

¿QUÉ INDICADORES ECONÓMICOS ADELANTAN LAS RECESIONES EN ESPAÑA? (*)

Máximo CAMACHO

Salvador RAMALLO

Universidad de Murcia

Resumen

En este trabajo usamos árboles de clasificación para examinar cuáles son los mejores indicadores económicos para anticipar las recesiones en la economía española a partir de una muestra de 270 indicadores mensuales desde enero de 1971 hasta el mes de marzo de 2020. La probabilidad de recesión predicha por los modelos a horizontes de tres y seis meses concuerda con las recesiones establecidas por el Comité de Fechado de Ciclos. En un ejercicio que simula la predicción en tiempo real, encontramos que los indicadores financieros y de la construcción cobraron especial protagonismo para predecir la Gran Recesión. Sin embargo, la recesión derivada de la crisis sanitaria de la COVID-19 se infiere mejor con indicadores adelantados de la tendencia del PIB, del desempleo y, en menor medida, de las ventas de coches.

Palabras clave: ciclos económicos, indicadores económicos, árboles de clasificación, *gradient boosting*.

Abstract

We use classification trees to evaluate the usefulness of 270 monthly leading indicators to perform early inferences on business cycle recessions in Spain from 1971.01 to 2020.03. In the in-sample analysis, we find that the indicators give warning signals of recessions 3 and 6 months ahead with significant accuracy. In the pseudo real-time forecasting analysis, we find that financial indicators and indicators for measuring performance of construction played a special role to predict the Great Recession. However, to anticipate the severe economic consequences of the ongoing COVID-19 pandemic are better anticipated with trend indicators of output, with labor market indicators, and, to a lesser extent, with car sales indicators.

Keywords: business cycles, economic indicators, classification trees, gradient boosting.

JEL classification: C22, E32, E27.

I. INTRODUCCIÓN

EL crecimiento económico español, como ocurre en el resto de los países industrializados, no supone un aumento continuado del nivel de actividad económica. Por el contrario, como señala el National Bureau of Economic Research (NBER), el crecimiento económico está caracterizado por una secuencia de expansiones y recesiones. De acuerdo con la visión de esta institución, una recesión, que comienza en un pico y acaba en un valle, implica una caída significativa del nivel de actividad económica común a los sectores económicos, más duradera que unos pocos meses y que normalmente se aprecia en indicadores como el PIB, la renta, el empleo, la producción industrial y las ventas. El NBER define las expansiones de manera simétrica.

La Gran Recesión, que supuso un retroceso generalizado del nivel de actividad económica durante la primera década del siglo XXI, dejó patente el poder devastador que pueden llegar a tener las recesiones en España. Las caídas en la bolsa, la pérdida de confianza de los agentes económicos, los problemas en los balances bancarios, el aumento incontrolado del desempleo y el derrumbe

del precio de la vivienda son solo algunos de los efectos que tuvo esa intensa y prolongada recesión en la economía española (1).

Por este motivo, en los últimos años han aparecido numerosos trabajos que analizan las características del ciclo económico en España. Por un lado, se han propuesto diversos indicadores económicos como media ponderada de indicadores individuales, como en Camacho y Pérez Quirós (2011), Camacho y Doménech (2012), Cuevas y Quilis (2012) y Cuevas, Pérez Quirós y Quilis, (2017). También, se ha analizado recientemente la sincronización de la economía española a nivel internacional, como país, en Camacho, Caro y López-Buenache (2020) y por regiones en Gadea-Rivas, Gómez-Loscós y Bandrés (2017) y Gadea-Rivas, Gómez-Loscós y Leiva-León (2019). Por último, se ha analizado la sincronización de las comunidades autónomas con el ciclo nacional, como en Cancelo (2004), Gadea-Rivas, Gómez-Loscós y Montañés (2012), y Camacho, Pacce y Ulloa (2018).

A pesar de los avances en conocer las características del ciclo económico español y su sincronización a nivel internacional y regional, pensamos

que falta por realizar un análisis exhaustivo de cuáles son los mejores indicadores para anticipar las recesiones en la economía española. Aunque la Asociación Española de Economía decidió crear un Comité de Fechado del Ciclo Económico Español en 2012 encargado de proporcionar un fechado histórico del ciclo de referencia en España, la determinación de los puntos de cambio en las fases del ciclo se hace *a posteriori* y el fechado de los picos y los valles se realiza con un retraso considerable. Conocer los indicadores que anticipan estos cambios de fase del ciclo económico español resultará de gran utilidad para que los agentes económicos tomen sus decisiones de forma óptima.

El marco de trabajo que hemos elegido para realizar este análisis es el de los árboles de clasificación, que se encuentran dentro de los algoritmos de aprendizaje supervisados. Hay varios motivos que justifican esta elección. El primer motivo es porque son una herramienta no paramétrica de clasificación de observaciones muy simple y flexible, permitiendo el uso de datos categóricos y sin que afecten mucho los datos atípicos. El segundo motivo es que permiten operar con bases de datos muy grandes ya que funcionan con algoritmos muy eficientes desde el punto de vista computacional. Estas técnicas son muy estables incluso en casos donde el número de variables explicativas es superior al tamaño muestral. El tercer motivo es debido a que son muy fáciles de interpretar ya que los resultados suelen ir acompañados de representaciones gráficas muy intuitivas. El cuarto motivo es porque proporcionan un marco idóneo para examinar la importancia relativa de los indicadores en la clasificación de las recesiones. Además, los árboles de clasificación permiten examinar la evolución de la importancia relativa en el tiempo de los indicadores por lo que ayudan a discriminar cuáles son los indicadores que mejor predicen cada una de las recesiones.

Recientemente, algunos trabajos han examinado la habilidad de los árboles de clasificación basados en *boosting* para la predicción de recesiones económicas, como Ng (2014) para EE.UU., Döpke, Fritsche y Pierdzioch (2017) para Alemania. Complementariamente, Ward (2017) usa árboles de clasificación que se combinan mediante *random forest* para identificar las crisis financieras internacionales. Piger (2020) realiza una comparativa de ambos (y otros) procedimientos para predecir recesiones en EE.UU. y concluye que la técnica de *boosting* produce mejores resultados. Por esta razón, nosotros usaremos árboles de clasificación que se estiman

usando el algoritmo *stochastic gradient boosting* que se describe en Friedman (2002).

Los principales resultados obtenidos en la aplicación empírica son los siguientes. Primero, la capacidad de los árboles de clasificación para anticipar las recesiones de la economía española a tres y seis meses es muy alta, tanto en el ejercicio dentro de la muestra como en la predicción recursiva. Segundo, en general es destacable la habilidad de los indicadores adelantados de la tendencia del PIB y de las ventas de coches, así como de la serie de paro registrado. Tercero, la habilidad predictiva aumenta cuando se añade la información que contienen los indicadores de confianza, bursátiles y tipos de interés. Cuarto, en un análisis que simula la predicción en tiempo real, encontramos que, para predecir la Gran Recesión, los indicadores financieros y los indicadores adelantados de la evolución de la construcción tomaron especial protagonismo. Sin embargo, se ha producido un cambio en los indicadores mejor posicionados para inferir la recesión derivada de la crisis sanitaria de la COVID-19. En concreto, la mayor importancia relativa aparece en los indicadores adelantados del PIB, en los indicadores relacionados con el paro registrado y, en menor medida, en los indicadores de ventas de coches. Este hecho evidencia el cambio en la naturaleza de la recesión COVID-19 en comparación con la Gran Recesión.

La estructura del trabajo es la siguiente. En la sección segunda, presentamos los árboles de clasificación aplicados a la clasificación de un momento del tiempo en recesión o expansión a partir de la información en un momento previo de un conjunto de indicadores económicos. Para lectores no iniciados en *boosting*, presentamos el algoritmo *AdaBoost* y el (*stochastic*) *gradient boosting*, ambos basados en mejorar iterativamente la clasificación de árboles sencillos poniendo el foco en las observaciones que han sido difíciles de clasificar hasta la iteración anterior. En la sección tercera, se presentan los resultados de la aplicación de estas técnicas a la predicción de las recesiones de la economía española. Las conclusiones se presentan en la última sección.

II. ÁRBOLES DE CLASIFICACIÓN PARA PREDECIR RECESIONES

Esta sección está dedicada a presentar la notación y el funcionamiento de los árboles de clasificación, así como a adaptar esta técnica al análisis del ciclo

económico con series temporales. Para facilitar este propósito, se incluye un ejemplo económico muy sencillo. Los lectores más familiarizados con estas técnicas pueden saltar directamente a la aplicación empírica.

1. Introducción a los árboles de clasificación

Aplicado a la predicción de las fases del ciclo económico con h períodos de adelanto, el problema de clasificación consiste en predecir en cada momento del tiempo t si la economía se encontrará en el grupo de las recesiones o en el de las expansiones en $t + h$, a partir de la observación de K indicadores $x_t = (x_{1t}, \dots, x_{Kt})$, con $t = 1, \dots, T$. La clasificación en los árboles de decisión consiste en particionar el conjunto de los valores posibles de los indicadores en regiones que se clasifican en expansiones o recesiones para t, h tratando de minimizar una función de pérdida. Como resultado, el momento t, h se clasificará en la fase del ciclo de la región a la que x_t pertenece.

El problema de la clasificación se puede ver como un modelo jerárquico de variables latentes. Etiquetemos que el ciclo económico mediante una variable latente no observada s_t que toma valores en el momento t en el conjunto $\{E, R\}$, donde E se refiere a las expansiones y R a las recesiones (2). Llamemos $S = \{s_{h-1}, \dots, s_{T-h}\}$ al conjunto de realizaciones de la variable latente. Por tanto, cuando en el momento t la variable latente toma un valor determinado, sabemos la fase del ciclo económico en el que se encuentra la economía: si $s_t = E$ la economía está en expansión mientras que si $s_t = R$ la economía está en recesión.

Para describir el funcionamiento de los árboles de clasificación, necesitamos precisar algunos conceptos. Llamemos X al conjunto que contiene todos los valores de x_t posibles. En el proceso de clasificación, necesitamos hacer una partición del espacio X en J regiones disjuntas, $\{A_1, \dots, A_J\}$, de tal forma

que $X = \bigcup_{j=1}^J \{A_j\}$. La clasificación consiste en asignar cada vector x_t a una región A_j y a proponer una estimación para la probabilidad de que las observaciones de esa región estén en recesión. La determinación de una región como recesión económica se realizará cuando la probabilidad de recesión de la región supere un determinado umbral (3).

Supongamos que conocemos s_t para todo $t = h+1, \dots, T+h$. Podemos calcular la probabilidad de una recesión para cualquier valor de los indicadores que pertenecen a una región como la proporción de las observaciones de la región que vienen de una recesión. Si definimos $I(\bullet)$ como un indicador que vale 1 si la condición dentro del paréntesis se satisface y 0 en caso contrario, proponemos la estimación de la probabilidad de recesión para una región A_j como

$$p_j^R = \frac{1}{T_j} \sum_{x_t \in A_j} I(s_{t+h} = R), \quad [1]$$

donde T_j es el número de observaciones en la región j . La probabilidad de expansión para cualquier observación de esa región será $p_j^E = 1 - p_j^R$.

Una vez determinada la probabilidad de recesión y expansión de cada región, podemos hacer inferencia sobre la fase del ciclo en la que se encontrarán las observaciones de los indicadores que pertenecen a cada región. Una opción muy extendida en el análisis de los ciclos económicos es presentar la inferencia que hace un modelo sobre la probabilidad de recesión en el momento $t + h$, una vez observado x_t . Con los árboles de clasificación, si conocemos x_t , podemos determinar a qué región pertenece y haremos inferencia sobre la probabilidad de recesión a la que da lugar como $p(s_{t+h} = R | x_t \in A_j) = p_j^R$.

Con esta información, el usuario puede determinar el umbral ϑ a partir del cual clasificará el momento $t + h$ como una recesión a partir del valor de x_t . A la inferencia que hagamos sobre la fase del ciclo una vez conocidos los datos, x_t , y los parámetros que usamos en el modelo, c , la llamaremos $\hat{s}_{t+h}(x_t, c)$ (4). Por tanto, la inferencia será $\hat{s}_{t+h} = R$ si $p(s_{t+h} = R | x_t \in A_j) \leq \vartheta$ mientras que será $\hat{s}_{t+h} = E$ si $p(s_{t+h} = R | x_t \in A_j) > \vartheta$. De esta manera, podemos inferir el valor de las realizaciones de la variable latente $\hat{S} = (\hat{s}_{h+1}, \dots, \hat{s}_{T+h})$. En árboles de clasificación con dos estados posibles, se suele tomar el umbral $\vartheta = 0,5$, es decir, clasificamos la región en la fase del ciclo más probable según las observaciones que pertenecen a esa región.

Este método permite hacer análisis de resultados y predicción de manera sencilla. En primer lugar, los árboles de clasificación nos van a permitir entender las razones por las que el momento del tiempo $t + h$ se ha clasificado como recesión o expansión analizando las características de la región en la que se localiza el vector de observaciones de los indicadores x_t . En segundo lugar, podemos predecir la

probabilidad de recesión para cualquier momento τ , una vez conocido el valor de los indicadores x_{τ} , a partir de la probabilidad de recesión de la región donde pertenezca x_{τ} , $p(s_{\tau+h} = R | x_{\tau} \in A_j) = p_j^R$, incluso si $\tau > T$.

2. Creación del árbol de clasificación

Breiman *et al.* (1984) describen un procedimiento para llevar a cabo las particiones de los árboles de decisión que se conoce como *árboles de clasificación y regresión (CART, por sus siglas en inglés)*. Este procedimiento parte de una regla de clasificación o función definida en X tal que, para cada x_{τ} , indica a qué región pertenece la observación, y en nuestro caso realiza inferencia sobre la asignación de cada región a una fase del ciclo económico.

Supongamos que conocemos el conjunto de realizaciones de la variable latente S . Para la creación de las regiones es necesario identificar una función de pérdida que nos permita comparar unas particiones con otras. Como la inferencia sobre el estado en el que se clasifica una región está en función del porcentaje de observaciones de la región que están en ese estado, las particiones se realizan con el objetivo de encontrar regiones lo menos impuras que sea posible (5). Para realizar las particiones necesitamos una medida del grado de impureza de un árbol y una regla que nos permita realizar particiones.

Una medida que nos permite determinar el grado de impureza de una región A_j es el Índice de Gini (6),

$$G_j = \sum_{s \in E, R} p_j^s (1 - p_j^s) = 2p_j^E p_j^R. \quad [2]$$

El índice toma el valor máximo de 0,5 cuando la distribución de las fases de ciclo en la región es muy impura, es decir, la región contiene el mismo número de recesiones que de expansiones. Sin embargo, el índice toma valores más cercanos a cero cuanto más predominante sea uno de los dos estados en la región. El grado de impureza del árbol se medirá como la media ponderada del grado de impureza de cada una de sus regiones

$$G = \sum_{j=1}^J \frac{T_j}{T} G_j, \quad [3]$$

donde las ponderaciones reflejan el peso de las observaciones de la región sobre el total. Este indi-

cador es una medida de la varianza total del árbol de clasificación.

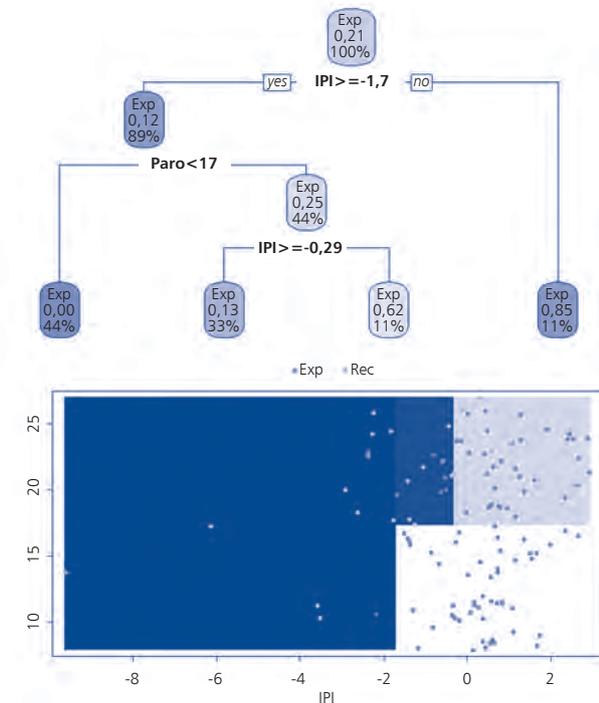
Como no es factible considerar todas las particiones posibles del espacio X , se utiliza el algoritmo *recursive binary splitting*, que se basa en realizar una partición recursiva. Al comienzo del algoritmo, todas las observaciones están en una única región, A , que coincide con todo el espacio X . Las particiones que vamos a considerar son rectangulares y se van a formar a partir de umbrales $c = (c_1, \dots, c_k)$ que compararemos con los valores que pueden tomar los indicadores (x_1, \dots, x_k) (7).

Para realizar la primera partición seleccionaremos un indicador x_k y un umbral c_k^1 que darán lugar a dos regiones rectangulares $A_1 = \{X | x_k < c_k^1\}$ y $A_2 = \{X | x_k \geq c_k^1\}$. Para la primera región, el indicador x_k tomará valores por debajo del umbral c_k^1 . Para la segunda región, el indicador será mayor o igual que el umbral. Para cada indicador x_k y cada umbral c_k^1 calcularemos el índice de Gini y elegiremos el indicador y el umbral que dan lugar a la partición que minimiza el grado de impureza del árbol resultante.

Para la segunda partición, necesitaremos dividir una de las dos regiones anteriores en dos, dando lugar a una segmentación de X en tres regiones. Las dos posibilidades son o bien $A_1^2 = \{X | x_k < c_k^1, x_j < c_j^2\}$, $A_2^2 = \{X | x_k < c_k^1, x_j \geq c_j^2\}$ y $A_3^2 = \{X | x_k \geq c_k^1\}$, o bien $A_1^2 = \{X | x_k < c_k^1\}$, $A_2^2 = \{X | x_k \geq c_k^1, x_j < c_j^2\}$, y $A_3^2 = \{X | x_k \geq c_k^1, x_j \geq c_j^2\}$ (8). De nuevo, para cada indicador x_j y cada umbral c_j^2 calcularemos el índice de Gini en cada una de las dos posibilidades anteriores. Nos quedaremos con aquella partición tal que el nuevo indicador y el nuevo umbral de lugar a una partición con el menor valor del índice.

En este procedimiento recursivo, la partición generada en un paso dará lugar a dos nuevas regiones en el siguiente paso. Este proceso se repite de forma iterativa hasta que se alcanza algún criterio que detiene al algoritmo. La idea es que árboles con muchas particiones presentarán un bajo grado de impureza, pero tendrán poca capacidad de clasificar bien nuevas observaciones. Algunos de los criterios más habituales para limitar el número de particiones son que ninguna región contenga menos de un número mínimo de observaciones, que el espacio X tenga un máximo de particiones o que la realización de una nueva partición no suponga mucha reducción en el índice de Gini del árbol resultante. La opción más habitual es que se parta de un árbol que da lugar a muchas regiones y «se

GRÁFICO 1
EJEMPLO DE ÁRBOL DE CLASIFICACIÓN



Notas: Análisis del árbol de clasificación aplicado a la agregación trimestral de los datos de producción industrial y de paro registrado (EPA). En el gráfico superior, cada nodo muestra la clasificación, el porcentaje de recesiones y el porcentaje de expansiones. En el gráfico inferior las áreas sombreadas son más oscuras cuanto mayor sea la probabilidad estimada de recesión.

Fuente: Elaboración propia a partir de OCDE.

poda» el árbol con el objetivo de minimizar el índice de Gini penalizado por el número de regiones usando procedimientos de valoración cruzada para determinar el grado de penalización.

El gráfico 1 permite entender la idea que hay detrás de los árboles de regresión mediante un ejemplo simplificado, en el que hemos supuesto $h = 0$. El objetivo del ejemplo es crear un árbol de clasificación para determinar si en un momento del tiempo la economía española está en recesión o en expansión. Para determinar la variable de estado del ciclo económico, $S = \{s_1, \dots, s_T\}$, se ha utilizado el ciclo económico de referencia trimestral que determina el Comité de Fechado del Ciclo Económico Español (CFCEE) en el período entre febrero de 1990 y febrero de 2019 (9). Por tanto, $s_t = Rec$ indica que Comité determinó que en el trimestre t la economía española estaba en recesión, mientras que $s_t = Exp$ indicaría que la economía estaba en expansión.

Los indicadores económicos x_t que se han tenido en cuenta para elaborar el árbol de clasificación son la tasa de paro desestacionalizada de la *Encuesta de población activa* y la tasa de crecimiento del índice de producción industrial, que se ha trimestralizado tomando la media del trimestre. Usando estos indicadores y la variable de estado, hemos estimado un árbol de decisión imponiendo cuatro particiones y hemos representado el resultado en el gráfico 1.

En este gráfico se pueden identificar los elementos que integran la terminología habitual de los árboles de decisión. El nodo raíz representa a todo el espacio X y está particionado secuencialmente en nuevas regiones hasta las cuatro regiones finales que aparecen en la parte inferior. Las diversas ramas del árbol surgen a partir del nodo raíz, dos nuevas en cada uno de los nodos de decisión. Por último, los nodos terminales u hojas del árbol representan las cuatro regiones en las que el espacio ha quedado dividido. Para facilitar la interpretación, los nodos informan del porcentaje de recesiones y el porcentaje de expansiones que incluyen las regiones que determinan.

El panel superior del gráfico 1 representa gráficamente la estructura del árbol con sus nodos, ramas y hojas. Al comienzo, el árbol asigna las observaciones a expansión porque solo el 21 por 100 de esta muestra eran recesiones. La primera partición asigna el 11 por 100 de las observaciones que tienen una tasa de crecimiento del IPI por debajo de 1,73 por 100 a la rama derecha y las clasifica como recesión ya que el 85 por 100 de estas observaciones son recesiones. Esta región no se vuelve a particionar. El 89 por 100 de las observaciones cuya producción industrial crece por encima de 1,73 por 100 se asignan a la rama izquierda del árbol y es la rama que sufre nuevas particiones. La probabilidad de recesión para esta región es solo del 12 por 100 ya que ese es el porcentaje de recesiones que contiene.

En esta rama izquierda aparece un segundo nodo que da lugar a una nueva partición que diferencia a los datos con una tasa de paro inferior a 17,31 por 100 de aquellos con una tasa de paro superior a esta cifra. En el primer caso, el árbol acaba en un nodo terminal que se refiere a una región clasificada como expansión económica. En esta región, caracterizada por un crecimiento del IPI superior al 1,73 por 100 y un paro inferior al 17,31 por 100, la probabilidad de recesión es de 0 porque no contiene ninguna ob-

servación de recesiones. En el segundo caso, cuando la tasa de paro está por encima del 17,31 por 100, se realiza una partición adicional, distinguiendo una región de recesiones cuando la producción industrial crezca por debajo de -0,29 por 100 y otra de expansiones cuando la producción industrial crezca por encima de esa cifra.

En el panel inferior del gráfico 1 se representan las cuatro particiones del espacio X a las que ha dado lugar el árbol de clasificación. Cada uno de los puntos representados en el plano hace referencia a las $T = 117$ observaciones de $x_t = (IPI_t, Paro_t)$. Las 93 observaciones que han sido calificadas por el Comité de Fechado de Ciclos como expansión aparecen en color rojo, mientras que las 24 observaciones calificadas como recesión por el Comité aparecen en color azul. La primera región, situada a la izquierda, se caracteriza por el valor que toma la producción industrial, $A_1^3 = \{X | IPI < -1,73\}$, y cualquier observación que pertenezca a ella será calificada por el árbol como una recesión.

La segunda región, situada en la parte inferior derecha, se caracteriza por una combinación de crecimiento de la industria no muy bajo y una tasa de paro baja, $A_2^3 = \{X | IPI < -1,73, Paro < 17,31\}$, y el árbol infiere para ella una expansión. Las dos regiones de la parte superior derecha se caracterizan por un paro elevado (igual o superior a 17,31). La situada más a la izquierda tiene además un crecimiento muy bajo de la producción industrial y se califica como recesión $A_3^3 = \{X | -1,73 < IPI < -0,29, Paro \geq 17,31\}$, mientras que la situada más a la derecha ofrece un crecimiento de la producción industrial no tan bajo $A_4^3 = \{X | IPI \geq -0,29, Paro \geq 17,31\}$ y se califica como expansión. Por tanto, las recesiones en la economía española se caracterizan por una caída brusca de la producción industrial o una caída algo más moderada de la producción industrial combinada con alto desempleo.

Para medir el grado de pureza del árbol podemos usar el índice de Gini. La región A_2^3 es muy pura en el sentido de que solo contiene observaciones que han sido calificadas por el CFCEE como expansiones, por lo que el índice de Gini es mínimo y vale $G_4 = 0$. La pureza de las regiones A_1^3 y A_4^3 también es muy alta: para A_1^3 10 de las 12 observaciones venían de recesiones ($G_1 = 0,28$), mientras que para A_4^3 , 34 de 39 observaciones vienen de expansiones ($G_4 = 0,22$). La región más impura es A_3^3 , ya que solo 8 de 13 observaciones vienen de recesiones

($G_3 = 0,47$). El índice de Gini total, como media ponderada de los anteriores, es $G = 0,16$.

Por último, es interesante medir no solo el grado de pureza del árbol, sino también la capacidad del mismo para formar regiones que no contengan observaciones mal clasificadas. Una manera sencilla de medir la habilidad de clasificación del árbol consiste en comparar el valor de las realizaciones de la variable latente $S = \{s_1, \dots, s_T\}$ con su inferencia usando el árbol de clasificación $\hat{S} = \{\hat{s}_1, \dots, \hat{s}_T\}$, de tal modo que $\hat{s}_1, \dots, \hat{s}_T(x_t | x_t \in A_m)$ es la clasificación que se hace para los valores de los indicadores que pertenecen a la región A_m (10).

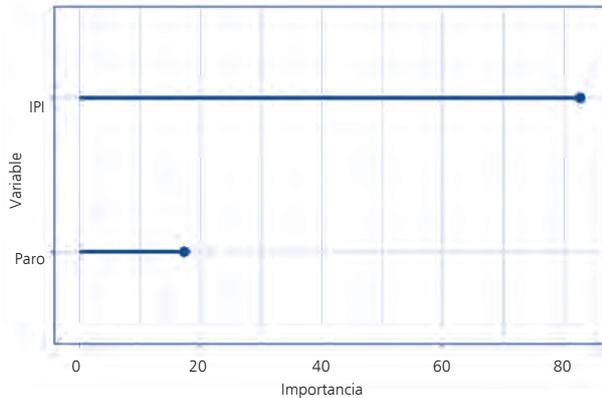
Así, la región A_1^3 , para la que se infiere recesión $\{\hat{s}_t = Rec | x_t \in A_1^3\}$ porque la mayoría de los momentos del tiempo t (con valores de los indicadores x_t) en esa región fueron determinados por el Comité como recesiones económicas, contiene dos observaciones de expansiones. En la región A_2^3 se infieren expansiones, $\{\hat{s}_t = Exp | x_t \in A_2^3\}$, y no contiene ninguna observación de recesiones. En la región A_3^3 se clasifica como recesión, $\{\hat{s}_t = Rec | x_t \in A_3^3\}$, aunque aparecen cinco observaciones de expansiones. Por último, la región A_4^3 contiene cinco observaciones de recesión cuando es una región de expansiones, $\{\hat{s}_t = Exp | x_t \in A_4^3\}$. Por tanto, el porcentaje de observaciones mal clasificadas por el árbol sobre el total de observaciones es del 10 por 100.

El gráfico resulta muy intuitivo porque se han sombreado las regiones resultantes en función de la probabilidad de recesión de las observaciones que se encuentren en cada una de estas áreas. Las áreas que aparecen más oscuras indican una mayor probabilidad de recesión. Por tanto, de manera muy visual podemos determinar visualmente la probabilidad de recesión para cualquier combinación de IPI y de paro.

Por último, Breiman *et al.* (1983) proponen una medida muy útil para seleccionar la importancia relativa de los indicadores para clasificar las observaciones entre expansiones y recesiones. La importancia relativa del indicador x_k en la clasificación final viene determinada por el número de veces que ese indicador se ha utilizado para realizar las particiones ponderadas por la reducción en el índice de Gini que proporcionan las particiones en las que participa.

Supongamos que en un árbol se hacen un total de P subparticiones en cada nodo que no sea ter-

GRÁFICO 2
EJEMPLO DE IMPORTANCIA RELATIVA



Notas: Las importancias relativas se han escalado para que sumen 100. Las variables son agregación trimestral de los datos de producción industrial y de paro registrado de la EPA.
Fuente: Elaboración propia a partir de OCDE.

minal y que el indicador que se usa para hacer la partición es v_p , con $p = 1, \dots, P$. Si llamamos ΔG_p a la reducción en el índice de Gini que se produce en esa partición, la importancia relativa del indicador x_k en la clasificación es

$$I_k = \sum_p^P \Delta G_p I(v_p = x_k). \quad [4],$$

donde $k = 1, \dots, K$. Para facilitar la interpretación, el indicador se suele normalizar para que las importancias relativas sumen 100 (11). El gráfico 2 muestra que la importancia relativa de la producción industrial para realizar la clasificación es mucho mayor que la del paro, ya que sus importancias relativas son 83 y 17, respectivamente.

3. Evaluación de los resultados

Para conocer la verdadera capacidad de clasificación del árbol debemos examinar hasta qué punto es capaz de clasificar 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 el período muestral en dos subperíodos. El primer subperíodo es el de entrenamiento, para el que supondremos conocida la clasificación $S_{T_1} = \{s_{h+1}, \dots, s_{T_1+h}\}$ y los datos de los indicadores económicos x_t para $t = 1, \dots, T_1$, con $T_1 < T$ (12).

El segundo subperíodo es el de evaluación, en el que se examina la habilidad de clasificación del modelo para las observaciones que no se han usado para generar el árbol. En este caso, se va a suponer desconocida la clasificación $S_{T_2} = \{s_{T_1h+1}, \dots, s_T\}$ y usaremos el árbol estimado con observaciones hasta T_1 para hacer una clasificación de las observaciones del período de evaluación $\hat{S}_{T_2} = \{\hat{s}_{T_1h+1}, \dots, \hat{s}_T\}$ a partir de los indicadores x_t para $t = T_1+1, \dots, T$. La comparativa entre S_{T_2} y \hat{S}_{T_2} nos informará de la verdadera capacidad del árbol para clasificar las observaciones.

4. Árboles de clasificación y boosting

En las aplicaciones de los árboles de clasificación se han detectado dos problemas. El primero es que los árboles presentan en numerosas ocasiones una habilidad de clasificación limitada en el período de evaluación. El segundo es que suelen dar lugar a resultados poco robustos, en el sentido de que pequeños cambios en los datos producen grandes modificaciones en el árbol de clasificación estimado. Para paliar estos problemas, una de las técnicas más extendidas es el *boosting*, que comenzó con un conjunto de técnicas de clasificación binaria propuestas por Freund y Schapire (1996), basadas en los trabajos anteriores de Schapire (1990) y Freund (1995), y que se conocen como algoritmo *AdaBoost* (*adaptive boosting*) (13).

La idea que está detrás de la aplicación del *boosting* a los árboles de clasificación consiste en crear secuencialmente árboles en los que cada árbol nuevo supone una modificación del anterior, de tal manera que en cada nueva iteración el algoritmo va aprendiendo de los errores cometidos hasta la iteración anterior. En el algoritmo se emplean árboles con pocas particiones por lo que son muy impuros, aunque con alta capacidad de clasificar nuevas observaciones, pero al ir ajustando secuencialmente los árboles se va reduciendo el grado de impuridad sin perder habilidad de clasificación. Esto se consigue porque en cada iteración el algoritmo centra en predecir correctamente las observaciones que las iteraciones anteriores no han sido capaces de hacerlo.

4.1. Discrete AdaBoost

Para aplicar el *AdaBoost* en árboles de clasificación, Ng (2014) describe la siguiente versión del algoritmo conocida como *discrete AdaBoost*, que es uno de los algoritmos *AdaBoost* más usuales en ár-

boles de clasificación. Por conveniencia, estas técnicas codifican la variable de estado s_{t+h} en el conjunto $\{-1, 1\}$, donde $s_{t+h} = -1$ hace referencia a las expansiones y $s_{t+h} = 1$ a las recesiones con $t = 1, \dots, T-h$. El algoritmo se puede resumir en los siguientes pasos (14):

1. Partimos de unos pesos iguales para todos los valores de los indicadores $w_t^1 = \frac{1}{T-h}$. Definamos la clasificación hasta la iteración j como S_{t+h}^j , con $t = 1, \dots, T-h$, a la que le asignamos un valor inicial neutro de $S_{t+h}^j = 0$ para todo $t = 1, \dots, T-h$.
2. Para un conjunto de iteraciones $m = 1, \dots, M$, se actualizan las clasificaciones realizadas hasta la iteración anterior \hat{S}_{t+h}^{m-1} con las de la iteración actual \hat{S}_{t+h}^m ajustadas por un factor corrector α^m

$$\hat{S}_{t+h}^m = \hat{S}_{t+h}^{m-1} + \alpha^m \hat{S}_{t+h}^m \quad [5]$$

El corrector α^m se define en el algoritmo, mediante la iteración de los siguientes pasos:

- 2.1. En la iteración m estimamos un árbol de decisión con pocas particiones que darán lugar a una clasificación \hat{S}_{t+h}^m , con $t = 1, \dots, T-h$, de tal manera que se minimice el error que se comete con este árbol. En este caso, se define el error como la suma ponderada de las veces que las observaciones están mal clasificadas,

$$e^m = \sum_{t=1}^{T-h} w_{t+h}^m I(s_{t+h} \neq \hat{S}_{t+h}^m) \quad [6]$$

El algoritmo solo actualiza \hat{S}_{t+h}^m si $e^m < 0,5$, ya que $e^m = 0,5$ sería una clasificación aleatoria (15).

- 2.2. En tal caso, calculamos el peso que le daremos a la clasificación que obtenemos con este árbol en la clasificación final mediante una función que decrece con el error cometido

$$\alpha^m = \frac{1}{2} \ln \left(\frac{1-e^m}{e^m} \right) \quad [7]$$

Si el error tiende al máximo de 0,5, α^m tiende a cero y ese árbol que genera las clasificaciones \hat{S}_{t+h}^m no tendrá casi influencia en la clasificación final. La ponderación del árbol estimado en el paso m será tanto mayor cuanto menor sea el error de clasificación del ese árbol.

- 2.3. Actualizamos el peso que le daremos a las observaciones en la siguiente iteración de la siguiente manera

$$w_{t+h}^{m+1} = \frac{w_{t+h}^m}{2\sqrt{e^m(1-e^m)}} \exp(-\alpha^m s_{t+h} \hat{S}_{t+h}^m) \quad [8]$$

donde se cumple que $\sum_{t=1}^{T-h} w_{t+h}^{m+1} = 1$. Cuando el error de clasificación del árbol se acerca al error máximo de 0,5, entonces $2\sqrt{e^m(1-e^m)}$ tiende a 1 y α^m tiende a cero. Esto implica que $w_{t+h}^{m+1} \approx w_{t+h}^m$ y los pesos de las observaciones que se usarán en la siguiente iteración casi no se actualizan con ese árbol. El algoritmo otorgará más importancia en la iteración $m + 1$ a las observaciones que no se han clasificado bien en la iteración m . Si la observación $t + h$ está bien clasificada, el signo de $s_{t+h} \hat{S}_{t+h}^m$ será positivo y el peso se multiplicará por un número menor que 1, que será menor cuanto menor sea el error del árbol. En caso contrario, el peso de esa observación mal clasificada se multiplicará por un número mayor que 1, que será mayor cuanto menor sea el error del árbol.

3. La clasificación final estará en función de la media de las clasificaciones de los árboles construidos en el proceso ponderadas de forma decreciente por el error que han cometido

$$\hat{S}_{t+h} = \text{sign} \left(\hat{S}_{t+h}^M = \sum_{m=1}^M \alpha_m \hat{S}_{t+h}^m \right) \quad [9]$$

donde $\text{sign}(z) = 1$ si $z > 0$ $\text{sign}(z) = -1$ si $z < 0$. Por tanto, podemos interpretar que la clasificación de una observación en t se determinará por lo que decida la mayoría cualificada de clasificaciones en las iteraciones. Si la mayoría de veces hemos clasificado a la observación $t + h$ como una expansión ($\hat{S}_{t+h}^m = -1$) y hemos acertado (α^m grandes) el signo que predominará para la media ponderada será el negativo y, por tanto, quedará finalmente calificado como expansión ya que $\hat{S}_{t+h}^m = -1$.

En las aplicaciones de este algoritmo se ha observado que la capacidad de clasificación se mantiene más o menos constante para valores de M relativamente pequeños cuando los árboles usados para generar \hat{S}_{t+h}^m en el paso 2.1 son el resultado de minimizar el error de clasificación de árboles de solo dos regiones (16). Para árboles más complejos que generan más regiones, el algoritmo tiende a

converger mucho antes, aunque a costa de incrementar la complejidad computacional. En cualquier caso, podemos determinar el error de clasificación que se obtiene en cada iteración del algoritmo *discrete AdaBoost* hasta un valor M^* grande y obtener el número M de árboles óptimo.

Friedman, Hastie y Tibshirani (2000), proporcionan argumentos estadísticos para entender las razones por las que los algoritmos *AdaBoost* daban tan buenos resultados empíricos. En esta influyente contribución, proponen que es más sencillo entender el proceso de clasificación de las observaciones entre recesiones y expansiones como la búsqueda del clasificador S_t que minimiza la esperanza de una función de pérdida exponencial (17):

$$E(FP(S_{t+h})) = E\left[\exp(-s_{t+h}S_{t+h})\right] \quad [10]$$

Cuando S_{t+h} clasifica bien, los signos de s_{t+h} y S_{t+h} coinciden y la función de pérdida es pequeña. Sin embargo, cuando S_{t+h} clasifica mal, la función de pérdida es grande.

Derivando la expresión anterior e igualando a cero, se puede demostrar que el valor del clasificador S_{t+h}^* que minimiza la anterior esperanza es

$$S_{t+h}^* = \frac{1}{2} \ln \left[\frac{p(s_t = 1 | x_t)}{p(s_t = -1 | x_t)} \right] \quad [11]$$

que coincide con la mitad del logaritmo del *odds ratio* o cociente entre la probabilidad de que ocurra una recesión y la probabilidad de que ocurra una expansión. Por similitud con los modelos logísticos, la probabilidad condicionada de que ocurra una recesión es

$$p(s_{t+h} = 1 | x_t) = \frac{\exp(2S_{t+h})}{1 + \exp(2S_{t+h})} = \frac{\exp(S_{t+h})}{\exp(-S_{t+h}) + \exp(S_{t+h})} \quad [12]$$

En los modelos logísticos, el logaritmo del *odds ratio* se estima como una suma ponderada de las observaciones por los coeficientes de la regresión. Friedman *et al.* (2000) proponen estimar S_{t+h}^* mediante una aproximación no paramétrica basada en modelos de regresión aditiva:

$$\hat{S}_{t+h}^M = \sum_{m=1}^M \alpha_m \hat{s}_{t+h}^m(c_m) \quad [13]$$

donde se ha hecho explícito que el clasificador, \hat{s}_{t+h}^m , depende de los umbrales, c^m , que se necesitan estimar para realizar las particiones.

El problema de optimización implica encontrar numéricamente el valor \hat{S}_{t+h}^M que minimice la función de coste exponencial [10], lo que implica encontrar la secuencia de pesos y umbrales óptimos $\{\alpha_m, c_m\}_{m=1}^M$. A partir de un valor dado de \hat{S}_{t+h}^{m-1} , los algoritmos de optimización secuencial aplicados en modelos de regresión aditiva se basan en encontrar secuencialmente la pareja $\{\alpha_m, c_m\}$ que minimiza

$$\{\alpha_m, c_m\} = \arg \text{Min} \sum_{m=1}^M \exp(-s_{t+h}(\hat{S}_{t+h}^{m-1} + \alpha_m \hat{s}_{t+h}^m(c_m))) \quad [14]$$

Friedman *et al.* (2000) muestran que el clasificador óptimo, $\hat{s}_{t+h}^m(c_m)$, es el que minimiza el error ponderado que aparece en la expresión [6], que las ponderaciones que minimizan la función de pérdida exponencial, α_m , coinciden con las que aparecen en [7] y que la actualización óptima de los pesos de los errores es la proporcionada por [8] (18).

4.2. (Stochastic) gradient boosting

Friedman (2001) propone un algoritmo conocido como *gradient boosting* que generaliza los algoritmos *AdaBoost* propuestos por Friedman *et al.* (2000). Recordemos que *AdaBoost* se centra en aplicar árboles sencillos a observaciones ponderadas en función del error cometido hasta la iteración anterior. Por el contrario, *gradient boosting* usa el algoritmo *functional gradient descent* para encontrar iterativamente el mínimo de una función de forma que, en cada iteración, se busca la dirección y el tamaño de actualización encaminada a optimizar la reducción del valor de la función. En concreto, en cada iteración la función se actualiza en proporción al negativo del gradiente de la función de pérdida que marca la dirección de la actualización. Por esta razón, en la práctica, *gradient boosting* aplica árboles sencillos para tratar de aproximar el negativo del gradiente de la función de pérdida evaluada en la clasificación de la iteración anterior.

Si usamos una función de pérdida exponencial como en [10], el objetivo es encontrar S_t que minimice la función de pérdida mediante una aproximación no paramétrica basada en modelos de regresión aditiva \hat{S}_{t+h}^M como en [13]. Siguiendo a Schapire y Freund (2012), el algoritmo consiste en la aplicación iterativa de dos pasos. Partiendo de una clasificación \hat{S}_{t+h}^{m-1} , el primer paso en cada iteración m debemos calcular la dirección de la actualización que viene marcada por el negativo de

la derivada de la función de pérdida exponencial evaluada en \hat{S}_{t+h}^{m-1}

$$g_{t+h} = - \left. \frac{\partial FP(S_{t+h})}{\partial S_{t+h}} \right|_{S_{t+h} = \hat{S}_{t+h}^{m-1}} = s_{t+h} \exp(-s_{t+h} \hat{S}_{t+h}^{m-1}) \quad [15]$$

Los valores de esta expresión se conocen como pseudoresiduos y serán mayores en las observaciones mal clasificadas. Omitiendo la normalización, podemos encontrar el árbol $\hat{S}_{t+h}^m(c_m)$ que mejor se ajusta a los pseudoresiduos como el que maximiza la función

$$PE = \frac{1}{T-h} \sum_{t=1}^{T-h} s_{t+h} \hat{S}_{t+h}^m(c_m) \exp(-s_{t+h} \hat{S}_{t+h}^{m-1}) \quad [16]$$

Maximizar esta función es equivalente a minimizar el error [6] en *AdaBoost* (19).

Gradient boosting también pone énfasis en clasificar bien en la iteración m las observaciones que han sido mal clasificadas en la iteración anterior. Una observación mal clasificada en la iteración $m-1$ supondrá un valor alto para $\exp(-s_{t+h} \hat{S}_{t+h}^{m-1})$, y si volviera a clasificarse mal el signo de $s_{t+h} \hat{S}_{t+h}^m(c_m)$ sería negativo por lo que implicaría un lastre para maximizar [16].

Una vez estimada la dirección de la mejora con \hat{S}_{t+h}^m , el segundo paso consiste en determinar el tamaño del ajuste. En concreto, se elige α_m como el valor que minimiza la función de pérdida exponencial y sabemos por Friedman *et al.* (2000) que coincide con la expresión [7]. Alternativamente, Friedman (2001) propone que, usando el criterio de minimización de la función de pérdida, se podría encontrar un peso distinto α_{mj} para cada región j del árbol estimado en la iteración m . Otra opción es suponer que el peso de los árboles en todas las iteraciones es una constante, $\alpha_m = \alpha$ para todo m .

Por último, Friedman (2001), incorpora un parámetro de ajuste que controla el tamaño de los saltos que se dan en el algoritmo

$$\hat{S}_{t+h}^m = \hat{S}_{t+h}^{m-1} + \eta \alpha_m \hat{S}_{t+h}^m(c_m) \quad [17]$$

donde $0 < \eta \leq 1$. 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 $\eta \geq 0,1$.

Finalmente, Friedman (2002) incorporara un elemento adicional en el algoritmo *stochastic gradient boosting*: el muestreo aleatorio de un porcentaje v de observaciones que forman parte del período de entrenamiento. En concreto, en cada iteración del algoritmo, el nuevo árbol de decisión se ajusta empleando únicamente una fracción θ de datos del período de entrenamiento, extraída de forma aleatoria y sin reemplazo. Este procedimiento mejora la capacidad de clasificación y agiliza la computación. En la práctica, para reducir tiempo de computación, θ será menor cuanto mayor sea el número de indicadores disponibles.

5. Importancia relativa y efecto interacción

Con el objetivo de que los árboles de clasificación se puedan usar para seleccionar de manera automática los indicadores más influyentes a la hora de anticipar el ciclo económico, Friedman (2001) propone una medida de la importancia relativa en el algoritmo *boosting*. Llamemos I_k^m a la importancia relativa del indicador x_k en cada árbol m de los M que se generan en el *boosting*, que se calcula usando la expresión [4]. La importancia relativa del indicador x_k en la clasificación obtenida con el *boosting* será la media de su importancia relativa en los árboles generados:

$$IB_k = \frac{1}{M} \sum_p I_k^m \quad [18]$$

donde $k = 1, \dots, K$ (20). Los indicadores económicos que nunca se usen para realizar particiones serán descartados como variables indicadoras del ciclo económico.

Por último, es importante disponer de alguna herramienta que permita esclarecer la caja negra que podría parecer el resultado de los árboles de decisión estimados con *boosting*. Para examinar cuál es el papel que juegan los indicadores en la construcción del árbol, Friedman (2001) propone usar los gráficos de dependencia parcial. Estos gráficos muestran el valor que toma el resultado del árbol de clasificación \hat{S}_{t+h}^M para distintos valores de algunos indicadores económicos (21). Recordemos que \hat{S}_{t+h}^M es una estimación de la mitad logaritmo del *odds ratio* y que, por tanto, un mayor valor en el eje de ordenadas está relacionado con una mayor probabilidad de recesión. Por tanto, estos gráficos nos van a ayudar a examinar si la relación entre la probabilidad de recesión y los indicadores económicos es nula, lineal, o más compleja.

Sin pérdida de generalidad, supongamos que queremos medir el efecto del primer indicador x_1 sobre $\hat{S}_{t+h}^M(x_1, x_{2t}, \dots, x_{Kt})$. El primer paso, consiste en construir el árbol de decisión usando *boosting* para la base de datos original. En el segundo paso, sustituimos el valor de x_1 por una sucesión de valores posibles de del indicador $x_{1i} = \{x_{11}, x_{12}, \dots, x_{1N}\}$. En el tercer paso, estimamos $\hat{S}_{t+h}^M(x_{1i}, x_{2t}, \dots, x_{Kt})$ para todo $t=1, \dots, T-h$ y para cada $i=\{1, \dots, N\}$. Finalmente, calculamos la media

$$\bar{\hat{S}}_{t+h}^M(x_{1i}) = \frac{1}{T-h} \sum_{t=1}^{T-h} \hat{S}_{t+h}^M(x_{1i}, x_{2t}, \dots, x_{Kt}) \quad [19]$$

con $i=\{1, \dots, N\}$. El gráfico de $\{x_{1i}, \bar{\hat{S}}_{t+h}^M(x_{1i})\}$ para $i=1, \dots, N$ es el gráfico de dependencia parcial para x_1 .

Usando técnicas de *bootstrap* se pueden estimar intervalos de confianza para los gráficos de dependencia parcial. Para ello, se pueden construir aleatoriamente un número B submuestras que incluyan un porcentaje μ de los datos originales. Para cada una de esas muestras se calcula la respuesta media $\hat{S}_{t+h}^{Mb}(x_{1i})$ con $b=1, \dots, B$. Si trabajamos a un nivel de confianza del λ por 100, la dependencia parcial será la mediana de las observaciones y aparece en el gráfico junto a los cuantiles $0,5\lambda$ por 100 y $(100-0,5\lambda)$ por 100.

De forma sencilla, se puede generalizar el gráfico de dependencia parcial para cualquier grupo de indicadores de x_i . Aunque gráficos de combinaciones complejas de indicadores son difíciles de interpretar, la combinación más frecuente es la que mide el efecto interacción entre dos indicadores a y b , se suele representar gráficamente la secuencia $\{x_{ai}, x_{bi}, \bar{\hat{S}}_{t+h}^M(x_{ai}, x_{bi})\}$ para $i=\{1, \dots, N\}$.

III. APLICACIÓN EMPÍRICA

Para el análisis empírico hemos seleccionado una muestra de 270 indicadores económicos mensuales. Después de ajustar las observaciones, el período muestral efectivo va desde enero de 1971 hasta marzo de 2020. Sin embargo, muchos de los indicadores económicos españoles se empiezan a construir desde fechas más recientes, lo que impide hacer el análisis con todos los indicadores para todo el período muestral completo. Por eso, hemos dividido el análisis en cuatro estudios parciales, que usan respectivamente indicadores que empiezan en

1971, en 1978, en 1988 y en 2004. Nos referiremos a estos cuatro estudios como 71s, 78s, 88s y 04s, respectivamente.

Los indicadores económicos se han obtenido de la base de datos de la OCDE (*Main Economic Indicators, MEI*), los tipos de interés del Fondo Monetario Internacional (*International Financial Statistics, IFS*), y los datos de agregados monetarios de Bloomberg. Cuando ha sido necesario, los indicadores se han transformado convenientemente para trabajar con sus versiones estacionarias. El número de indicadores que se usan para llevar a cabo estos análisis es creciente. En el análisis 71s, se usan 36 indicadores entre los que se encuentran los indicadores adelantados de la OCDE relacionados con producción, ventas de coches, capacidad de utilización, empleo, precios, tipos de cambio y comercio (22).

En el análisis 78s se incorporan 39 nuevos indicadores de la OCDE relacionados con la construcción y tipos de interés, además de otras series de producción, empleo y precios. En la ampliación que da lugar a 88s, se incorporan 58 nuevos indicadores de confianza del consumidor, bursátiles y tipos de interés, además de nuevos indicadores de producción, precios, comercio y empleo. La mayor ampliación de indicadores aparece en la ampliación de 04s con 137 indicadores nuevos, sobre todo de empleo, precios y tipos de interés. La importancia del análisis realizado con estos 270 indicadores es que podemos hacer inferencia sobre la probabilidad de recesión durante la Gran Recesión y de la recesión provocada por la crisis sanitaria de la COVID-19 con una amplia batería de indicadores.

Para establecer los períodos de recesión hemos usado el fechado del ciclo económico de referencia español que proporciona el Comité de Fechado del Ciclo Económico Español creado por la Asociación Española de Economía en 2012. En la muestra utilizada, el Comité clasifica cinco períodos de recesión que suponen el 13 por 100 de las observaciones de esta muestra. En los gráficos de esta sección los períodos de recesión aparecerán como áreas sombreadas para facilitar la inspección visual de los resultados.

El análisis de la capacidad predictiva de los indicadores se ha hecho a horizontes de predicción h de tres y seis meses. La estimación se realiza usando el paquete GBM de R. Hemos usado la función de pérdida «adaboost». El número máximo de árboles permitido es de 2000 y se ha elegido el óptimo M por validación cruzada (*five-fold cross validation*).

El número mínimo de observaciones por nodo es de cinco y en los árboles se construyen seis regiones finales para permitir efecto interacción entre los indicadores. El parámetro de ajuste de *gradient boosting* es $\eta = 0,005$, mientras que en *stochastic gradient boosting* la fracción θ de datos del período de entrenamiento es del 50 por 100. Para hacer inferencia en los gráficos de dependencia parcial hemos usado un total de $B = 1000$ replicaciones *bootstrap*.

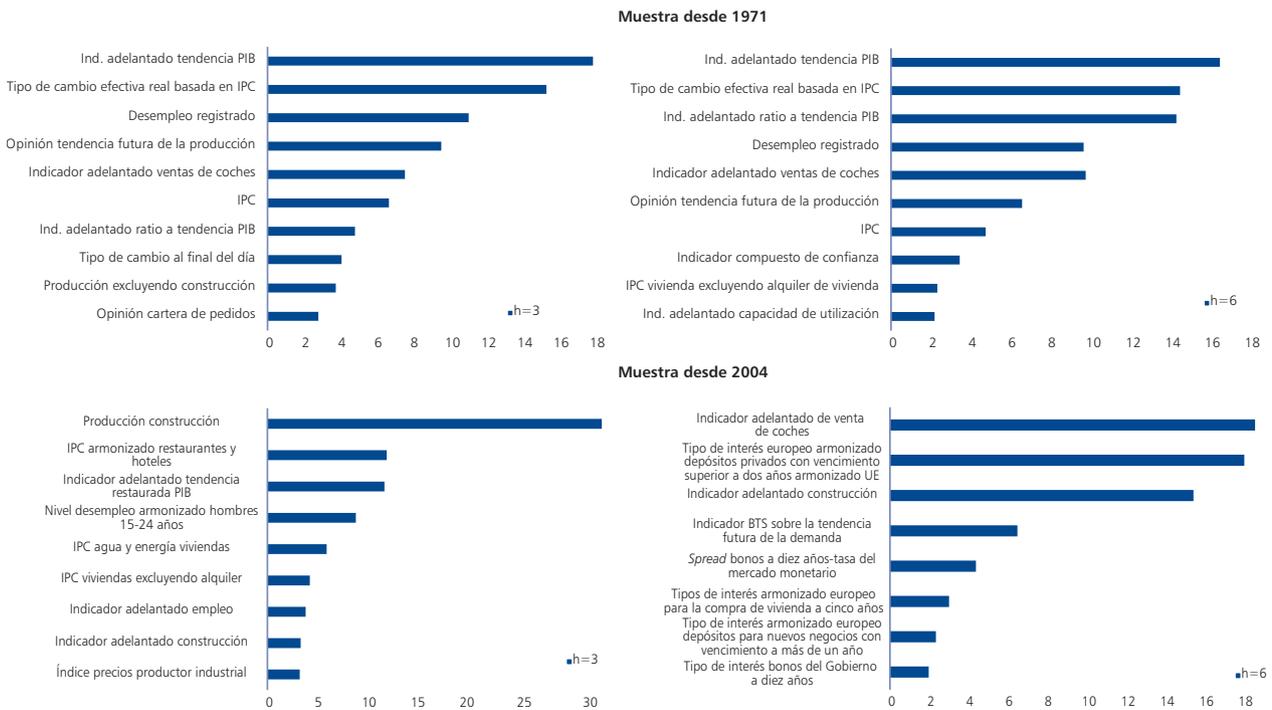
1. Inferencia dentro de la muestra

La importancia relativa de los indicadores económicos se encuentra en el gráfico 3. Como ilustración, vamos a centrarnos en los casos que incluyen más período muestral (71s) y más indicadores económicos (04s). Para facilitar la interpretación, las importancias se han escalado para que sumen 100 y se muestran solo los 10 indicadores con mayor importancia relativa. Usando el análisis 71s, los principales indicadores son los de la OCDE que adelantan la evolución del PIB, el desempleo y las

ventas de coches. Si nos centramos en el análisis 04s, que incluyen las dos recesiones que se conocen como la Gran Recesión y la caída en la actividad económica provocada por la crisis sanitaria de la COVID-19, cobran protagonismo la producción y el indicador adelantado de la construcción, índices de precios, e indicadores financieros. Sin embargo, las ventas de coches siguen siendo importantes.

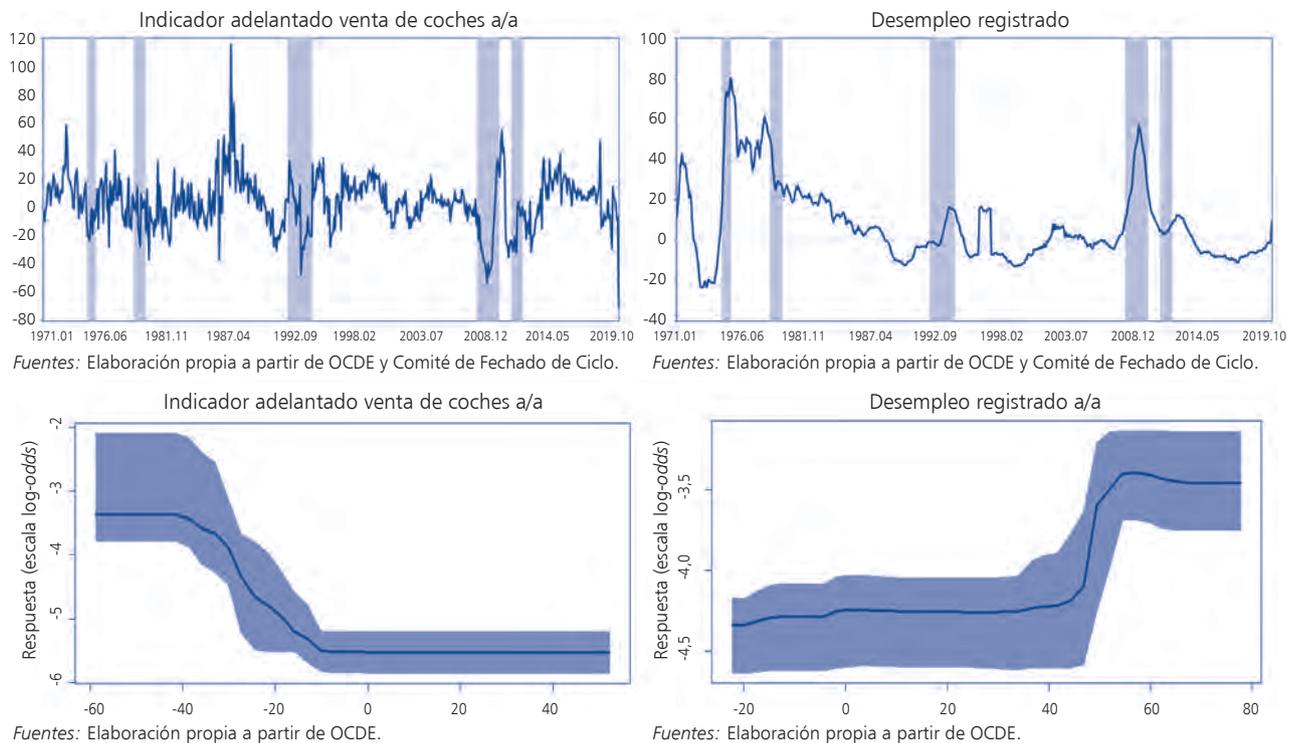
Usando el criterio de la importancia relativa del análisis 70s, el gráfico 4 muestra la evolución (gráficos superiores) y la dependencia parcial (gráficos inferiores) de dos de los indicadores económicos más importantes: las tasas de crecimiento interanuales del indicador adelantados de la venta de coches y de desempleo registrado. Para facilitar la interpretación de los resultados, los gráficos superiores muestran áreas sombreadas que marcan los períodos que el Comité de Fechado de Ciclos clasifica como recesiones junto a la evolución de los indicadores. En los gráficos inferiores, junto a la dependencia parcial se incluyen los intervalos de confianza al 95 por 100 como áreas sombreadas.

GRÁFICO 3
IMPORTANCIA RELATIVA DENTRO DE LA MUESTRA



Notas: Las importancias relativas se han escalado para que sumen 100.
Fuente: Elaboración propia a partir de OCDE.

GRÁFICO 4
DEPENDENCIA PARCIAL

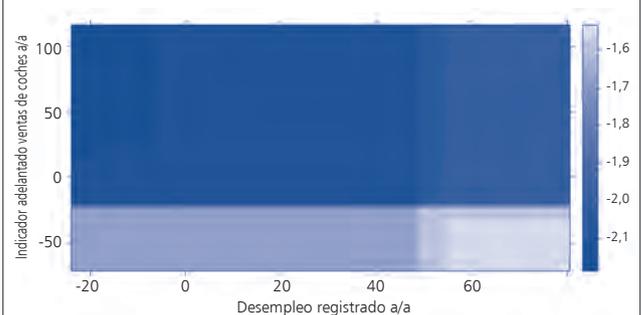


Notas: Gráficos superiores: tasas de crecimiento interanuales del indicador adelantado de la venta de coches y del desempleo registrado; áreas sombreadas son recesiones según Comité de Fechado de Ciclos. Gráficos inferiores: efecto sobre el logaritmo del *odds ratio* en seis meses; áreas sombreadas muestran los intervalos de confianza al 95%. Período muestral: 1971.01-2020.03.

En los gráficos de dependencia parcial, el valor de los indicadores económicos se muestra en el eje de abscisas mientras que la mitad de los logaritmos del *odds ratio* a horizonte de predicción de seis meses para esos valores y sus respectivos intervalos de confianza se muestran en el eje de ordenadas. Como cabría esperar, la relación entre el valor de los indicadores y la probabilidad de recesión es negativa en el caso del indicador adelantado de la venta de coches, mientras que la relación es positiva en el caso del indicador del desempleo. En todos los casos se muestra un claro efecto no lineal de los indicadores en la probabilidad de recesión estimada. En el caso del indicador de la venta de coches, el gráfico inferior izquierdo muestra una caída de la probabilidad de recesión (medida en escala del *odds ratio*) abrupta cuando la tasa interanual del indicador cae por debajo del 20 por 100. En gráfico superior izquierdo muestra que esas tasas negativas del indicador coinciden con las recesiones del Comité. En el caso del desempleo, la subida abrupta de la probabilidad de recesión aparece para tasas cercanas al 40 por 100 en el gráfico

inferior derecho. En el gráfico superior derecho se muestra que esos valores presentan una correlación positiva con las recesiones del Comité.

GRÁFICO 5
EFECTO INTERACCIÓN



Notas: Efecto conjunto sobre el logaritmo del *odds ratio* en seis meses de las tasas de crecimiento interanuales del indicador adelantado de venta de coche y el aumento interanual del desempleo registrado.

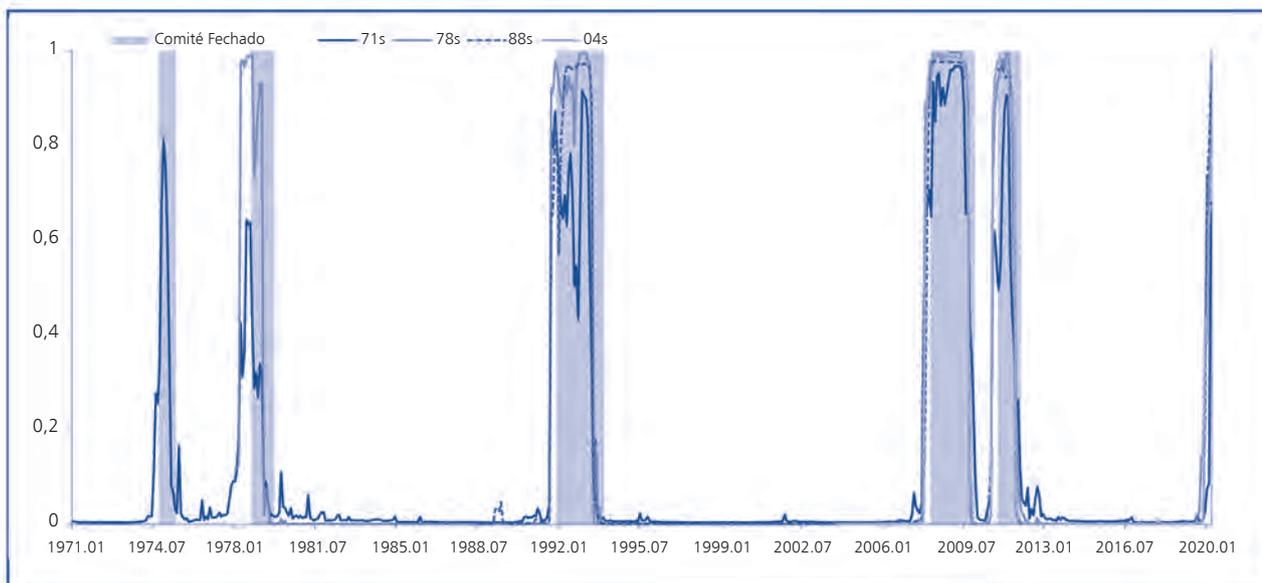
Fuente: Elaboración propia a partir de OCDE.

Para ilustrar la importancia del efecto de interacción entre los indicadores económicos, el gráfico 5 muestra el efecto sobre el logaritmo del *odss ratio* (escala derecha) a horizonte de predicción de seis

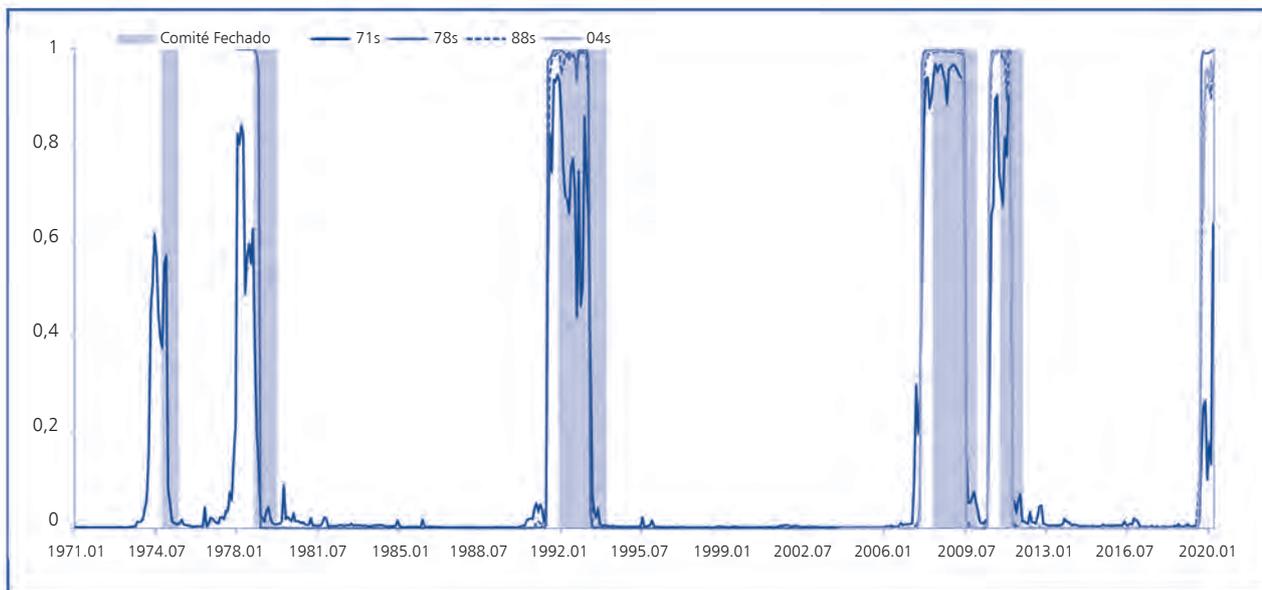
meses los valores conjuntos que toman las tasas de crecimiento del indicador adelantado de venta de coches (eje de ordenadas) y del desempleo registrado (eje de abscisas). La interpretación del grá-

GRÁFICO 6
PROBABILIDADES DE RECESIÓN DENTRO DE LA MUESTRA

Predicción a 3 meses



Predicción a 6 meses



Notas: Estimación con muestras de indicadores que comienzan en 1971, 1978, 1988 y 2004. Las áreas sombreadas identifican a las recesiones del Comité de Fechado de Ciclos.

Fuentes: Elaboración propia a partir de OCDE y Comité de Fechado de Ciclo.

fico es muy sencilla: cuanto más clara es el área sombreada, mayor es la probabilidad de recesión (medida en escala del *odds ratio*) de la combinación de valores del indicador adelantado y del empleo. Para un valor dado de la evolución del desempleo, el color del área se hace más claro cuando el valor del indicador adelantado de la venta de coches baja, mientras que, para un valor dado de la variación interanual del indicador adelantado de la venta de coches, el área es más clara cuanto mayor sea el crecimiento del desempleo. Sin embargo, la combinación de valores bajos de la tasa interanual del indicador adelantado de la venta de coches y altos de la variación del indicador del desempleo son especialmente alarmantes en términos de probabilidad de recesión.

El gráfico 6 muestra la inferencia sobre la estimación de la probabilidad de recesión $p(s_{t+h} = 1 | x_t)$ obtenida por los modelos 71s, 78s, 88s y 04s para horizontes de predicción h de tres y seis meses. Además, se incluyen áreas sombreadas que marcan las recesiones de la economía española, según el Comité de Fechado de Ciclos. Se observa que la capacidad del árbol de clasificación para hacer inferencia sobre la probabilidad de recesión aumenta conforme se añaden indicadores económicos, especialmente a partir del modelo 78s donde básicamente entran en juego los indicadores de producción, precios y tipos de interés. Usando la información de estos indicadores, la estimación de la probabilidad de recesión sube hasta casi a uno en momentos clasificados como recesión por el Comité. Esto supone una significativa ganancia de la eficiencia relativa en la clasificación, comparado con la inferencia que conseguían los indicadores más largos usados en la predicción de las recesiones desde 1971. Además, los cuatro modelos infieren con probabilidad prácticamente uno que la economía española podría haber empezado una nueva recesión a partir de los datos de marzo de 2020 ocasionada por la crisis sanitaria de la COVID-19.

2. Inferencia en pseudotiempo real

No es posible evaluar en tiempo real cómo funcionaría el modelo para predecir a tres y seis meses las recesiones de la economía española porque no podemos obtener los datos de los indicadores en tiempo real de todas las variables que se incluyen en el modelo en cada momento en el que se hicieran las predicciones. Sin embargo, usaremos una aproximación a este ejercicio tomando submuestras

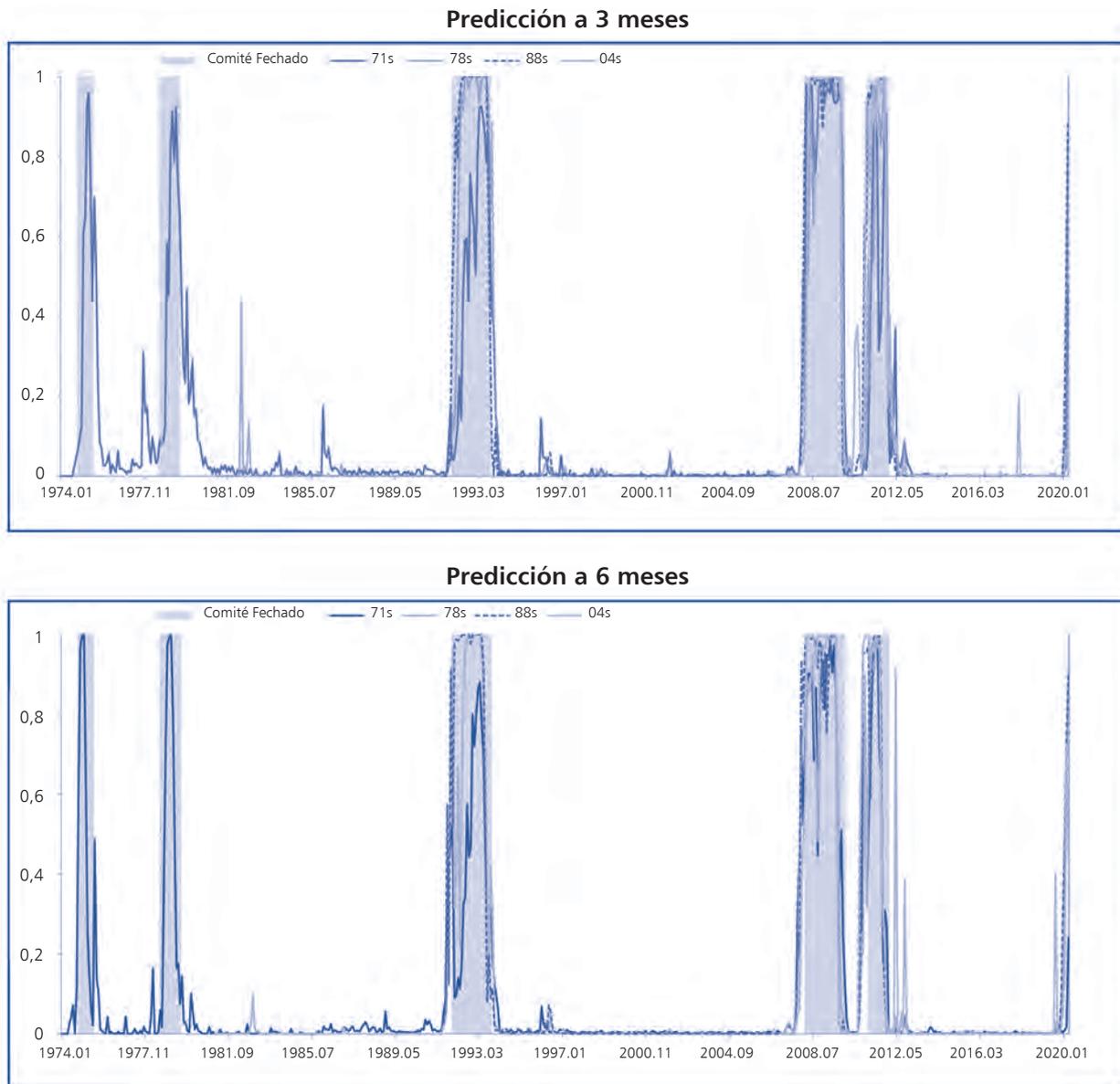
a partir de la base de datos más larga y realizando predicciones recursivas con esas submuestras.

Tomemos como ejemplo de cómo se construyen las predicciones recursivas al ejercicio de predicción a tres meses del análisis 71s. Para realizar la primera predicción se parte de la base de datos en mayo de 1974, se construye el árbol de clasificación con el algoritmo *stochastic gradient boosting*, y se hace una predicción de la probabilidad de recesión en agosto de 1974. Para realizar la segunda predicción en septiembre de 1974, se repite la operación, pero usando datos hasta junio de 1974. Siguiendo este proceso iterativamente, la última base de datos que se utiliza es la que se trunca en marzo de 2020 con la que se predice la probabilidad de junio de 2020.

Las predicciones recursivas de los cuatro análisis, 71s, 78s, 88s y 04s se muestran en el gráfico 7, junto a las recesiones del Comité de Fechado de Ciclos que se identifican mediante áreas sombreadas. El ejercicio fuera de la muestra corrobora los resultados que se han obtenido con el análisis dentro de la muestra. Como primer comentario, es importante destacar la alta capacidad de los árboles de clasificación para predecir probabilidades de recesión que suben de casi cero a casi uno en las recesiones de la economía española. La habilidad de clasificación del ciclo económico aumenta con el modelo 88s cuando se añade la información que contienen los indicadores de confianza, bursátiles y tipos de interés. Para anticipar la Gran Recesión, los indicadores financieros y los de construcción cobraron un papel protagonista.

Como el análisis fuera de la muestra implica una estimación recursiva del modelo, podemos analizar la evolución de la importancia relativa de los indicadores. El gráfico 8 muestra la evolución de la importancia relativa de los indicadores con mayor importancia relativa y representantes los indicadores relacionados con la tendencia del PIB, ventas de coches, desempleo, construcción y financieros con el conjunto de datos 78s. Como en casos anteriores, las recesiones del Comité de Fechado de Ciclos se representan con áreas sombreadas. Aunque la importancia relativa al final de la muestra es similar para todos los indicadores, la importancia se diferencia mucho en su evolución a lo largo de este período. Al comienzo de la muestra, los indicadores de empleo, venta de coches y de PIB eran más importantes. Sin embargo, a mitad de los noventa, el *spread* cobra una importancia creciente y se convierte en el indicador adelantado más importante de

GRÁFICO 7
PROBABILIDADES DE RECESIÓN FUERA DE LA MUESTRA



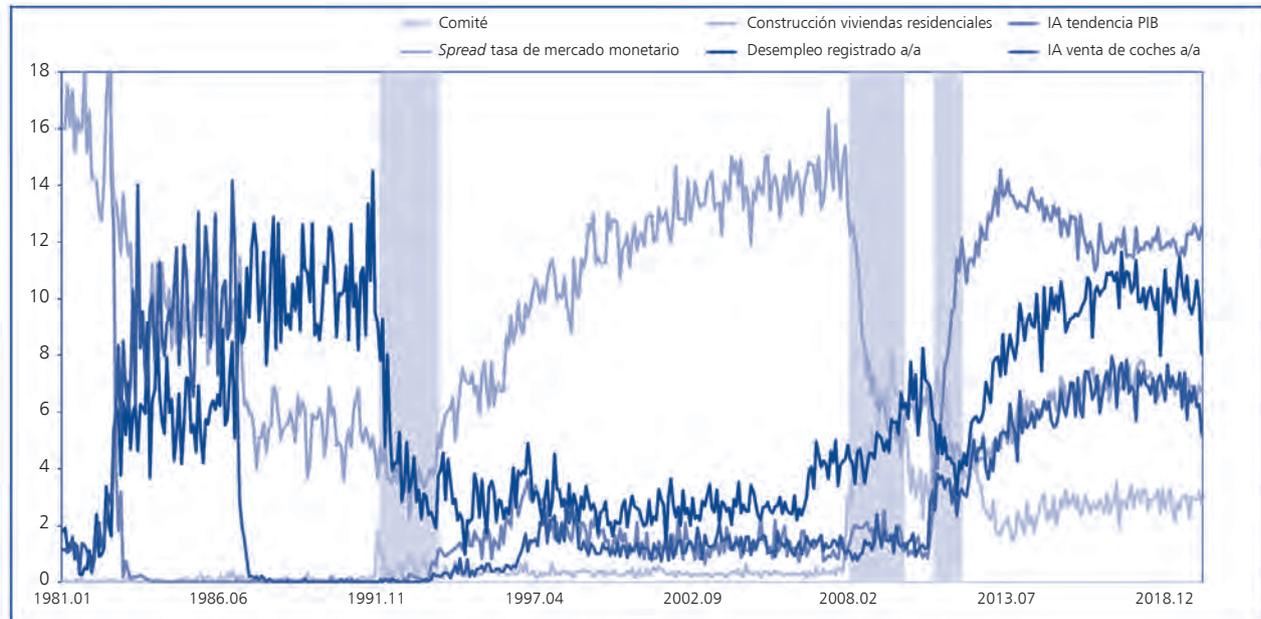
Notas: Estimación recursiva del árbol de clasificación con muestras de indicadores que comienzan en 1971, 1978, 1988 y 2004. Las áreas sombreadas identifican a las recesiones del Comité de Fechado de Ciclos.
 Fuentes: Elaboración propia a partir de OCDE y Comité de Fechado de Ciclo.

la Gran Recesión, mientras que los indicadores de construcción en los años 2000 pasan a ser uno de los principales indicadores coincidentes.

Un resultado interesante es que hemos encontrado un cambio en la importancia relativa de los indicadores económicos con los que se forma

inferencia sobre la recesión derivada de la crisis sanitaria de la COVID-19. En este caso, los indicadores con mayor importancia relativa vuelven a ser los indicadores de PIB, producción, consumo y desempleo, desbancando a los indicadores financieros y de la construcción que eran los protagonistas de la formación de inferencia en la Gran Re-

GRÁFICO 8
EVOLUCIÓN DE LA IMPORTANCIA RELATIVA



Notas: Resultados obtenidos con la base de datos desde 1978. Las áreas sombreadas hacen referencia a las recesiones del Comité de Fechado de Ciclos.

Fuentes: Elaboración propia a partir de OCDE y Comité de Fechado de Ciclo.

cesión. Este hecho constata que los fundamentos económicos que provocaron esas dos recesiones son diferentes.

IV. CONCLUSIONES

En este trabajo se presentan los árboles de regresión como una alternativa capaz de sintetizar la información que contienen grandes bases de datos de indicadores económicos para anticipar las recesiones de la economía española. Para una muestra de 270 indicadores comprendida entre enero de 1971 y marzo de 2020, las predicciones recursivas de las probabilidades de recesión con tres y seis meses de adelanto son casi cero excepto en los momentos identificados por el Comité de Fechado del Ciclo Económico Español, donde suben a casi uno.

Usando los gráficos de dependencia parcial hemos identificado cuáles son los indicadores económicos más influyentes para predecir las recesiones en España. Durante todo el período, entre los mejores indicadores para predecir las recesiones se encuentran los indicadores adelantados de tendencia del PIB y ventas de coches, y el paro registrado. La

capacidad de anticipar recesiones aumenta cuando se añaden los indicadores de confianza, bursátiles y tipos de interés. En la Gran Recesión, ganaron protagonismo los indicadores de la construcción y algunos financieros relacionados con la tenencia de depósitos. Sin embargo, en la recesión generada por la COVID-19, los indicadores de producción, desempleo y consumo vuelven a ser los indicadores más importantes, constatando los diferentes fundamentos económicos de esas dos recesiones.

NOTAS

(*) Agradecemos los comentarios de MANUEL RUIZ. Los autores agradecen a los proyectos PID2019-107192 GB-I00 (AEI/10.13039/501100011033) y al 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. *Corresponding author:* Máximo Camacho, Universidad de Murcia, Facultad de Economía y Empresa, Departamento de Métodos Cuantitativos para la Economía, 30100, Murcia, España. E-mail: mcamacho@um.es

(1) CAMACHO, GADEA-RIVAS y PÉREZ QUIRÓS (2019) sitúan a la economía española entre las que más efectos negativos tuvo la Gran Recesión.

(2) En este trabajo suponemos que solo hay dos estados posibles en el ciclo económico, expansiones y recesiones. El modelo podría extenderse sin dificultad si suponemos más estados posibles.

(3) Por ejemplo, para clasificar en función del estado más probable se usará el umbral de 0,5.

(4) Para simplificar la notación, se omitirá la dependencia de \hat{S}_{t+h} en x_t y c .

(5) Por ejemplo, si la región A_j se acaba clasificando como región de recesiones, se procurará que incluya el mayor porcentaje de observaciones en recesión que sea posible.

(6) Otras medidas habituales del grado de impureza son el error de clasificación y la entropía.

(7) Ejemplos de particiones no rectangulares se pueden encontrar en PÁEZ *et al.* (2019).

(8) Podría ocurrir que las dos particiones se determinen por el mismo indicador, $x_k = x_j$.

(9) La Asociación Española de Economía creó el Comité en el año 2012 para determinar el ciclo económico español utilizando la metodología de otros comités como el CEPR en la zona del euro y el NBER en EE.UU.

(10) Mientras que S_t viene determinada por el Comité y no depende de x_t , $\hat{S}_t = \hat{S}_t(x_t | x_t \in A_m)$ se infiere a partir de la región en la que se encuadran los indicadores x_t .

(11) En ocasiones se hace 100 la importancia relativa del indicador más importante y se relativizan las importancias relativas del resto de indicadores respecto al más importante.

(12) Sin pérdida de generalidad, vamos a suponer que el período de entrenamiento está al principio de la muestra.

(13) Otras técnicas usadas para dotar de robustez a los árboles de clasificación son el *bagging* y el *random forest*. A diferencia del *boosting*, están basadas en generar copias de los datos originales usando *bootstrap*. En cada copia se estima un árbol de clasificación y se combinan los resultados para dar lugar a una clasificación final. Por tanto, se generan árboles profundos con poca impureza, pero con poca habilidad de clasificar correctamente observaciones nuevas. Sin embargo, agregando el resultado de muchos árboles generados se consigue mejorar la habilidad para clasificar observaciones nuevas sin incrementar mucho la impureza.

(14) Generalmente el *AdaBoost* se aplica a la submuestra de entrenamiento y se evalúa en la submuestra de evaluación. Para facilitar la exposición, vamos a suponer que se entrena en toda la muestra.

(15) SCHAPIRE (1990) demuestra que, aunque los árboles de cada iteración solo clasifiquen un poco mejor que la clasificación aleatoria, la clasificación resultante del algoritmo *boosting* es muy buena.

(16) En estos sencillos árboles se eligen un indicador y un umbral dando lugar a una partición con dos regiones.

(17) FRIEDMAN *et al.* (2000) y contribuciones posteriores proponen otras funciones de pérdida. Usaremos la exponencial por su relación con *AdaBoost*.

(18) El capítulo 10 de libro HASTIE *et al.* (2001) ofrece una explicación detallada.

(19) Como criterio de ajuste se podría usar una función de pérdida cuadrática, $\sum_{t=1}^{T-h} (g_{t+h} - \hat{S}_{t+h}^m)^2$, para elegir el árbol que minimiza distancia con el gradiente. Aunque daría lugar a clasificaciones \hat{S}_{t+h}^m distintas, no tendría tanto impacto en \hat{S}_{t+h}^M .

(20) La importancia relativa I_k^m podría medirse en términos de la reducción del error cuadrático si así es como se ajusta el árbol en cada iteración al negativo del gradiente.

(21) Alternativamente, se puede analizar el efecto sobre la probabilidad que aparece en [11]. Sin embargo, el rango de variación es mucho menor y los resultados no son tan intuitivos.

(22) El listado de los 270 indicadores con el período disponible y la transformación estacionaria de los mismos se encuentra disponible bajo petición a los autores.

BIBLIOGRAFÍA

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. y STONE, C. (1984). *Classification and regression trees*. Wadsworth, New York: Chapman & Hall.

CAMACHO, M., CARO, A. y LÓPEZ-BUENACHE. 2020. The two-speed Europe in business cycle synchronization. *Empirical Economics*, en prensa.

CAMACHO, M. y DOMÉNECH, R. (2012). MICA-BBVA a factor model of economic and financial indicators for short-term GDP forecasting. *SERIEs: Journal of the Spanish Economic Association*, 3, pp. 475-497.

CAMACHO, M., GADEA-RIVAS, M. y PÉREZ QUIRÓS, G. (2019). The Great Recession. Worse than ever? *Revista de Economía Aplicada*, 76, pp. 73:100.

CAMACHO, M., PACCE, M. y ULLOA, C. (2018). Regional business cycle phases in Spain. *Estudios de Economía Aplicada*, 36, pp. 875-896.

CAMACHO, M. y PÉREZ QUIRÓS, G. (2011). Spain-STING: Spain Short Term Indicator of Growth. *The Manchester School*, 79, pp. 594-616.

CANCELO, J. (2004). El ciclo del empleo en España. Un análisis regional. *Revista económica de Castilla-La Mancha*, 4, pp. 107-138.

CUEVAS, A., PÉREZ QUIRÓS, G. y QUILLIS, E. (2017). Integrated model of short-term forecasting of the Spanish Economy (Mipred Model). *Revista de Economía Aplicada*, 25, pp. 5-25.

CUEVAS, A. y QUILLIS, E. (2012). A factor analysis for the Spanish economy. *Journal of the Spanish Economic Association*, 3, pp. 311-338.

DÖPKE, J., FRITSCH, U. y PIERDZIOCH, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting*, 33, pp. 745-759.

FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 121, pp. 256-285.

FREUND, Y. y SCHAPIRE, R. (1996). Experiments with a new boosting algorithm. *Proceedings of ICML*, 13, pp. 148-156.

FRIEDMAN, J., HASTIE, T. y TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28, pp. 337-407.

FRIEDMAN, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, pp. 1189-1232.

- (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38, pp. 367-378.
- GADEA-RIVAS, M., GÓMEZ-LOSCOS, A. y BANDRÉS, E. (2018). Clustering regional business cycles. *Economic Letters*, 162, pp. 171-176.
- GADEA-RIVAS, M., GÓMEZ-LOSCOS, A. y MONTAÑÉS, A. (2012). Cycles inside cycles: Spanish regional aggregation. *Journal of the Spanish Economic Association*, 3, pp. 423-456.
- GADEA-RIVAS, M., GÓMEZ-LOSCOS, A. y LEIVA-LEÓN, D. (2019). Increasing linkages among European regions. The role of sectoral composition. *Economic Modelling*, 80, pp. 222-243.
- HASTIE, T., TIBSHIRANI, R. y FRIEDMAN, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- MEASE, D., WYNER, A. y BUJA, A. (2007). Cost-weighted boosting with jittering and over/undersampling: Jous-boost. *Journal of Machine Learning Research*, 8, pp. 409-439.
- 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.
- SCHAPIRE, R. (1990). The strength of weak learnability. *Machine Learning*, 5, pp. 197-227.
- SCHAPIRE, R. y FREUND, Y. (2012). *Boosting: Foundations and algorithms*. MIT Press.
- WARD, F. (2017). Spotting the danger zone: Forecasting financial crisis with classification tree ensembles and many predictors. *Journal of Applied Econometrics*, 32, pp. 359-378.