CAPÍTULO V

Éxitos y retos de *big data* en análisis económico: un recorrido a través de ejemplos

Pilar Poncela*
Eva Senra

La ingente cantidad de información disponible presenta soluciones y retos a problemas existentes en análisis económico. Cada vez son más las aplicaciones de éxito, bien basadas en actualización de la metodología estadística disponible, bien en la utilización de nuevas bases de datos. El reto pendiente es pasar de aplicaciones puntuales de éxito del uso de big data a su utilización generalizada por parte de los responsables del análisis económico. Presentamos diversos ejemplos (integración financiera, *nowcasting* y generación de nuevos indicadores de innovación y movilidad), señalando algunas oportunidades que big data proporciona y apuntando algunos retos que quedan por resolver.

Palabras clave: análisis de conglomerados, calidad, indicadores, procesamiento natural del lenguaje.

* Se agradece el apoyo financiero del Ministerio de Ciencia, proyectos números MINECO/FEDER PID2019-107161GB-C32 y PID2019-108079GB-C22/AEI/10.13039/501100011033.

1. INTRODUCCIÓN

La revolución del *big data* nos proporciona una ingente cantidad de información generada por los propios individuos a partir de sus interacciones con las redes sociales (texto, fotos y vídeos), publicaciones en blogs, texto de búsquedas en internet o repositorios, datos empresariales como las transacciones comerciales (*e-comercio,* tarjetas bancarias, móviles), información generada por máquinas, a través del conocido como Internet de las Cosas (IOT) que nos permite disponer de datos de sensores (tráfico, tiempo, cámaras de seguridad), o dispositivos de rastreo (GPS, localización del móvil), entre otros. Esta revolución también ha dado lugar a la necesidad de adaptar las técnicas estadísticas para su aplicación a bases de datos masivos.

Las profesiones relacionadas con el análisis de datos se encuentran entre las más demandadas en estos momentos, entre otros motivos, debido al big data. Las grandes empresas ya incorporan estos perfiles en sus compañías, las convierten en *data-driven* y reconocen la necesidad de almacenar y preservar sus datos, organizarlos y definir el tipo de objetivos e indicadores que necesitan. Y las que no lo hacen tienen en mente la necesidad de gestionar sus datos y no perder oportunidades de negocio a corto, medio y largo plazo.

Si bien el problema de la toma de decisiones dentro del contexto de una empresa es complejo, la construcción y utilización de indicadores basados en big data que puedan ser utilizados en la estadística oficial y sean útiles en la toma de decisiones de política económica es todavía un reto pendiente.

Este capítulo tiene dos objetivos: primero, presentar varios ejemplos donde se han utilizado técnicas de big data en análisis económico y, segundo, reflexionar cuándo esto no constituye una moda pasajera, o una colección de casos de éxito, sino una revolución que ha venido para quedarse e incorporarse de manera sistemática en su aplicación en política económica.

El primer objetivo ilustra, por un lado, la necesidad de adaptación de las técnicas estadísticas en un contexto de mayor información mediante una aplicación a la integración financiera utilizando datos diarios de bolsa a través del análisis de conglomerados o *cluster*. Dada la gran abundancia de datos, es necesario restringir la comparación entre los distintos índices bursátiles solamente a las dinámicas relevantes, ilustrando cómo se puede modificar el análisis clásico de conglomerados de series temporales a grandes conjuntos de datos.

Por otro lado, se ilustra la construcción de indicadores a partir de bases de datos alternativas a través de tres ejemplos, que responden a necesidades diferentes. Así, el segundo ejemplo analiza si la incorporación de la información contenida en las noticias es útil para generar predicciones de corto plazo, o incluso del trimestre actual o pasado, antes de la publicación del dato oficial de los principales indicadores macroeconómicos (*nowcasting*). En tercer lugar, se ilustra un procedimiento de generación de nuevos indicadores de innovación basados en la minería de textos. Esto es de gran necesidad, puesto que la rápida innovación en algunos sectores impide disponer de información oficial sobre esta actividad. Finalmente, se presenta

un ejercicio experimental de integración de los datos de geolocalización de los teléfonos móviles, por parte del Instituto Nacional de Estadística (INE), para poder estimar en tiempo real la movilidad de forma alternativa a como se venía haciendo mediante respuestas a encuestas ligadas al Censo cada 10 años. A consecuencia de la pandemia, desde el sector privado, Google y Apple han hecho públicos indicadores de movilidad basados en la información de sus usuarios. El ejercicio emprendido por el INE ha permitido dar respuesta igualmente a la necesidad de proporcionar indicadores de movilidad en tiempo real desde el sector público.

Los ejemplos anteriores no pretenden ser una lista exahustiva del uso de datos masivos en análisis económico, sino que ilustran diversos aspectos de su utilización tales como la modificación de técnicas estadísticas existentes, la incorporación de nuevas bases de datos o la generación de nuevos indicadores ante la presencia de datos masivos en casos concretos, siendo todos ellos casos de éxito.

El segundo objetivo de este capítulo es analizar indicios de calidad necesarios para que los ejemplos anteriores puedan conducir a una práctica sistemática de la utilización de big data en análisis económico. Los distintos indicadores presentados en las aplicaciones se enfrentan a diferentes retos de cara a su validación. Aquellos dirigidos a monitorizar la actividad en tiempo real y nowcasting tienen una clara metodología de validación basada en la evaluación de los errores de predicción con los métodos habituales de exactitud predictiva, una vez publicado el indicador oficial que pretenden adelantar. Por el contrario, la validación no es inmediata en aquellas situaciones en las que no existe un indicador de referencia oficial al que asemejarlo, como es el caso de la generación de nuevos indicadores de innovación. Por último, este también sería el caso si el Censo de 2021 no incluye las preguntas referentes a movilidad que permitan comparar las estimaciones obtenidas con big data con las oficiales.

Este capítulo se organiza de la siguiente forma. Las secciones 2 a 5 presentan diversos ejemplos de aplicación de big data. Así, la sección 2 analiza la adaptación de técnicas tradicionales a grandes conjuntos de datos a través de un estudio de integración financiera entre países. La sección 3 muestra la utilidad de bases de datos basadas en noticias para la monitorización de la actividad en tiempo real y la predicción a corto plazo. La sección 4 presenta cómo la necesidad de generar nuevos indicadores de fenómenos ágiles y muy específicos, como los relacionados con la innovación, lleva a considerar la utilización de bases de datos alternativas. La sección 5 recoge los resultados de un estudio experimental llevado a cabo por el INE en relación con la movilidad a través del uso de datos de geolocalización. La sección 6 analiza las ventajas de los indicadores presentados e introduce algunos de los retos a los que nos enfrentamos para validar la utilización de los mismos de manera sistemática en análisis económico.

2. INTEGRACIÓN FINANCIERA: ADAPTACIÓN DE TÉCNICAS TRADICIONALES

A través del análisis de la integración financiera a nivel europeo, se ilustra la modificación de una técnica estadística tradicional, el análisis de conglomerados o cluster, para

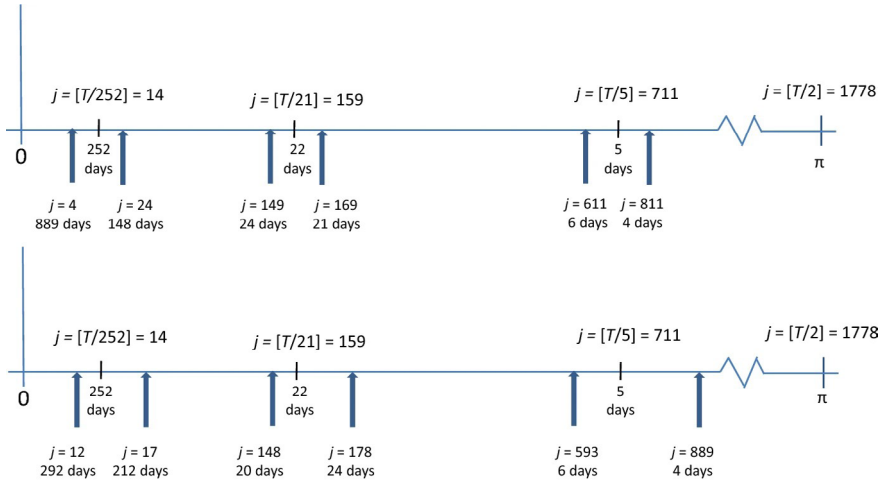
adaptarla al incremento del tamaño de la base de datos (Caiado, Crato y Poncela, 2020). El problema económico que justificó su adaptación es la necesidad de evaluar el grado de integración financiera en la Unión Europea (UE). La UE ha sido un gran catalizador para la liberación de los mercados en Europa. No obstante, la integración financiera no se ha conseguido aún y el sesgo nacional en las decisiones de inversión es notable. Tanto la Comisión Europea (CE) como el Banco Central Europeo (BCE) monitorizan el grado de integración financiera de manera continua (véase, por ejemplo, BCE, 2018, o Nardo *et al.,* 2017). Existen diversas formas de abordar dicha integración financiera, dependiendo de qué dimensión de la misma se desee analizar. Una de las formas es comprobar si se cumple la ley del precio único. Por ejemplo, si esta se da, el precio de un determinado activo debería ser el mismo en todos los mercados bursátiles. Para medir el grado de integración financiera a nivel europeo se analizan los grupos de índices financieros de acuerdo a su proximidad utilizando el análisis de conglomerados.

El análisis de conglomerados, aquí aplicado a series temporales, permite agrupar las mismas en base a su proximidad o semejanza. Esta se mide eligiendo una característica o propiedad P que defina el comportamiento de las series temporales y calculando la distancia entre las series atendiendo a esta propiedad. Piccolo (1990) propuso utilizar la distancia euclídea entre los coeficientes de la representación autorregresiva de las series. Desde una perspectiva no parámetrica, es decir, sin necesidad de estimar un modelo para las series, Galeano y Peña (2000) introdujeron medidas de distancia basadas en la función de autocorrelación de las mismas y, recientemente, Alonso y Peña (2019) han incluido la información sobre la dependencia lineal para formar los grupos. En la línea de las propuestas no paramétricas, Caiado *et al.* (2006) introdujeron los métodos en el dominio de la frecuencia y propusieron utilizar la distancia entre los periodogramas de las series. El periodograma mide la variabilidad (más precisamente el cuadrado de la amplitud de la onda) asociada a cada frecuencia. Es decir, descompone la varianza muestral asociada a una serie como la suma de las varianzas asociadas a cada frecuencia. Caiado *et al.* (2006) calculan la proximidad entre dos series calculando la distancia entre sus periodogramas. Dadas las observaciones de una serie temporal de longitud T, se define el periodograma para frecuencias angulares $w_j = (2\pi j)/T$, $j = 1, ..., [T/2]$, donde [z] denota la parte entera de z. Cuando el número de datos es muy elevado, para disminuir el número de cálculos que hay que realizar, Caiado *et al.* (2020) propusieron realizar el análisis de conglomerados calculando la distancia entre periodogramas no para todas las frecuencias sino sólo para aquellas asociadas a las principales fluctuaciones (aquellas donde la varianza es mayor). Esto puede ser de utilidad en macroeconomía para el análisis del ciclo de negocios o en finanzas. En este último caso es conocido que las series financieras diarias pueden presentar oscilaciones a las frecuencias diaria, semanal, mensual y anual. La propuesta de Corsi (2009) sobre modelos autorregresivos heterogéneos para volatilidad estocástica es un buen reflejo de este hecho, por lo que bastaría medir la distancia entre periodogramas sólo para ciertas frecuencias. Para ilustrar cómo se seleccionarían las frecuencias elegidas, supongamos, por ejemplo, que estamos analizando series diarias y estamos interesados solamente en el ciclo anual. Para series de tiempo de longitud T = 3556 (tamaño muestral de las series que posteriormente analizaremos en la aplicación sobre integración financiera), seleccionaremos ciclos de alrededor de 252 días laborables, es decir, fluctuaciones correspondientes a la frecuencia anual que, en este caso,

correspondería a la abscisa del periodograma j_s = T/252 = 14. Como es posible que exista cierta heterogeneidad en los ciclos anuales (por ejemplo, el número de días festivos no es el mismo en todos los países), nos gustaría seleccionar un intervalo alrededor de la frecuencia de interés, en este caso, I_{14}. Si este intervalo es simétrico alrededor de dicha frecuencia, por ejemplo, $I_{14} \pm 10 = [I_4; I_{24}]$ seleccionaríamos ciclos entre 148 y 889 días laborables. Aunque el intervalo es simétrico alrededor de la frecuencia de interés, es asimétrico en el número de días que consideramos alrededor de 252. Nuestra propuesta es usar intervalos que nos den ciclos de $\pm h$ días alrededor de un número de días dado, aunque resulten en intervalos asimétricos en la frecuencia. El mismo razonamiento se puede seguir para seleccionar el intervalo de frecuencias para captar el ciclo mensual y semanal. En la figura[1] 1 se han representado los ejes del periodograma. En el eje de abscisas, se señalan las principales frecuencias de variación de los índices considerados, así como intervalos simétricos en frecuencia (panel superior) o en el tiempo (panel inferior) alrededor de estas frecuencias. Únicamente compararíamos el periodograma para estos intervalos.

Figura 1.

Intervalos de frecuencias de muestreo en el periodograma. Panel superior: intervalos simétricos en frecuencia. Panel inferior: intervalos simétricos en el tiempo

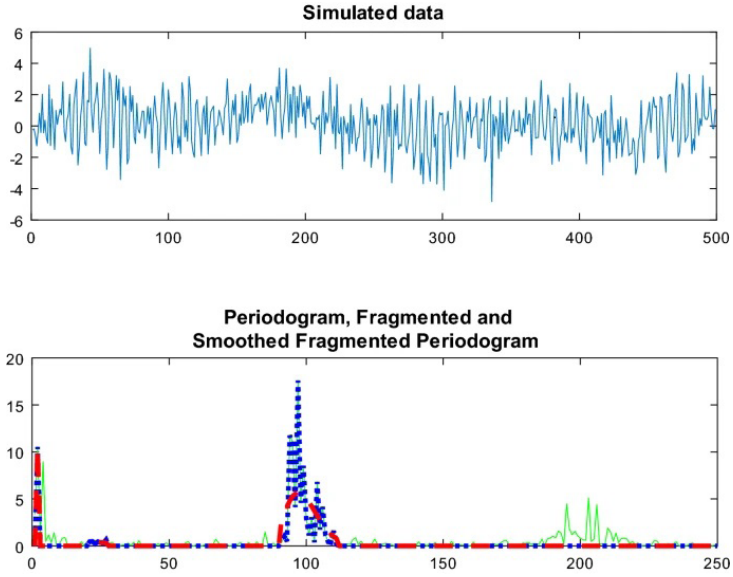


Para disminuir la varianza en la estimación del periodograma, Caiado *et al.* (2020) propusieron suavizarlo antes de proceder al cálculo de la distancia. La figura 2 muestra muestra en el panel superior un conjunto de datos simulados con el modelo de Corsi (2009) y en el panel inferior, el periodograma completo, el periodograma fragmentado para las frecuencias de interés y su suavizado. Mediante simulaciones, Caiado *et al.* (2020) comprobaron que usar el periodograma fragmentado suavizado aumenta la tasa de series correctamente clasificadas.

[1] Todos los gráficos de esta sección provienen del artículo de Caiado *et al.* (2020) y se reproducen aquí bajo licencia de Creative Commons.

Figura 2.

Panel superior: serie simulada con variación estacional semanal, mensual y anual. Panel inferior: periodograma (línea continua verde), periodograma fragmentado (línea punteada azul) y periodograma fragmentado suavizado (línea discontinua roja)



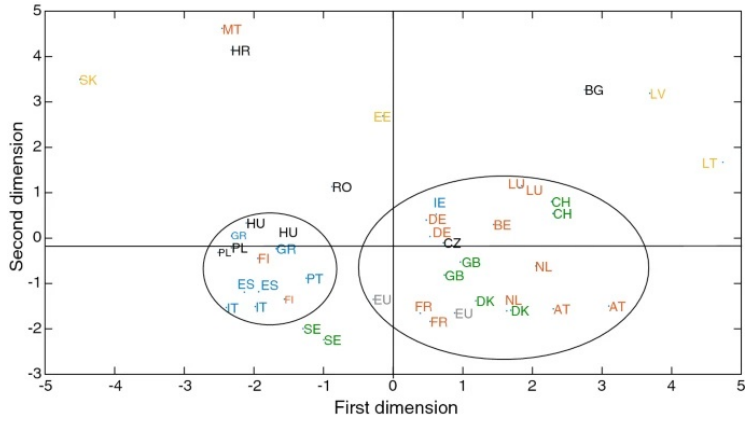
Caiado *et al.* (2020) aplican la propuesta anterior de clasificación para grandes conjuntos de datos a precios diarios de cierre de 44 mercados bursátiles[2] en Europa correspondientes al período comprendido entre el 2 de enero de 2003 y el 31 de diciembre de 2016. Para analizar el efecto que tuvo la crisis de deuda soberana en la integración financiera en Europa, se dividen las series en dos subperíodos, antes de la crisis (2 de enero de 2003 a 30 de junio de 2011) y después de la crisis de deuda soberana (1 de julio de 2011 a 31 de diciembre de 2016) y se analizan las series de retornos definidas como las tasas de variación relativas de los precios que se aproximan por las primeras diferencias del logaritmo. Se repite el siguiente ejercicio con cada submuestra para ver cómo cambian los resultados: se calcula el periodograma fragmentado suavizado de cada una de las 44 series y se calcula la matriz de distancias entre dichos periodogramas. Después de aplicar componentes principales a la matriz de distancias, se hace un gráfico de los 44 retornos financieros en función de las dos primeras componentes principales (escalado multidimensional). En la figura 3 se muestran los resultados correspondientes al período anterior a la crisis de deuda soberana y en la figura 4, después de la crisis. Se identifican las series con el acrónimo del país al que pertenece el índice bursátil[3]. Se recogen con distintos colores los clubs de países habitualmente analizados

[2] De algunos países, se consideran varios índices. Por ejemplo, para España, se dispone de las series correspondientes al IBEX 35 y al Índice General de la Bolsa de Madrid.

[3] Así, tanto el IBEX 35 como el Índice General de la Bolsa de Madrid aparecen como ES en las gráficas.

Figura 3.

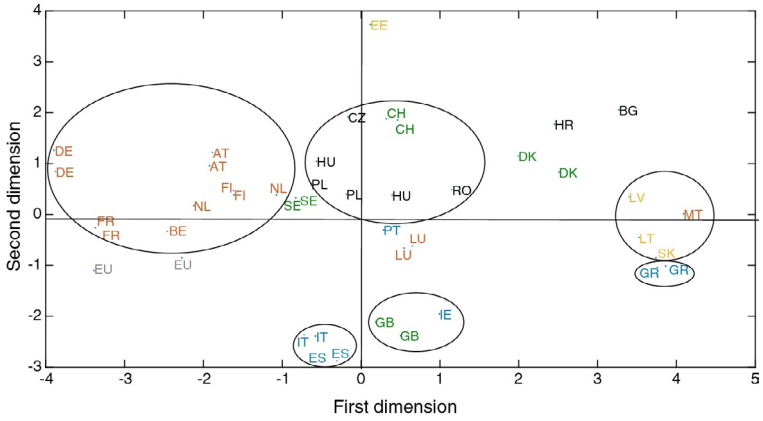
Escalado multidimensional de índices financieros en Europa antes de la crisis de deuda soberana. Muestra: 2/1/2003 a 30/6/2011



en integración financiera (Nardo *et al.,* 2017), rodeándose con un círculo los distintos grupos obtenidos mediante análisis de conglomerados jerárquico. En azul se muestran los países en dificultades o que han experimentado un deterioro significativo de la zona del euro (Irlanda, Grecia, España, Italia y Portugal); en rojo, los del núcleo de la zona del euro (Austria, Bélgica, Finlandia, Francia, Alemania, Luxemburgo y los Países Bajos); en verde, los del centro y oeste de Europa no pertenecientes a la zona del euro (Dinamarca, Gran Bretaña, Suecia y Suiza);

Figura 4.

Escalado multidimensional de índices financieros en Europa después de la crisis de deuda soberana. Muestra: 1/7/2011 a 31/12/2016



en naranja, los de la zona del este del euro (Estonia, Letonia, Lituania, Eslovaquia); en negro, los de la zona no euro oriental (Bulgaria, República Checa, Hungría, Polonia, Rumania y Croacia) y, finalmente, en gris, los índices europeos globales (eurostock50 y stxe600).

Desde el punto de vista de análisis económico, dichas figuras representan cómo la crisis de deuda soberana cambió el mapa de la integración financiera en Europa. Aunque hay indicios en la figura 3 de que existen dos grupos de países, los que no muestran dificultades financieras y los que sí lo hacen, en la figura 4 la separación entre países es mucho más clara. En el primer período aparecen en un mismo grupo los países sin dificultades, pertenezcan o no al área del euro, mientras que en la figura 4 los países del área euro sin dificultades forman un único grupo.

3. INDICADORES PARA LA MONITORIZACIÓN DE LA ACTIVIDAD ECONÓMICA EN TIEMPO REAL A TRAVÉS DE NOTICIAS

La generación de predicciones macroeconómicas, cada vez a horizonte temporal más corto, es clave en cuestiones de politica económica. Uno de los principales caballos de batalla de los modelos tradicionales de predicción macroeconómica es el retraso en la publicación de los datos oficiales. Esto ha dado lugar al desarrollo de modelos que incorporen información a distinta frecuencia mezclando, por ejemplo, datos mensuales y trimestrales para la generación de predicciones del producto interior bruto, y que tengan en cuenta los distintos calendarios de publicación de los datos incorporando la información tan pronto como está disponible, en vez de esperar a tener paneles balanceados de datos. Véase, por ejemplo, Camacho, Pérez-Quirós y Poncela (2013) y Banbura *et al.* (2013) para una revisión de los métodos de predicción de corto plazo y nowcasting. El uso de variables provenientes de encuestas (denominadas *soft data* en el argot), disponibles con anterioridad a las variables cuantitativas de actividad económica real (o *hard data*), sirve para paliar en parte la falta de estos últimos antes de su publicación en los modelos de predicción. Así, por ejemplo, Camacho y Pérez-Quirós (2010) utilizan un modelo factorial para sintetizar la información de diversos indicadores económicos a fin de generar predicciones de PIB de la eurozona que incorporan datos trimestrales (distintas estimaciones del PIB y desempleo) y mensuales, tanto *hard* (producción industrial,...) como *soft*. La versión para la economía española, Spain -STING, se recoge en Camacho y Pérez-Quirós (2011). Véase, por ejemplo, Stock y Watson (2016) o Bok *et al.* (2018) para una revisión de la literatura sobre el modelo factorial para la predicción macroeconómica y nowcasting o Poncela, Ruiz y Miranda (2021) quienes señalan la predicción en tiempo real y monitorización de la actividad económica como una de las principales aplicaciones del filtro de Kalman en análisis factorial dinámico revisando la literatura empírica sobre el tema.

No obstante, recientemente, han aparecido nuevas fuentes de datos alternativos, masivos, derivados del uso de tarjetas de crédito, noticias en prensa, en redes sociales, búsquedas por internet (por ejemplo, *Google trends*) o reseñas en Twitter (Loureiro y Alló, 2020) que están disponibles casi en tiempo real. Estas bases alternativas de datos masivos no son generadas por las agencias oficiales de estadística ni están diseñadas para reflejar el comportamiento de un

determinado agregado macroeconómico. Sin embargo, son de muy alta frecuencia, producidas en tiempo real, por lo que pueden ayudar a paliar las carencias de los modelos de predicción macroeconómica que se usan hoy en día cuando todavía no se dispone del dato oficial.

Una de las mencionadas bases de datos masivos alternativas lo constituyen las noticias en prensa. Estas han sido utilizadas de distintas maneras en análisis económico. Por una parte, se cree que reflejan la incertidumbre de la situación económica. La literatura actual argumenta que la incertidumbre afecta la actividad económica, véase, por ejemplo, Baker, Bloom y Davis (2016) quienes construyen un indicador de incertidumbre política y económica contando los artículos que contienen un determinado término. De manera análoga, el Banco de España ha construido indicadores de incertidumbre para la economía española y las principales economías latinoamericanas. (Véase, Ghirelli, Pérez y Urtasun, 2019; Ghirelli, Pérez y Urtasun, 2020, respectivamente). Otro uso de las noticias es explotar su contenido predictivo sobre el estado de la economía. Diversos bancos centrales, instituciones e investigadores han comprobado el poder predictivo de las noticias para los principales agregados macroeconómicos, en especial PIB, inflación y paro. Véase, para el Reino Unido, Rambaccussing y Kwiatkowski (2020), y Kalamara *et al.* (2020). Estos últimos encuentran que la información contenida en las noticias de los tres principales periódicos (*Daily Mirror, Daily Mail* y *The Guardian*) contiene poder predictivo sobre un simple modelo autorregresivo en predicciones con horizonte temporal de hasta nueve meses. La mejora predictiva de estos modelos disminuye considerablemente al incluir en los mismos factores extraídos de series de actividad económica, aunque siguen siendo útiles, sobre todo, en la vecindad de los puntos de giro, es decir, cuando cambia la fase del ciclo de negocios en las series de actividad real (no tanto para la inflación que, como apuntan Stock y Watson [2007], es difícil de predecir). El Banco de España encuentra resultados similares para predecir el PIB de España (véase Aguilar *et al.,* 2020) y Thorsrud (2020) encuentra el mismo tipo de resultados para el PIB trimestral de Noruega incorporando series temporales diarias extraídas de noticias en un modelo factorial. Para EE. UU., Barbaglia, Consoli y Manzan (2020) encuentran resultados análogos para la predicción de los principales agregados macroeconómicos utilizando indicadores de sentimiento construidos a partir de noticias de los seis principales periódicos de EE. UU. usando un total de 6,6 millones de artículos y 4, 2×10^9 palabras.

La utilización de las bases de datos basadas en noticias para la predicción en tiempo real y la monitorización de la actividad económica se lleva a cabo en tres pasos. El primer paso es construir la base de datos de noticias. El segundo paso consiste en pasar de estas bases de datos de noticias a series temporales que puedan ser utilizadas en los modelos de predicción. Para ello se pueden utilizar diversos algoritmos tales como contar las veces que aparece un determinado término, por ejemplo, en la prensa diaria o aquellos basados en diccionarios donde se da un valor entre -1 y +1 a un término específico dependiendo de si su connotación es negativa o positiva para lo que se analiza el entorno en el que aparece dicho término[4]. Por último, se utilizan modelos de aprendizaje automático (Kalamara *et al.,* 2020) o modelos econométricos más o menos sencillos para generar predicciones a muy corto plazo y

[4] En la siguiente sección referida a EURITO, se explican muy brevemente, algunas particularidades de los algoritmos de búsquedas de términos, por ejemplo, en relación a la presencia de sinónimos.

monitorizar el estado de la economía. Dentro de estos últimos, Barbaglia, Consoli y Manzan (2020) utilizan simples regresiones predictivas donde los regresores son los retardos de la variable a predecir que pueden ser aumentados con otros indicadores de actividad económica, mientras que Thosrud (2020) utiliza un modelo factorial dinámico.

En resumen, las noticias están disponibles a muy alta frecuencia, mucho antes de que dispongamos de los datos de estadísticas oficiales y son útiles para la predicción en tiempo real de la actividad económica y la monitorización de la economía. Para usarlas en modelos de previsión macroeconómica hay que pasar de los datos de texto a series temporales que podamos introducir en los modelos de predicción. Aunque una vez que se publican las estadísticas oficiales mensuales (datos construidos para reflejar el comportamiento de una determinada variable económica), los indicadores basados en noticias pueden perder su capacidad predictiva para el PIB trimestral, siguen siendo útiles en la vecindad de los puntos de cambio (Kalamara *et al.,* 2020), sobre todo en las recesiones (Barbaglia, Consoli y Manzan, 2020) y se han mostrado de gran utilidad en la recesión debida a la COVID-19 donde los indicadores basados en encuestas (los primeros disponibles informativos sobre el estado de la economía) no han reflejado a tiempo ni correctamente el *shock* económico (Aguilar *et al.* 2020).

4. EURITO: GENERACIÓN DE INDICADORES EN ÁREAS VÍRGENES

EURITO Research and Innovation Indicators (EURITO, 2018) es un proyecto financiado por la Comisión Europea, realizado por un consorcio de cuatro organizaciones: Nesta (Reino Unido), Fraunhofer (Alemania), Danmark Tekniske Universitet (DTU, Dinamarca) y la Fundación Cotec para la Innovación (España). El objetivo de EURITO es elaborar indicadores de desarrollo de la investigación e innovación a partir de la huella digital de sus actividades, en un área en la que los indicadores disponibles (European Innovation Scoreboard) están principalmente basados en encuestas o resultados como patentes y publicaciones que generalmente ofrecen una visión demasiado agregada y retardada de su desarrollo. EURITO es el acrónimo formado a partir del título del proyecto "EU Relevant, Inclusive, Timely, Trusted and Open Research Innovation Indicators" que resume los objetivos del mismo: indicadores EURopeos de innovación relevantes que puedan proporcionar la información necesaria para el conocimiento y desarrollo de políticas de investigación y desarrollo; Inclusivos permitiendo que la cobertura de los indicadores se extienda más allá de la investigación en disciplinas de las STEM o de empresas altamente tecnológicas, incluyendo la innovación en servicios o sectores menos tecnológicos, así como la consideración de redes informales; confiables en Tiempo real, garantizando la representatividad, calidad, oportunidad y pronta publicación, indicando sus limitaciones y validados por los investigadores y la industria; Abiertos (Open), de forma que tanto las fuentes de datos originales como la metodología y códigos de programación empleados permitan reproducir y multiplicar las oportunidades para su mejora, extensión y aplicación.

A modo de ejemplo, describimos la elaboración de indicadores sobre tecnologías emergentes desarrollado en el primer estudio piloto del proyecto (Nesta, 2019). La medición

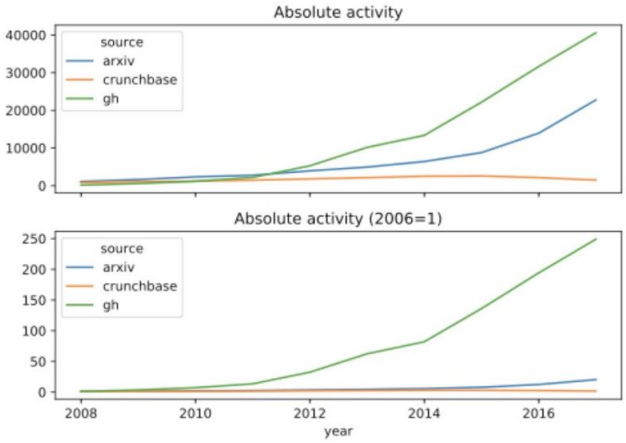
de la aparición y desarrollo de una tecnología emergente es relevante en sí misma, así como su impacto en la aplicación a la creación de nuevos productos, servicios o actividades de investigación. Prácticamente por definición una tecnología emergente, es susceptible de caer fuera de taxonomías preestablecidas utilizadas en el análisis industrial, científico y tecnológico lo que dificulta y retrasa su medición mediante indicadores tradicionales, Bakhshi y Mateos-García (2016). Sin embargo, las descripciones en formato texto incluidas en los documentos en repositorios de investigación científica, información sobre financiación, patentes, páginas webs de las compañías, plataformas de colaboración, entre otros, pueden considerarse como nuevas bases de datos no estructuradas que permitan de una forma ágil medir el desarrollo de una tecnología emergente.

El estudio piloto se centra en el desarrollo de la *inteligencia artificial,* entendida como el conjunto de tecnologías que usan datos y métodos de aprendizaje automático o *machine-learning* (Mateos-García, 2018). Los indicadores que se persiguen son el nivel de investigación en inteligencia artificial en la Unión Europea y su evolución en el tiempo, comparado con sus principales competidores en agregado y entre los Estados miembros, entre otros. A tal fin, se puede encontrar información en distintas bases de datos que recogen repositorios de artículos de investigación (arXiv, Microsoft Academic Graph), software (GitHub), financiación (CORDIS), patentes (PATSTAT), nuevas *startups* y compañías tecnológicas (CRUNCHBASE), redes de colaboración (Meetup) o habilidades demandadas y ofertadas (webs universidades y anuncios de empleo). Las bases de datos son heterogéneas y muestran importantes diferencias y problemas para la construcción de un indicador de actividad basado en el contenido de los textos. Una primera dificultad se encuentra, por ejemplo, en la distinta longitud de las descripciones entre las fuentes de datos lo que puede llevar a que las medidas de relevancia no sean comparables. También es preciso tener en cuenta la necesidad de disponer de los metadatos, y conocer indicadores que identifiquen cada observación como el sexo o la localización geográfica. En EURITO esta información estaba disponible en algunas bases de datos, pero donde no era el caso, se ha aproximado mediante algoritmos que identifican, por ejemplo, la localización geográfica a partir del nombre de la institución del investigador.

Como una primera aproximación para desarrollar el indicador de actividad, el estudio piloto selecciona tres bases de datos de acceso abierto con cobertura europea: Crunchbase, arXiv y GitHub. El objetivo es identificar las entradas relacionadas con el concepto de inteligencia artificial mediante la aplicación de técnicas de procesamiento natural del lenguaje. El procedimiento se inicia con una palabra clave (en nuestro caso *Artificial Intelligence*) y la búsqueda de sinónimos en un espacio multidimensional mediante similaridad semántica (algoritmo word2vec, Mikolov, Yih and Zweig, 2013). Esta búsqueda selecciona aquellas palabras que aparecen en un contexto similar a la palabra clave original. En el desarrollo del indicador se corrige la posibilidad de que algunos sinónimos sean demasiado genéricos y produzcan un elevado número de documentos irrelevantes. Para ello se eliminan de la lista de palabras claves aquellas con bajo *TF-IDF (Term-Frequency Inverse-Document Frequency)* que normaliza el número de veces que un término aparece en un documento por el número de veces que aparece en un *corpus*.

Figura 5.

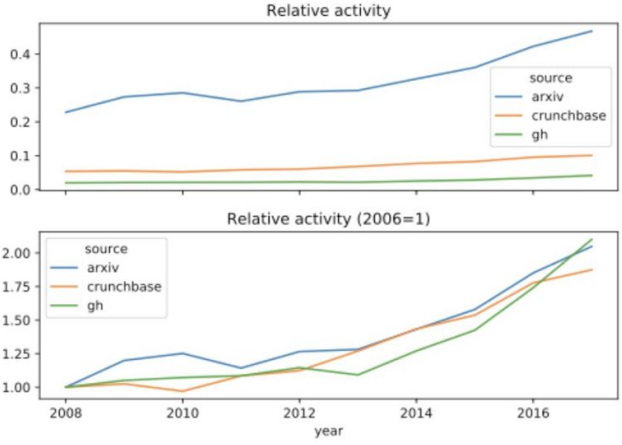
Número de documentos relacionados con actividad artificial en números absolutos e índice (2006=1)



Las figuras 5 y 6 muestran la evolución del nivel de actividad absoluta y relativa en inteligencia artificial en las tres bases de datos seleccionadas: Crunchbase, arXiv y GitHub (denotada como gh en el gráfico)[5]. Mientras que el nivel de actividad absoluto cuenta

Figura 6.

Proporción de documentos, relacionados con actividad artificial en números absolutos e índice (2006=1)



[5] Estos gráficos se toman del entregable relacionado con el primer estudio piloto (Nesta, 2019) y se reproducen con consentimiento de los autores.

directamente el número de entradas relacionadas con la inteligencia artificial, el nivel en términos relativos normaliza esta medida comparando con el total de entradas en cada base de datos.

Tal y como se observa, el desarrollo de la inteligencia artificial ha crecido tanto en números absolutos como en términos relativos en las tres fuentes de datos consideradas, siendo la más representativa arXiv, donde el crecimiento acumulado desde 2008 es del 250 % llegando a suponer el 40 % de las publicaciones totales en esta base de datos en el año 2017.

5. MOVILIDAD A PARTIR DE LA GEOLOCALIZACIÓN DE TELÉFONOS MÓVILES: NUEVOS INDICADORES PARA VIEJAS PREGUNTAS

Desde un punto de vista de política económica y social, el conocimiento de las cifras de población es necesario en muchos aspectos relacionados como por ejemplo la asignación de recursos públicos. El análisis de la movilidad cotidiana (por motivos laborales o educativos) y de la movilidad estacional (relacionado con el turismo nacional o internacional) resulta igualmente necesario para reconocer en cada momento las necesidades reales de los distintos territorios. El Censo de Población y Viviendas es una operación estadística elaborada por el Instituto Nacional de Estadística que se realiza cada diez años y permite conocer las características de las personas, hogares, edificios y viviendas. El Censo de 2011 fue novedoso pues incluyó, por primera vez, registros administrativos y encuestas dirigidas al 10 % de la población. Nuevamente, el INE se enfrenta a un reto metodológico, elaborando el Censo de Población y Viviendas de 2021 basado totalmente en el uso de registros. Dentro de los trabajos preparatorios del Censo 2021, el INE, bajo la categoría de *Estadística Experimental,* ha realizado un estudio de movilidad (EM1) a partir de registros georeferenciados de telefonía móvil, como fuente de datos alternativa a los cuestionarios censales (INE, 2020a).

Los resultados obtenidos provienen del análisis de la posición de más del 80 % de los teléfonos móviles en toda España, con la colaboración de los tres principales operadores de telefonía móvil (Orange, Telefónica y Vodafone). En relación con el ámbito de la investigación, se considera a la población residente en España (se excluyen los teléfonos de numeración extranjera) para el total del territorio español dividido en 3.214 *áreas de movilidad INE* (agrupación de municipios con menos de 5.000 empadronados hasta que superan esta cantidad, municipios entre 5.000 y 50.000 empadronados, o divisiones de aquellos municipios con más de 50.000 empadronados en barrios -*SCD, sub-city districts*).

A partir de los datos de posición de los teléfonos durante una semana laboral de referencia (los días 18 a 21 de noviembre de 2019), el INE proporciona la matriz de movilidad cotidiana que permite conocer la población residente en el área (según Padrón a 1 de enero de 2019), la población residente que se mantiene en su área, la población que llega al área, la población detectada durante el día en el área, la variación de población, los destinos a los que se desplazan sus residentes y el origen de los que llegan al área. De cara a la movilidad, se define el área de residencia del teléfono móvil como aquella donde el teléfono se encuentra con mayor frecuencia durante el período previo al considerado (entre dos y tres meses según

el operador). Asimismo, el área de destino se define como la más frecuente fuera del área de residencia en la que se encuentra el terminal al menos durante cuatro horas al día en la franja horaria de 10:00 a 18:00 y al menos dos días de los cuatro observados de la semana laboral normal de referencia. Esta información permite conocer casi en tiempo real, los movimientos de la población a un elevado nivel de desagregación geográfica, así como los trayectos que realiza.

La primera columna de la tabla 1 recoge las estimaciones de movilidad diurna obtenidas en el área de *Madrid (SCD Sol)*. SCD Sol es un área de Madrid central que cuenta con 7.309 residentes, de los cuales 2.026, el 27,7 %, se mantienen en el área y, al mismo tiempo, fue área de destino de 14.790 personas no residentes, más del doble de la población que procedían de más de 190 orígenes identificables. Estas estimaciones señalan esta área como fuerte receptora de población diurna de manera cotidiana.

Tabla 1.

Movilidad en Madrid - SCD Sol

	Semana referencia	*Fechas estacionales*			
Variable	*18-21 noviembre*	*20 julio*	*15 agosto*	*24 noviembre*	*25 diciembre*
Población residente	7.309	7.309	7.309	7.309	7.309
que se mantiene en el área	2.026	3.100	2.665	3.652	2.466
Población total					
detectada durante el día	16.816				
que pernocta en el área		16.839	15.710	31.390	30.609
Variación de población	12.653	9.530	8.401	24.081	23.300

Adicionalmente, el INE proporciona información sobre la población que pernocta fuera de su área de residencia, lo que permite complementar la información sobre los movimientos diurnos por motivos laborales o educativos con la estimación de la variación debida a causas estacionales. Se proporciona información sobre dos fechas de agosto (20 de julio y 15 de agosto), un domingo normal (24 de noviembre, continuidad de la semana de referencia considerada en la movilidad diurna) y el día de Navidad (25 de diciembre). Se consideran en esta ocasión todos los teléfonos presentes en territorio español en esas fechas y se determina el área de pernoctación a partir del lugar más frecuente en el que se encuentra el aparato desde las 22:00 horas del día anterior hasta las 6:00 horas en esa fecha.

Las columnas 2 a 5 de la tabla 1 recogen las estimaciones de movilidad estacional en las fechas señaladas. La población residente no cambia, sigue siendo la correspondiente a fecha 1 de enero de 2019 del Padrón. Estas estimaciones permiten señalar *Madrid Sol* como un destino fuertemente receptor de turistas nacionales tanto en las fechas consideradas de verano como de invierno. Se identifica el mayor atractivo del área de *Madrid Sol* en las fechas de invierno frente al verano, puesto que mientras que en las fechas del 20 de julio y 15 de

agosto las pernoctaciones son de algo menos del doble de la población residente, el 24 de noviembre[6] y el 25 de diciembre prácticamente las multiplican por 4.

5.1. La incidencia de la COVID-19 en la movilidad

La metodología empleada en la generación de la matriz de movilidad cotidiana ha permitido al INE, en colaboración con los tres principales operadores de telefonía móvil, estimar el movimiento de la población desde la irrupción de la pandemia de la COVID-19 y la declaración del estado de alarma desde el 16 de marzo hasta la actualidad, INE (2020b, 2020c). En una primera fase, durante el estado de alarma, el INE utilizó las mismas áreas de movilidad cotidiana y redefinió el área de residencia como aquella donde el teléfono ha pasado la mayor parte del tiempo en el horario de 0:00 a 6:00 horas y el área de destino como la más frecuente en el período de 10:00 a 16:00 horas con un mínimo de dos horas, respectivamente. Los resultados obtenidos, elevados al total de la población, han permitido conocer la variación de la movilidad frente a un día de una semana de referencia *normal* coincidente con la utilizada en el estudio de la movilidad cotidiana. Desde el 24 de junio al 30 de diciembre, se ha venido recogiendo la información de movilidad los miércoles y los domingos, recuperando las definiciones de área de residencia y destino de la matriz de movilidad cotidiana.

La figura 7 recoge la evolución del porcentaje de población que sale del área de residencia en el total nacional y en el área de *Madrid Sol* a partir de estos datos. El primer punto señala el porcentaje de movilidad en la semana de referencia de 2019[7] que era, tanto en el total nacional como en *Madrid Sol* cercana al 30 %, cifra que no se ha recuperado hasta la fecha, siendo el máximo a nivel nacional del 22 % en los días laborables de junio y julio y del 23 % el domingo 29 de noviembre en *Madrid Sol.* Destaca en la figura 7 la dualidad entre los días laborables y los fines de semana a nivel nacional durante el estado de alarma, que también se mantiene en la segunda mitad de 2020 situándose la movilidad los miércoles (línea punteada azul superior) alrededor del 20 % frente a los domingos (línea punteada azul inferior) que se reduce entre el 10 y el 15 %.

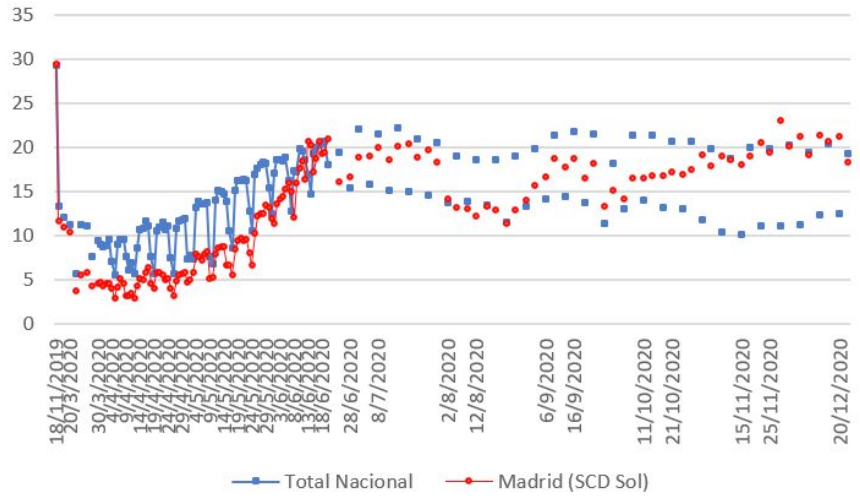
La figura 7 también permite analizar el distinto comportamiento en *Madrid Sol.* Por una parte, durante el primer período del confinamiento, donde la movilidad fue mucho menor que en el total nacional, manteniendo cifras alrededor del 3 % durante el mes de abril y no alcanzando el porcentaje de movilidad nacional hasta el mes de junio. Con posterioridad la diferencia se manifiesta principalmente en la inexistencia de comportamiento dual entre día laborable y domingo.

[6] Las cifras no son comparables con las de movilidad cotidiana de la primera columna, puesto que antes se recogía movilidad diurna y teléfonos españoles y ahora son pernoctaciones y todos los teléfonos presentes.

[7] En la figura, se presenta con una línea continua la información diaria durante el estado de alarma hasta el 23 de junio y, mediante puntos espaciados la información correspondiente a los miércoles y domingos desde el 24 de junio hasta el final de la muestra.

Figura 7.

Movimiento de personas por áreas de movilidad
(Porcentaje)



5.2. Otros indicadores de movilidad: Google y Apple

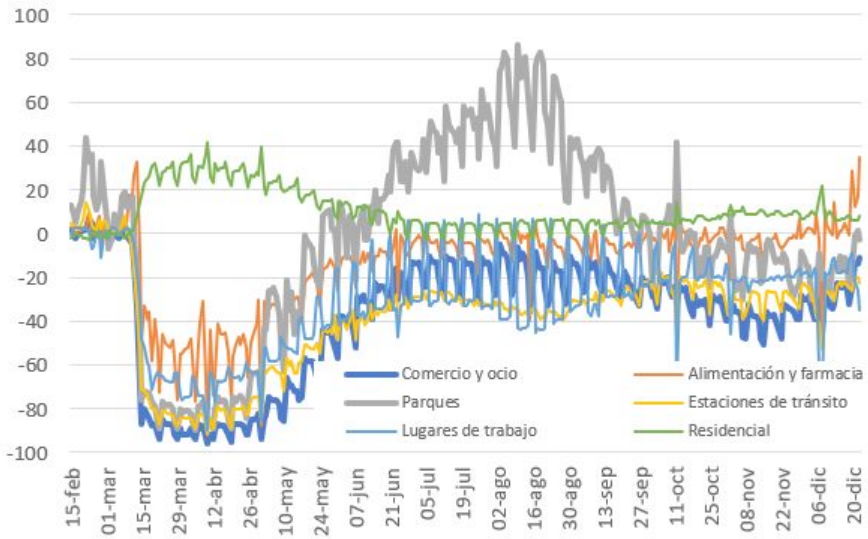
Alternativamente, Google (2020) y Apple (2020) están utilizando la información que proporcionan sus usuarios para generar indicadores de movilidad. Estas fuentes de datos ya están siendo utilizadas en el análisis económico, véase por ejemplo, Woloszko (2020) que monitoriza la actividad económica utilizando los indicadores de movilidad de Google para Kartal, Depren y Depren (2020) que analizan la reacción de los principales índices bursátiles de los países de Asia del este ante la COVID-19 utilizando los datos de Apple, entre otros.

Google proporciona estadísticas a partir de los datos de los usuarios que han habilitado el historial de ubicaciones de su cuenta. En concreto, informa sobre las tendencias de movimiento a lo largo del tiempo ordenadas por zonas geográficas y clasificadas en diversas categorías de lugares atendiendo a su actividad: *Comercio y Ocio, Alimentación y farmacias, Parques, Estaciones de tránsito, Lugares de trabajo y Residencial.* Los datos recogen el cambio en el número de visitas a dichos lugares, clasificados por su actividad, y su duración en comparación al valor medio de cada día de la semana durante un período de cinco semanas desde el 3 de enero hasta el 6 de febrero de 2020. Los datos están disponibles diariamente desde el 15 de febrero de 2020.

La figura 8 presenta los datos proporcionados por Google por actividad en tipo de destino.

Figura 8.

Tendencias de movilidad de Google
(Total España)



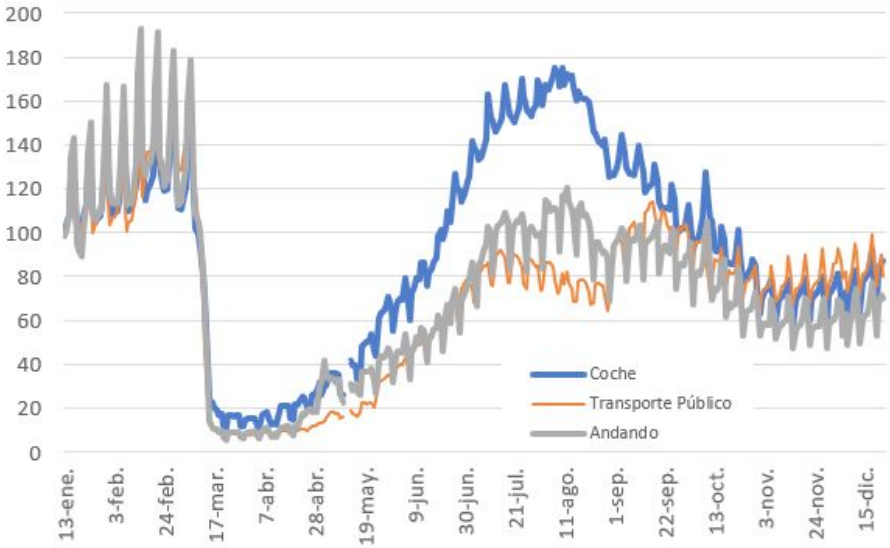
La figura 8 refleja claramente la falta de movilidad del confinamiento, siendo la categoría *Residencial* la única con valores positivos, en relación con el período de referencia pre-COVID-19, y la posterior relajación de las medidas que mantienen esta categoría en niveles positivos pero de menor magnitud. Igualmente, la figura 8 nos recoge la consecuente reducción en las visitas al resto de lugares durante el confinamiento, siendo el que menor descenso sufre la categoría de *Alimentación y Farmacia,* que aún así ve reducidas las visitas hasta valores alrededor del 60 % con respecto al período de referencia. Con posterioridad al confinamiento, la única categoría que alcanza niveles superiores al período de referencia han sido las correspondiente a *Parques* durante el período estival.

Apple también proporciona informes diarios de tendencias de movilidad a partir de las solicitudes de indicaciones en *Mapas de Apple.* En concreto, publica información diaria sobre el porcentaje del número de consultas realizadas en comparación con las realizadas el 13 de enero. Se dispone de datos sobre solicitudes de traslados en coche, transporte público o a pie, para el total nacional, por comunidades autónomas y las cuatro ciudades más grandes de España: Madrid, Barcelona, Valencia y Sevilla. La figura 9 muestra la evolución de la movilidad según la información proporcionada por Apple.

Los datos recogidos en la figura 9, al igual que las cifras de INE y las de Google, indican que existe una fuerte estacionalidad semanal a nivel nacional y un fuerte impacto de la pandemia en la movilidad. En el caso de Apple la información proporcionada muestra una

Figura 9.

Tendencias de movilidad de Apple
(Total España)



mayor recuperación de la movilidad mediante el uso del automóvil, siendo la única categoría que llega a alcanzar cifras superiores al 13 de enero de 2020 (pre-COVID), durante los meses de verano.

6. ALGUNOS RETOS Y VENTAJAS A PARTIR DE LOS EJEMPLOS

Los ejemplos anteriores muestran cómo las técnicas estadísticas y las nuevas bases de datos resultan prometedoras en el análisis económico y, en especial, en la generación de indicadores. Hemos visto que mediante técnicas de big data se pueden abordar problemas económicos relevantes, bien actualizando técnicas estadísticas existentes (integración financiera), generando nuevos indicadores en materias donde no existía información oficial (EURITO), adelantando indicadores oficiales (noticias) o proporcionando información en tiempo real de variables que antes se estimaban mediante encuestas cada diez años (movilidad).

No obstante, para su generalización y utilización en el análisis económico los indicadores deben validar su capacidad para representar fielmente la realidad que persiguen identificar. Un problema que es diferente dependiendo de cada una de las aplicaciones. En el caso de los indicadores basados en noticias y, en general, todos aquellos dirigidos a la predicción económica o el nowcasting, la validación viene de la mano de su utilidad para adelantar la información oficial que será publicada posteriormente. EURITO y los indicadores de

movilidad proporcionan nuevas variables en materias donde no existe información oficial publicada al respecto con la que comparar. La validez o credibilidad de estos datos deben fundamentarse en su adecuación a estándares de calidad, como los recogidos en el Código de Buenas Prácticas del Sistema Estadístico Europeo (ESS, 2017). Algunos de estos estándares son: Relevancia; Exactitud y Fiabilidad; Prontitud y puntualidad en la publicación de los datos; Coherencia y Comparabilidad, Accesibilidad y Claridad.

La ausencia de sesgos y la fiabilidad de los datos de partida y de las estadísticas generadas es uno de los retos pendientes. Realmente, con la irrupción del big data, ha cambiado el paradigma clásico de muestreo. La oportunidad de obtener información mediante bases de datos alternativas, no diseñadas para el propósito del estudio, puede apartar la posibilidad de mantener un esquema de inferencia tradicional. Por una parte, los tamaños de muestra tan grandes nos llevan a pensar que estamos prácticamente considerando la población y que los errores de muestreo son negligibles. No obstante, se pueden presentar dos problemas. El primero es que, al no disponer de un diseño muestral, estemos incurriendo en errores de cobertura y exista un importante sesgo por selección de muestra en los posibles análisis derivados. El segundo es que los errores de medida de los datos no sean completamente aleatorios y se propaguen con el tamaño de muestra en vez de compensarse. Además, es preciso que las modificaciones de las técnicas estadísticas que llevan a cabo los procedimientos de aprendizaje automático para adaptarse a las nuevas necesidades del big data, garanticen que cubren los términos que se refieren al problema de interés. Así, por ejemplo, en la aplicación sobre integración financiera, si dejamos fuera alguna frecuencia de fluctuación importante, podemos omitir información relevante para la comparación entre índices. O en el caso de EURITO, si el procedimiento de similaridad semántica utilizado no es capaz de recoger el espectro semántico relevante para el problema en cuestión.

El reto del análisis de la coherencia de los indicadores obtenidos depende del ejemplo analizado. Los indicadores construidos para monitorizar la actividad en tiempo real tienen una clara metodología de validación de su coherencia basada en la evaluación de los errores de predicción, con los métodos habituales de exactitud predictiva, una vez publicado el indicador oficial que pretenden adelantar. La evaluación de la coherencia no es inmediata en aquellas situaciones donde no existe un indicador de referencia oficial. En el caso de EURITO, el alto nivel de desagregación sectorial y geográfico y las características de los fenómenos de innovación estudiados, no permiten una validación cuantitativa clara más allá de la comparación con grandes agregados de I+D+i y el análisis de su coherencia geográfica y temporal. En el caso de las variables de movilidad, la publicación de distintos indicadores a partir de fuentes de información diferentes, aunque a distinto nivel de desagregación sectorial y geográfico, quizás pueda servir para estudiar la coherencia entre las series a partir de un estudio de características comunes, entre aquellas que se aproximen al mismo fenómeno.

Otro tipo de cuestiones de carácter institucional no tratadas en este capítulo vienen recogidas en Salgado (2017) y Salgado y Oancea (2020). Entre estas se encuentran el acceso al dato (privacidad, continuidad y coste de las fuentes de información), los recursos tecnológicos necesarios (humanos y de capital), la independencia profesional en la generación de los datos, la coordinación y la cooperación público-privada, entre otras.

Este recorrido a través de los ejemplos no pretende, ni mucho menos, ser exhaustivo sobre los retos a los que hay que enfrentarse, sino que simplemente intenta señalar alguno de los más relevantes que trascienden tras el simple análisis de estos cuatro ejemplos.

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.

Alonso, A. M. y Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing,* 29, pp. 655–676.

Apple (2020). Informes de tendencias de movilidad. https://covid19.apple.com/mobility

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

Bakhshi, H. y Mateos-García, J. (2016). New data for innovation policy. *Working Paper.* London: Nesta.

Banbura, M., Giannone, D. M., Modugno, M. y Reichlin, L. (2013). Now-casting and the real-time data flow. En: *Handbook of Economic Forecasting,* 2, pp. 195-237.

Barbaglia, L., Consoli, S. y Manzan, S. (2020). *Forecasting with Economic News.* Manuscrito disponible en SSRN: https://ssrn.com/abstract=3698121

BCE (2018). *Financial Integration in Europe.* op.europa.eu/en/publication-detail/-/publication/c7c3826a-526a-11e8-be1d-01aa75ed71a1/language-en

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M. y Tambalotti, A. (2018). Macroeconomic now-casting and forecasting with big data. *Annual Review of Economics,* 10, pp. 615-643.

Caiado, J., Crato, N. y Peña, D. (2006). A periodogram-based metric for time series classification. *Comput Stat Data Anal,* 50, pp. 2668–2684.

Caiado, J., Crato, N. y Poncela, P. (2020). A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification,* 14, pp. 117-146.

Camacho, M. y Perez-Quirós, G. (2010). Introducing the EURO-STING: Short Term INdicator of Euro Area Growth. *Journal of Applied Econometrics,* 25, pp. 663-694.

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

Camacho, M., Perez-Quirós, G. y Poncela, P. (2013). Short-term forecasting for empirical economists. A survey of the recently proposed algorithms. *Foundations and Trends in Econometrics,* 6, pp. 101-161.

Corsi, F. (2009). Heterogeneous autoregressive model of realized volatility (HAR-RV). *J Financ Econom,* 7, pp. 174–196.

EURITO (2020). *EU Relevant, Inclusive, Timely, Trusted, and Open Research Innovation Indicators.* http://www.eurito.eu/

Galeano, P. y Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas,* 4, pp.383–404

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

Ghirelli, C., Pérez, J. J. y Urtasun, A. (2020). Economic Policy Uncertainty in Latin America. *Documento de Trabajo,* No. 2024, Banco de España.

Google (2020). *Informes de movilidad local sobre COVID-19.* https://www.google.com/covid19/mobility/

INE (2020a). *Estudio EM-1 de movilidad a partir de la telefonía móvil.* Diciembre 2020. http://www.ine.es/experimental/movilidad/experimental_em.htm

INE (2020b). *Análisis de la movilidad de la población durante el estado de alarma por COVID-19 a partir de la población de los teléfonos móviles.* Junio 2020. https://www.ine.es/covid/exp_movilidad_covid_proyecto.pdf

INE (2020c). *EM-3 - Estudio de movilidad a partir de la telefoníma móvil durante el período julio-diciembre 2020 (EM-3).* Noviembre 2020. https://www.ine.es/experimental/movilidad/exp_em3_proyecto.pdf

Kalamara, E., Turrell, A., Kapetanios, G., Kapadia, S. y Redl, C. (2020). Making text count: economic forecasting using newspaper text. *Bank of England Staff Working Paper,* No. 865.

Kartal, M. T., Depren, S. K. y Depren, Ö. (2020). How Main Stock Exchange Indices React to Covid-19 Pandemic: Daily Evidence from East Asian Countries. *Global Economic Review.* DOI: 10.1080/1226508X.2020.1869055.

Loureiro, M. y Alló, M. (2020). Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain. *Energy Policy,* 143, 111490.

Mateos-García, J. (2018). *The Complex Economics of Artificial Intelligence.* Disponible en SSRN 3294552.

Mikolov, T., Yih, W. y Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. En: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/N13-1090

Nardo, M., Ndacyayisenga, N., Papanagiotou, E., Rossi, E. y Ossola, E. (2017). Measures and drivers of Financial Integration in Europe. European Commission, Joint Research Centre. *Report EUR 28469 EN.* doi:10.2760/92134.

Nesta (2019). *Pilot 1: Emerging Technology Ecosystems.* http://www.eurito.eu/pilots-and-indicators/

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *J Time Ser Anal,* 11, pp. 152–164.

Poncela, P., Ruiz, E. y Miranda, K. (2020). Factor extraction using Kalman filter and smoothing: this is not just another survey. *International Journal of Forecasting* (en prensa).

Rambaccussing, D. y Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting,* 36, pp. 1501-1516.

Salgado, D. (2017). Big Data en la Estadística Pública: Retos ante los primeros pasos. *Revista de Economía Industrial.* 3 trimestre, pp. 121-129.

Salgado, D. y Oancea, B. (2020). On new data sources for the production of official statistics. *Working Paper,* 01/2020. Instituto Nacional de Estadística (INE).

Stock, J. H. y Watson, M. W. (2007). Why Has U.S. Inflation Become Harder to Forecast? Journal of Money, Credit and Banking, 39, Issue s1.

Stock, J. H. y Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *Handbook of Macroeconomics,* 2, pp. 415-525.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of business cycles. *Journal of Business and Economic Statistics,* 38(2), pp. 393-409.

Woloszko, N. (2020). Tracking activity in real time with Google Trends. *OECD Economics Department Working Papers,* N. 1634. https://doi.org/10.1787/18151973