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

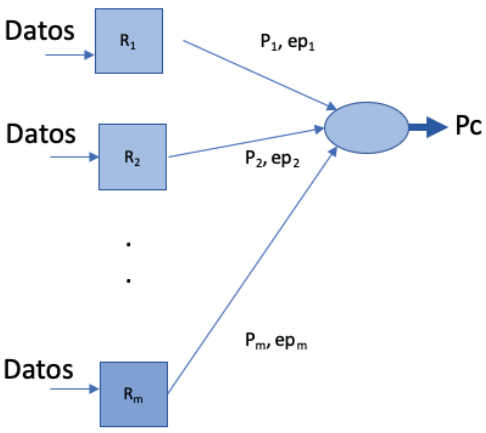
proporcionan los métodos clásicos de regresión paso a paso, ya que la regla final de predicción con boosting no corresponde a una estimación de un modelo de regresión múltiple por mínimos cuadrados. En estos métodos de selección de variables tradicionales vamos escogiendo nuevas variables por su capacidad 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:

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

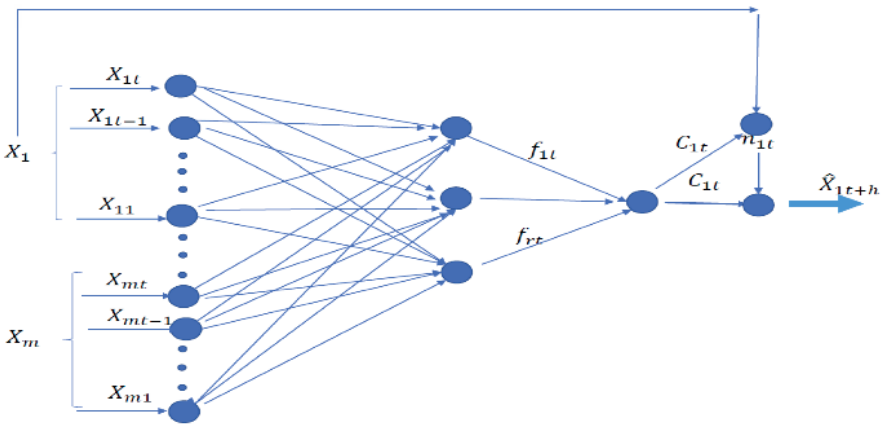
donde el valor de la serie i en el instante t es la suma de los efectos de los factores sobre esa serie en ese instante, más su término específico. Agrupando todas las series y todos los instantes temporales en una matriz de datos, X, que tendrá dimensiones T x m, los valores en esta matriz de datos 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

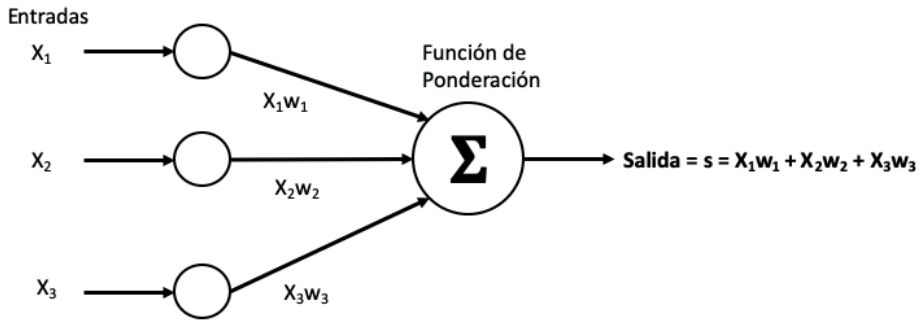
REPRESENTACIÓN GRÁFICA DE UN MODELO FACTORIAL DINÁMICO



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.

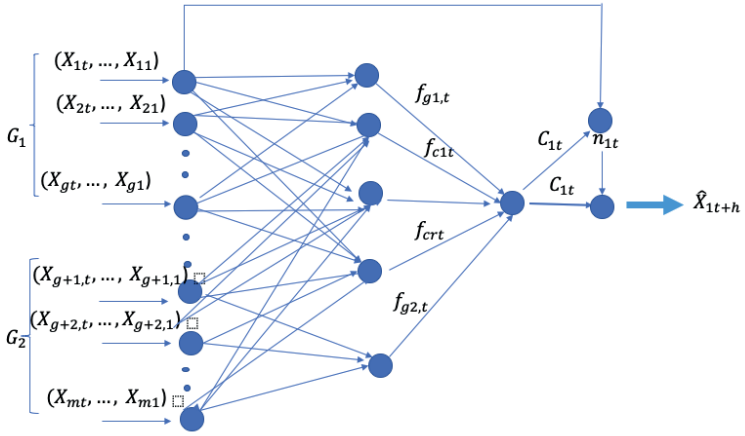
Los modelos factoriales dinámicos pueden incluir heterogeneidad como atípicos o estructura de grupos, véase Alonso, Galeano y Peña (2020). La figura 4 presenta un ejemplo de este tipo de modelo con dos grupos. Las series $1,...g$ pertenecen al primer grupo y las $g+1,...,m$ al segundo. Por ejemplo, corresponden 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



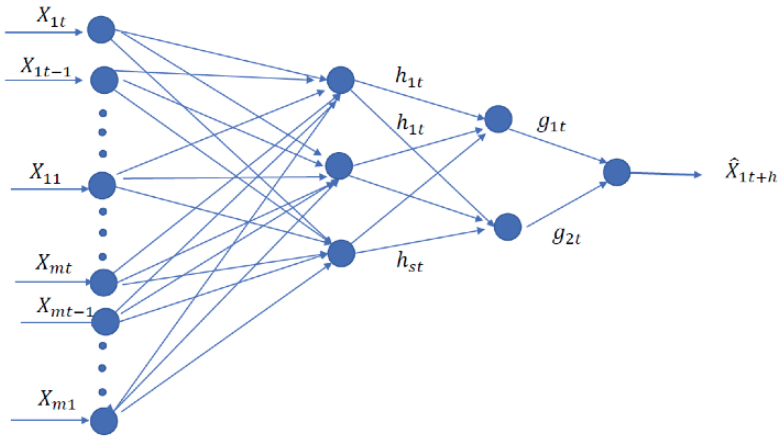
Fuente: Elaboración propia.

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

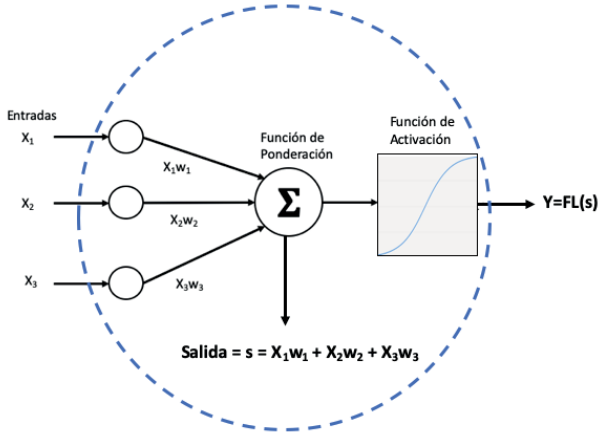
UNA RED NEURONAL CON DOS CAPAS INTERMEDIAS CON S Y 2 NUDOS



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.

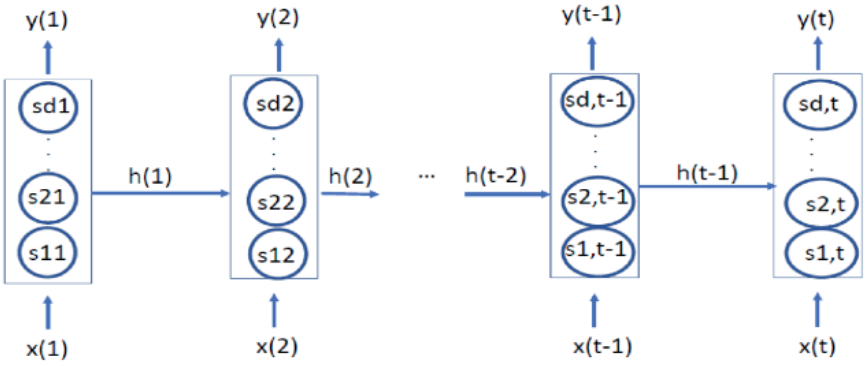
La figura 5 representa una red neuronal o perceptrón con dos capas, la primera con s nudos y la segunda con 2 nudos. En cada nudo se realizan las operaciones que se detallan en la figura 6: una combinación lineal de la entrada con ciertos pesos y una transformación no lineal, que suele ser la función 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,

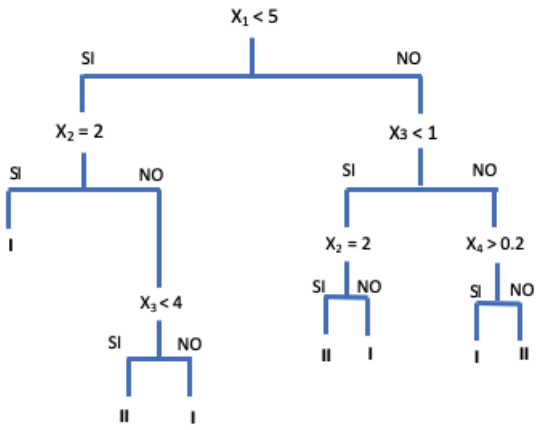
y el proceso se continúa mientras haya reducciones claras en el MSE. Si continuásemos hasta el final podríamos dividir la muestra hasta llegar a N grupos donde cada observación se prevería con su valor y el MSE sería cero. Obviamente, este resultado no es muy útil y generalmente el proceso finaliza cuando 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, $\sum_{j=1}^{p}|\beta_j|$, 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).

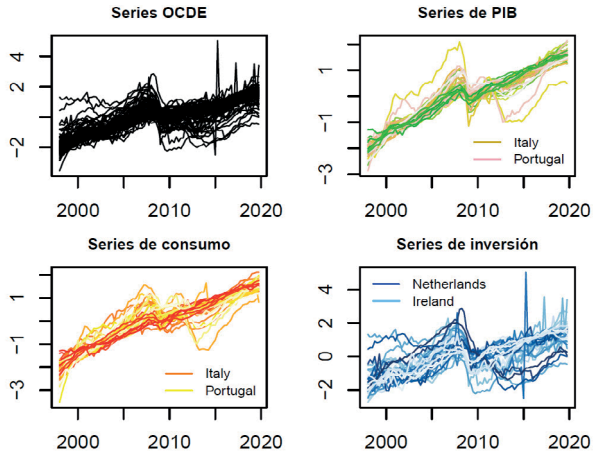
5. APLICACIÓN EMPÍRICA: PREDICCIÓN DE VARIABLES ASOCIADAS AL CICLO ECONÓMICO

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)

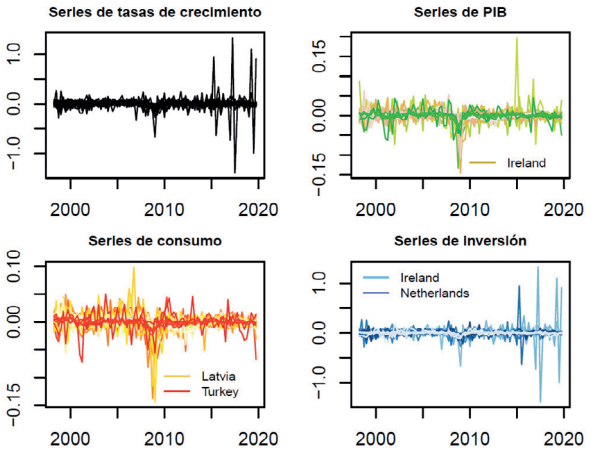


Fuente: Elaboración propia.

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

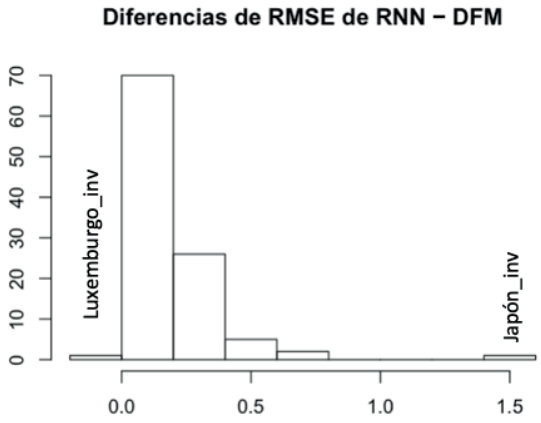
PRECISIÓN DE LAS PREDICCIONES CUANDO EL HORIZONTE DE PREDICCIÓN ES A UNO Y TRES PASOS, $h=1,3$

	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

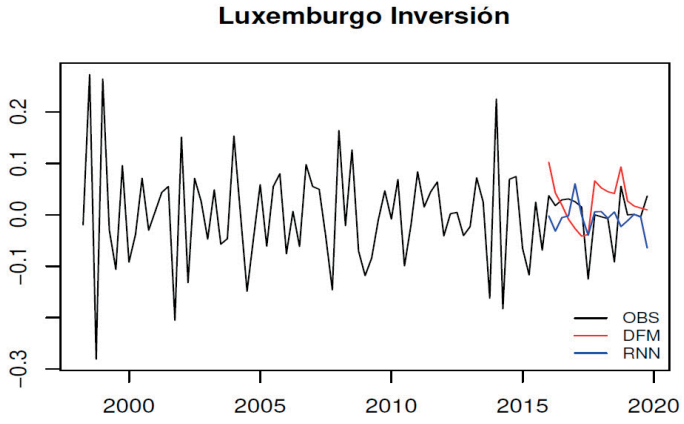


Fuente: Elaboración propia.

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

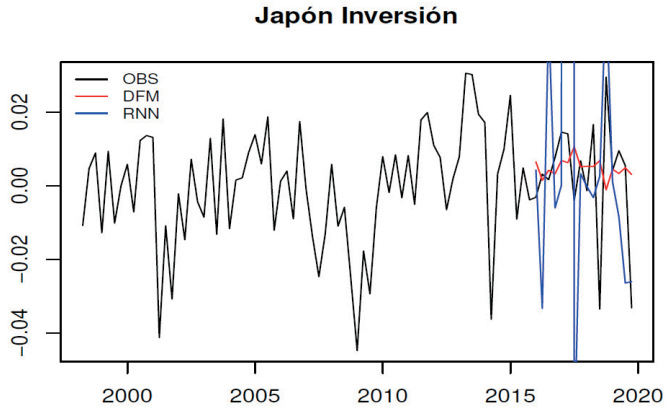
PREDICCIÓN DE LA TASA DE CRECIMIENTO DE LA INVERSIÓN DE LUXEMBURGO (NEGRO) USANDO UN MODELO DE FACTORES (AZUL) Y UNA RED NEURONAL RECURRENTE (ROJO)



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)

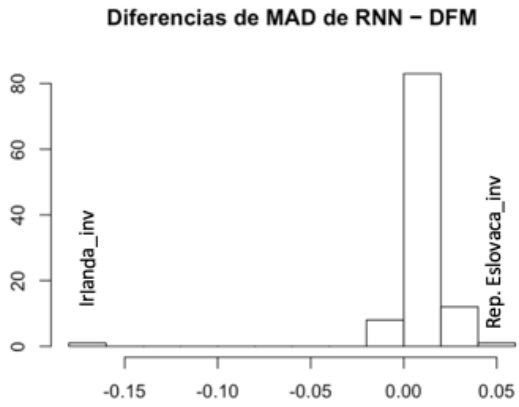


Fuente: Elaboración propia.

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

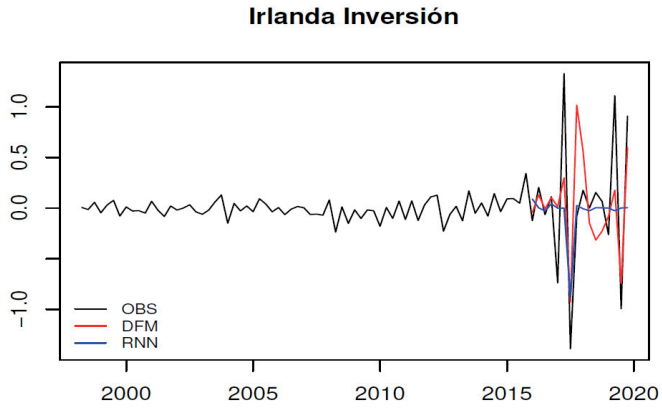
HISTOGRAMA DE LAS DIFERENCIAS ENTRE LA MAD DE LOS ERRORES DE LA RED NEURONAL RECURRENTE Y LA DEL MODELO DE FACTORES



Fuente: Elaboración propia.

FIGURA 15

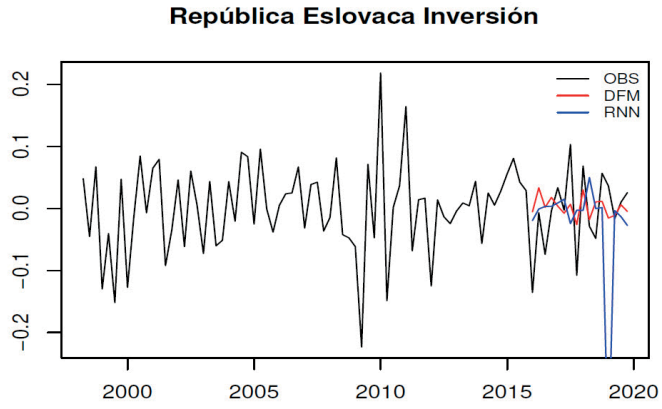
PREDICCIÓN DE LA TASA DE CRECIMIENTO DE LA INVERSIÓN DE IRLANDA (NEGRO) USANDO UN MODELO DE FACTORES (AZUL) Y UNA RED NEURONAL RECURRENTE (ROJO)



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.