CAPÍTULO VII

Desarrollos con *big data* para el análisis coyuntural en los bancos centrales

Corinna Ghirelli
Samuel Hurtado
Javier J. Pérez
Alberto Urtasun

Los bancos centrales utilizan intensivamente datos estructurados (micro y macro) para el desarrollo de sus funciones, entre las que destaca el seguimiento en tiempo real de la actividad económica. El desarrollo tecnológico está permitiendo integrar nuevas fuentes de datos masivos, más granulares y disponibles con mayor frecuencia, en muchos casos no estructuradas. Esto supone un desafío importante desde el punto de vista de la gestión, el almacenamiento, la seguridad y la confidencialidad. Este capítulo analiza las ventajas y los retos de estas nuevas fuentes, y describe algunos casos de éxito de su incorporación en el ámbito del análisis económico y la previsión.

Palabras clave: big data, datos masivos, predicción económica, análisis textual, incertidumbre económica, datos de prensa.

1. INTRODUCCIÓN[1]

El desarrollo de las nuevas tecnologías y las redes sociales ha abierto la posibilidad de utilizar nuevas fuentes de datos, que presentan características específicas en términos de volumen, nivel de detalle, frecuencia y (o falta de) estructura. En los últimos años se han desarrollado una gran cantidad de aplicaciones que explotan estas nuevas fuentes de datos en las áreas de economía y finanzas, particularmente en los bancos centrales (BC). En el área específica del análisis económico, las nuevas fuentes de datos presentan un potencial significativo para los BC, incluso teniendo en cuenta que estos ya venían haciendo un uso muy intensivo de datos estadísticos, tanto individuales (microdatos) como agregados (macroeconómicos) para realizar sus funciones.

En particular, estas nuevas fuentes están permitiendo a los BC, entre otras áreas de aplicación:

- La ampliación de la información utilizada para llevar a cabo las funciones de estabilidad financiera, supervisión bancaria y de sistemas de pagos (Broeders y Prenio, 2018; Fernández, 2019; Alonso y Carbó, 2020; Moreno Bernal y González Pedraz, 2020; Nyman, Kapadia, Tuckett *et al.,* 2018; Carlsen y Storgaard, 2010; Duarte, Rodrigues y Rua, 2017; Gil, Pérez, Sánchez y Urtasun, 2017).

- La aplicación de nuevas metodologías de análisis económico (véase, por ejemplo, Fernández-Villaverde, Hurtado y Nuño, 2019).

- Un análisis más profundo (datos más detallados) y un seguimiento más preciso (disponibilidad de datos casi en tiempo real) de la actividad económica (Kapetanios y Papailias, 2018; Thorsrud, 2018; Aprigliano, Ardizzi y Monteforte, 2017; Duarte, Rodrigues y Rua, 2017; Gotz y Knetsch, 2019; D'Amuri y Marcucci, 2017; Ferrara y Simoni 2019; Carriére-Swallow y Labbé, 2013) o alguna faceta de interés específico para los BC, como el seguimiento del crédito (Petropoulos *et al.,* 2019).

- Una mejor cuantificación de la confianza y la incertidumbre de los agentes, y sus expectativas de inflación o crecimiento (Baker, Bloom, y Davis, 2016; Ghirelli, Pérez y Urtasun, 2019 y 2020; Aguilar *et al.,* 2020).

- Una mejor valoración de la política económica y un aumento de la capacidad de simular medidas alternativas, debido principalmente a la disponibilidad de microdatos que pueden utilizarse para caracterizar de manera precisa la heterogeneidad de los agentes y, por tanto, llevar a cabo un análisis más profundo y

[1] Este capítulo descansa en gran medida en las ideas y resultados desarrollados en Ghirelli *et al.* (2019) y Ghirelli *et al.* (2020).

preciso de su comportamiento (Chetty, Friedman y Rockoff, 2014; Pew Research Center, 2012).

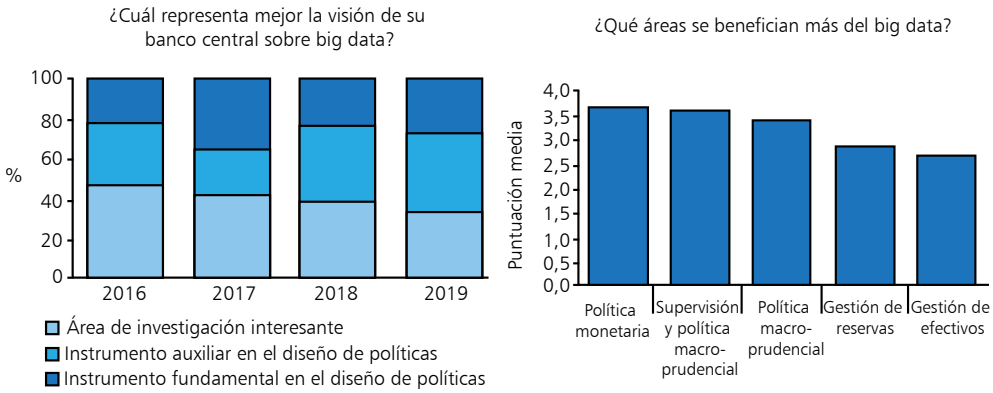
- La mejora de las estadísticas disponibles, como las relativas al turismo (Lacroix, 2019; Artola y Galán, 2010) o al mercado de la vivienda (Loberto, Luciani y Pangallo, 2018).

Según la encuesta anual del portal de información Central Banking[2], en 2019 más del 60% de los BC utilizaron fuentes de *big data* en sus operaciones, y dos tercios de ellos las utilizaron como instrumento principal o auxiliar en el proceso de diseño de sus políticas (véase la figura 1).

El resto de este capítulo está estructurado como sigue. En la sección segunda se proporciona una breve descripción de las principales ventajas y desafíos relacionados con la utilización de estas nuevas fuentes de datos por parte de los BC. En la tercera sección se describen algunos casos de éxito en que los BC han utilizado nuevas fuentes de datos para llevar a cabo sus funciones. En particular, nos centramos en la medición de la incertidumbre económica con artículos de prensa (apartado 3.1.), el uso de los informes regulares de un BC como herramienta de comunicación sobre el estado de la economía (apartado 3.2.), y algunas aplicaciones en el ámbito de la predicción macroeconómica (apartado 3.3.).

FIGURA 1

ALGUNOS RESULTADOS DE LA ENCUESTA DE CENTRAL BANKING (2019) SOBRE EL USO DE BIG DATA EN LOS BANCOS CENTRALES (ECONOMÍAS AVANZADAS Y EMERGENTES)



Fuente: Central Banking (2019) (https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central-banks-2019-survey-results).

[2] Disponible en https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central-banks-2019-survey-results. A la encuesta respondieron 58 bancos centrales, con una muestra importante tanto de países avanzados como de emergentes.

2. NUEVAS FUENTES DE DATOS: OPORTUNIDADES PARA LOS BANCOS CENTRALES[3]

Los BC utilizan de manera intensiva bases de datos estructurados para llevar a cabo sus funciones, en las áreas de supervisión bancaria, estabilidad financiera o política monetaria. Se utilizan datos de naturaleza micro y datos macro. Algunos ejemplos de micro datos son los balances de las empresas (véanse, por ejemplo, Menéndez y Mulino, 2018; Banco de España, 2018), la información relativa al volumen de crédito concedido por las instituciones financieras a personas y empresas, o los datos relacionados con las decisiones financieras de los agentes (véase, por ejemplo, Banco de España, 2017). En cambio, las principales fuentes de información en el ámbito de la macroeconomía suelen ser las cuentas nacionales o datos elaborados por los propios BC, así como un conjunto amplio de información publicada por otros organismos como, por ejemplo, datos de seguridad social (Ministerio de Seguridad Social), precios de las acciones (Bloomberg) o precios de la vivienda (portales inmobiliarios).

Gracias a los avances tecnológicos las fuentes de información se están ampliando de manera significativa, en particular en lo que se refiere a su granularidad y frecuencia. En muchos casos se puede obtener información casi en tiempo real sobre decisiones individuales realizadas por individuos o empresas, y la mayoría de las veces a frecuencias más altas que con fuentes de datos tradicionales. Por ejemplo, los datos de transacciones de tarjetas de crédito, que se pueden utilizar para aproximar las decisiones de consumo de los hogares, están potencialmente disponibles en tiempo real con un coste reducido, especialmente si se compara con el coste de realizar encuestas de hogares en todo el país.

La disponibilidad de grandes cantidades de información supone enfrentarse a retos importantes en cuanto a su gestión, a las necesidades y capacidades de almacenamiento, a los costes, seguridad y confidencialidad de la infraestructura necesaria, o a la calidad de los datos (sobre este último punto, véase Einav y Levin, 2014). Además, la gestión óptima de grandes conjuntos de datos estructurados y no estructurados requiere la integración de nuevos perfiles profesionales (científicos e ingenieros de datos) en los BC e implica la necesidad de su completa transformación digital. Asimismo, la diferente naturaleza de las nuevas fuentes de información requiere la asimilación y el desarrollo de técnicas para transformar y sintetizar los datos brutos en formatos que puedan incorporarse al análisis económico. Por ejemplo, las técnicas de análisis textual permiten procesar la información contenida en textos y convertirla en datos estructurados. Este ocurre, por ejemplo, con la información que se obtiene, entre otros, de noticias de prensa, redes sociales (Facebook y Twitter), portales de búsqueda web (por ejemplo, de vivienda o trabajo).

[3] La discusión en esta sección presenta las principales ideas de nuestro trabajo conjunto con Juan Peñalosa en Ghirelli *et al.* (2019).

Las nuevas fuentes de datos están ampliando la frontera de la estadística, en particular en el campo de las estadísticas no financieras. Este es el caso, por ejemplo, de las iniciativas para adquirir mejores medidas de precios en la economía utilizando técnicas de *web-scraping*[4] (Loberto, Luciani y Pangallo, 2018), o determinados elementos del comercio exterior, como la estimación de los movimientos turísticos mediante el seguimiento de redes móviles u otras fuentes (véase Hardy *et al.,* 2017; Lacroix, 2019; Artola y Galán, 2010). Los países en desarrollo, que se enfrentan a mayores dificultades para establecer infraestructuras estadísticas sólidas, están comenzando a utilizar las nuevas fuentes de datos, incluso para realizar estimaciones de algunos agregados de las cuentas nacionales (véase Hammer, Kostroch y Quirós, 2017). Asimismo, la ampliación de las fronteras de la estadística también está afectando al "monopolio público" sobre la información, dado que gran parte de la nueva información son propiedad de empresas privadas, y mucha de la información es incluso de libre acceso a través de la web.

3. ALGUNAS APLICACIONES EN EL ÁMBITO DE LOS BANCOS CENTRALES

3.1. Uso de la prensa para medir la incertidumbre

Las aplicaciones que involucran el análisis de texto han adquirido una importancia especial en el área del análisis económico. Con estas técnicas, se puede obtener la información relevante de un texto y luego sintetizarla y codificarla en forma de indicadores cuantitativos. Primero, el texto se prepara (preprocesamiento), eliminando los elementos del texto que no son relevantes (artículos, palabras no relevantes, números) y cortando las palabras a su raíz (es decir, quitando las terminaciones). Segundo, la información relevante se sintetiza principalmente mediante el cálculo de la frecuencia de palabras o grupos de palabras. Intuitivamente, la frecuencia relativa de grupos de palabras relacionados con un determinado tema permite evaluar la importancia relativa de este tema en el texto. Los datos de texto son una nueva fuente de información valiosa, ya que reflejan los principales acontecimientos actuales que afectan a las decisiones de los agentes económicos y están disponibles en tiempo real.

Una rama reciente de la literatura se centra en construir indicadores de incertidumbre económica a partir de artículos de prensa. El trabajo más influyente y seminal de esta literatura es Baker *et al.* (2016). Estos autores construyeron un índice de incertidumbre acerca de las políticas económicas para Estados Unidos (*Economic Policy Uncertainty Index, EPU*), basado en el volumen de artículos de periódicos que contienen palabras relacionadas con los conceptos de incertidumbre, economía y política. A raíz de este artículo, muchos investigadores han utilizado en sus análisis indicadores de incertidumbre basados en textos, proporcionando evidencia empírica de los efectos negativos de aumentos de la incertidumbre sobre la actividad económica como, por ejemplo,

[4] El web-scraping es un proceso en el que se navega automáticamente en sitios web para extraer contenido y datos de esas páginas.

Meinen y Roehe (2017) para Alemania, Francia, Italia y España;Fontaine, Didier y Razafindravaosolonirina (2017) para China, o Colombo (2013) y Azqueta-Gavaldon *et al.* (2020) para la eurozona.

El resto de esta sección presenta dos estudios que, siguiendo estas metodologías, desarrollan indicadores de incertidumbre acerca de las políticas económicas para: (1) la economía española y (2) para los principales países de América Latina (Argentina, Brasil, Chile, Colombia, México, Perú y Venezuela). Estos indicadores se han construido a partir de la prensa española y se utilizan actualmente en las tareas de seguimiento y previsión económica del Banco de España.

3.1.1. Incertidumbre acerca de las políticas económicas en España

Si bien Baker, Bloom y Davis (2016) también elaboraron un índice EPU para España basado en los dos principales periódicos generalistas españoles (*El País* y *El Mundo*), Ghirelli, Pérez y Urtasun (2019) desarrollan un nuevo índice de incertidumbre acerca de las políticas económicas para España, que se basa en la metodología de Baker, Bloom y Davis (2016) pero amplía la cobertura de prensa de dos a siete periódicos[5], extiende su cobertura temporal (a partir de 1997 en lugar de 2001), y enriquece las palabras clave utilizadas en las expresiones de búsqueda.

El indicador aumenta o disminuye en momentos relacionados con eventos asociados con un aumento o una disminución de la incertidumbre económica *a priori,* como los ataques terroristas del 11 de septiembre de 2001 en Estados Unidos, el colapso de Lehman Brothers en septiembre de 2008, la solicitud de ayuda financiera de Grecia en abril de 2010, la solicitud de ayuda financiera para reestructurar el sector bancario y los bancos de ahorro en España en junio de 2012, el referéndum del *brexit* en junio de 2016, o los episodios de tensión política en la región española de Cataluña en octubre de 2017. Ghirelli, Pérez y Urtasun (2019) muestran la existencia de una relación dinámica significativa entre este indicador y las principales variables macroeconómicas, de manera que incrementos inesperados en el indicador de incertidumbre acerca de las políticas económicas tienen efectos macroeconómicos adversos. Específicamente, un aumento de la incertidumbre reduce significativamente el PIB, el consumo y a la inversión. Este resultado está en línea con los resultados de la literatura empírica acerca de la incertidumbre económica.

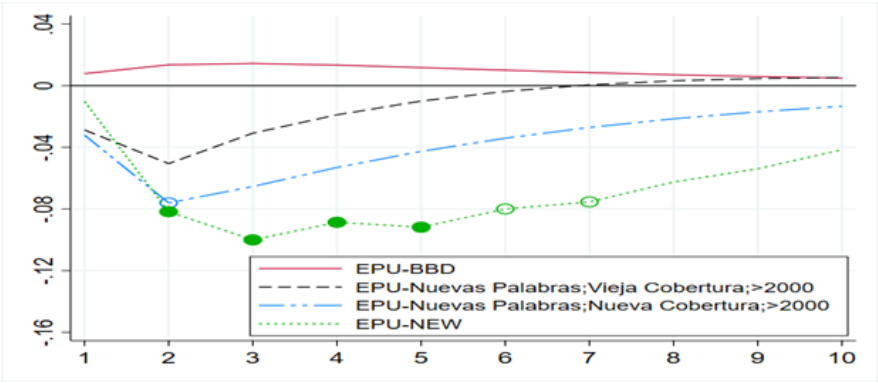
Además, en ese artículo se muestra la importancia de enriquecer el conjunto de palabras clave utilizado en las expresiones de búsqueda y ampliar la cobertura de prensa y temporal en la construcción del índice. Estas conclusiones se muestran en la figura 2,

[5] El nuevo índice se basa en los cuatro periódicos de carácter generalista más leídos en España y los tres principales periódicos de negocios: *El País, El Mundo, El Economista, Cinco Días, Expansión, ABC, La Vanguardia.*

que compara las respuestas macroeconómicas a perturbaciones de incertidumbre, usando versiones alternativas del índice EPU, en que se varían una por una las dimensiones mencionadas anteriormente (riqueza de las palabras, cobertura temporal y cobertura de prensa), pasando del índice EPU construido por Baker, Bloom, y Davis (2016) al nuevo índice construido por Ghirelli, Pérez y Urtasun (2019). Todas estas dimensiones son importantes ya que todas contribuyen a obtener el signo negativo esperado en las respuestas. Sin embargo, ampliar la cobertura temporal es clave para mejorar la precisión de las estimaciones y obtener resultados significativos, así como la cobertura de prensa.

FIGURA 2

RESPUESTA DEL PIB DE ESPAÑA ANTE DISTINTAS PERTURBACIONES DE INCERTIDUMBRE CONSTRUIDAS CON VERSIONES ALTERNATIVAS DEL ÍNDICE EPU



Nota: La figura muestra la función impulso-respuesta de la tasa de crecimiento del PIB español hasta 10 trimestres después de un shock positivo de una desviación estándar en el índice EPU para España. El eje horizontal representa los trimestres desde el choque. El eje vertical mide la tasa de crecimiento del PIB español (en puntos porcentuales). Los círculos completos (vacíos) indican significatividad estadística al 5% (10%); la línea continua indica la ausencia de significatividad estadística. EPU-BBD: índice EPU para España proporcionado por Baker, Bloom y Davis (2016). EPU-NEW: índice EPU para España elaborado por Ghirelli, Pérez y Urtasun (2019). Los modelos de autorregresión vectorial (VAR) incluyen el índice EPU, el spread entre los bonos alemanes y los bonos españoles, la tasa de crecimiento del PIB y la tasa de crecimiento del índice de precios al consumidor (IPC); el EPU global se incluye como variable exógena.

Fuentes: Banco de España y Policy-Uncertainty-web.

3.1.2. Incertidumbre acerca de la política económica en América Latina

La literatura también demuestra que la incertidumbre económica en un país puede tener efectos indirectos en otros países, así como ramificaciones globales. En este marco, Ghirelli, Pérez y Urtasun (2020) desarrollan índices de incertidumbre acerca de las políticas económicas para los principales países de América Latina (AL): Argentina,

Brasil, Chile, Colombia, México, Perú y Venezuela. El objetivo es doble. Primero, medir la incertidumbre acerca de las políticas económicas en los países de AL, y obtener una narrativa de los "choques de incertidumbre" y sus efectos potenciales sobre la actividad económica en estos países. Segundo, explorar en qué medida esos choques en AL tienen el potencial de extenderse a España. Este último país representa un caso relevante para estudiar este tipo de contagio internacional, dados sus importantes vínculos económicos con la región latinoamericana.

Los indicadores de incertidumbre se construyen siguiendo la misma metodología descrita en la sección anterior, es decir, contando los artículos de los siete diarios españoles más importantes que contienen palabras relacionadas con los conceptos de economía, política e incertidumbre. Sin embargo, además, se modifican las búsquedas para que se adapten a cada uno de los países latinoamericanos de interés[6]. Estos indicadores se basan en la prensa española y, por tanto, reflejan puramente la variación en la incertidumbre en los países de AL que es relevante para la economía española. El supuesto es que la prensa española refleja con precisión la situación política, social y económica de la región de AL, dados los estrechos vínculos económicos y culturales existentes, incluido el idioma común para la mayoría de países. Por esta razón, se puede afirmar que los índices basados en la prensa española proporcionan medidas relevantes acerca de la incertidumbre política en esos países. Esta metodología también es coherente con una rama de la literatura que utiliza la prensa internacional para calcular indicadores de texto para un amplio conjunto de países (véanse, por ejemplo, Ahir, Bloom y Furceri, 2019, y Mueller y Rauh, 2018).

En el estudio de referencia se muestra como una mayor incertidumbre en América Latina afecta a las empresas españolas más expuestas a la región. En particular, los resultados empíricos muestran que una perturbación positiva inesperada en el índice EPU agregado para AL, genera una caída significativa en la tasa de variación de la cotización de dichas empresas. El resultado se mantiene cuando se realizan ejercicios individuales para todos los países de AL considerados en el trabajo. Y lo confirman las pruebas de placebo, que consideran empresas españolas que cotizan en la bolsa española pero que no tienen intereses económicos en la región latinoamericana. En un segundo conjunto de resultados, Ghirelli, Pérez y Urtasun (2020), muestran como los *shocks* en los índices EPU de AL afectan a las siguientes variables macroeconómicas españolas: las exportaciones, la inversión extranjera directa de España a América Latina y el PIB español.

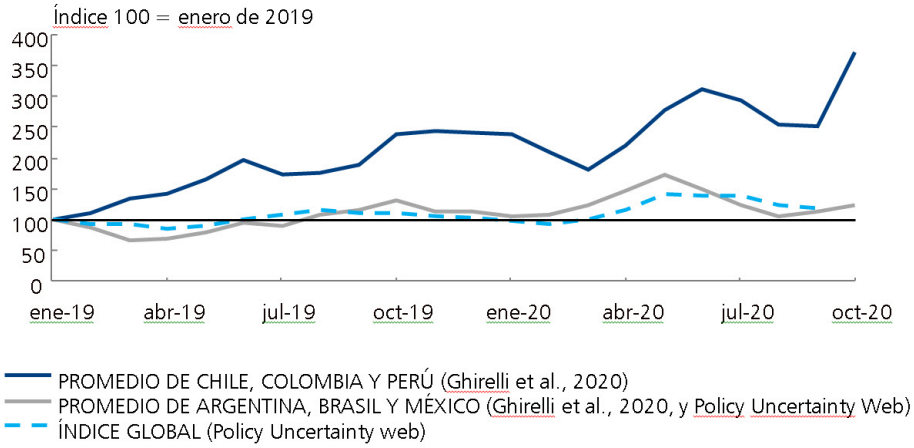
Además de los resultados discutidos, de un carácter más estructural, la disponibilidad de estos índices en tiempo real permite su uso para el seguimiento de la incertidumbre

[6] En particular, (1) se requiere que cada artículo contenga el nombre del país de interés de AL; (2) entre el conjunto de palabras clave relacionadas con la política, se incluye también el nombre del banco central y el nombre del lugar de trabajo del gobierno del país de interés. Para obtener más detalles, véase Ghirelli, Pérez y Urtasun, (2020).

sobre las políticas económicas en los países de referencia de América Latina, lo que, dados los resultados apuntados en el párrafo anterior, resulta relevante para la estimación de corto plazo *(nowcasting)* de las principales macromagnitudes españolas, pero también de estos países. En la figura 3 se muestra la evolución reciente de estos índices, desde enero de 2019 hasta octubre de 2020, agrupando los países analizados y normalizando el valor de los distintos índices a 100 en enero de 2019. Como puede observarse, la crisis asociada a la COVID-19 causó un aumento generalizado de la incertidumbre en todos los países de AL y a nivel global. No obstante, los países en los que también se vienen registrando tensiones sociales e institucionales presentaron aumentos mucho más marcados, como se refleja con claridad en el agregado de Chile, Perú y Colombia, al compararlo con el de Argentina, Brasil y México.

FIGURA 3

EVOLUCIÓN RECIENTE DE LA INCERTIDUMBRE SOBRE LAS POLÍTICAS ECONÓMICAS EN ALGUNAS ECONOMÍAS DE AMÉRICA LATINA



Nota: Medias móviles de tres meses.

Fuentes: Banco de España y Policy-uncertainty-web.

3.2. La narrativa sobre la economía como previsión sombra: un análisis con informes trimestrales del Banco de España

Otra de las técnicas de minería de texto consiste en el uso de métodos de diccionario para el análisis de sentimiento. Un diccionario es una lista de palabras asociadas con sentimientos positivos y negativos. Estas listas se pueden construir de varias formas, desde técnicas puramente manuales hasta técnicas de aprendizaje automático. El análisis de sentimiento a su vez se basa en búsquedas en bases de datos de texto y requiere que el investigador tenga acceso a los textos. En su versión más simple, las búsquedas permiten calcular la frecuencia de términos positivos y negativos en un texto. El índice

de sentimiento se define como la diferencia (con algunos pesos) entre las dos frecuencias, es decir, un texto tiene un sentimiento positivo (negativo) cuando la frecuencia de términos positivos es mayor (menor) que la de los términos negativos.

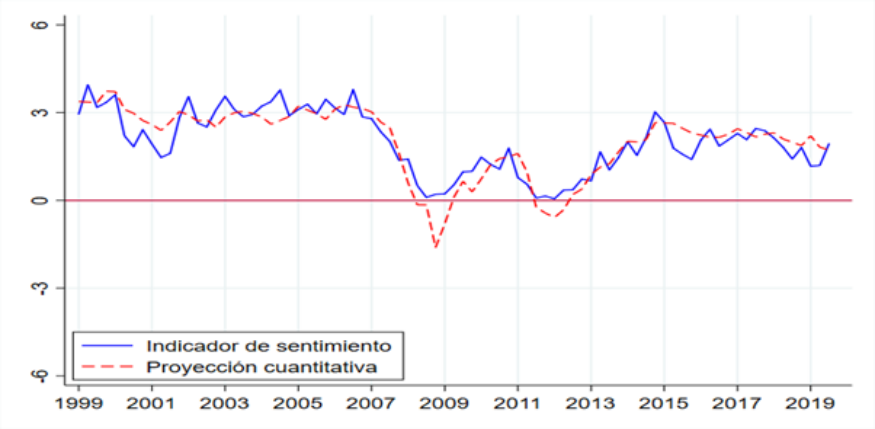
En este arpartado, se proporciona un ejemplo de análisis de sentimiento para mostrar la utilidad de los datos de texto (Diaz-Sobrino *et al.,* 2020). Para ello, se utiliza la técnica más básica de análisis de sentimiento, es decir, el simple conteo de palabras contenidas en un diccionario creado por los autores de este capítulo. El ejercicio se basa en el *Boletín Económico Trimestral de la economía española* del Banco de España, publicado *online* desde el primer trimestre de 1999. El objetivo es construir un indicador (a partir del primer trimestre 1999) que refleje el sentimiento de los informes de perspectivas económicas del Banco de España. El análisis muestra que el indicador refleja correctamente las previsiones cuantitativas del PIB del Banco de España. Esto significa que la narrativa cualitativa de los informes contiene información similar a la transmitida por las previsiones cuantitativas.

Para realizar el análisis, primero se crea un diccionario en español de términos positivos y negativos que se usan típicamente en el lenguaje económico para describir la economía. Por ejemplo, palabras como "crecimiento" o "aumento" entre términos positivos, y "disminución" o "reducción" entre términos negativos. Para disminuir los casos de signos incorrectos, se ignoran estos términos cuando aparecen alrededor (dentro de nueve palabras antes o después) de palabras que tienen un significado opuesto, como "desempleo" o "déficit". Por último, se asigna un peso de +1 y -1 a las cuentas de términos positivos y negativos en cada texto, se suma las cuentas ponderadas de los términos del diccionario y se divide el número resultante por la longitud del texto. A continuación, se compara el índice con las previsiones de crecimiento del PIB realizadas cada trimestre por el Banco de España, que en la mayor parte de la muestra considerada fueron registradas internamente pero no publicadas *online*.

Los resultados muestran una relación dinámica significativa entre ambas series: el indicador de sentimiento sigue fielmente el ciclo español, así como la evolución de las previsiones cuantitativas. Además, la comparación muestra que los informes son informativos no solo en el horizonte de previsión a corto plazo, sino también en el horizonte a uno a dos años. El indicador de sentimiento muestra la correlación más alta con las previsiones a dos años. La figura 4 muestra el indicador textual (línea azul continua) frente a la previsión de crecimiento del PIB realizada por el Banco de España para el horizonte a dos años (línea roja discontinua). Esta evidencia sugiere que la narrativa contenida en los informes del Banco de España refleja de manera muy fiel las previsiones cuantitativas de crecimiento del PIB obtenida por esta misma institución. Esto significa que un lector "sofisticado" podría haber inferido las previsiones de crecimiento del PIB del Banco de España a partir de sus informes.

FIGURA 4

PREVISIONES CUANTITATIVAS E INDICADOR DE SENTIMIENTO DE LOS INFORMES DEL BANCO DE ESPAÑA



Nota: La figura muestra el indicador de sentimiento (línea azul continua) frente a las previsiones cuantitativas del Banco de España (línea roja discontinua).

Fuente: Banco de España.

3.3. Previsiones con nuevas fuentes de datos

Normalmente, los ejercicios de previsión de los bancos centrales se llevan a cabo combinando indicadores cualitativos *(soft)* con indicadores cuantitativos *(hard),* que representan el conjunto de información publicado por las instituciones de estadística. La principal limitación de los datos cuantitativos es que se publican con cierto retraso y con baja frecuencia (por ejemplo, trimestralmente). En cambio, los indicadores cualitativos proporcionan información cualitativa (por lo tanto, de menor calidad que los datos cuantitativos) acerca de la coyuntura económica con una frecuencia más alta que los datos cuantitativos, como por ejemplo las encuestas de confianza de empresas y consumidores. La utilidad de los indicadores cualitativos es máxima al inicio del trimestre, cuando falta información macroeconómica, y disminuye tan pronto como se publican indicadores *hard* (Ferrara y Simoni, 2019).

Los indicadores de texto son un nuevo tipo de indicadores cualitativos. En comparación con los tradicionales, basados en encuestas, los indicadores textuales muestran las siguientes características: (1) suponen un menor coste, ya que no se basan en encuestas mensuales sino en suscripciones a servicios de prensa; (2) proporcionan más flexibilidad, ya que se pueden seleccionar las palabras clave en función de las necesidades específicas y obtener la serie temporal completa (mirando a los textos pasados), mientras que en una encuesta, se debería incluir una nueva pregunta y en consecuencia la serie temporal empezaría a partir de ese momento.

El resto de esta sección presenta tres aplicaciones en que se muestra como las nuevas fuentes de datos pueden mejorar los ejercicios de previsión económica. La primera se basa en el análisis de sentimiento. La segunda aplicación muestra cómo el aprendizaje automático puede mejorar la precisión de las técnicas de previsión disponibles. Finalmente, la tercera aplicación valora la importancia relativa de indicadores basados en nuevas fuentes de datos, como Google Trends y transacciones de tarjetas de crédito.

3.3.1. Un método "supervisado"

Este ejercicio se centra en construir un indicador textual para mejorar el seguimiento de la actividad económica (para detalles, véase Aguilar *et al.,* 2020). Para ello, se utiliza un procedimiento similar a lo que se ha descrito anteriormente para la elaboración del indicador de incertidumbre acerca de las políticas económicas, es decir, descansa en el número de artículos en la prensa española que contienen palabras clave específicas, y se usan los mismos periódicos para ello. Además, en este caso, se construye un diccionario de palabras positivas y negativas que se suelen utilizar cuando se describe la evolución de la tasa de crecimiento del PIB, la variable objetivo de interés, para identificar correctamente el tono de los artículos de prensa y, en particular, hasta qué punto se está tratando de repuntes o desaceleraciones de la economía. En concreto, el indicador se construye en la siguiente manera:

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones positivas. Se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recuperacion*" o una de las siguientes palabras (aceler*, crec*, increment*, recuper*, aument*, expansi*, mejora*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de 5 palabras entre sí. Se ignora "crecimiento económico" dado que tiene un tono neutro (se usa para describir indiferentemente un crecimiento negativo o positivo).

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones negativas: se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recesión*" o "crisis" o una de las siguientes palabras (descen*, ralentiz*, redu*, disminu*, contraccion*, decrec*, desaceler*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de cinco palabras entre sí.

Sucesivamente, para cada periódico, se calcula la diferencia entre el total de artículos relacionados con las recuperaciones y las recesiones, y si divide el valor resultante por

el número total de artículos económicos publicados en el mismo periódico en cada mes. Esta proporción, primero, se estandariza, a continuación, se computa una serie agregada tomando la media entre las series de los distintos periódicos y, finalmente, se le quita la media para que tenga media 0. El panel derecho de la figura 5 muestra el indicador textual (línea azul continua) contra la tasa de crecimiento del PIB (línea roja y discontinua). Aguilar *et al.* (2020) muestran que el indicador textual tiene poder predictivo para la previsión de la tasa de crecimiento del PIB español, a través de un ejercicio de previsión de PIB (datos del PIB no revisados) a corto plazo en (pseudo) tiempo real.

Una de las principales ventajas de los indicadores basados en periódicos es que se pueden actualizar en tiempo real y son de alta frecuencia. Estas ventajas han sido extremadamente valiosas desde el brote de la COVID-19, cuando los indicadores de confianza tradicionales basados en encuestas no han proporcionado señales correctas sobre la actividad económica[7]. Como ejemplo, el panel derecho de la figura 5 se muestra el indicador textual con una frecuencia semanal alrededor del confinamiento (14 de marzo de 2020). Su evolución refleja correctamente la drástica reducción de la actividad económica española en esa época.

3.3.2. Un método "no supervisado"[8]

Este ejercicio se compone de dos partes: en una primera parte se utiliza el método *Latent Dirichlet Allocation (LDA)* (Blei, Ng y Jordan, 2003) para extraer un conjunto de artículos de prensa indicadores cuantitativos que representen la importancia de determinados temas a lo largo del tiempo. La segunda parte utiliza los datos resultantes del método LDA para mejorar las previsiones del PIB español a través de un modelo de aprendizaje automático.

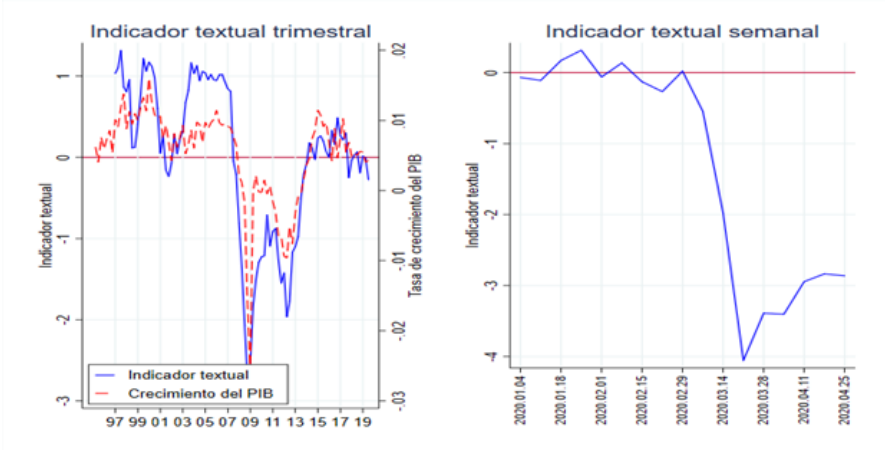
El LDA es un método de aprendizaje no supervisado, lo que significa que la definición de los temas no la decide el investigador, sino que es el resultado de ejecutar el modelo sobre los datos. El primer paso del proceso es construir un *corpus* con datos de texto. En este caso, se trata de una base de datos de más de 780.000 observaciones que contienen todas las noticias publicadas por *El Mundo* (uno de los principales periódicos españoles) entre 1997 y 2018, extraídas del repositorio de prensa española Dow Jones. A continuación, se tiene que procesar la base de datos de manera que se obtenga una versión del *corpus* que excluya puntuación, números o caracteres especiales, en que todo el texto esté en minúscula y que excluya las palabras más comunes como artículos y conjunciones. También es útil reducir las palabras a su raíz básica (la parte de la palabra que captura su significado central) eliminando algunas variaciones de las palabras como, por ejemplo, los tiempos verbales.

[7] En Aguilar *et al.* (2020), se compara este indicador textual con el indicador de sentimiento económico (ESI) de la Comisión Europea y se muestra que, para España, el primero mejora significativamente la previsión a corto plazo del PIB en comparación con el ESI.

[8] Véase Ghirelli *et al.* (2020).

FIGURA 5

UN INDICADOR TEXTUAL PARA ANTICIPAR LA EVOLUCIÓN DEL PIB DE ESPAÑA



Nota: La figura de la derecha muestra el indicador textual trimestral de la economía (línea azul y sólida) frente a la tasa de crecimiento del PIB español (línea roja y discontinua) hasta junio de 2019. La figura de la izquierda muestra el indicador textual semanal de enero a marzo de 2020.

Fuentes: Banco de España e Instituto Nacional de Estadistica (INE).

El segundo paso es representar el conjunto de textos con un modelo "bolsa-de-palabras" *(bag-of-words)*: en términos prácticos, una tabla con una fila para cada texto y una columna para cada posible palabra. Cada celda contiene números que indican cuántas veces aparece cada palabra en cada texto (notar que por construcción la matriz contiene muchos ceros porque la mayoría de las palabras de un diccionario extenso no aparecen en la mayoría de los textos).

A continuación, el algoritmo LDA procesa esta representación "bolsa-de-palabras" para tratar de identificar 128 temas diferentes de que se trata en la base de datos[9] y asignar a cada texto la probabilidad de que pertenezca a cada uno de los esos temas. Intuitivamente, el algoritmo analiza los textos y determina qué palabras tienden a aparecer juntas y cuáles no, asignándolas de manera óptima a diferentes temas para minimizar la distancia entre textos asignados a un tema determinado y maximizar la distancia entre textos asignados a diferentes temas. El resultado final es una base de datos que contiene, para cada trimestre de 1997 a 2018, el porcentaje de artículos que tratan cada uno de los 128 temas identificados por el modelo de aprendizaje no supervisado. Además, cada artículo se procesa también con un diccionario de términos positivos y negativos, y los resultados se agregan en indicadores trimestrales que

[9] En los modelos LDA, el investigador debe elegir el número de temas que quiere extraer. En general, la cantidad de temas se elige minimizando medidas de la bondad del modelo LDA.

representen en qué medida los artículos relacionados con cada tema tienen un tono positivo o negativo.

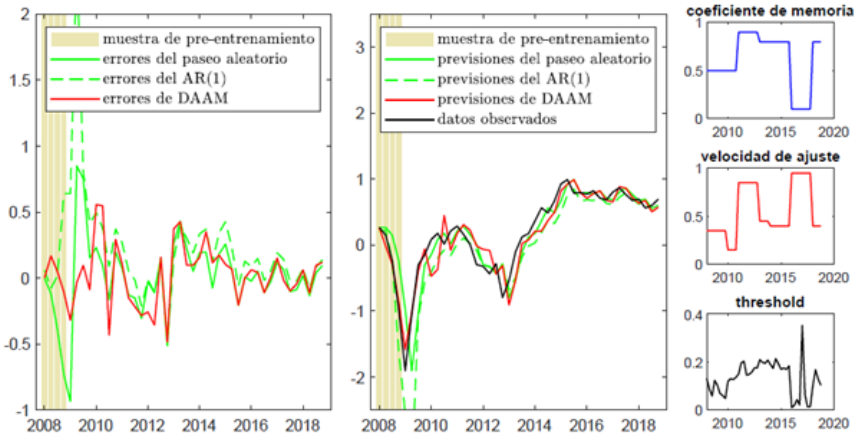
En la segunda parte del ejercicio, se recurre a un modelo de aprendizaje automático para explotar los datos resultantes del método LDA y obtener previsiones del PIB español. El aprendizaje automático se refiere a una gama muy amplia de métodos y algoritmos utilizados en diferentes campos, como la visión artificial. En el contexto de la economía, se utilizan máquinas de vectores de soporte, bosques aleatorios y redes neuronales para analizar microdatos sobre millones de consumidores o empresas y encontrar correlaciones, patrones de comportamiento e incluso relaciones causales. Para la previsión de series temporales, se utilizan técnicas *ensemble* (como *boosting* y *bagging*) para construir modelos de previsión sólidos combinando de manera óptima una gran cantidad de modelos más débiles (para más detalles sobre estas técnicas, véase Barrow y Crone, 2016).

Para este ejercicio se utiliza uno de estos modelos ensemble para hacer la previsión del PIB: un modelo de agregación doblemente adaptativo que utiliza los resultados del método LDA (denominado DAAM-LDA, de su acrónimo en inglés). Una ventaja de este modelo es que puede adaptarse a cambios en las relaciones entre los datos a lo largo del tiempo. Los ingredientes del modelo son un conjunto de 128 modelos de series temporales simples y de bajo desempeño; en particular, cada modelo es la regresión del crecimiento trimestral del PIB español sobre su valor retrasado y el peso de cada tema en cada trimestre, y su tono prevalente, representado por un indicador de positividad y negatividad. La estimación se hace en tiempo real, de manera que los modelos se estiman cada trimestre y se guarda la previsión del primer trimestre fuera de la muestra. La mayoría de los estos modelos muestran un desempeño débil fuera de la muestra por dos razones: (1) el peso de cada tema en la prensa y su positividad o negatividad son indicadores con una relación señal/ruido relativamente baja; (2) la mayoría de los temas identificados por el método LDA no están realmente relacionados con la economía. Efectivamente, solo cuatro de los 128 modelos muestran un mejor desempeño que un simple paseo aleatorio. La ventaja de los métodos ensemble es que están diseñados específicamente para construir modelos sólidos a partir de un conjunto de modelos débiles. Además, el investigador no tiene que decidir qué temas son útiles y cuáles no: en cada periodo el modelo descarta automáticamente cualquier tema que no proporciona buenas previsiones en los períodos anteriores. En este caso, se ha decidido combinar las previsiones a través de una función umbral a un parámetro que varía a lo largo del tiempo[10]. Es decir, el modelo "ensemble" analiza el desempeño reciente de cada modelo de partida para decidir si utilizarlo en el cómputo del promedio del siguiente trimestre o no (pero todos los modelos que se consideran en el promedio tienen igual peso).

[10] Una manera más sofisticada seria construir una función de ponderación no lineal que trasforme el desempeño de cada modelo en el trimestre actual en su ponderación óptima para el trimestre sucesivo. En este caso, la función de ponderación óptima es muy parecida a una función umbral a un parámetro, lo que explica que se haya elegido esta simplificación en este ejercicio.

FIGURA 6

RESULTADOS DEL EJERCICIO DE PREVISIÓN EN TIEMPO REAL DEL CRECIMIENTO DEL PIB ESPAÑOL TRIMESTRAL



Nota: El DAAM-LDA es el modelo de agregación doblemente adaptativo con datos LDA presentado en esta sección.

Fuente: Banco de España.

La figura 6 resume los resultados de este experimento y también muestra el umbral que se utiliza en cada momento del tiempo, así como el parámetro de memoria y la velocidad de ajuste del umbral óptimo cada año. El modelo DAAM-LDA proporciona mejores previsiones respecto a un modelo de paseo aleatorio, incluso si solo cuatro de los 128 modelos débiles que utiliza como ingredientes realmente lo hacen (véase tabla 1). Si nos centramos en los últimos cuatro años de la muestra (2015-2018), los resultados se pueden comparar también con los más recientes modelos de previsión del PIB a corto plazo actualmente en uso en el Banco de España (la previsión oficial del Banco de España y el modelo Spain-STING, véase Camacho y Perez-Quiros, 2011). Para este período de muestra, el modelo DAAM-LDA funciona mejor que el paseo aleatorio, el

TABLA 1

PREVISIONES DEL PIB ESPAÑOL: RAÍZ CUADRADA DEL ERROR CUADRÁTICO MEDIO EN EJERCICIOS FUERA DE MUESTRA EN TIEMPO REAL

	RW	AR (1)	BdE	DAAM-LDA	Spain-Sting
2008-2018	0,29	0,476	0,082	0,24	---
2015-2018	0,11	0,155	0,76	0,097	0,121

Nota: La tabla muestra la raíz cuadrada del error cuadrático medio fuera de muestra para las previsiones de crecimiento del PIB trimestral español obtenidas por los siguientes modelos (es orden): la caminata aleatoria, el modelo AR (1) simple, la previsión oficial del Banco de España, el modelo de agregación doblemente adaptativo con datos LDA y el modelo Spain-STING.

Fuente: Banco de España.

modelo AR (1) simple y el modelo Spain-STING. No obstante, las previsiones oficiales del Banco de España muestran un desempeño superior en comparación con los métodos estadísticos considerados en esta sección.

3.3.3. Previsión del consumo privado con Google-trends, tarjetas de crédito e indicadores de incertidumbre

El ejercicio presentado en esta sección resume el trabajo de Gil *et al.* (2018). El objetivo del trabajo es averiguar si las nuevas fuentes de información pueden ayudar a predecir el consumo privado de los hogares. Normalmente, los datos oficiales para la medición de las decisiones de gasto de los hogares privados son los datos de las cuentas nacionales que están disponibles con una frecuencia trimestral (datos hard). Por esta razón, los datos cuantitativos se suelen combinar con indicadores cualitativos, de naturaleza más cualitativa, pero con frecuencia más elevada (véase la discusión a principio del apartado 3.3. de este capítulo). El objetivo de este ejercicio es testar el poder predictivo de nuevas fuentes de datos juntos con los datos más tradicionales, tanto hard como soft.

En particular, se consideran las siguientes fuentes de datos mensuales: (1) datos de cajeros automáticos (ATM), que comprenden a la retirada de efectivo en terminales de cajeros automáticos y pagos en puntos de venta (POS) con tarjetas de débito y crédito; (2) indicadores Google-trends, que proporcionan indicadores del comportamiento de consumo basados en patrones de búsqueda en Internet proporcionados por Google; (3) medidas de incertidumbre económica y de política[11].

Estos indicadores se combinan con otros tradicionales cuantitativos (afiliados a la seguridad social; Índice de comercio minorista; Índice de actividad en los servicios) y cualitativos (índice PMI de servicios; Índice de confianza del consumidor) en un modelo de múltiples frecuencias en el que los indicadores entran en frecuencia mensual, y la variable objetivo, el consumo privado de la contabilidad nacional, lo hace en trimestral. Para calcular el desempeño de cada grupo de indicadores, se consideran distintos modelos, que se diferencian en el conjunto de indicadores incluidos en cada uno. Los modelos estimados incluyen indicadores de cada grupo por separado, varios grupos al mismo tiempo y diferentes combinaciones de modelos individuales. Como referencia, se considera un modelo de paseo aleatorio en el cual se incluye en los trimestres futuros la última tasa de crecimiento trimestral observada para el consumo privado. Se evalúa el desempeño de cada modelo para la previsión en el horizonte a corto plazo (trimestre actual, *nowcast*), y también de a 1 a 4 trimestres.

Los principales resultados se muestran en la tabla 1, y pueden resumirse como sigue. En primer lugar, entre los modelos que utilizan únicamente indicadores de cada grupo,

[11] Medido alternativamente por el índice de volatilidad bursátil IBEX y el índice EPU basado en texto proporcionado por Baker, Bloom y Davis (2016) para España.

los que utilizan indicadores cuantitativos y tarjetas de pago ofrecen un mejor desempeño que los demás en las previsiones a corto plazo y, algo menos, en las previsiones a 1 y a 4 trimestres (véase Panel A del cuadro). Los errores cuadráticos medios relativos (RMSE) son en casi todos los casos menores que uno, aunque solo en algunos casos las previsiones de los modelos son estadísticamente diferentes de las obtenidas por el modelo de paseo aleatorio trimestral. En general, los otros modelos no mejoran sistemáticamente la previsión de un paseo aleatorio trimestral. Las dos principales excepciones son el modelo con indicadores cualitativos para los horizontes de previsión a corto plazo y los basados en Google Trends para las previsiones a más largo plazo. Estos resultados son coherentes con el supuesto que los indicadores de Google Trends proporcionan información actual sobre los proyectos de compras futuras.

En segundo lugar, el Panel B del cuadro muestra los resultados de la estimación de modelos que incluyen indicadores cuantitativos agregando, a la vez, indicadores que pertenecen a los otros grupos (indicadores cualitativos, tarjetas de pago, indicadores de incertidumbre, indicadores de Google Trends). En general, la precisión de las previsiones a corto plazo no mejora cuando se incluyen más indicadores, excepto para los indicadores soft. No obstante, las previsiones más a largo plazo mejoran significativamente cuando se expande el modelo de referencia, especialmente cuando se añaden los indicadores de incertidumbre y los de Google trends para las previsiones a 4 trimestres.

En tercer lugar, se destaca que la combinación (promedio) de modelos con grupos individuales de indicadores mejora el desempeño de la previsión en todos los casos y a todos los horizontes (ver Panel C del cuadro). En particular, la combinación de las previsiones de modelos que incluyen indicadores cuantitativos con aquellos con tarjetas de pago ofrece, en general, el mejor desempeño de previsión para todos los horizontes. Al mismo tiempo, agregar las previsiones obtenidas con indicadores cualitativos parece añadir valor en las previsiones a corto plazo. Además, combinar un amplio conjunto de modelos proporciona un RMSE menor respecto al obtenido por el paseo aleatorio trimestral en la previsión a cuatro trimestres.

En conclusión, Gil, Pérez, Sánchez y Urtasun (2018) muestran que aunque los indicadores tradicionales proporcionen una buena previsión del consumo privado en tiempo real, las nuevas fuentes de datos añaden valor, sobre todo aquellas relacionadas con tarjetas de crédito, pero también, en menor medida, los indicadores de Google Trends y los indicadores de incertidumbre, cuando se combinan con otras fuentes.

TABLA 1

ESTADÍSTICOS RMSE (RAÍZ CUADRADA DEL ERROR CUADRÁTICO MEDIO) RELATIVOS: RATIO DE CADA MODELO RESPECTO A UN PASEO ALEATORIO TRIMESTRAL [A]

Panel A: Modelos con indicadores de cada grupo:									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
Indicadores cuantitativos (``hard'') [b]	0.84	0.75 *	0.79	0.75 **	0.81	0.80	0.98	0.97	1.00
Indicadores cualitativos (``soft'') [c]	1.01	0.85	0.85	1.11	1.05	1.05	1.09	1.10	1.19 *
Tarjetas de pago (cuantía, cu) [d]	0.79	0.82	0.88	0.65 ***	0.84	0.69 **	0.74 **	0.84	0.83
Tarjetas de pago (numeros) [d]	1.05	1.15	1.13	0.90	1.10	0.98	0.75 **	0.81	0.79
Indicadores de incertidumbre [e]	1.06	0.97	0.99	1.00	1.05	1.06	0.94	1.00	1.02
Google: agregado de todos los indicadores	1.04	1.06	1.06	0.85	1.03	1.03	0.71 **	0.79	0.79
Google: bienes duraderos (retrasado)	1.04	0.97	0.98	0.96	1.04	1.04	0.85 *	0.93	0.93
Panel B: Modelos con indicadores de grupos diferentes:									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
cuantitativos \& cualitativos	0.69 **	0.78	0.77	0.67 ***	0.76 *	0.72 *	0.79 *	0.82 *	0.80 *
cuantitativos \& Tarjetas de pago (cu) [d]	0.90	0.82	0.91	0.67 ***	0.79	0.78	0.86	0.89	0.91
cuantitativos \& Incertidumbre	0.88	0.86	0.75	0.74 **	0.91	0.93	0.69 **	0.76	0.76
cuantitativos \& Google (agregado)	0.85	0.76	0.77	0.81 *	0.94	0.89	0.77 **	0.81 *	0.82
cuantitativos \& Google (duraderos)	0.91	0.95	0.87	0.69 **	0.83	0.88	0.72 **	0.76 *	0.77 *
Panel C: Combinación de modelos:									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
Todos los modelos [f]	0.66 **	0.71 **	0.69 **	0.65 ***	0.77 *	0.65 **	0.73 **	0.78 *	0.78 *
Hard \& Tarjetas de pago (cu) [d]	0.62 **	0.69 **	0.71 **	0.53 ***	0.69 **	0.52 ***	0.79 *	0.86	0.84
Hard \& Tarjetas de pago (cu) [d] \& Soft	0.65 **	0.67 **	0.67 **	0.65 ***	0.74 **	0.59 ***	0.83 *	0.89	0.92
Hard \& Soft	0.65 **	0.66 **	0.66 **	0.77 **	0.75 **	0.69 **	0.91	0.94	1.02
Hard \& Google (duraderos)	0.77 **	0.78 **	0.76 **	0.74 ***	0.83	0.78 *	0.85	0.91	0.90

Notas: Los asteriscos denotan los resultados del test de Diebold-Mariano, cuya hipótesis nula es que dos métodos de previsión proporcionan resultados de igual precisión. Se utiliza una función de pérdida cuadrática. El número en cada celda representa el diferencial de pérdida del método mencionado en su línea horizontal en comparación con la alternativa de paseo aleatorio trimestral. * (**) [***] Indica el rechazo de la hipótesis nula al nivel del 10% (5%) [1%] de significatividad. [a] Errores de previsión calculados como la diferencia con la primera publicación de los datos de consumo privado. Las previsiones se generan de forma recursiva durante la ventana móvil de 2008T1 (m1) a 2017T4 (m3). [b] Datos del registro de afiliados a la Seguridad Social; Índice de comercio minorista; Índice de actividad del sector servicios. [c] PMI de servicios; índice de confianza del consumidor. [d] Serie agregada de tarjetas de crédito vía POS y cajeros automáticos. [e] Volatilidad del mercado de valores (IBEX); índice de incertidumbre de la política económica (EPU). [f] Combinación de los resultados de 30 modelos, incluyendo modelos en los que se incluyen los indicadores de cada bloque por separado, modelos que incluyen el bloque cuantitativo y cada uno de los otros bloques, y versiones de todos los modelos mencionados que incluyan además valores retrasados de las variables.

Referencias

Aguilar P., Ghirelli, C., Pacce, M. y Urtasun, A. (2020). Can news help to measure economic sentiment? An application in Covid-19 times. *Documento de Trabajo,* No. 2027. Banco de España.

Ahir, H., Bloom, N. y Furceri, D. (2019). The World Uncertainty Index. *Working Paper,* 19–027. Standford Institute for Economic Policy Research.

Alonso, A. y Carbó, J. M. (2020). Machine Learning in Credit Risk: Measuring the Dilemma Between Prediction and Supervisory Cost. *Documento de Trabajo*, No. 2032. Banco de España.

Aprigliano V., Ardizzi, G. y Monteforte, L. (2017). Using the payment system data to forecast the Italian GDP. *Working paper,* No. 1098. Bank of Italy.

Artola C. y Galán, E. (2012). Tracking the future on the web: construction of leading indicators using internet searches. *Documento de Trabajo,* No. 1203. Banco de España.

Azqueta-Gavaldon A., Hirschbühl, D. Onorante, L. y Saiz, L. (2020). Sources of economic policy uncertainty in the euro area: an unsupervised machine learning approach. *Working Paper,* No. 2359. European Central Bank.

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

Banco de España (2017). *Survey of Household Finances, 2014: methods, results and changes since 2011.* Artículo Analítico No. 1/2017. Banco de España.

— (2018). Central Balance Sheet Data Office. Annual results of non-financial corporations 2017. Banco de España.

Barrow, D. K. y Crone, S. (2016). A comparison of AdaBoost algorithms for time series forecast combination. *International Journal of Forecasting,* 32(4), pp. 1103-1119.

Bhattarai, S., Chatterjee, A. y Park, W. Y. (2019). Global spillover effects of US uncertainty. *Journal of Monetary Economics*. https://doi.org/10.1016/j.jmoneco.2019.05.008

Biljanovska, N., Grigoli, F. y Hengge, M. (2017). Fear Thy Neighbor: Spillovers from Economic Policy Uncertainty. *Working Paper* No., 17/240. International Monetary Fund.

Blei, D. M., Ng, A. Y. y Jordan, M. I. (January 2003). Lafferty, John (ed.). Latent Dirichlet Allocation. *Journal of Machine Learning Research.*

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives,* 28(2), pp. 153-176.

Bodas, D., García, J., Murillo, J. Pacce, M., Rodrigo, T., Ruiz, P., Ulloa, C. Romero, J. y Valero, H. (2018). Measuring Retail Trade Using Card Transactional Data. *Working Paper,* No. 18/03. BBVA Research.

Broeders, D. y Prenio, J. (2018). Innovative Technology in Financial Supervision (Suptech) - The Experience of Early Users. Financial Stability Institute Insights on Policy Implementation. *Working paper,* No. 9. Bank for International Settlements, July.

Camacho, M. y Perez-Quirós, G. (2011). Spain-Sting: Spain Short-Term Indicator of Growth. *The Manchester School,* 79, pp. 594–616.

Carlsen, M. y Storgaard, P. E. (2010). Dankort payments as a timely indicator of retail sales in Denmark. *Working paper,* No. 66. Bank of Denmark.

Carriére-Swallow Y. y Labbé, F. (2013). Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting,* 32, pp. 289-298.

Central banking (2019). *Big data in central banks: 2019 survey results.* https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central- banks-2019-survey-results

Colombo, V. (2013). Economic policy uncertainty in the US: Does it matter for the euro area? *Economics Letters,* 121(1), pp. 39-42.

Chetty, R., Friedman, J. y Rockoff, J. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *The American Economic Review,* 104(9), pp. 2633-2679.

D'Amuri F. y Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting,* 33(4), pp. 801-816.

Diaz-Sobrino, N., Ghirelli, C., Hurtado, S., Pérez, J. J. y Urtasun, A. (2020), The narrative about the economy as a shadow forecast: an analysis using Bank of Spain quarterly reports. *Documento de Trabajo* (próxima aparición). Banco de España.

Duarte, C., Rodrigues, P. M. y Rua, A. (2017). A mixed frequency approach to the forecasting of private consumption with ATM/POS data. *International Journal of Forecasting,* 33(1), pp. 61-75.

Einav, L. y Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy,* 14, pp. 1-24.

Farné, M. y Vouldis, A. T. (2018) A Methodology for Automatised Outlier Detection in High-Dimensional Datasets: An Application to Euro Area Banks' Supervisory Data. *Working Paper,* No. 2171. European Central Bank.

Fernández, A. (2019). Artificial intelligence in financial services. Analytical Articles. *Boletín Económico,* No. 2/2019. Banco de España.

Fernández-Villaverde, J., Hurtado, S. y Nuño, G. (2019). Financial Frictions and the Wealth Distribution. *Working Paper,* No. 26302. National Bureau of Economic Research. September.

Ferrara, L. y Simoni, A. (2019). When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage. *Working paper,* No. 2019-04. Center for Research in Economics and Statistics.

Fontaine, I., Didier, L. y Razafindravaosolonirina, J. (2017). Foreign policy uncertainty shocks and US macroeconomic activity: Evidence from China. *Economics Letters,* 155, pp. 121-125.

Ghirelli, C., Hurtado, S., Pérez, J. J. y Urtasun, A. (2020). New data sources for central banks. En: S. Consoli, D. Reforgiato Recupero, y M. Saisana, *Data Science for Economics and Finance: Methodologies and Applications.* Springer (próxima aparición).

Ghirelli, C., Peñalosa, J., Pérez, J. J. y Urtasun, A. (2019). Some implications of new data sources for economic analysis and official statistics. *Boletín Económico,* No. 2/2019. Banco de España. Mayo.

Ghirelli, C., Pérez, J. J. y Urtasun, A. (2019). A new economic policy uncertainty index for Spain. *Economics Letters,* 182, pp. 64-67.

— (2020). Economic Policy Uncertainty in Latin America. *Documento de Trabajo,* No. 2024. Banco de España.

Gil, M., Pérez, J. J., Sánchez, A. J. y Urtasun, A. (2018). Nowcasting Private Consumption: Traditional Indicators, Uncertainty Measures, Credit Cards and Some Internet Data. *Documento de Trabajo,* No. 1842. Banco de España.

Gotz, T. B. y Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting,* 35, pp. 45-66.

Hammer, C. L., Kostroch, D. C. y Quirós, G. (2017). Big Data: Potential, Challenges and Statistical Implications. *IMF Staff Discussion Note,* 17/06. International Monetary Fund.

Hardy, A., Hyslop, S., Booth, K. B., Aryal, J., Gretzel, U. y Eccleston, R. (2017). Tracking tourists' travel with smartphone-based GPS technology: a methodological discussion. *Information Technology & Tourism,* 17, pp. 255-274.

Kapetanios, G. y Papailias, F. (2018). Big Data & Macroeconomic Nowcasting: Methodological Review. *ESCoE Discussion Paper,* 2018-12. Economic Statistics Centre of Excellence.

Lacroix R. (2019). The Bank of France datalake. En: Bank for Inter- national Settlements (ed.), The use of big data analytics and artificial intelligence in central banking, *IFC Bulletins,* vol. 50. Bank for International Settlements.

Loberto, M., Luciani, A. y Pangallo, M. (2018). The potential of big housing data: an application to the Italian real-estate market. *Working paper,* No. 1171. Bank of Italy.

Menéndez, A. y Mulino, M. (2018). Results of non-financial corporations in the first half of 2018. *Boletín Económico,* No. 3/2018. Banco de España.

Meinen, P. y Roehe, O. (2017). On measuring uncertainty and its impact on investment: Cross-country evidence from the euro area. *European Economic Review,* 92, pp. 161-179.

Moreno Bernal, A. y González Pedraz, C. (2020). Análisis de sentimiento del informe de estabilidad financiera. *Documento de Trabajo,* No.2011. Banco de España.

Mueller, H. y Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review,* 112(2), pp. 358-375.

Nyman R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P. y Smith, R. (2018). News and narratives in financial systems: exploiting big data for systemic risk assessment. *Staff Working Paper,* No. 704. Bank of England.

Pew Research Center (2012). Assessing the Representativeness of Public Opinion Surveys. Mimeo. https://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

Petropoulos, A., Siakoulis, V., Stavroulakis, E. y Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. En: Bank for International Settlements (ed.), The use of big data analytics and artificial intelligence in central banking, *IFC Bulletins,* vol. 50. Bank for International Settlements.

Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of business cycles, *Journal of Business & Economic Statistics,* pp.1-17.

Trung, N. B. (2019). The spillover effect of the US uncertainty on emerging economies: a panel VAR approach. *Applied Economics Letters*, 26(3), pp. 210-216.

Apéndice

EXPLICACIONES ADICIONALES SOBRE EL CONTENIDO DE LAS FIGURAS

Figura 2. Respuesta del PIB de España ante distintas perturbaciones de incertidumbre construidas con versiones alternativas del índice EPU

La figura muestra la función impulso-respuesta de la tasa de crecimiento del PIB español hasta diez trimestres después de un *shock* positivo de una desviación estándar en el índice EPU para España, y se ha elaborado a partir del modelo ilustrado en Ghirelli, Pérez y Urtasun (2019). Se trata de un modelo VAR estimado por MCO que incluye las siguientes variables endógenas: el indicador EPU (en niveles), el diferencial de la deuda soberana española a 10 años sobre el bund alemán, el PIB real (en frecuencia trimestral y en tasas de crecimiento) [a su vez, el consumo real agregado de los hogares o la inversión real de bienes de capital], la tasa de inflación (tasas de crecimiento trimestrales del IPC) y una variable que controla por la incertidumbre global representada por el índice EPU global de Baker, Bloom y Davis (2016), que se asume exógena. En este modelo se incluyen un numero óptimo de retardos que se identifican a partir del criterio de la información de Schwarz. Para la identificación de los impulsos-respuestas de una variable sobre las otras variables endógenas se utiliza la descomposición de Cholesky de la forma reducida de la matriz de varianza y covarianza.

En la figura 2 se investiga el papel relativo de (1) la riqueza de palabras clave, (2) la cobertura de prensa y (3) cobertura temporal en la conducción de los resultados. La figura compara las respuestas del PIB al impulso de las versiones alternativas del EPU en las que se modifica una de las dimensiones mencionadas anteriormente a la vez, pasando de la versión EPU-BBD (el EPU construido como en Baker, Bloom y Davis, 2016) al nuevo índice EPU, propuesto en Ghirelli, Pérez y Urtasun (2019). La línea *"EPU-Nuevas palabras, vieja Cobertura; >2000"* se refiere a un índice que se construye usando las palabras clave de Ghirelli, Pérez y Urtasun (2019), pero manteniendo la cobertura de prensa y de tiempo con las del índice EPU-BBD. Por tanto, comparar esta respuesta con la de EPU-BBD permite comprobar la relevancia de enriquecer las palabras clave en las búsquedas. Del mismo modo, para la línea *"EPU-Nuevas palabras, nueva Cobertura; >2000"* se utilizan las palabras clave y la cobertura de prensa de Ghirelli, Pérez y Urtasun (2019), pero se mantiene la cobertura de tiempo como en el índice EPU-BBD. Comparando las respuestas de los choques en *"EPU-Nuevas palabras, vieja Cobertura; >2000"* y *"EPU-Nuevas palabras, nueva Cobertura; >2000",* se puede apreciar la contribución de ampliar la cobertura de prensa al construir el índice. Finalmente, comparando los resultados de *"EPU-Nuevas palabras, nueva Cobertura; >2000"* con los obtenidos con EPU-NEW se puede comprobar la importancia de incrementar el periodo temporal. De acuerdo con la figura, todas las dimensiones (i) - (iii) son importantes, ya

que todas contribuyen a obtener el signo negativo esperado en las respuestas del PIB. Además, la definición de la cobertura del periodo es clave para mejorar la precisión de las estimaciones y obtener resultados significativos.

Figura 4. Previsiones cuantitativas e indicador de sentimiento de los informes del Banco de España

La figura 4 muestra el indicador de sentimiento del *Boletín Económico Trimestral* del Banco de España frente a las previsiones cuantitativas elaboradas por el Banco de España, y se ha elaborado en Ghirelli, Hurtado, Pérez y Urtasun (2020). Para ello, se considera el *Boletín Económico Trimestral de la economía española* publicado online por el Banco de España desde el primer trimestre de 1999. En concreto, se considera el apartado inicial que contiene los principales mensajes. Con esta información se construye un indicador de sentimiento del primer trimestre de 1999.

El índice de sentimiento se construye de la siguiente manera. Primero, se crea un diccionario de términos positivos y negativos en castellano. Para construir la lista de palabras, se lee una muestra de los informes considerados y se identifican los términos que se usan más frecuentemente para describir la situación macroeconómica (considerando adjetivos, adverbios, verbos y sustantivos). Se seleccionan 47 términos (raíces de palabras), eliminando las terminaciones de las palabras. En segundo lugar, se asigna una puntuación igual a +1 (-1) a las palabras que expresan un sentimiento positivo (negativo). En tercer lugar, se cuenta cuántas veces aparece cada palabra del diccionario en cada texto y se pondera cada resultado con su puntuación. En quinto lugar, se suman todas las apariciones ponderadas de cada texto y se divide el número resultante por la longitud total del texto.

Figura 5. Un indicador textual para anticipar la evolución del PIB de España

La imagen de la derecha del figura 5 muestra el indicador textual de la economía con frecuencia trimestral frente a la tasa de crecimiento del PIB español hasta junio de 2019. La figura de la izquierda muestra el mismo indicador textual con frecuencia semanal de enero a marzo de 2020. El indicador textual de la economía se ilustra en Aguilar *et al.* (2020).

Dicho indicador textual es un indicador de sentimiento basado en el análisis de texto de artículos publicados en periódicos españoles desde 1997. Se computa utilizando la base de datos de Factiva y se tienen en cuenta siete periódicos nacionales *(ABC, El País, El Mundo, La Vanguardia, Expansión, Cinco Días, El Economista).* Básicamente, este indicador captura el tono económico de los artículos de noticias publicados en la prensa

española, reflejando el equilibrio entre el número de noticias que contienen palabras clave relacionadas con repuntes y recesiones en el ciclo económico español. Para su construcción se siguen los siguientes pasos.

En primer lugar, con frecuencia mensual, se hacen tres tipos de búsquedas en cada uno de los periódicos mencionados:

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones positivas (pos): se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recuperacion*" o una de las siguientes palabras (aceler*, crec*, increment*, recuper*, aument*, expansi*, mejora*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de cinco palabras entre sí. Se ignora "crecimiento económico" dado que tiene un tono neutro (se usa para describir indiferentemente un crecimiento negativo o positivo).

- Se cuenta el número de artículos en el que se habla del crecimiento de la economía con connotaciones negativas (neg): se consideran los artículos que satisfacen simultáneamente los siguientes criterios: (1) hablan sobre noticias económicas o financieras; (2) mencionan la palabra "Españ*"; (3) mencionan "recesión*" o "crisis" o una de las siguientes palabras (descen*, ralentiz*, redu*, disminu*, contraccion*, decrec*, desaceler*) siempre que aparezcan acompañadas de los términos "economía" o "económic*" a una distancia máxima de cinco palabras entre sí.

- Se cuenta el número de artículos que hablan sobre noticias económicas o financieras (total).

En segundo lugar, para cada periódico, se calcula la siguiente proporción: (pos-neg)/total. A esta proporción se la estandariza utilizando el periodo comprendido entre enero 1997 y febrero 2020. De esta manera la volatilidad de las distintas series es comparable entre los distintos periódicos. Por último, se computa una serie agregada tomando la media entre las series de los distintos periódicos y se le quita la media del periodo comprendido entre enero 1997 y febrero 2020.