BANCO DE ESPAÑA
Eurosistema

BOLETÍN ECONÓMICO 2/2019
ARTÍCULOS ANALÍTICOS

22 de mayo de 2019

Algunas implicaciones de las nuevas fuentes de datos para el análisis económico y la estadística oficial

Corinna Ghirelli, Juan Peñalosa, Javier J. Pérez y Alberto Urtasun

Resumen

Las nuevas tecnologías están permitiendo la aparición de fuentes de datos de muy alta frecuencia, de mayor granularidad que las tradicionales y de acceso generalizado, en muchos casos, por parte de los distintos agentes económicos. Estos desarrollos abren nuevas vías y nuevas oportunidades a la estadística oficial y al análisis económico. Desde el punto de vista de un banco central, el uso y la incorporación de estos datos a sus tareas tradicionales plantea retos significativos, derivados de su gestión, almacenamiento, seguridad y confidencialidad. Además, existen problemas con su representatividad estadística. Por su parte, dado que estos datos están a disposición de muchos agentes, y no exclusivamente de las instituciones estadísticas oficiales, existe el riesgo de que se generen distintas medidas sobre el mismo fenómeno con estándares de calidad heterogéneos, lo que puede crear confusión en la opinión pública. Algunas de estas fuentes, con formato no estructurado, como las textuales, requieren el uso de nuevas técnicas de procesamiento para que puedan incorporarse en el análisis económico con un formato adecuado (cuantitativo). Su uso, además, determina la incorporación de técnicas de aprendizaje automático, entre otras, al conjunto tradicional de metodologías de análisis. En este artículo se revisan, desde el punto de vista de un banco central, algunas de las potencialidades e implicaciones de este nuevo fenómeno para el análisis económico y la estadística oficial, y se aportan ejemplos de trabajos recientes.

Palabras clave: nuevas fuentes de información económica, *big data,* ciencia de datos, aprendizaje automático, análisis de texto.

Códigos JEL: C10, C18, C50, C80 y D80.

ALGUNAS IMPLICACIONES DE LAS NUEVAS FUENTES DE DATOS PARA EL ANÁLISIS ECONÓMICO Y LA ESTADÍSTICA OFICIAL

Este artículo ha sido elaborado por Corinna Ghirelli, Juan Peñalosa, Javier J. Pérez y Alberto Urtasun, de la Dirección General de Economía y Estadística.

Introducción

En la última década, el desarrollo de las nuevas tecnologías y de las redes sociales ha propiciado la aparición de nuevas fuentes de datos, comúnmente denominadas *big data.* Estos nuevos datos presentan una serie de características particulares en cuanto a su volumen y su grado de detalle (muy superiores a las fuentes tradicionales), su disponibilidad a altas frecuencias y su carácter en muchas ocasiones no estructurado (esto es, no necesariamente numérico ni organizado, como, por ejemplo, los datos provenientes de textos o imágenes). En los últimos años han aparecido numerosas aplicaciones de estas nuevas fuentes de información en el ámbito económico-financiero —y de los bancos centrales en particular—, tanto en lo relativo a la ampliación de la información de base utilizada para el desarrollo de sus funciones como en lo concerniente al análisis económico-prudencial y la supervisión bancaria que estos desarrollan [véase, por ejemplo, Broeders y Prenio (2018) y Fernández (2019)], así como desarrollos metodológicos de interés [véase Fernández-Villaverde *et al.* (2019)].

En el ámbito concreto del análisis económico, estas nuevas fuentes de datos presentan un potencial significativo, incluso teniendo en cuenta que los bancos centrales ya realizan un uso especialmente intensivo de datos estadísticos, tanto individuales (microdatos) como agregados (macroeconómicos), para el desarrollo de sus funciones. En particular, estas fuentes permiten:

— Un mejor conocimiento y un seguimiento más ágil de la realidad económica, con un menor desfase de los fenómenos que se quieren analizar y con un mayor grado de detalle. En concreto, se dispone de información relevante y, en ocasiones, en tiempo real para la predicción económica de variables clave, como el PIB, el consumo privado o el empleo.

— Recoger información sobre variables difíciles de medir (como el sentimiento o las expectativas), pero que son fundamentales para la toma de decisiones de los agentes económicos.

— Una mejor evaluación de la política económica y más posibilidades de simular medidas alternativas, debido, especialmente, a que se dispone de microdatos que eventualmente podrían permitir una mejor caracterización de la heterogeneidad de los agentes y, por tanto, un análisis más profundo y preciso de su comportamiento.

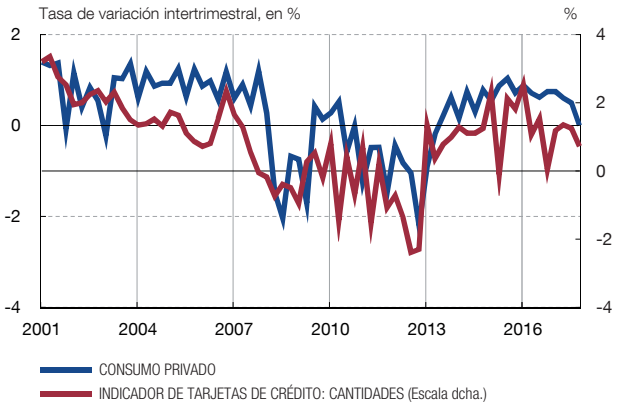
El resto del artículo se divide en dos partes. En la primera, que incluye los recuadros 1 y 2, se presentan las principales aportaciones de la literatura reciente, y se destacan algunas aplicaciones novedosas. En la última parte, se examinan las implicaciones de este nuevo fenómeno para la estadística oficial.

Nuevas fuentes de datos para el análisis económico

Los bancos centrales utilizan de manera intensiva bases de datos estructurados en el desarrollo de sus funciones. Los datos estructurados son aquellos que ya están clasificados y ordenados y que pueden utilizarse de manera inmediata por parte de los economistas

CONSUMO PRIVADO Y OPERACIONES CON TARJETAS DE CRÉDITO GRÁFICO 1

1 INDICADOR DE TARJETAS DE CRÉDITO/DÉBITO: CANTIDADES (a)



2 INDICADOR DE TARJETAS DE CRÉDITO/DÉBITO: NÚMERO DE OPERACIONES (a)

FUENTES: Contabilidad Nacional Trimestral (INE) y Gil, Pérez, Sánchez y Urtasun (2018).

a Agregado de gasto en puntos de venta y retirada de efectivo en cajeros.

Descargar

de los bancos centrales. Algunos ejemplos del ámbito de los datos individuales (microeconómicos) son los relativos a los balances de las empresas [por ejemplo, de la Central de Balances del Banco de España; véase Menéndez y Mulino (2018) o Banco de España (2018)], la información acerca del volumen de los créditos otorgados por las instituciones financieras a particulares y empresas (por ejemplo, los trabajos que usan la Central de Información de Riesgos del Banco de España[1]), o los datos relativos a las decisiones financieras de los agentes recogidos en la Encuesta Financiera de los Hogares [véase Banco de España (2017)]. La frecuencia de publicación de estas bases de datos suele ser anual o trimestral. En el ámbito de la macroeconomía, la principal fuente de información la constituye la Contabilidad Nacional, que en el caso español es elaborada y difundida por el Instituto Nacional de Estadística y el Banco de España, aunque se publica numerosa información relativa a la situación económica y financiera con una frecuencia más alta, principalmente mensual, pero también superior.

El desarrollo de las nuevas tecnologías permite expandir las fuentes de información, con una mayor granularidad y una mayor frecuencia. Por un lado, la mayor granularidad está asociada al volumen de la información disponible. Gracias a las nuevas tecnologías se puede obtener información acerca de cada acción realizada por parte de personas o empresas (es decir, el nivel más desagregado posible). Un ejemplo lo constituyen las transacciones realizadas con tarjetas bancarias. Por otro lado, la mayor frecuencia hace referencia a que esos datos se actualizan con alta velocidad, diariamente o incluso en tiempo real. En este sentido, siguiendo con el ejemplo anterior, las transacciones con tarjetas de crédito, que se pueden utilizar para aproximar las pautas de consumo de las familias, se encuentran disponibles, potencialmente, en tiempo real, con un coste muy limitado de explotación de la información, en particular si se compara con el coste asociado a la realización de encuestas a hogares a escala nacional. A título de ejemplo, el gráfico 1 permite apreciar cómo las operaciones con tarjeta de crédito evolucionan de

1 La Central de Información de Riesgos del Banco de España (CIRBE) proporciona datos confidenciales e información muy detallada sobre casi la totalidad de los préstamos otorgados por las entidades de crédito en España (todos los préstamos de más de 6.000 euros).

manera muy parecida a cómo lo hace el consumo de las familias para el caso español [véanse Gil *et al.* (2018) y Bodas *et al.* (2018)].

La disponibilidad de cantidades enormes de nueva información presenta retos significativos para la infraestructura de gestión, almacenamiento, seguridad y uso confidencial de estos datos:

— En primer lugar, una gestión ordenada de datos de distinta naturaleza requiere una infraestructura flexible que permita almacenar tanto datos estructurados (en formato de tabla) como datos no estructurados (sin formato específico), y al mismo tiempo disponer de herramientas analíticas para visualizar la información relevante que contienen.

— Es necesario también el desarrollo de sistemas de seguridad adecuados para proteger la privacidad de las diferentes bases de datos.

— Por su alta frecuencia, el volumen de los datos no estructurados aumenta de manera exponencial, lo que requiere buscar un sistema que minimice los costes de almacenaje, sin perjuicio de los requisitos de seguridad.

— Por último, una gestión óptima de los datos no estructurados supone la integración de nuevos perfiles profesionales en los bancos centrales (científicos de datos) y exige una colaboración más estrecha entre diferentes áreas de estas instituciones; por ejemplo, las dedicadas a sistemas de información, estadística y economía.

Por su parte, la distinta naturaleza de algunas de las nuevas fuentes de información hace necesaria la asimilación y el desarrollo de técnicas que trasformen y sinteticen los datos, trasladándolos a formatos que se puedan incorporar al análisis económico. En especial, en el caso de los datos no estructurados, la información recogida no tiene especial valor hasta que no se ha procesado. A título de ejemplo, las técnicas de análisis textual permiten procesar la información contenida en textos y convertirla en datos estructurados (para una breve descripción de esas metodologías, véase el recuadro 2). Entre estos nuevos tipos de información, se pueden destacar los siguientes:

— *Tendencias de búsqueda de Google (Google Trends).* Se trata de una herramienta ofrecida por Google que permite acceder al volumen de las búsquedas realizadas por los usuarios en la Red acerca de determinados términos de interés.

— *Bases de datos on line de medios de comunicación.* Numerosos servidores en la red almacenan textos de distintas fuentes de información, como artículos de periódicos y de revistas, entre otros. De aquí se puede extraer información acerca de los temas publicados en la prensa y del tratamiento realizado.

— *Redes sociales* (por ejemplo, Facebook y Twitter). A partir de los mensajes de los usuarios en las redes sociales, se pueden extraer cuáles son las opiniones prevalentes y el tono general de las discusiones en la Red.

— *Portales de búsqueda en la Red.* En este grupo, por ejemplo, se encontrarían los portales concebidos para la búsqueda de viviendas o de ofertas de

empleo, que permiten extraer información acerca del mercado inmobiliario o laboral, respectivamente.

— *Datos de teléfonos móviles.* A través de sus aplicaciones y de la combinación de los datos de geoposicionamiento, el uso del móvil permite aprender sobre las costumbres, actividades y movimientos de los usuarios.

— *Datos de satélites.* Las imágenes satelitales facilitan, por ejemplo, la medición de las zonas agrícolas en países poco desarrollados o del consumo nocturno de energía eléctrica[2].

No obstante, las nuevas fuentes de datos presentan algunos problemas que aconsejan utilizarlas de modo informado. En particular [véase Einav y Levin (2014)], pueden existir problemas de representatividad de las muestras, por ejemplo, de aquellas que reflejan información acerca de los usuarios que manejan las nuevas tecnologías, que no necesariamente representan la población de interés. Además, la disponibilidad masiva de información no estructurada requiere el uso de técnicas estadísticas y de aprendizaje automático que permitan resumirla; de este modo, ganan prominencia los modelos predictivos automáticos, frente a la aproximación del análisis empírico tradicional en el que la teoría económica suele guiar a los analistas en la elección de las variables. Así, en modelos predictivos con *big data,* suele ser el modelo el que selecciona las variables que son más relevantes[3]. Por último, estas fuentes son novedosas y es difícil determinar en qué medida su representatividad va a mantenerse en el tiempo. Esto plantea riesgos a la inversión en términos de la investigación que puede hacerse para explotar una fuente específica.

En el recuadro 1 se realiza una revisión de un conjunto amplio de trabajos recientes en los que se aplican nuevas técnicas y nuevos datos al análisis económico.

Implicaciones de las nuevas fuentes de datos para la estadística oficial

Las nuevas fuentes de datos también abren posibilidades hasta ahora sin explotar para la elaboración de estadísticas[4]. Es cierto que su capacidad para afectar al ámbito más específico de los bancos centrales, que es el de las estadísticas financieras, es limitada, dado que estas se basan en gran medida en registros administrativos (como los balances bancarios o la declaración de las empresas de sus transacciones con el exterior), que difícilmente pueden ser suplidos por fuentes alternativas y cuya disponibilidad está asegurada al estar los agentes obligados a enviar regularmente esa información a las autoridades de regulación o supervisión. Frente a estas fuentes estructuradas, utilizadas desde hace mucho tiempo con fines estadísticos, la nueva información (en gran medida, desestructurada) desempeñaría, sobre todo, un papel complementario, orientado posiblemente a fines muy concretos o de corte más cualitativo (como la detección de carencias de información en determinados segmentos.

Es importante destacar que en el ámbito de la estadística oficial se viene trabajando con información granular desde hace mucho tiempo. Existen estadísticas en las que se puede descender a un alto grado de detalle, como en el caso de las emisiones y de las carteras

2 El consumo nocturno de energía eléctrica se considera como una buena *proxy* del desarrollo de un país y se ha demostrado que es particularmente útil para los países en desarrollo con baja disponibilidad de datos tradicionales de contabilidad nacional.

3 Véanse Belloni *et al.* (2012) y Belloni, Chernozhukov y Hansen (2014) como ejemplos de aplicaciones en las que modelos de aprendizaje automático mejoran estudios de inferencia causal, identificando el número óptimo de variables de control o de instrumentos a partir de un vasto conjunto de variables potenciales.

4 Véase, entre otros, Bean (2016).

de valores, donde hay información con el detalle valor a valor, los préstamos individuales concedidos por el sistema bancario, las transacciones y posiciones de los agentes nacionales frente al resto del mundo, los balances individuales de las sociedades no financieras, etc. En la estadística de los bancos centrales hay, además, una tendencia creciente a la obtención y explotación de estos microdatos, que pueden contribuir a mejorar la calidad de las estadísticas oficiales, al permitir un contraste más claro de los datos en la fuente y un análisis de consistencia con las variables agregadas. Además, la disponibilidad de esa información permite analizar distintas características de la información (instrumentos, plazos, contrapartidas o monedas de denominación, entre otras muchas) sin necesidad de volver a requerirla, con el consiguiente ahorro de costes para los informantes, pues los microdatos suelen presentar múltiples dimensiones válidas para el análisis. Con todo, este aumento de la disponibilidad de microdatos exige, al mismo tiempo, un desarrollo de nuevos medios técnicos para que su explotación derive en una mayor calidad de las estadísticas.

Un efecto positivo para la estadística oficial del auge de estas nuevas fuentes de datos procede del desarrollo del instrumental técnico que se está produciendo para manejar esa ingente información. Estas nuevas herramientas —que incluyen técnicas de inteligencia artificial, aprendizaje automático y análisis de datos— pueden ser aprovechadas por la estadística oficial para el manejo de los microdatos estructurados, especialmente para la mejora de su calidad (por ejemplo, para la detección y depuración de valores anómalos) o para conciliar información que se recibe de fuentes distintas y con diferente frecuencia.

Posiblemente, en el ámbito de las estadísticas no financieras ha habido más iniciativas para explotar las nuevas fuentes de información. Entre estas destacan las relativas a la medición de precios (mediante técnicas de *web scraping)* o de algunas partidas del comercio exterior (estimando, por ejemplo, los movimientos de turistas mediante el seguimiento de las redes de telefonía móvil). En los países en vías de desarrollo, que tienen más dificultades para establecer una infraestructura estadística sólida, estas nuevas fuentes están comenzando a utilizarse incluso para realizar estimaciones de algunos agregados de la Contabilidad Nacional [véase Hammer *et al.* (2017)].

Finalmente, debe reseñarse que la abundancia de información disponible supone también un reto para la estadística oficial, pues permite que, a partir de esos datos, cualquier usuario sea capaz de generar su propia medición de algún fenómeno económico y de difundirla. Frente a esta eventual competencia del sector privado en la generación de información caracterizada como «estadística», y que puede venir de fuentes tan poco fiables como las redes sociales, es necesario reforzar el marco de calidad y de transparencia de las estadísticas oficiales. La estadística oficial responde a una metodología consolidada y comparable internacionalmente, que es la base para poder evaluar de un modo objetivo tanto la situación económica como la respuesta adecuada de la política económica. En este sentido, las autoridades estadísticas deberían mostrar de modo transparente los métodos de elaboración de las estadísticas oficiales y contrarrestar el potencial efecto de las *fake news* sobre la medición de variables económicas, diseñando una política de comunicación que permita desmontar los bulos antes de que arraiguen.

22.5.2019.

BIBLIOGRAFÍA

ACCORNERO, M., y M. MOSCATELLI (2018). *Listening to the Buzz: Social Media Sentiment and Retail Depositors,* Working Paper, n.º 1165, Banca d'Italia, febrero.

APRIGLIANO, V., G. ARDIZZI y L. MONTEFORTE (2017). *Using the Payment System Data to Forecast the Italian GDP,* Working Paper 1098, Banca d'Italia.

ARTOLA, C., y E. GALÁN (2012). *Las huellas del futuro están en la web: construcción de indicadores adelantados a partir de las búsquedas en internet,* Documentos Ocasionales, n.º 1203, Banco de España.
BAKER, S. R., N. BLOOM y S. J. DAVIS (2016). «Measuring Economic Policy Uncertainty», *The Quarterly Journal of Economics,* 131, pp. 1593-1636.
BANCO DE ESPAÑA (2018). *Central de Balances. Resultados anuales de las empresas no financieras, 2017.*
— (2017). *Encuesta Financiera de las Familias (EFF) 2014: métodos, resultados y cambios desde 2011,* Artículos Analíticos, Banco de España, enero.
BEAN, C. (2016). *Independent Review of UK Economic Statistics,* ONS, Reino Unido.
BELLONI, A., D. CHEN, V. CHERNOZHUKOV y C. HANSEN, (2012). «Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain», *Econometrica,* 80 (6), pp. 2369-2429.
BELLONI, A., V. CHERNOZHUKOV y C. HANSEN (2014). «Inference on Treatment Effects After Selection Among High-Dimensional Controls», *The Review of Economic Studies,* 81 (2), pp. 608-650.
BHOLAT, D., S. HANSEN, P. SANTOS y C. SCHONHARDT-BAILEY (2015). *Text mining for central banks: handbook,* Centre for Central banking Studies (33), pp. 1-19.
BODAS, D., J. GARCÍA, J. MURILLO, M. PACCE, T. RODRIGO, P. RUIZ, C. ULLOA, J. ROMERO y H. VALERO (2018). *Measuring Retail Trade Using Card Transactional Data,* Working Paper 18/03, BBVA Research.
BROEDERS, D., y J. PRENIO (2018). «Innovative Technology in Financial Supervision (Suptech) - The Experience of Early Users», *Financial Stability Institute Insights on policy implementation,* n.º 9, Bank for International Settlements, julio.
CENTRAL BANKING (2019). *UK Statistics Body to Launch High-Speed Indicators,* disponible en https://www.centralbanking.com/central-banks/economics/data/4094151/uk-statistics-body-to-launch-high-speed-indicators.
CHOI, H., y H. VARIAN (2012). «Predicting the Present with Google Trends», *Economic Record,* 88(1), pp. 2-9.
EINAV, L., y J. LEVIN (2014). «The Data Revolution and Economic Analysis», *Innovation Policy and the Economy,* 14, pp. 1-24.
FERNÁNDEZ, A. (2019). «Inteligencia artificial en los servicios financieros», Artículos Analíticos, *Boletín Económico,* Banco de España, de próxima aparición.
FERNÁNDEZ VILLAVERDE, J., S. HURTADO y G. NUÑO (2019). *Financial Frictions and the Wealth Distribution,* mimeo, Banco de España.
GARCÍA-URIBE, S. (2018). *The Effects of Tax Changes on Economic Activity: a Narrative Approach to Frequent Anticipations,* Documentos de Trabajo, n.º 1828, Banco de España.
GHIRELLI, C., J. J. PÉREZ y A. URTASUN (2019). *A New Economic Policy Uncertainty Index for Spain,* Documento de Trabajo, n.º 1906, Banco de España. Versión revisada próxima a aparecer en Economics Letters.
GIL, M., J. J. PÉREZ, A. J. SÁNCHEZ y A. URTASUN (2018). *Nowcasting Private Consumption: Traditional Indicators, Uncertainty Measures, Credit Cards and Some Internet Data,* Documentos de Trabajo, n.º 1842, Banco de España.
HAMMER, C. L., D. C. KOSTROCH y G. QUIRÓS (2017). *Big Data: Potential, Challenges and Statistical Implications,* IMF Staff Discussion Note, 17/06.
LOBERTO, M., A. LUCIANI y M. PANGALLO (2018). *The Potential of Big Housing Data: an Application to the Italian Real-Estate Market,* Working Paper, n.º 1171, Banca d'Italia.
MENÉNDEZ, Á., y M. MULINO (2018). «Resultados de las empresas no financieras hasta el segundo trimestre de 2018», *Boletín Económico,* 3/2018, Banco de España.
NYMAN, R., S. KAPADIA, D. TUCKETT, D. GREGORY, P. ORMEROD y R. SMITH (2018). *News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment,* Staff Working Paper n.º 704, Bank of England.
THORSRUD, L. (2018). «Words are the New Numbers: A Newsy Coincident Index of the Business Cycle», *Journal of Business & Economic Statistics.*
TURRELL, A., B. SPEIGNER, J. DJUMALIEVA, D. COPPLE y J. THURGOOD (2018). *Using Job Vacancies to Understand the Effects of Labour Market Mismatch on UK Output and Productivity,* Staff Working Paper n.º 737, Bank of England, julio.

ALGUNOS EJEMPLOS DE APLICACIONES RECIENTES CON NUEVAS FUENTES DE DATOS EN EL ÁMBITO DEL ANÁLISIS ECONÓMICO

RECUADRO 1

Textos de prensa y otros textos financieros. Baker, Bloom and Davis (2016) construyen un indicador de incertidumbre política *(Economic Policy Index,* EPU) para Estados Unidos, calculado como el volumen de los artículos publicados en la prensa que contienen palabras relacionadas con los conceptos de incertidumbre, economía y política [para España, véase Ghirelli *et al.* (2019)]. Por su parte, Thorsrud (2018), a partir de los artículos publicados en el principal periódico económico de Noruega, construye un indicador de los temas económico-financieros relevantes y demuestra que esta información, junto con variables tradicionales, permite mejorar las predicciones del PIB. Por otro lado, García-Uribe (2018), a partir de datos sobre las noticias de Estados Unidos, construye un indicador de anuncios relativos a la aprobación de cambios impositivos, demostrando que las noticias asociadas a anuncios de reducciones de impuestos estimulan la actividad económica. Finalmente, Nyman *et al.* (2018) construyen un índice relacionado con el tono de los mercados financieros («ansiedad» y «excitación»), a partir de comunicaciones del Banco de Inglaterra, informes de agencias financieras y noticias de Reuters. Este indicador presenta poder explicativo para anticipar movimientos en variables económicas y financieras.

Twitter. Accornero y Moscatelli (2018) construyen un indicador de sentimiento económico a partir de tuits que mencionan las mayores instituciones financieras italianas, mostrando cómo el análisis del tono de los tuits es útil para mejorar la predicción de la evolución de los depósitos minoristas.

Google Trends. Por su parte, Choi y Varian (2012) muestran cómo indicadores de Google Trends mejoran las predicciones a corto plazo del consumo privado en aquellos casos en que las compras se planifican con adelanto (por ejemplo, los bienes duraderos). Gil *et al.* (2018) usan esta herramienta para construir índices de consumo de bienes duraderos y no duraderos que presentan capacidad de anticipación de las decisiones de consumo de los agentes. Artola y Galán (2012) utilizan la misma herramienta para construir un indicador del flujo de turistas británicos en España —los principales clientes de la industria turística española—. Este indicador permite mejorar la predicción de variables de demanda y actividad en España.

Tarjetas de crédito. Aprigliano *et al.* (2017), a partir de datos sobre el sistema de pago electrónico, muestran cómo la información acerca de los flujos electrónicos de pago mejora las predicciones del PIB frente a modelos con indicadores tradicionales del ciclo económico. Por su parte, Gil *et al.* (2018) valoran la contribución para la predicción a corto plazo del consumo de los hogares de un conjunto amplio de nuevas fuentes de datos. Entre sus resultados destaca el relativo a los indicadores sobre operaciones con tarjetas de crédito, que presentan una alta capacidad predictiva (véase gráfico 1 del texto principal). Entre los trabajos disponibles en la literatura se encuentra también Bodas *et al.* (2018). Estos autores utilizan datos de todas las transacciones con tarjetas de débito y de crédito del BBVA, efectuadas en España, para replicar el índice tradicional de ventas minoristas del INE. Esto les permite, además, construir índices de consumo minorista en ámbitos en los que no se dispone de estadística oficial, como el regional o el sectorial.

Portales de búsquedas online. Loberto *et al.* (2018), a partir de anuncios de alquiler y de compra de viviendas del principal portal de servicios inmobiliarios de Italia, estudian la evolución de los precios de las viviendas en segmentos específicos del mercado inmobiliario, e identifican las características de las viviendas que son más relevantes para explicar su precio. En el ámbito del mercado de trabajo, Turrell *et al.* (2018), a partir de datos de anuncios de ofertas de trabajo de un portal de contratación de empleados, analizan la existencia de desajustes en el mercado laboral, en el aumento en el ámbito tanto ocupacional como regional, en la evolución reciente del mercado laboral británico.

BREVE DESCRIPCIÓN DE LAS TÉCNICAS DE ANÁLISIS DE TEXTOS Y UNA APLICACIÓN PARA EL CASO ESPAÑOL

RECUADRO 2

Las aplicaciones con técnicas de análisis de textos *(text mining)* han adquirido una importancia especial en el ámbito del análisis económico. Estas técnicas permiten extraer información relevante de los textos, que se puede sintetizar y codificar bajo la forma de indicadores cuantitativos. En un primer momento se prepara el texto *(pre-processing),* en particular quitando la parte del texto que no es informativa para el análisis (artículos, palabras no relevantes, números, caracteres raros), así como las terminaciones de las palabras, para quedarse con su raíz. En una segunda etapa, se sintetiza la información contenida en las palabras mediante indicadores cuantitativos, obtenidos, básicamente, del cálculo de la frecuencia de las palabras o de los grupos de palabras. Intuitivamente, la frecuencia relativa de grupos de palabras que pertenecen a un determinado asunto permite valorar la importancia relativa de dicho asunto en el texto.

Las técnicas de explotación de textos se pueden clasificar de la siguiente manera:

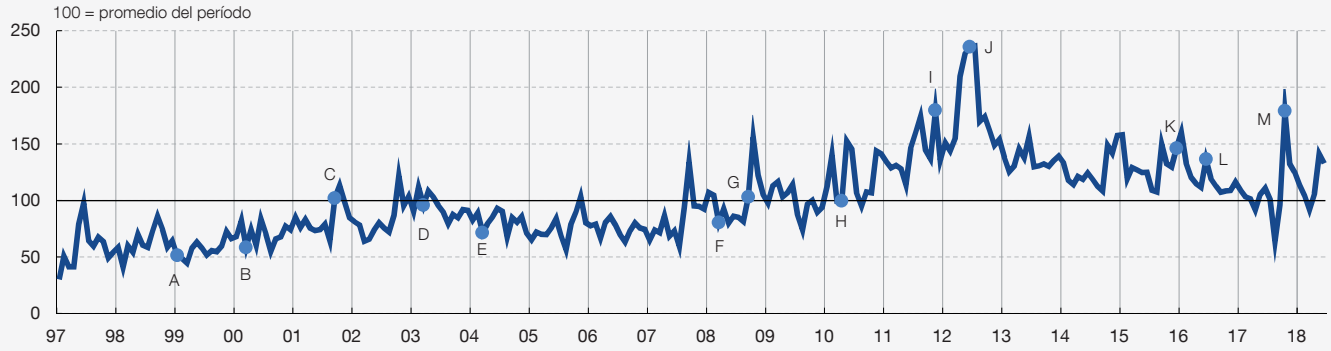
— *Búsquedas con operadores lógicos* (booleanos)[1]. Se trata de búsquedas de palabras clave en textos con operadores lógicos (entre los que destacan «y» y «o»), que establecen relaciones simples entre los términos de búsqueda. El resultado es el total de los textos que cumplen los requisitos de la búsqueda. Estas se pueden efectuar a través de los motores de búsqueda o directorios de bases de datos, y no necesitan que el investigador tenga acceso a los textos completos.

1 Operadores básicos de búsqueda lógica o *booleana* son, en inglés, «and», «or» y «not».

Gráfico 1
INCERTIDUMBRE ECONÓMICA Y SOBRE LA POLÍTICA ECONÓMICA: EVENTOS DESTACADOS

El indicador muestra mayores movimientos en las fechas de los eventos asociados a variaciones más elevadas de la incertidumbre en España.

INDICADOR DE INCERTIDUMBRE ECONÓMICA Y SOBRE LAS POLÍTICAS ECONÓMICAS



Selección de eventos destacados

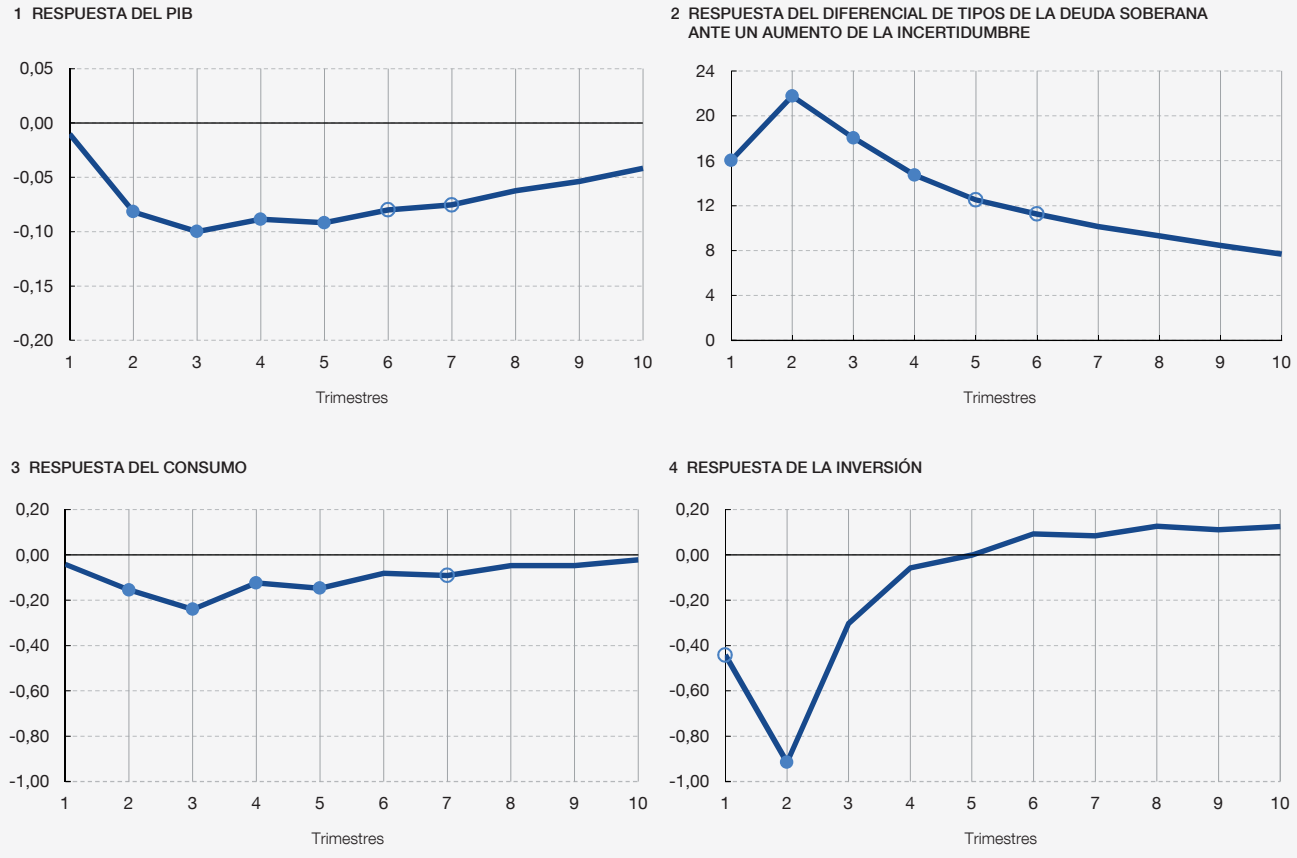
A	Enero de 1999	Constitución de la UEM
B	Marzo de 2000	Elecciones generales en España
C	Septiembre de 2001	Ataques terroristas del 11-S
D	Marzo de 2003	Invasión de Irak
E	Marzo de 2004	Ataques terroristas del 11-M y elecciones generales en España
F	Marzo de 2008	Rescate del Bear Sterns y elecciones generales en España
G	Septiembre de 2008	Quiebra de Lehman Brothers
H	Abril de 2010	Grecia solicita ayuda financiera
I	Noviembre de 2011	Elecciones generales en España
J	Junio de 2012	Ayuda financiera a España
K	Diciembre de 2015	Elecciones generales en España
L	Junio de 2016	Referéndum en el Reino Unido *(brexit)* y elecciones generales en España
M	Octubre de 2016	Crisis catalana

FUENTE: Ghirelli, Pérez y Urtasun (2019).

— *Análisis de diccionario.* También el análisis de diccionario se basa en búsquedas sobre bases de datos de textos, pero requiere que el investigador tenga acceso a los textos. En este caso, las búsquedas proporcionan la frecuencia relativa de las palabras clave en cada texto, que es una *proxy* de la relevancia de los temas que estas representan. Este análisis se utiliza en los estudios de sentimiento económico. En concreto, se necesita un listado de palabras positivas y negativas, y con este se calcula la frecuencia de los términos positivos y de los negativos en un texto o porción de texto. El índice de sentimiento se define como la diferencia entre esas frecuencias: es decir, el texto tiene un sentimiento positivo (negativo) cuando la frecuencia de los términos positivos es superior (inferior) a la de los términos negativos.

— *Técnicas de «temas latentes»* (Topic Modelling). Permiten descubrir los temas principales en un cuerpo de textos sin que el investigador tenga que especificar las palabras clave de interés. Este último es el límite del análisis de diccionario, porque las palabras clave se restringen a las ideas previas que el investigador tenía antes de hacer el análisis de los temas. En cambio, los modelos de temas latentes permiten el descubrimiento de temas que el investigador no había identificado, necesariamente, *ex ante.* El supuesto de base es que cada texto es representado por una mezcla de temas no observados, que el investigador tiene que identificar. Para ello, se estudia la frecuencia de las palabras dentro del cuerpo. Intuitivamente, patrones repetidos en que las mismas palabras coinciden en los mismos textos identifica los temas. Por

Gráfico 2
EFECTOS DE UN AUMENTO DE LA INCERTIDUMBRE ECONÓMICA Y SOBRE LAS POLÍTICAS ECONÓMICAS (a)



FUENTE: Ghirelli, Pérez y Urtasun (2019).
NOTA: ● indica significatividad estadística al 5 % y ○ indica significatividad estadística al 10 %.

a En el modelo VAR se incluyen como variables endógenas la incertidumbre medida por los indicadores sintéticos de los mercados financieros, de desacuerdo y de incertidumbre sobre las políticas económicas, el PIB/consumo/inversión, el diferencial de la deuda soberana española respecto al bono alemán y un índice de precios, y como variables exógenas, la volatilidad del EUROSTOXX-50, el EPU para el conjunto de la UE y un indicador sintético de incertidumbre europea (calculado de manera similar a como se calculan los índices sintéticos para España).

ejemplo, si palabras como «hospital», «enfermo» y «salud» aparecen en los mismos textos, el modelo reconoce ese como un tema en sí, representado por esas palabras[2].

La medición de la incertidumbre económica inducida por la política económica es uno de los ámbitos de relevancia para el análisis macroeconómico que ha estudiado la literatura reciente. Para el caso español, el trabajo más reciente es el de Ghirelli *et al.* (2019). En este artículo, siguiendo la metodología de Baker *et al.* (2016), los autores construyen un indicador de incertidumbre acerca de las políticas económicas del país y encuentran una relación dinámica significativa entre este indicador y las principales variables macroeconómicas. En concreto, a partir de búsquedas de palabras clave en la prensa española, se cuenta cada mes el total de los artículos de periódico que contienen al mismo tiempo términos relacionados con los conceptos de incertidumbre, economía y política económica. El índice se basa en siete periódicos nacionales, de entre los más leídos, lo que ofrece una cobertura de prensa muy extensa: los cuatro periódicos generalistas más leídos en el país y los tres periódicos financieros de mayor difusión.

El indicador presenta aumentos o reducciones significativos en torno a eventos asociados, *ex ante,* con subidas o reducciones de la incertidumbre económica. Cabe destacar que el indicador recoge los principales eventos de las últimas décadas que se podrían asociar a variaciones significativas de la incertidumbre (véase gráfico 1). Por ejemplo, entre otros, los ataques terroristas del 11 de septiembre de 2001 en Estados Unidos, la quiebra de Lehman Brothers en septiembre de 2008, la solicitud de ayuda financiera por parte de Grecia en abril de 2010, la petición de ayuda financiera para la reestructuración del sector bancario y de las cajas de ahorros por parte de España en junio de 2012, el referéndum del *brexit* en junio de 2016, o los episodios de tensionamiento político en la Comunidad Autónoma de Cataluña en octubre de 2017.

Asimismo, se estima que aumentos inesperados del indicador de incertidumbre sobre las políticas económicas conllevan efectos macroeconómicos adversos. Los gráficos 1 y 2 ilustran los resultados principales de los ejercicios realizados y proporcionan la respuesta de las principales variables macroeconómicas (el PIB, la prima de riesgo española, el consumo privado y la inversión en equipo) frente a una perturbación inesperada de la incertidumbre (véase gráfico 2). En particular, un aumento no esperado de la incertidumbre causaría una reducción significativa del PIB, el consumo y la inversión, mientras que la prima de riesgo se incrementaría.

2 Dentro de la familia *Topic Modelling* hay muchos modelos: el más popular es el modelo LDA *(Latent Dirilichet Allocation);* para una panorámica de los modelos, véase Bholat *et al.* (2015).