CUADERNOS ECONÓMICOS ICE INFORMACIÓN COMERCIAL ESPAÑOLA

Investigación económica y datos masivos: mercados, fines sociales y colaboración público-privada*

Economic research and big data: Markets, socio-economic research and public-private partnerships

José García Montalvo
Universitat Pompeu Fabra

Resumen

Este artículo presenta un análisis de las perspectivas abiertas por la creciente disponibilidad de datos masivos para realizar investigación económica, así como los riesgos asociados a las mismas. A diferencia de otras contribuciones, la aproximación adoptada se enfoca en los generadores de la información distinguiendo entre los datos generados por empresas privadas para buscar soluciones de mercado, los datos originados en las administraciones públicas y las experiencias recientes de colaboración público-privada que se están abriendo en el campo del uso de datos masivos (big data) y la aplicación de técnicas de aprendizaje automático (machine learning).

Palabras clave: *datos masivos, aprendizaje automático, colaboración público-privada.*
Códigos JEL: *C10, C18, C55, C81.*

Abstract

This article presents an analysis of the opportunities opened by the increasing availability of big data for economic research as well as the risks associated with them. Unlike other contributions, the approach adopted here focuses on the generators of the information, distinguishing between data generated by private companies to seek market solutions, the data originating in public administrations used for research, and the recent experiences of public-private partnerships that are opening large datasets of private companies for research in empirical economics using big data and machine learning techniques.

Keywords: *big data, machine learning, public-private partnership.*

* Este artículo se basa fundamentalmente en una conferencia impartida en la Fundación Ramón Areces titulada «*Big data* y economía en la era de las *fake news* y las verdades imprecisas».

DOI: https://doi.org/10.32796/cice.2021.102.7311
Recibido: Septiembre 2021 · Aceptado: Octubre 2021

1. Introducción

Este artículo presenta un recorrido por las oportunidades abiertas por la creciente disponibilidad de datos masivos para realizar investigación económica y los riesgos asociados a las mismas. La aproximación adoptada distingue entre los datos generados por empresas privadas para buscar soluciones de mercado, los datos originados en las administraciones públicas y las experiencias recientes de colaboración público-privada que se están abriendo en el campo del uso de datos masivos (*big data*) y la aplicación de técnicas de aprendizaje automático (*machine learning*). Por datos masivos entendemos tanto lo que normalmente se describe como *big data* (datos no tradicionales con formato muy heterogéneo como sensores, GPS, *clicks*, *logs* de servidores, correos electrónicos, imágenes, voz, texto de redes sociales, etc.) así como datos estructurados de fuentes administrativas o empresas. Este artículo es, en parte, continuación de García-Montalvo (2014) donde se realiza una panorámica de la utilización de datos masivos con objetivos predictivos en el sistema bancario y financiero. En el mismo se describen sistemas de *scoring* basados en información de las redes sociales, y procedimientos innovadores de calificación crediticia fundamentados en el análisis textual de las solicitudes digitales. Asimismo, se describen procedimientos de detección de fraude en tarjetas de crédito, de personalización de las tarifas de las compañías de seguros, etc. García-Montalvo (2021) describe un conjunto amplio de experiencias personales en el trabajo de investigación económica utilizando grandes bases de datos administrativos y técnicas de aprendizaje automático.

2. Datos privados para soluciones de mercado

El dominio de las bases de datos masivas comienza en los años noventa con la mejora de los sistemas de información y el aumento de la potencia de los ordenadores. Las primeras aplicaciones de datos masivos se producen en las grandes infraestructuras científicas como los sincrotrones. De hecho, aunque se piensa que en la actualidad es Internet el mayor generador de datos, esta idea es errónea. El Large Hadron Collider del CERN produce 600 TB/segundo con sus 15 millones de sensores y, después de un filtrado que desecha la mayor parte de la información, necesita almacenar 25 PB/año. De hecho, el problema fundamental en la actualidad para el avance del trabajo con bases de datos cada vez más grandes no es la capacidad de cálculo, sino que la creación de información avanza a más velocidad que la capacidad de almacenarla y la energía necesaria para mover dicha información entre los procesadores y los dispositivos de almacenaje es cada vez mayor.

La mayoría de las primeras aplicaciones basadas en datos masivos en el campo de la economía y las finanzas tienen que ver con datos generados en mercados financieros. Es la época del surgimiento de los *quants*[1]. En general, los objetivos son poner precio a productos financieros complejos, analizar correlaciones entre

[1] Para una descripción muy interesante de esta época, consúltese Derman (2004).

enormes conjuntos de activos financieros o buscar estrategias para aprovechar pequeñas disfuncionalidades de precios que se producían entre mercados o activos. Pero también se podían utilizar para hacer evaluaciones de la implementación de medidas para mejorar la liquidez de un mercado. Por ejemplo, García Montalvo (2003) analiza, usando decenas de millones de observaciones de *bids*, *asks* y operaciones cerradas en el mercado de futuros de deuda pública, el impacto de la introducción de creadores de mercado en 1995 en el Mercado Oficial de Futuros y Opciones Financieras de España (MEFF) sobre la liquidez del mercado.

La llegada de procedimientos de aprendizaje automático más avanzados[2], favorecidos por la creciente potencia de los ordenadores, facilitó la extensión de la utilización de los datos para la solución de problemas empresariales. La utilización de datos masivos para la solución de problemas de mercado viene marcada por varios hitos. La sustitución de los críticos y editores de Amazon por un sistema automático de recomendación basado en el algoritmo *item-by-item* fue uno de los primeros grandes logros de la recopilación masiva de información sobre compras de los clientes de Amazon. Pero el primer gran hito es el caso Target. Imaginémonos a principios de los años 2000. Un coche recorre las nevadas llanuras de Minnesota hasta llegar a un centro comercial donde para enfrente de una tienda de la cadena Target. Un hombre sale del coche y se dirige al interior donde solicita hablar con el encargado. Cuando este llega, el cliente desparrama sobre el mostrador un montón de vales con ofertas de pañales de niños, carritos de bebé, muñecos, potitos, etc. El cliente, muy enfadado, grita: «¡Cómo se atreven a enviarle estos vales a mi hija! ¡Solo tiene 15 años!». El encargado intenta tranquilizar al cliente y le pide disculpas argumentando que a veces los sistemas fallan y que siente el error. El cliente sale de la tienda, todavía muy enfadado, y sube al coche desapareciendo en el horizonte.

A los pocos días, el encargado telefonea al cliente para volver a disculparse por el error. Al sonar el teléfono e identificarse, el cliente detiene su intento de disculpa. «Quizá quien tendría que disculparse soy yo. Bajo el techo de mi casa estaban sucediendo cosas que yo mismo desconocía.» En efecto, su hija adolescente estaba embarazada y Target lo supo antes que su propio padre[3]. El procedimiento utilizaba la combinación de productos en una cesta de la compra para elaborar tipologías de clientes. Las mujeres embarazadas son un objetivo frecuente de los modelos predictivos, puesto que la captación en esta etapa garantiza compras futuras y fidelidad temporal. Este tipo de predicción fue posible por la enorme cantidad de información disponible a partir de la utilización de los lectores de códigos de barras y las mejoras en los métodos de aprendizaje automático.

---

[2] La generalización de procedimientos como el *support vector machine,* las técnicas de *random forest* y el redescubrimiento de las redes neuronales que, después de las limitaciones mostradas en los años sesenta, ahora pasan a denominarse *deep learning* para intentar evitar el estigma asociado a estas técnicas, se ve favorecida por la creciente capacidad de cálculo de los ordenadores.

[3] Esta historia se conoce porque Andrew Pole, el estadístico a cargo de los primeros modelos predictivos de Target, la contó en una conferencia muchos años después, y el *New York Times* se hizo eco, apareciendo también en Duhigg (2011).

El segundo hito importante es el premio de Netflix. La competición premiaba con un millón de dólares al mejor algoritmo colaborativo para predecir los *ratings* de las películas de los clientes que superara en un 10 % la capacidad predictiva del modelo utilizado por Netflix en el momento. La competición comenzó en octubre de 2006. Netflix proporcionó unos 100 millones de *ratings* de unos 500.000 usuarios sobre 17.770 películas. La dificultad consistía en que el modelo tenía que predecir únicamente utilizando la información sobre los *ratings* pasados sin ninguna información sobre usuarios o películas. En septiembre de 2009, el grupo BellKor's Pragmatic Chaos triunfó en la competición, en la que participaron 2.000 equipos de 150 países, incluidos equipos universitarios de los mejores centros del mundo. La mejora del grupo sobre el algoritmo original fue del 10,06 %.

La multiplicación de la utilización de datos para solucionar problemas de mercado es interminable. En servicios financieros las nuevas Fintech utilizan procedimientos de calificación basados en datos de redes sociales. Por ejemplo, quiénes son tus amigos en Facebook y cuántos están en listas de morosos, quiénes son tus contactos en LinkedIn (número de jefes en la empresa actual o directivos de otras empresas), el análisis textual de solicitudes de crédito o incluso el hecho de capitalizar o no capitalizar el nombre y apellidos en una solicitud. En el campo de los seguros de auto la aceptación de sensores en el vehículo permite un perfilado muy preciso del nivel de riesgo de un conductor. Pero sin duda en el campo de la publicidad en internet es donde más se ha avanzado con subastas de posicionamientos casi instantáneas.

La aproximación a la solución de estos problemas también ha cambiado. Una forma de visualizar la evolución de estos procedimientos consiste en analizar como la inteligencia artificial ha ido destronando a los mejores jugadores de ajedrez y Go. En 1997, y después de algunos intentos, Deep Blue consiguió vencer a Kasparov zanjando de forma definitiva la disputa de los ajedrecistas con las máquinas. La siguiente frontera era un juego mucho más complejo: el Go. El AlphaGo de Google venció en 2015 a un segundo dan del Go por 5 a 0. Pero el momento definitivo llegó con la victoria sobre Lee Sedol, héroe nacional de Corea y mejor jugador de la historia. AlphaGo ganó a Sedol en el Google DeepMind Challenge de 2016 por 4 a 1. Lo importante no es la demostración del dominio de las máquinas en estos juegos sino la evolución de los procedimientos. Mientras en los sistemas de inteligencia artificial clásicos se introducían las reglas y miles de jugadas de partidas históricas, AlphaGo funciona de una forma muy diferente. Sus procedimientos de redes neuronales aprenden jugando millones de partidas… contra sí mismo. Este sistema permite acumular una cantidad ingente de jugadas, muchas de ellas nunca utilizadas por humanos, así como los resultados de las diferentes estrategias. Los procedimientos de aprendizaje reforzado consiguen evitar los vicios y servidumbres de las jugadas humanas del pasado y los programas anteriores.

3. Datos masivos para el avance social

Los datos masivos y los procedimientos de aprendizaje automático han pasado también a ser la base de una creciente cantidad de investigación social y económica. Un problema de los métodos tradicionales de recopilación de datos, como las encuestas, es que cada vez tienen un nivel de no respuesta superior (Meyer *et al*., 2015) y los encuestados mienten con más frecuencia, en particular en preguntas relacionadas con magnitudes económica importantes como la renta. Además, los datos de encuestas retrospectivas tienen problemas con la fidelidad de las respuestas dado que las preguntas retrospectivas suelen tener un alto nivel de error de medida. Card *et al*. (2010) describen los datos administrativos como muy superiores a las encuestas porque son muestras mucho más grandes (toda la población), son longitudinales por naturaleza y tienen menos problemas de medida, imputaciones o pérdida de muestra[4]. Finalmente, las operaciones estadísticas basadas en encuestas suponen disponer de la información con bastante retraso.

Los datos administrativos y algunos de los datos obtenidos en Internet no presentan estos problemas[5]. Por ejemplo, Davidowitz (2017) define las búsquedas en Internet como el suero de la verdad. Todo el mundo miente… excepto cuando busca en Internet[6]. Por ejemplo, cuando se escribe en el buscador «es mi marido» el buscador completa con el término «homosexual» (al menos en el momento que Davidowitz escribió el libro) en lugar de «infiel» que sería quizás lo primero que nos vendría a la cabeza. Uno podría pensar que existe una gran correlación entre el desempleo en Estados Unidos y las búsquedas de «oficina de empleo» o «vacante». Sin embargo, en momentos de mayor desempleo crecen las búsquedas de «porno» y de juegos de solitario. Se busca con más frecuencia «tamaño de pene» o «alargamiento de pene» que como cambiar una rueda. Se busca mucho más «mi hijo es superdotado» que «mi hija es superdotada» y, sin embargo, se busca más «mi hija es obesa/guapa/fea» que «mi hijo» cuando los niños tienen estadísticamente una probabilidad superior de ser obesos que las niñas.

La disponibilidad creciente de enormes bases de datos, en muchos casos geocodificadas, de procedencia muy diversa hace de la economía una disciplina cada vez más científica. Un ejemplo destacado de este tipo de estudios es el Billion Prices Project (Cavallo & Rigobon, 2016), que estima la evolución en tiempo real de la evolución de los precios utilizando millones de precios de tiendas *online*. Más allá de los resultados estrictamente científicos, el proyecto permitió comparar la evolución

[4] Esta visión contrasta con la aproximación en Barcellan *et al*. (2017) donde el mensaje básico es que no existe ningún tipo de dato superior a otro.

[5] En este sentido, las redes sociales son menos útiles. Por ejemplo, multitud de investigadores han intentado utilizar análisis del sentimiento en Twitter y Facebook para predecir el resultado de elecciones con muy poco éxito. Han tenido más éxito las aplicaciones del análisis textual a las actas de comités de los bancos centrales (Hansen *et al*., 2018).

[6] Este es el motivo por el que Google Trends es una herramienta cada vez más utilizada para hacer todo tipo de predicciones.

de la inflación oficial y de los precios *online* en diversos países de Latinoamérica. Con esta información se pudo comprobar que los patrones eran similares en Brasil, Colombia o Chile pero no en Argentina, donde la diferencia acumulada de la inflación entre 2007 y 2011 fue un 65 % (Cavallo, 2013).

Otros datos que se utilizan cada vez más intensivamente son las imágenes de satélites. Diversos estudios han comprobado la elevada correlación de la luz nocturna con el nivel de desarrollo (Chen & Nordhaus, 2011), la bondad relativa del PIB per cápita y la renta media de las encuestas de presupuestos familiares como medida del desarrollo económico (Pinkovskiy & Sala-i-Martin, 2016) o la desigualdad (García Montalvo *et al*., 2021). Más recientemente, también se está avanzando en complementar la medida a partir de la luz nocturna con imágenes de día en línea con las propuestas originales en Jean *et al*. (2016). Además, los datos de luz nocturna se usan rutinariamente para medir el nivel de desarrollo en áreas pequeñas que no coinciden con unidades administrativas y no tienen información sobre renta o riqueza (García Montalvo & Reynal-Querol, 2021).

Otra fuente de datos que genera una cantidad ingente de información son los teléfonos móviles. Durante el periodo epidémico la geolocalización de los móviles ha servido para construir índices de movilidad para analizar el impacto de la COVID-19 en el movimiento de las personas (Chen *et al*., 2020). Pero quizás una de las propuestas más innovadora sea la que se presenta en Athey *et al*. (2020). Usando datos de GPS obtenidos a partir de teléfonos móviles Athey *et al*. (2020) analiza una nueva medida de segregación denominada aislamiento experimentado, que analiza la exposición de individuos de distintas razas, nivel socioeconómico, etc. en los distintos sitios que visitan a lo largo del día. Con esta información muestran que el aislamiento experimentado es menor que la segregación residencial, aunque ambas medidas están muy correlacionadas por ciudades y áreas geográficas.

Finalmente, otra fuente reciente de datos administrativos masivos muy útiles para la economía es la información de las agencias tributarias. Estos organismos han sido generalmente muy refractarios a compartir cualquier tipo de información, pero en fechas recientes la situación está cambiando. Probablemente este cambio esté en parte relacionado con el éxito del programa de investigación de Chetty y coautores sobre el análisis de la movilidad social y la igualdad de oportunidades en Estados Unidos[7]. Aunque inicialmente hubo cierta polémica sobre si la utilización de los datos se ajustaba a la solicitud original, los espectaculares resultados obtenidos acallaron estas críticas. A partir de esta información es posible analizar en áreas geográficas muy pequeñas, la relación entre la posición en la distribución de la renta de los hijos y sus padres. Uno de los resultados más impactantes se refiere a la proporción de hijos que ganan más que ganaban sus padres en Estados Unidos: mientras en 1940 este porcentaje alcanzaba el 92 %, en 1984 era solo el 50 %, lo que muestra una caída espectacular (Chetty *et al*., 2016).

[7] https://opportunityinsights.org/

El dominio creciente de los datos está cambiando incluso la forma de explicar la economía en los cursos básicos. La puesta en marcha del proyecto CORE *(Curriculum Open-access Resources in Economics)* en 2013 supuso un cambio significativo en cómo se enseñaba economía en muchas universidades. No obstante, en muchas de las grandes universidades los cursos introductorios seguían guiándose por la visión clásica de la enseñanza de los principios de economía. Pero la enseñanza de los principios básicos de la economía basados únicamente en análisis empíricos está llegando a los cursos introductorios. Un ejemplo sintomático es el curso EC10 de la Universidad de Harvard. Este es el curso básico de principios de economía que realizan muchos universitarios en Harvard, desde biólogos hasta físicos y, por supuesto, estudiantes que quieren hacer el *major* en economía. Hasta el año 2019 el curso tuvo una estructura habitual, con muchas curvas de indiferencia, restricciones presupuestarias y cálculo elemental. La sustitución de Greg Mankiw por Raj Chetty como instructor del curso supuso una revolución absoluta en la forma de enseñar el curso. Las curvas de indiferencia dejaron paso a los datos y las cuestiones relativas a la distribución de la renta incrementaron su presencia sustancialmente.

4. La colaboración público-privada

Tradicionalmente la mayor parte de los datos se generaban y almacenaban en instituciones públicas (institutos de estadística, organismos oficiales, etc.) donde el acceso estaba claramente definido[8]. En la actualidad la mayor parte de la generación de datos se encuentra en el sector privado. De esta forma es necesario establecer colaboraciones con las empresas generadoras de datos para poder aprovechar su potencial en la evaluación de políticas públicas, el seguimiento de la economía, etc.

El advenimiento de la pandemia de COVID-19 y la necesidad de contar con información de muy alta frecuencia para hacer un seguimiento de la evolución económica y del impacto de las políticas adoptadas para mitigar el efecto de la pandemia ha hecho más importante la disponibilidad rápida de grandes bases de datos, impulsando la colaboración público-privada. En este caso el calificativo público puede referirse a universidades, organismos públicos y centros de investigación trabajando con grandes bases de datos de empresas privadas, aunque también se puede referir a bases de datos públicas construidas a partir de microdatos de empresas privadas.

Un ejemplo de esta colaboración es el *economic tracker* del *Opportunity Insight*. Chetty *et al*. (2020) describen los datos que utilizan para hacer un seguimiento del impacto de la COVID-19 en diferentes dimensiones. Entre las empresas privadas que colaboran aportando datos están Affinity Solutions (gasto en tarjetas de crédito y débito), CoinOut (transacciones en efectivo), Womply (utilización de tarjetas de crédito en pequeños negocios), Paychex (datos sobre empleo y nóminas de 670.000 pequeñas y medianas empresas), Intuit (ofrece servicios de pago de nóminas a

[8] Otro problema era la cuestión de la coordinación entre diferentes instituciones, los requerimientos para su utilización, etc. Estos temas serán comentados con posterioridad.

empresas), Earnin (ofrece servicios de avance de la nómina que permite acceder a la renta de los trabajadores que se han inscrito en la web), Kronos (servicio de gestión de trabajadores que comprende a 30.000 empresas y 3.2 millones de trabajadores), Homebase (herramienta de gestión de pequeñas empresas) y Zearn (aplicación de matemáticas utilizada por 925.000 estudiantes).

Otro ejemplo de colaboración público-privada son los acuerdos entre universidades y multitud de bancos para utilizar la información de las transacciones en sus cuentas. El listado en estos momentos incluye ya muchos países: Estados Unidos, Reino Unido, Francia, Japón, Dinamarca, Suecia, Islandia, Portugal, Países Bajos, China y España. Las empresas colaboradoras incluyen todo tipo de entidades financieras desde Fintechs hasta bancos tradicionales. El Cuadro 1 recoge un resumen de las investigaciones en marcha y las entidades privadas que les dan soporte.

CUADRO 1
EQUIPOS DE INVESTIGACIÓN Y ENTIDADES FINANCIERAS COLABORADORAS

Autores	País	Entidad privada	Usuarios
Agregadores financieros (aplicaciones financieras y Fintech)			
Kuchler *et al*. (2018)	Estados Unidos	Ready to Zero	516
Hacioglu *et al*. (2021)	Reino Unido	Money Dashboard	8.350
Baker *et al*. (2020)	Estados Unidos	SaverLife	5.746
Baker (2018)	Estados Unidos	Confidencial	156.606
Olaffson & Pagel (2018)	Islandia	Meniga	66.262
Olaffson *et al*. (2018)	Islandia	Meniga	55.545
G. Montalvo & Reynal (2020)	España	Fintonic	236.053
Gelman *et al*. (2014)	Estados Unidos	Check	23.000
Bancos tradicionales			
Bounie *et al*. (2021)	Francia	CIC (Credit Mutuel)/ CB (Carte bancaire)	300.000/1,8 millones de afiliados a CB
Sheridan *et al*. (2020)	Dinamarca/Suecia	Danske Bank	860.000
Aspachs *et al*. (2020)	España	Caixabank	3.028.204
Cox *et al*. (2020)	Estados Unidos	JP Morgan	5.014.672
Carvalho *et al*. (2021)	España	BBVA (tarjetas y TPV)	2.200.000 empresas
Kubota *et al*. (2021)	Japón	Mizuho Bank	2.800.000
Golec *et al*. (2020)	Países Bajos	ABN AMRO	2.000.000
Carvalho *et al*. (2020)	Portugal	Sociedade Interbancária de Serviços	

FUENTE: Elaboración propia.

El objetivo de estas colaboraciones es analizar teorías económicas como la renta permanente (Gelman *et al*., 2014; Olaffson *et al*., 2018), el efecto de los confinamientos (Chetty *et al*., 2020; Sheridan *et al*., 2020), el impacto de las políticas de transferencias de rentas frente a la COVID-19 (Kubota *et al*., 2021; Baker *et al*., 2020), la evolución del gasto durante la pandemia (Carvalho *et al*., 2021; Hacioglu *et al*., 2021; Bouni, 2021), el impacto distribucional de la epidemia sobre el gasto (García Montalvo & Reynal-Querol 2020; Hacioglu *et al*., 2021) o la evolución de la desigualdad mensual (Aspachs *et al*., 2021).

En el caso de España también se están produciendo ejemplos de colaboración entre entidades financieras y equipos de investigación. Un ejemplo es la colaboración entre un equipo de investigadores del departamento de Economía de la Universidad Pompeu Fabra y la unidad de investigación de Caixabank en el análisis en tiempo real de la desigualdad en España desde el comienzo de la epidemia de COVID-19 (Aspach *et al*., 2021)[9] o la colaboración de varios investigadores con el BBVA en la elaboración de información de muy alta frecuencia sobre la evolución del gasto en tarjeta de crédito por tipología de gasto (Carvalho *et al*., 2021). Asimismo, García Montalvo y Reynal-Querol (2020) utilizan los datos de la aplicación de agregación financiera Fintonic para analizar el efecto distribucional de la epidemia sobre el gasto. En todos estos casos se trata de bases de datos enormes. Por ejemplo, Carvalho *et al*. (2021) utilizan 1.300 millones de operaciones con tarjetas y terminales del BBVA. García Montalvo y Reynal-Querol (2020) usan 350 millones de transacciones de la Fintech española Fintonic.

Otro ejemplo de uso de datos privados se produce en el caso de la colaboración entre instituciones públicas u organismos oficiales y empresas. Por ejemplo, el Banco Central Europeo (BCE) usa la base de datos privada AnaCredit (*analytical credit data sets*), que contiene datos armonizados de la zona euro en una base de datos única, para sus análisis prudenciales y supervisorios. Asimismo, el Bank for International Settlements (BIS) recolecta y procesa información bancaria confidencial en colaboración con bancos centrales y autoridades nacionales para su International Data Hub (Doerr *et al*., 2021).

No obstante, sorprende que algunas de estas colaboraciones reciban críticas generalizadas a pesar de perseguir fines sociales. Un ejemplo claro en el caso español fue el anuncio del INE de comprar datos de las compañías telefónicas sobre geolocalización de los clientes para estudiar la movilidad urbana. Cuando se anunció el estudio los medios de comunicación titularon de forma muy crítica: «Así va a rastrear el INE tu móvil durante 8 días: a qué operadoras afecta y cómo evitarlo», «Las operadoras cobrarán medio millón de euros por ceder los datos al INE», «El INE va a espiar tu móvil 8 días: ¿Qué puedes hacer?». Esta reacción del público sorprende sobre todo cuando la mayoría de los usuarios de un teléfono móvil están dispuestos a ceder toda su información personal dando acceso a todos los datos de su móvil para instalarse una simple aplicación de linterna. Sin embargo, si una institución pública hace un

[9] https://inequality-tracker.caixabankresearch.com/

estudio con datos de usuarios de telefonía totalmente anonimizados y agregados por zonas, ponen el grito en el cielo. En este sentido falta claramente una mayor sensibilización de la población sobre la importancia de las bases de datos masivos para la consecución de fines sociales.

5. Limitaciones de los datos masivos y las técnicas de aprendizaje automático

En las secciones anteriores se ha mostrado el interés que tienen los datos masivos y las técnicas de aprendizaje automático para la solución de problemas de mercado, la investigación con objetivo social y la evaluación de políticas públicas. En particular, disponer de grandes bases de datos permite realizar análisis con un nivel de granularidad geográfica muy grande y, en muchos casos, con poco retraso temporal. No obstante, también tiene limitaciones y peligros (García Montalvo, 2014). Evidentemente, el primer peligro es injerencia en la privacidad de las personas con el objetivo de mejorar la capacidad predictiva de los modelos o su utilización para objetivos poco éticos[10]. Este problema es particularmente relevante respecto a los datos producidos por empresas privadas.

En segundo lugar, los algoritmos suelen reproducir los sesgos y la discriminación existente en la realidad *(lack of algorithm fairness)*. Por ejemplo, un algoritmo de *scoring* crediticio puede discriminar a las mujeres o las personas afroamericanas sin necesidad de incluir el identificador del grupo como una de las variables clasificatorias puesto que dicha característica está muy correlacionada con otras de las docenas de variables que se incluyen en el modelo. Un ejemplo se produjo con la investigación que el Departamento de Servicios Financieros del Estado de Nueva York a las tarjetas de crédito de Apple porque el algoritmo cargaba tipos de interés más altos a las mujeres[11]. Además, los procedimientos de aprendizaje automático suelen ser cajas negras que generan correlaciones y buenas predicciones, pero no explican causalidades. En este sentido es cada vez más necesario hacer auditorias de los algoritmos para entender y evitar estos sesgos[12].

En tercer lugar, la calidad de los datos es también un desafío importante. Es bien conocido que una característica habitual de las bases de datos masivas es la baja proporción señal/ruido. Por tanto, precisan de grandes esfuerzos para la limpieza de los datos. Un ejemplo claro son los datos obtenidos en las redes sociales. También hay problemas de representatividad, muestreo y fusión de datos de diferentes fuentes. La posibilidad de que los errores en la captura, fusión o limpieza de los datos

[10] Un ejemplo bien conocido de este problema es el caso de Cambridge Analytics.

[11] El departamento de servicios financieros justificaba su actuación señalando que «cualquier algoritmo que, intencionadamente o no, resulte en un tratamiento discriminatorio de la mujer viola las leyes de Nueva York».

[12] De hecho uno de los motivos tradicionales para utilizar procedimientos de aprendizaje automático es evitar las docenas de sesgos en las decisiones humanas que han sido reconocidos por la economía comportamental y la psicología.

generen consecuencias negativas para los ciudadanos a partir de la aplicación de técnicas de aprendizaje automático a datos de baja calidad es real. Por ejemplo, la industria de generación de calificaciones crediticias a partir de datos capturados en Internet puede provocar serios perjuicios a solicitantes de créditos si sus datos tienen un elevado grado de error. Yu *et al*. (2014) analizaron la información disponible por varias agencias de generación de calificaciones crediticias a partir de datos de Internet. Yu *et al*. (2014) seleccionaron a cinco compañías de *big data* y obtuvo los informes sobre 15 voluntarios para el estudio. Los informes recibidos tenían multitud de errores, estimaciones desmesuradas (salario doble del real del solicitante), direcciones incorrectas, multitud de información faltante (incluidas cuentas en redes sociales, etc.). Ciertamente los datos de las empresas tradicionales de generación de *credit score* de consumidores también son muy mejorables. Un estudio de 2013 de la Federal Trade Commission de Estados Unidos señalaba que el 20 % de los informes crediticios de estas compañías contienen errores y un 5 % de estos errores resultaron en una rebaja del *credit score* que impidió a los clientes conseguir un crédito o les supuso pagar un tipo superior. El problema de las agencias que se basan en *big data* es que los clientes no tienen forma de saber cómo se ha calculado su *credit score*. No hay forma de confirmar independientemente la capacidad predictiva del algoritmo utilizado. De esta forma el consumidor puede acabar siendo afectado negativamente por un *credit score* calculado a partir de datos erróneos aunque, a diferencia del caso de las agencias tradicionales, es más difícil estimar si esto ha afectado a su capacidad de conseguir un crédito o el tipo[13].

Además, incluso si los datos y los procedimientos de predicción son adecuados, es preciso hacer un seguimiento constante de la capacidad predictiva de los modelos pues la captura de datos puede cambiar en el tiempo y perder capacidad explicativa. Un buen ejemplo es lo sucedido con Google Flu trends. Como se ha comentado anteriormente las búsquedas en Google son muy útiles en modelos predictivos. Ginsberg *et al*. (2009) propusieron usar las búsquedas de una bolsa de palabras relacionadas con los síntomas de la gripe para hacer un seguimiento en tiempo real de la evolución de la gripe estacional con una enorme precisión geográfica. De hecho, el Centro de Control de Enfermedades Infecciosas (CDC) de Estados Unidos tardaba casi dos semanas en producir información sobre la evolución de la gripe y usando las búsquedas se podría proporcionar un indicador casi en tiempo real. La capacidad predictiva del procedimiento fue sorprendente en el periodo 2004-2011. Sin embargo, en 2012 la capacidad predictiva se deterioró enormemente como muestran Lazer *et al*. (2014). El motivo aparente de este problema es el cambio en el buscador de Google que sugiere palabras cuando se realiza una búsqueda.

[13] Garcia Montalvo (2014).

6. Investigación económica y acceso a datos masivos

Hace ya muchos años que, ante la previsible expansión de la utilización de bases de datos masivas, economistas académicos han analizado la necesidad de establecer procedimientos estandarizados para facilitar el trabajo de los investigadores. Card *et al*. (2010) presentan una propuesta para expandir el acceso a datos administrativos para investigación económica en Estados Unidos ante lo que describían como «erosión de la ventaja de EE. UU. en la creación de datos para la investigación económica». En aquel momento Card *et al*. (2010) ya se referían a los países escandinavos como ejemplo en la construcción, fusión y facilidad de acceso a grandes bases de datos administrativos. El llamado modelo nórdico ha sido desde siempre el espejo en el que muchos países han querido reflejarse a la hora de conseguir expandir el uso de datos administrativos para el uso en investigación. Tanto las oficinas estadísticas de Finlandia, Dinamarca, Noruega y Suecia permiten el acceso a los investigadores a grandes bases de datos administrativos ya fusionados. El acceso en el caso noruego también puede estar intermediado por el Norwegian Social Sciences Data Service. Más recientemente países como Alemania (IAB) y Francia (CASD) también han puesto en marcha organismos para facilitar el acceso a grandes bases de datos administrativos. Pero sin duda uno de los casos de éxito de los últimos años es la iniciativa del Reino Unido. Uno de los motivos fundamentales de esta iniciativa fue, de forma parecida a lo sucedido en Estados Unidos, la sensación de que el Reino Unido estaba perdiendo el tren de la investigación económica de frontera por las limitaciones que tenían los investigadores para acceder a grandes bases de datos administrativos. En 2014 se puso en marcha la Administrative Data Research Network del Reino Unido (ADRN) como una colaboración entre universidades, organismos gubernamentales, agencias estadísticas e investigadores para facilitar el acceso a datos administrativos ya fusionados entre diversas fuentes[14]. La iniciativa ha dado enormes frutos. En este sentido genera envidia sana que unos meses después de empezar la pandemia de COVID-19 investigadores del Reino Unido fueran capaces de fusionar 17 millones de historiales electrónicos de salud y calcular las probabilidad de mortalidad por COVID-19 en función de decenas de factores de riesgo (Williamson *et al*., 2020).

La situación en España, al menos hasta hace pocos años, fue de considerable retraso respecto a la situación de los países más avanzados en modelos de acceso a datos administrativos masivos. Esta situación de retraso es paradójica dado que los organismos públicos en España generan gran cantidad de datos administrativos de excelente calidad. Y más todavía si se piensa que la Agencia Estatal de Administración Tributaria (AEAT) española tiene una cantidad ingente de datos y es una de las más avanzadas del mundo en la utilización de técnicas de aprendizaje automático para la investigación tributaria. La base de datos Zujar contiene más de 20 billones

[14] Más recientemente, el Reino Unido ha puesto en marcha una estrategia nacional de datos (UK National Data Strategy, 2020).

de registros. Además, cuentan con Prometeo, Infonor, Midas y procedimientos avanzados de *web scraping*. Finalmente, los sistemas de inteligencia artificial de Electra o Teseo son capaces de generar resultados sorprendentes sobre relaciones comerciales de empresas y contribuyentes con imágenes de redes que permiten identificar de forma rápida si el patrón se corresponde con alguna de las tipologías habituales de fraude tributario.

Ciertamente las agencias tributarias han sido siempre muy refractarias a compartir sus datos incluso si se trata de realizar investigación con fines sociales. La AEAT no ha sido una excepción. Evidentemente la naturaleza de los datos recomienda ser muy cuidadosos en su gestión, aunque esto no debería impedir la realización de investigación imponiendo todo tipo de condiciones para garantizar que se cumplen estrictamente unos protocolos de anonimización y uso responsable de los datos. Pero en los últimos años la AEAT ha ido mostrándose más receptiva a compartir sus datos. La colaboración entre la AEAT y el INE en la construcción del Atlas de la Distribución de la Renta de los Hogares es un ejemplo de esta apertura. La Agencia Tributaria también participó muy activamente, y fue fundamental, en la construcción del Sistema Estatal de índices de Alquiler de Vivienda, que proporciona información de precios de alquiler hasta nivel de distrito censal. Otro signo de este nuevo tiempo es que la AEAT haya permitido a la AIReF utilizar las declaraciones de renta anonimizadas (35,5 millones) que se han cruzado con la Encuesta de Presupuestos Familiares para analizar la fiscalidad conjunta de IRPF e IVA de las familias (AIReF, 2020), aunque estos datos no son públicos.

En los últimos años también se han hecho avances en otros datos administrativos. La tradicional muestra continua de vidas laborales (MCVL) se ha enlazado con el IRPF. El PET (panel de datos de empresa-trabajadores) proporciona una visión similar a la MCVL pero desde la perspectiva de la empresa a partir del enlace de los registros sobre las empresas y las vidas laborales de los trabajadores. Asimismo, la creación del laboratorio de datos del Banco de España permite acceder a los datos de la Central de Balances. Dentro de las colaboraciones entre organismos públicos e instituciones de investigación el DataReSS, una colaboración entre el instituto de estadística de Cataluña (IDESCAT) y la Barcelona Graduate School of Economics, facilita el acceso a los registros que producen las administraciones públicas catalanas.

En todo caso, y con los avances, todavía queda mucho camino por recorrer. Como señala con acierto el documento de la AIReF (2020) «España no está aprovechando el potencial para la evaluación de las políticas públicas que se deriva de los datos administrativos en poder de las administraciones públicas». Experiencias recientes muestran que falta todavía mucho por hacer y que las propuestas de AIReF (2020) son un buen punto de partida para recorrer ese camino. De todas formas, en los últimos meses se ha producido dos hitos relevantes que potencialmente podrían ser muy transformadores. En primer lugar, se creó en agosto de 2020 la Oficina del Dato, dependiente de la Secretaria de Estado de Digitalización e Inteligencia Artificial. El artículo 2 del BOE de su creación señala cinco objetivos:

a) El diseño de las estrategias y marcos de referencia en materia de gestión de datos, la creación de espacios de compartición de datos entre empresas, ciudadanos y Administraciones Públicas de manera segura y con gobernanza (*sandboxes*, *data spaces* nacionales y europeos, ecosistemas de datos para uso sectorial tanto público como privado, etc.) y el empleo masivo de los datos en los sectores productivos de la economía mediante tecnologías *Big Data* e Inteligencia Artificial, entre otras, así como el desarrollo de mecanismos de acceso seguros a estas plataformas de datos, para la toma de decisiones públicas basadas en datos o para uso empresarial, garantizando su seguridad y gobernanza a través de arquitecturas API u otros mecanismos
b) El diseño de las políticas de Gobernanza y estándares en la gestión y análisis de datos que deben regir en la Administración General del Estado. Coordinación de modelos, recomendaciones y valoraciones sobre soluciones tecnológicas de codificación, anonimización y tratamiento de datos, sistemas de geolocalización, plataformas y modelos de intercambio, interacciones, modelizaciones y valoración de riesgos, seguridad en la gestión y almacenamiento de los datos, entre otros.
c) El desarrollo de un Centro de Competencia de analítica avanzada de datos que defina las metodologías y mejores prácticas y que asegure que se desarrollan las competencias tecnológicas y las herramientas necesarias para la toma de decisiones basadas en datos por parte de las Administraciones Públicas, permitiendo el desarrollo de políticas basadas en evidencia.
d) La formación y desarrollo de mecanismos de transferencia de conocimiento a los distintos ministerios y Administraciones Públicas.
e) La coordinación técnica de las iniciativas en materia de datos de los distintos departamentos ministeriales y Administraciones Públicas en el marco de las estrategias y programas de la Unión Europea.

El segundo hito relevante es el comunicado institucional del INE, la Agencia Tributaria, la Seguridad Social y el Banco de España (2021) para trabajar conjuntamente en el desarrollo de un sistema de acceso a sus bases de datos con fines científicos de interés público.

Conclusiones

La utilización de datos masivos y algoritmos de aprendizaje automático tienen cada vez un papel más relevante en la aproximación empresarial a las soluciones de mercado y la investigación con fines sociales. La colaboración entre autoridades públicas también puede favorecer la utilización y fusión de datos administrativos de diversas administraciones. Por su parte la creciente participación del sector privado en la generación de datos útiles para la investigación económica (evaluaciones de políticas públicas, seguimiento de la economía a alta frecuencia, etc.) hace cada

vez más importante la colaboración público-privada en el aprovechamiento de estas bases de datos. En este contexto, el acceso a microdatos bancarios proporciona una de las fuentes de información con mayor potencial. Así lo muestran multitud de estudios recientes que utilizan este tipo de datos para analizar, con gran granularidad y alta frecuencia, fenómenos económicos muy relevantes como las consecuencias de la pandemia de COVID-19 o el impacto de las políticas destinadas a atenuar sus efectos.

Hasta hace pocos años el aprovechamiento de los datos administrativos de las instituciones públicas españolas para la evaluación de políticas públicas ha estado muy alejado del enorme avance que se estaba produciendo en otros países. En los últimos tiempos se observa una mayor sensibilización de las instituciones públicas respecto a la importancia de los datos administrativos para la investigación económica, aunque todavía queda mucho camino por recorrer en la sensibilización del público y los medios de comunicación. Ejemplos de este cambio son la creciente colaboración de la Agencia Tributaria en la construcción de estadísticas derivadas de sus datos, la creación de la Oficina del Dato y el reciente anuncio de INE, la Agencia Tributaria, la Seguridad Social y el Banco de España de desarrollar un sistema de acceso a sus bases de datos con fines científicos de interés público. Aunque llevamos bastante retraso respecto a otros países estas iniciativas abren una ventana de oportunidad que, dependiendo de cómo se concrete, puede producir un avance muy significativo en la calidad y relevancia de la investigación económica en España.

Referencias bibliográficas

AIReF, Autoridad Independiente de Responsabilidad Fiscal (2020, octubre). *Opinión para una estrategia de acceso a datos administrativos*. Opinión 1/20.

Aspachs, O., Durante, R., Graziano, A., Mestres, J., Reynal-Querol, M., & Montalvo, J. G. (2021). Tracking the impact of COVID-19 on economic inequality at high frequency. *PloS one*, *16*(3). https://doi.org/10.1371/journal.pone.0249121

Athey, S., Ferguson, B. A., Gentzkow, M., & Schmidt, T. (2020). *Experienced segregation* (NBER Working paper No. 27572). National Bureau of Economic Research.

Baker, S. R. (2018). Debt and the response to household income shocks: Validation and application of linked financial account data. *Journal of Political Economy*, *126*(4), 1504-1557.

Baker, S. R., Farrokhnia, R. A., Meyer, S., Pagel, M., & Yannelis, C. (2020). *Income, liquidity, and the consumption response to the 2020 economic stimulus payments* (NBER Working paper No. 27097). National Bureau of Economic Research.

Barcellan, R., Nielsen, P., Calsamiglia, C., Camerer, C., Cantillon, E., Crépon, B., ... Wright, L. (2017). Developments in Data for Economic Research. In L. Matyas, R. Blundell, E. Cantillon, B. Chizzolini, M. Ivaldi, W. Leininger et al. (Eds.), *Economics without borders: Economic Research for European Policy Challenges* (pp. 568-611). Cambridge University Press. doi:10.1017/9781316636404.015

Bounie, D., Camara, Y., Fize, E., Galbraith, J., Landais, C., Lavest, C., & Savatier, B. (2020). Consumption dynamics in the covid crisis: real time insights from French transaction bank data. *Covid Economics*, *59*, 1-39.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. *American Economic Association, ten years and beyond: Economists answer NSF's call for long-term research agendas*. National Science Foundation

Carvalho, V. M., Hansen, S., Ortiz, A., Garcia, J. R., Rodrigo, T., Rodriguez Mora, S., & Ruiz de Aguirre, P. (2020). *Tracking the COVID-19 crisis with high-resolution transaction data*. CEPR Discussion Paper No. DP14642.

Carvalho, P., Peralta, S. & dos Santos, J. P. (2020). *What and how did people buy during the Great Lockdown? Evidence from electronic payments*. ECARES Working Paper 2020-20.

Cavallo A., & Rigobon R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, *30*(2), 151-78.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152-165.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, *108*(21), 8589–8594.

Chen, S., Igan, D. O., Pierri, N., Presbitero, A. F., Soledad, M., & Peria, M. (2020). Tracking the economic impact of COVID-19 and mitigation policies in Europe and the United States. *IMF Working Papers*, *2020*(125).

Chetty, R., Friedman, J. N., Hendren, N., & Stepner, M. (2020). *Real-time economics: A new platform to track the impacts of COVID-19 on people, businesses, and communities using private sector data* (NBER Working Paper No 27431). National Bureau of Economic Research.

Chetty, R., Grusky D., Hell M., Hendren N., Manduca R., & Narang J. (2016). The Fading American Dream: Trends in Absolute Income Mobility Since 1940. *Science* 356(6336), 398-406.

Cox, N., Ganong, P., Noel, P., Vavra, J., Wong, A., Farrell, D., ... Deadman, E. (2020). Initial impacts of the pandemic on consumer behavior: Evidence from linked income, spending, and savings data. *Brookings Papers on Economic Activity*, *2020*(2), 35-82.

Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really*. Dey Streets Books

Derman, E. (2004). *My life as a quant*. John Wiley & Sons.

Doerr, S., Gambacorta, L., & Garralda, J. M. S. (2021). Big data and machine learning in central banking. *BIS Working Papers*, (930).

Duhigg, C. (2011). *The power of habit*. Random House.

García Montalvo, J. (2003). Liquidity and market makers: a pseudo-experimental analysis with ultrahigh frequency data. *The European Journal of Finance*, *9*(4), 358-378.

García Montalvo, J. (2014). El impacto del big data en los servicios financieros. *Papeles de Economía Española*, *Número extraordinario*, 43-59.

García Montalvo, J., & Reynal-Querol M. (2020). *Distributional Effects of COVID-19 on Spending: A First Look at the Evidence from Spain*, Barcelona GSE Working Paper 1201.

García Montalvo, J. (2021). Data science y sus aplicaciones económicas: una perspectiva personal. En D. Peña, P. Poncela & E. Ruiz (Eds.), *Análisis Econométrico y Big Data* (pp. 5-24). FUNCAS.

García Montalvo, J., Reynal-Querol M., & Muñoz J.C. (2021). *Measuring inequality from above*. Barcelona GSE Working Paper No. 1252.

García Montalvo, J., & Reynal-Querol, M. (2021). Ethnic diversity and growth: Revisiting the evidence. *Review of Economics and Statistics*, *103*(3), 521-532.

Gelman, M., Kariv, S., Shapiro, M. D., Silverman, D., & Tadelis, S. (2014). Harnessing naturally occurring data to measure the response of spending to income. *Science*, *345*(6193), 212-215.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.

Golec, P., Kapetanios, G., Neuteboom, N., Ritsema, F., & Ventouri, A. (2020). *Disentangling the effect of government restrictions and consumers' reaction function to the Covid-19 pandemic: evidence from geo-located transactions data for The Netherlands*. DAFM Working Paper Series, No. 2020/4.

Hacıoğlu-Hoke, S., Känzig, D. R., & Surico, P. (2021). The distributional impact of the pandemic. *European Economic Review*, *134*, 103680.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801-870.

Instituto Nacional de Estadística, Seguridad Social & Banco de España. (2021, 13 de abril). Comunicado oficial por el cual el Instituto Nacional de Estadística, la Seguridad Social y el Banco de España acuerdan trabajar conjuntamente en el desarrollo de un sistema de acceso a sus bases de datos con fines científicos de interés público. [comunicado de prensa].

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Kubota, S., Onishi, K., & Toyama, Y. (2021). Consumption responses to COVID-19 payments: evidence from a natural experiment and bank account data. *Journal of Economic Behavior & Organization*, *188*, 1-17.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Olafsson, A., & Pagel, M. (2018). The liquid hand-to-mouth: Evidence from personal finance management software. *The Review of Financial Studies*, *31*(11), 4398-4446.

Pinkovskiy, M., & Sala-i-Martin, X. (2016). Lights, camera… income! Illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, *131*(2), 579-631.

Sheridan, A., Andersen, A. L., Hansen, E. T., & Johannesen, N. (2020). Social distancing laws cause only small losses of economic activity during the COVID-19 pandemic in Scandinavia. *Proceedings of the National Academy of Sciences*, *117*(34), 20468-20473.

Williamson, E., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., ... Goldacre, B. (2020). OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. *MedRxiv*.

Yu, P., McLaughlin, J., & Levy, M. (2014). Big Data: A Big Disappointment for Scoring Consumer Credit Risk. *NCLC, National Consumer Law Center, Boston, MA*, *14*.