunciadas en tiempos de crisis, cuando las predicciones son generalmente más importantes en la toma de decisiones. Ghirelli *et al.* (2021) describen también varios casos de éxito en la utilización de ML en la predicción económica dentro del Banco de España entre los que se encuentra la previsión a corto plazo de magnitudes macroeconómicas basada en Google trends y noticias de prensa, y la construcción de índices de sentimiento, entre otras.

Dependencias no-linealidades

Sin embargo, no es necesario utilizar datos de variables no-estándar y/o tener grandes bases de datos para recurrir a procedimientos de ML para la predicción. Coulombe *et*

al. (2020) realizan un análisis exhaustivo de las propiedades de los procedimientos de ML para la predicción macroeconómica y concluyen que el factor determinante del éxito de estos procedimientos es la existencia de no-linealidades (que, en el caso de las series macroeconómicas, suelen aparecer en presencia de crisis); ver también Kim y Swanson (2014) que concluyen que los procedimientos ML tienen ventaja predictiva en épocas de crisis[6]. Además también observan que la ventaja de los procedimientos de ML frente a procedimientos tradicionales aumenta con el horizonte de predicción.

Combinación de predicciones

Finalmente, se han realizado varias aplicaciones en las que los procedimientos ML se han utilizado para la combinación de predicciones. Recientemente, Diebold y Shin (2019) proponen utilizar LASSO para estimar los pesos de la combinación de predicciones[7]. Es conocido que, cuando se combinan predicciones, la combinación que parece tener mejores propiedades es la media. Sin embargo, al combinar un número muy grande de predicciones, algunas de ellas pueden ser redundantes o, al menos, no tener suficiente información como para ser incluidas. Diebold y Shin (2019) proponen estimar los coeficientes de la combinación de predicciones utilizando LASSO y centrando la restricción alrededor de $\frac{1}{K}$, donde K es el número de predicciones que se combinan, de forma que algunas de las predicciones de la combinación desaparezcan y los coeficientes del resto se reduzcan hacia la media. El procedimiento es denominado *partially-egalitarian LASSO* (pe-LASSO). Diebold y Shin (2019) analizan el comportamiento de pe-LASSO para predecir el crecimiento en la eurozona basado en la combinación de predicciones del *Survey of Professional Forecasters* del Banco Central Europeo y concluyen que obtienen EPCMs menores que con la media muestral de las predicciones.

Recientemente, Montero-Manso *et al.* (2020) también han propuesto utilizar ML, en concreto, Gradient Boosting para combinar las predicciones de diversos modelos.

3.2. Predicción microeconómica

Los procedimientos ML también han sido utilizados a nivel microeconómico para la predicción de la demanda de productos concretos; ver, por ejemplo, Huber y Stuckenschmit (en prensa) que utilizan Boosting para la predicción de la demanda de una distribuidora de pan y Fildes, Ma y Kolassa (en prensa) para una revisión de la literatura sobre demanda al por menor. Bose *et al.* (2017), dentro de Amazon, proponen una plataforma para la predicción probabilística de la demanda que permite trabajar con catálo-

[6] Véase también la lista de referencias con aplicaciones de procedimientos de ML para la predicción de series macroeconómicas.

[7] Bajo una función de pérdida cuadrática, el problema de la combinación óptima de predicciones es un problema de regresión lineal.

gos de millones de productos. Taieb y Hyndman (2014) predicen la demanda de energía eléctrica mediante un procedimiento híbrido en el que se utiliza Boosting en los residuos de un modelo lineal. Finalmente, Camacho, Ramallo y Ruiz Marín (2021) predicen los precios de la vivienda en Madrid basados en árboles de decisión en un contexto de datos de sección cruzada.

Además, Choi y Varian (2012) ilustran cómo utilizar los índices diarios y semanales publicados en tiempo real por Google sobre el volumen de consultas que los usuarios hacen en internet (Google trends) para predecir variables como, por ejemplo, las ventas de vehículos a motor, las reclamaciones del subsidio de desempleo, el nivel de confianza de los consumidores o el número de turistas en Australia.

En un contexto microeconómico, ML también se ha utilizado con éxito para la predicción de quiebras. En un artículo muy interesante, Petropoulos *et al.* (2020) analizan un panel de datos bancarios para determinar los determinantes de quiebra y, simultáneamente, proporcionar una señal adelantada para una quiebra potencial. En este análisis utilizan redes elásticas para seleccionar las variables determinantes de la quiebra y comparan las predicciones de varios procedimientos ML para la predicción, concluyendo que RF y ANN tienen un comportamiento parecido y superior a los demás métodos considerados.

Finalmente, en el contexto de predicción microeconómica, los procedimientos ML también se han implementado para la combinación de predicciones. Por ejemplo, Barrow y Crone (2016a) utilizan ML para la combinación de predicciones y analizan el comportamiento del procedimiento propuesto en el contexto de un subconjunto de series de ventas de la competición M3.

3.3. Predicción financiera

Lo mismo que en el caso de la predicción macroeconómica, las principales aplicaciones de predicción ML en el contexto de variables financieras, pueden agruparse en aquellas que tratan de reducir la dimensión, las que utilizan predictores no estándar y las que tratan de funciones predictivas no lineales.

Regularización

Swanson, Xiang y Yang (en prensa) utilizan componentes principales con *targeted predictors* (basado en LASSO y elastic net) para predecir la estructura de los tipos de interés utilizando índices económicos. Concluyen que ML puede ser importante en la predicción cuando esta se realiza de manera individual. Sin embargo, cuando consideran la combinación de diferentes modelos predictivos, los modelos que sólo utilizan el pasado de los tipos de interés tienen EPCMs menores que aquellos que incorporan información de variables macroeconómicas.

Predictores no estándar

También existen aplicaciones muy interesantes de datos no estructurados para la predicción financiera, intentando relacionar el comportamiento de los mercados con noticias o búsquedas en Google. Una de las primeras contribuciones en esta dirección es Tetlock, Saar-Tsechansky y Macskassy (2008) que analizan la relación entre los medios de difusión y el mercado financiero. Recientemente, ha aparecido un concepto nuevo conocido como *Internet concern,* que trata de la utilización de datos de búsquedas para la cuantificación de la especulación en un mercado determinado. Wang *et al.* (2018) utilizan Internet concern para predecir la volatilidad en el mercado del crudo. Perlin *et al.* (2017) también han analizado la relación de las búsquedas en Google con mercados financieros. Chen *et al.* (en prensa) utilizan ANN y RF para predecir los tipos de cambio del bitcóin en función de varios factores económicos y tecnológicos entre los que está, por ejemplo, el número de búsquedas en Google y en la Wikipedia. El reciente artículo de Gentzkow, Kelly y Taddy (2019) realiza una excelente revisión sobre procedimientos ML aplicados a la utilización de textos en la predicción económica y financiera.

En el caso de datos de alta frecuencia (intradiarios), Yan *et al.* (2020) utilizan SVM para predecir volatilidad.

Finalmente, Arratia (2021) propone modelizar rendimientos financieros en función de sentimientos. Limongi y Ravazzolo (2019) también investigan cómo los sentimientos de los inversores afectan a los rendimientos financieros y evalúan el poder predictivo de los índices de sentimiento sobre rendimientos de mercados financieros en EE.UU. y UE.

Dependencias no lineales

En el área de finanzas, Israel, Kelly y Moskowitz (en prensa) describen las posibilidades de ML en relación a distintos problemas. En concreto, estos autores exponen los retos a los que se enfrentan los procedimientos ML en la predicción de rendimientos financieros, caracterizados por: (1) no tener (relativamente) grandes cantidades de datos, al estar estos limitados por el hecho de que no se puede experimentar para generar nuevas observaciones, y_i, de las que se pueda aprender; y (2) por tener la señal relativamente pequeña en relación al ruido, consecuencia de la eficiencia de los mercados financieros. Sin embargo, los procedimientos ML se han utilizado con éxito en algunos problemas de predicción de rendimientos financieros como, por ejemplo, cuando existen no-linealidades en las relaciones dinámicas entre los rendimientos financieros y variables de estado o dinámicas complejas en la exposición a los factores de riesgo; ver, por ejemplo, Gu, Kelly y Xiu (en prensa a, en prensa b) que muestran ganancias significativas, tanto desde el punto de vista financiero como estadístico, cuando se utilizan autoenconders (redes neuronales correspondientes a componentes principales), árboles o redes neuronales, respectivamente, para la predicción de precios de activos financieros individuales. Sin embargo, recientemente, De Nard, Hediger y Leippold (2020)

proponen utilizar procedimientos más sencillos basados en *subsampling,* con mejores resultados.

3.4. Comparación entre procedimientos

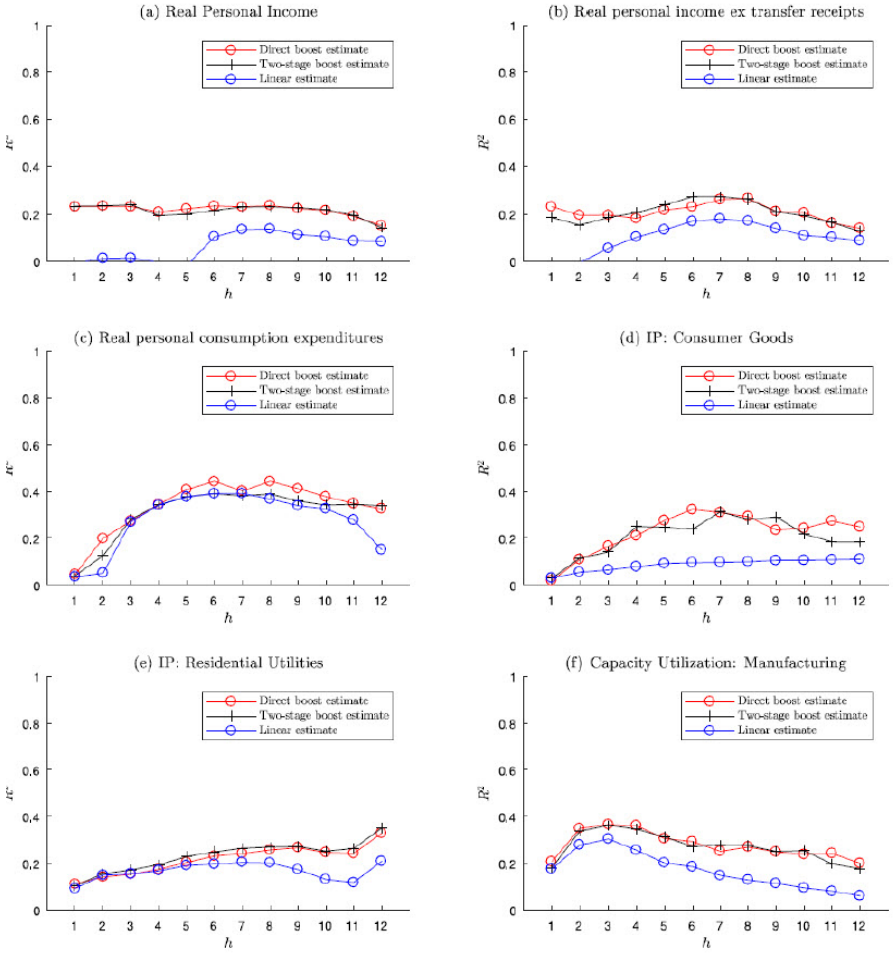
Los procedimientos ML han sido comparados entre sí y con otros procedimientos tradicionales en el contexto de la predicción de variables económicas. Las comparaciones se han realizado tanto en el contexto de competiciones de procedimientos predictivos en las que se comparan los procedimientos en la predicción de un gran número de variables como en el contexto de la predicción de una única o de un número reducido de variables.

Entre las comparaciones de predicciones en el contexto de un gran número de variables, una de las primeras comparaciones sistemáticas entre un amplio abanico de procedimientos ML fue realizada por Ahmed *et al.* (2010) en el contexto de las variables mensuales relacionadas con la economía y negocios de la competición M3; ver Makridakis y Hibon (2000) para una descripción de dicha competición. En dicho contexto y realizando predicciones un periodo-hacia-adelante univariantes, Ahmed *et al.* (2010) concluyen que los mejores procedimientos, en términos de minimizar el EPCM, se obtienen utilizando las redes neuronales y la regresión de procesos Gaussianos. Más recientemente, Smyl (2020) es el procedimiento ganador de la última competición M4. El procedimiento propuesto por Smyl (2020) se basa en una combinación entre suavizamiento exponencial y redes neuronales recurrentes. Kim y Swanson (2014) también comparan distintos procedimientos ML para la predicción de 144 variables macroeconómicas de EE.UU. en el contexto de los modelos de difusión basados en factores. Nuevamente, concluyen que la ventaja de los procedimientos ML se encuentra en presencia de no-linealidades con procedimientos basados en la combinación de factores y regularización obteniendo menores EPCMs. Kauppi y Virtanen (en prensa) también realizan una análisis exhaustivo comparando predictores lineales y Boosting para las superanalizadas 128 series macroeconómicas de McCraken y Ng (2016). La ventaja comparativa de Boosting frente a procedimientos lineales depende del tipo de no-linealidad que tengan los datos y, por lo tanto, puede cambiar dependiendo de la variable concreta que se vaya a predecir. Por ejemplo, la figura 4 que se ha tomado de Kauppi y Virtanen (en prensa) muestra los coeficientes de determinación en la predicción de varias magnitudes macroeconómicas para $h = 1,...,12$ cuando se utiliza Boosting y modelos lineales en la predicción. Las magnitudes que se predicen son ingresos y consumo individuales, precios de bienes de consumo y de residencias y la capacidad de utilización. Esta figura ilustra que las conclusiones sobre la comparación entre procedimientos tradicionales y ML pueden depender tanto de las propiedades de la variables que se predice como del horizonte de predicción. Por ejemplo, en el caso de la predicción del consumo individual, las diferencias solo se observan en el caso de predicciones 12 meses hacia adelante. Sin embargo, en el caso de las variables de ingresos, las diferencias se observan en las predicciones a corto plazo y se hacen menores cuando el horizonte de predicción se incrementa.

Como hemos comentado anteriormente, el comportamiento de los procedimientos de predicción ML también se ha comparado con procedimientos "tradicionales" en el contexto de la predicción de una única o un número reducido de variables. Muy recientemente, Joseph *et al.* (2021) comparan varios procedimientos para la predicción de la inflación en Reino Unido basada en la desagregación de los precios a nivel granular (581 precios desagregados). Los modelos que utilizan en su comparación se basan en: (1) Modelos de factores dinámicos para la reducción de la dimensionalidad; (2) modelos de regresión con regularización (*Ridge* y LASSO); y (3) ML (*Elastic Net,* SVM, ANN y

FIGURA 4

PREDICCIONES DE VARIABLES MACROECONÓMICAS TOMADAS DE KAUPPI Y VIRTANEN (EN PRENSA)



RF). Las predicciones obtenidas con estos procedimientos son comparadas en términos de EPCMs con respecto a un modelo autorregresivo. Concluyen que los procedimientos no-lineales de ML y, en particular, SVM, tienen un gran potencial en la predicción de los datos granulares y muy volátiles que se obtienen a nivel de precios microeconómicos, abriendo la puerta a que puedan ser utilizados para modelizar precios muy desagregados a nivel de productos o incluso de precios obtenidos de las páginas web. Los mejores resultados de los procedimientos de ML se observan al aumentar el horizonte de predicción y, en concreto, en la predicción a medio plazo (un año hacia adelante). La conclusión sobre que la utilización de procedimientos ML es más útil cuando el horizonte de predicción se incrementa también se ha obtenido por muchos otros autores como, por ejemplo, Kim y Swanson (2014) en el contexto de su comparación y D'Amuri y Marcucci (2017) en la predicción de desempleo en EE.UU. Kotchoni, Leroux y Stevanovic (2019) también concluyen que la utilización de procedimientos de regularización para predecir varias variables económicas tiene ventajas al aumentar el horizonte de predicción. Colombo y Pelagatti (en prensa) llegan a la misma conclusión al predecir tipos de cambio. La mejora en la capacidad predictiva de los procedimientos ML aparece con $h>3$.

Fornaro y Luomaranta (2020) utilizan procedimientos Boosting para proporcionar predicciones de la economía finlandesa. Comparan dichas predicciones con otras obtenidas con procedimientos "tradicionales" y concluyen que las predicciones basadas en BD se adelantan en el tiempo a las tradicionales lo que, evidentemente es una gran ventaja cuando dichas predicciones van a ser utilizadas en la toma de decisiones.

Finalmente, Medeiros *et al.* (2021) también predicen la inflación en USA y concluyen que los procedimientos ML con un gran número de predictores producen predicciones más precisas que otros modelos tradicionales de series temporales. Entre los procedimientos ML, ellos concluyen que el que merece más atención es RF, cuyo buen comportamiento es debido no únicamente a cómo se seleccionan los predictores si no también a la incorporación de relaciones dinámicas no lineales entre variables macroeconómicas relevantes y la inflación. Mediante RF se pueden reducir hasta un 30% el EPCM cuando se compara con predicciones basadas en procedimientos tradicionales de series temporales.

4. CONCLUSIONES

A lo largo de este capítulo, hemos ilustrado cómo los procedimientos ML pueden ser muy útiles en distintos contextos de la predicción relacionada con variables económicas. Los procedimientos basados en regularización han mostrado su ventaja en la predicción basada en factores comunes como una forma automática de seleccionar los predictores. Los procedimientos basados en árboles de decisión y redes neuronales se han mostrado especialmente ventajosos en el caso de la predicción de variables con dependencias no-lineales desconocidas. Además, los procedimientos de predicción ML tienen gran potencial cuando los datos son granulares y volátiles, así como en aplica-

ciones con predictores no estructurados. Finalmente, varios autores concluyen que las predicciones basadas en ML pueden adelantarse en el tiempo a las basadas en procedimientos tradicionales y que tienen ventajas comparativas a medida que el horizonte de predicción se incrementa.

Los procedimientos de predicción ML han sido originalmente diseñados para observaciones de sección cruzada independientes. Sin embargo, dichos procedimientos están aún en su infancia en el caso de datos temporales, con muchos temas importantes sin explorar o solo explorados parcialmente; ver la discusión en Barker (2020). Por ejemplo, algunos de los procedimientos ML diseñados para datos independientes, se han extendido a datos temporales estacionarios. Sin embargo, la predicción mediante ML en el caso de variables no estacionarias todavía está sin tratar en profundidad con muy pocas menciones en la literatura. Ahmed *et al.* (2010) han considerado el efecto de la diferenciación en el contexto de las variables económicas observadas mensualmente de la competición M3 cuando se utiliza ML para la predicción univariante y concluyen que los resultados son peores (mayores EPCM) que cuando se predicen las variables en niveles. Bontempi, Ben Taieb y Le Borgue (2013) han considerado la utilización de ML en el contexto de variables no estacionarias y concluyen que los procedimientos de aprendizaje local pueden aplicarse a variables no estacionarias. Finalmente, en relación con la presencia de no-estacionariedad, Colombo y Pelagatti (en prensa) han considerado la utilización de VSM en el contexto de modelos con mecanismo de corrección del error mientras que Escribano y Wang (en prensa) predicen inflación del precio de la gasolina en España mediante una combinación de procedimientos de series temporales clásicos basados en cointegración no-lineal y árboles aleatorios, mejorando a cada uno de estos procedimientos por separado en términos de EPCM.

En este capítulo, por cuestiones de espacio, nos hemos centrado en predicción un-periodo-hacia-adelante de una única variable. La predicción multivariante está todavía desarrollándose. Recientemente, De Stefani *et al.* (2019) han considerado predicción multivariante proponiendo una versión ML de los modelos de factores dinámicos (DFMs). Aunque la mayoría de la literatura se ha centrado en predicciones un-periodo-hacia-adelante, existen también varias aportaciones donde se han considerado predicciones varios-periodos-hacia-adelante; ver, por ejemplo, Bontempi, Taieb y Le Borgue (2013), Taieb y Hyndman (2014), De Stefani *et al.* (2019) y Kauppi y Virtanen (en prensa).

Las predicciones son realmente útiles para la toma de decisiones cuando van acompañadas de medidas de incertidumbre, es decir, cuando aparecen en términos probabilísticos. Evidentemente, cualquier predicción lleva asociada una incertidumbre por lo que no es sólo importante que dicha incertidumbre sea lo más pequeña posible sino también que, sea cuál sea su nivel, la podamos medir adecuadamente. La toma de decisiones será diferente dependiendo de la incertidumbre asociada con las predicciones que se utilicen. La predicción de las probabilidades asociadas con fenómenos futuros y la consideración de diversas formas de incertidumbre está siendo el foco de atención

de muchos trabajos relacionados con los procedimientos de ML descritos en la sección anterior; ver las referencias en la discusión de Januschowski *et al.* (2020) y Salinas *et al.* (2020). Sin embargo, este es un área en la que todavía es necesario realizar una investigación intensa. Hyndman y Athanasopoulos (2018) proponen utilizar boostrap basado en residuos para construir intervalos de predicción en el caso de redes neuronales autorregresivas.

Finalmente, queremos mencionar que, a pesar de que los procedimientos ML han demostrado su utilidad empírica, en muchos casos, se carece de una explicación analítica del porqué de estos buenos resultados. Solamente existen algunas explicaciones parciales que ayuden a entender mejor las propiedades de estos procedimientos. Por ejemplo, recientemente, Scornet, Biau y Vert (2015) han probado la consistencia del procedimiento Bagging mientras que Athey, Tibshirani y Wager (2019) proporcionan resultados asintóticos para RF; ver también Bülmann y Yu (2002) para algunos resultados asintóticos del Bagging y Subbagging.

Referencias

Ahmed, N. K., Atiya, A. F., El Gayar, N. y El Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews,* 29(5-6), pp. 594-621.

Aprigliano, V., Ardizzi, G. y Monforte, L. (2019). Using payment system data to forecast economic activity. *International Journal of Central Banking,* 15(4), pp. 55-80.

Aras, S. y Kocakoc, I. D. (2016). A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing,* 174, pp. 974-987.

Aromi, J. D. (2020). Linking words in economic discourse: Implications for macroeconomic forecasts. *International Journal of Forecasting,* 36(4), pp. 1517-1530.

Arratia, A. (2021). Predicciones financieras basadas en análisis de sentimiento de textos y minería de datos. En: D. Peña, P. Poncela y E. Ruiz (eds.), *Nuevos Métodos de Predicción Económica con Datos Masivos.* Madrid: Funcas.

Athey, S., Tibshirani, J. y Wager, S. (2019). Generalized random forests. *Annals of Statistics,* 47(2), pp. 1148-1178.

Awad, M. y Khanna, R. (2015). *Efficient Learning Machines.* Berkeley: Springer.

Babii, A., Ghysels, E. y Striankas, J. (2020). Machine learning time series regressions with an application to nowcasting, arXiv:2005.14057v2[econ.EM].

Bai, J. y Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics,* 146, pp. 304-317.

— (2009). Boosting diffusion indexes. *Journal of Applied Econometrics,* 24, pp. 607-629.

Barker, J. (2019). Machine learning in M4: What makes a good unstructured model? *International Journal of Forecasting,* 36(1).

Barrow, D. K. y Crone, S. F. (2016a), A comparison of AdaBoost algorithms for time series forecast combination. *International Journal of Forecasting,* 32, pp. 1103-1119.

— (2016b). Cross-validation aggregation for combining autoregressive neural network forecasts. *International Journal of Forecasting,* 32(4), pp. 1120-1137.

Biau, G. y Scornet, E. (2016). A random forest guided tour, with discussion. *TEST,* 25, pp. 197-227.

Bontempi, G., Ben Taieb, S. y Le Borgue, Y.-A. (2013). Machine learning strategies for time series forecasting. En: M.-A. Aufure y E. Zimanyi (eds.), *Business Intelligence.* Berlin: Springer-Verlag.

Bose, J.-H., Flunkert, V., Gasthaus, J., Januschowski, T., Large, D., Salinas, D., Schelter, S., Eeger, M. y Wang, Y. (2017). Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment,* 10(12).

Breiman, L. (2001). Random forest. *Machine Learning,* 45(1), pp. 5-32.

Bülmann, P. y Yu, B. (2002). Analyzing Bagging. *Annals of Statistics,* 30(4), pp. 927-961.

Camacho, M., Ramallo, S. y Ruiz Marín, M. (2021). Arboles de decisión en economía: una aplicación a la determinación del precio de la vivienda. En: D. P. Peña, P. Poncela y E. Ruiz (eds.). *Nuevos Métodos de Predicción Económica con Datos Masivos.* Madrid: Funcas.

Caro, A. y Peña, D. (2021). Predicción de series temporales económicas con modelos factoriales dinámicos y de Machine Learning. En: D. Peña, P. Poncela y E. Ruiz (eds.). *Nuevos Métodos de Predicción Económica con Datos Masivos.* Madrid: Funcas.

Chakraborty, C. y Joseph, A. (2017). Machine learning at central banks. *Staff Working Paper,* no. 674.

Chen, W., Xu, H., Jiu, L. y Gao, Y. (2021). Machine learning model for Bitcoin exchange rate predictions using economic and technology determinants. *International Journal of Forecasting*.

Choi, H. y Varian, H. (2012). Predicting the present with Google trends. *Economic Records,* 88, pp. 2-9.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B. y LeCun, L. (2015). The loss surfaces of multilayer networks. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS).*

Clements, M. P. y Fritsche, U. (2020).Text-based data and forecasting: Editors introduction. *International Journal of Forecasting,* 36(4), pp. 1476-1477..

Colombo, E. y Pelagatti, M. (2020). Statistical learning and exchange rate forecasting. *International Journal of Forecasting,* 36(4), pp.1260-1289..

Coulombe, P. G., Leroux, Stevanovic, M. D. y Surprenant, S. (2021). How is machine learning useful for macroeconomic forecasting? *International Journal of Forecasting.*

D'Amuri, F. y Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting,* 33, pp. 801-816.

De Nard, G., Hediger, S. y Leippold, M. (2020). *Subsample factor models for asset pricing: The rise of VASA.* Manuscrito.

De Stefani, J., Le Borgne, Y.-A., Caelen, O., Hattab, D. y Bontempi, G. (2019). Batch and incremental dynamic factor machine learning for multivariate multi-step-ahead forecasting. *International Journal of Data Science and Analytics,* 7(4), pp. 311-329.

Diebold, F. X. y Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting,* 35, pp. 1679-1691.

Diebold, F. X., Ghysels, E., Mykland, P. y Zang, L. (2019). Big data in dynamic predictive econometric models (editorial). *Journal of Econometrics,* 212(1), pp. 1-3.

Eklund, J. y Kapetanios, G. (2008). A review of forecasting techniques for large data sets. *National Institute Economic Review,* 203(1), pp. 109-115.

Escribano, A. y Wang, D. (en prensa). Forecasting gasoline prices with mixed random forest error correction models. *International Journal of Forecasting.*

Fan, J., Ke, Y. y Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics,* 216, pp. 71-85.

Fildes, R., Ma, S. y Kolassa, S. (en prensa). Retail forecasting: Research and practice. *International Journal of Forecasting.*

Fornaro, P. y Luomaranta, H. (2020). Nowcasting Finish real economic activity: a machine learning approach. *Empirical Economics,* 58(1), pp. 55-71.

Garboden, P. M. E. (2019). Sources and types of big data for macroeconomic forecasting. En: P. Fuleky (ed.). *Macroeconomic Forecasting in the Era of Big Data.* Springer.

Gentzkow, M., Kelly, B. T. y Taddy, M. (2019). Text as data. *Journal of Economic Surveys,* 57(3), pp. 535-574.

Ghirelli, C., Hurtado, S., Pérez, J. J. y Urtasun, A. (2021). Desarrollos con Big data para el análisis coyuntural en los bancos centrales. En: D. Peña, P. Poncela y E. Ruiz (eds.). *Nuevos Métodos de Predicción Económica con Datos Masivos,* Madrid: Funcas.

Gu, S., Kelly, B. y Xiu, D. (en prensa a). Autoenconder asset pricing models. *Journal of Econometrics*.

— (en prensa b). Empirical asset pricing via machine learning. *Review of Financial Studies.*

Hillebrand, E., Likas, M. y Wei, W. (en prensa). Bagging weak predictors. *International Journal of Forecasting.*

Hu, M. J. C. (1964), Application of the adaline system to weather forecasting. *Master Thesis, Technical report,* 6775-1. Stanford, CA: Stanford Electronic Laboratories,.

Huber, J. y Stuckenschmit, H. (en prensa), Daily retail demand forecasting using machine learning with emphasis on calendar special days. *International Journal of Forecasting.*

Hyndman, R. y Athanasopoulos, G. (2018). *Forecasting: Principles and Practice,* 2ª edición, OTexts, Melbourne (Australia). OTexts.com/fpp2. Descargado en 7/29/2020.

Israel, R., Kelly, B. y Moskowitz, T. (en prensa). Can machines "learn" finance? *Journal of Investment Management.*

Januschowski, T., Gasthaus, J., Wang, Y. Salinas, D., Flunkert, V., Bohlke-Scheider, M. y Lallot, C. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting,* 36, pp. 167-177.

Joseph, A., Kalamara, E. Potjagailo, G. y Kapetanios, G. (2021). Forecasting UK inflation bottom up. *International Journal of Forecasting.*

Jokubaitis, S., Celov, D. y Leipus, R. (en prensa). Sparse structures with LASSO through principal components: Forecasting GDP components in the short run. *International Journal of Forecasting*.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G. y Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. *Staff Working Paper,* no. 865. Bank of England.

Kanuppi, H. y Virtanen, T. (2020). Boosting nonlinear predictability of macroeconomic time series. *International Journal of Forecasting,* 37(1), pp. 151-170.

Kapetanios, G. y Zikes, F. (2018). Time varying LASSO. *Economics Letters,* 169, pp. 1-6.

Kim, H. H. y Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics,* 178, pp. 352-367.

— (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage. *International Journal of Forecasting,* 34(2), pp. 339-354.

Kotchoni, R., Leroux, M. y Stevanovic, D. (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics,* 34(7), pp. 1050-1072.

Limongi Conceto, C. y Ravazzolo, F. (2019). Optimism in financial markets: Stock market returns and investor sentiments. *Journal of Risk and Financial Management,* 12(141).

Lipton, Z. C., Berkowitz, J. y Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning, arXiv:1506.00019

Makridakis, S. y Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting,* 16, pp. 451-476.

McCracken, M. W. y Ng, S. (2016), FRED-MD: A monthly data base for macroeconomic research. *Journal of Business & Economic Statistics,* 34, pp. 574-589.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, A. y Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of ML methods. *Journal of Business and Economic Statistics,* 39(1).

Montero-Manso, P., Athanasopoulos, G., Hyndmann, R. J. y Talagala, T. S. (2020). FFORMA: feature-based forecast model averaging. *International Journal of Forecasting,* 36(1), pp. 86-92.

Palit, A. K. y Popovic, D. (2005). Computational Intelligence in Time Series Forecasting. *Theory and Engineering Applications,* Springer-Verlag, Secaucus.

Pan, L. y Politis, D. N. (2016). Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions (with discussion). *Journal of Statistical Planning and Inference,* 177, pp. 1-27.

Panagiotelis, A., Athanasopoulos, G., Hyndman, R. J., Juang, B. y Vahid, F. (2019). Macroeconomic forecasting for Australia using a large number of predictors. *International Journal of Forecasting,* 35(2), pp. 616-633.

Perlin, M. S., Caldeira, A., Santos, A. P. y Pontuschka, M. (2017), Can we predict the financial markets based on Google's search quaries? *Journal of Forecasting,* 36, pp. 454-467.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. y Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting,* 36(3), pp. 1092-1113.

Salinas, D., Flunkert, V., Gasthaus, J. y Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting,* 36(3), pp. 1181-1191.

Schumacher, C. (2010). Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters,* 107(2), pp. 95-98.

Scornet, E., Biau, G. y Vert, J.-P. (2015). Consistency of random forests. *Annals of Statistics,* 43, pp. 1716-1741.

Scott, S. L. y Varian, H. (2014), Predicting the present with Bayesian structural time series. *Journal of Mathematical Modelling and Numerical Optimization,* 5(1-2).

Simon, N., Friedman, J., Hastie, T. y Tibshirani, R. (2013), A sparse-group LASSO. *Journal of Computational and Graphical Statistics,* 22(2), pp. 231-245.

Smola, A. J. y Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing,* 4, pp. 199-222.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting,* 36(1), pp. 75-85.

Swanson, N. R., Xiang, W. y Yang, X. (en prensa). Predicting interest rates using shrinkage methods, real-time diffusion indexes, and model combinations. *Journal of Applied Econometrics,* 35(5), pp. 587-613.

Taieb, S. B. y Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting,* 30(2), pp. 382-394.

Tetlock, P. C., Saar-Tsechansky, M. y Macskassy, S. (2008). More than words: quantifying langiage to measure firms' fundamentals. *Journal of Finance,* 63(3), pp. 1437-1467.

Tsay, R. y Chen, R. (2019). *Nonlinear Time Series Analysis.* Wiley.

Umbach, S. L. (2020). Forecasting with supervised factor models. *Empirical Economics,* 58(1), pp. 169-190.

Vapnik, V. N. (2006). *Estimation of Dependences Based on Empirical Data.* New York: Springer.

Varian, H. R. (2014). Big Data: New tricks for econometrics. *Journal of Economic Prespectives,* 28(2), pp. 3-18.

Wang, J., Athanasopoulos, G., Hyndman, R. J. y Wang, S. (2018). Crude oil price forecasting based on internet concern using an extreme learning machine. *International Journal of Forecasting,* 34(4), pp. 665-677.

Yan, R., Yu, L., Yu, H., Xu, G., Wu, Y. y Liu, Z. (2020). Big data analytics for financial market volatility forecast based on support vector machine. *International Journal of Information Management,* 50, pp. 452-462.

Zhang, G., Patuwo, B. E. y Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting,* 14(1), pp. 35-62.

Funcas

C/ Caballero de Gracia, 28
Madrid, 28013, Spain
Tel. +34 91 5965481 +34 91 5965718
Email: publica@funcas.es
www.funcas.ceca.es

P.V.P.: Edición papel, 20€ (IVA incluido)
P.V.P.: Edición digital, gratuita

ISBN 978-84-17609-48-1