CAPÍTULO I

Data Science y sus aplicaciones económicas: una perspectiva personal

José García Montalvo*

En este trabajo se realiza un recorrido por diversas aplicaciones económicas basadas en *Big Data* y la aplicación de técnicas de *Machine Learning.* Las aplicaciones incluyen la ingeniería reversa de procedimientos de calificación crediticia, el seguimiento de la economía a muy alta frecuencia durante la pandemia de la COVID-19, la predicción del precio de la vivienda a nivel de código postal, la construcción de indicadores de crecimiento y desigualdad usando imágenes de satélites y la predicción electoral. El objetivo es destacar, utilizando estas aplicaciones, los aspectos que aportan más posibilidades en la utilización de dichas técnicas en el campo de la economía, así como matizar las excesivas expectativas que estos procedimientos puedan haber generado.

Palabras clave: árboles de decisión, imágenes de satélite, *scoring* crediticio.

* El autor agradece la colaboración, en las aplicaciones con datos bancarios, de Miguel Barcino, Estrella García, Lupina Iturriaga, Olga Pascual, José Carlos Pla y Guillermo Soler. Especial agradecimiento a los coautores de los estudios sobre los que se basa este trabajo: Oriol Aspach, Ruben Durante, Alberto Graziano, Josep Mestres, Omiros Papaspiliopoulos, J. M. Raya, Marta Reynal-Querol, y Timothée Stumpf-Fétizon.

1. INTRODUCCIÓN

La introducción de la ciencia de los datos *(Data Science)* en la economía, la empresa y las finanzas ha avanzado enormemente en los últimos años. En García-Montalvo (2014) se introducen los elementos fundamentales que caracterizan los proyectos de *Data Science,* entendido como la fusión del big data y las técnicas de *Machine Learning*[1], especialmente en contraste con la metodología habitual en econometría. Asimismo se analizan las primeras aplicaciones exitosas de técnicas de big data (BD) y *Machine Learning* (*ML*) en el campo de la economía (el proyecto Billion Prices de MIT o las predicciones híbridas utilizando indicadores adaptados de Google Trends) y, particularmente, en el sector financiero. García-Montalvo (2014) analiza la aplicación de técnicas de ML a las calificaciones crediticias de familias y empresas, tanto en su versión tradicional como en sus versiones basadas en la reputación social *(credit score social)*. Entre otras aplicaciones al ámbito financiero y bancario se incluían la protección contra el fraude en tarjetas de crédito, la aplicación al *peer-to-peer lending,* la segmentación de los clientes bancarios, el cumplimiento de la normativa regulatoria, la optimización del capital bancario o la creciente personalización de los precios de los seguros. El artículo acaba con algunos peligros de la utilización indiscriminada de algoritmos de ML a bases de datos cada vez más masivas. En particular, el peligro de pérdida de privacidad, la replicación de la discriminación observada en la realidad, la confianza en una caja negra o la necesidad de actualizar los procedimientos cuando varían las condiciones en las que se recogen los datos como demuestra el caso de la predicción de la gripe con el Global Flu Trends.

Este artículo es una continuación de aquel primer trabajo. Con la perspectiva del tiempo transcurrido he podido comprender mejor las dificultades de los alumnos de econometría para entender la aproximación adoptada en los proyectos de BD y las técnicas de ML. También he tenido multitud de experiencias en proyectos, cada vez más intensivos en la utilización de datos, que me han permitido mejorar la evaluación de qué cosas tienen sentido y qué cosas son meros fuegos de artificio sin ninguna implicación de mejora de nuestra comprensión de los fenómenos analizados. Por último, algunos de los peligros potenciales comentados en el artículo de 2014 se han manifestado como inconvenientes efectivos. El objetivo de este artículo es desarrollar estos tres temas actualizando la visión del artículo original y desarrollando nuevos temas.

2. *MACHINE LEARNING* PARA ECONÓMETRAS

En García-Montalvo (2014) se desarrollan los conceptos fundamentales que diferencian la aproximación basada en el ML de la aproximación de la econometría moderna. Fundamentalmente se enfatizaba el elevado nivel de ruido frente a señal del BD comparativamente

[1] En este trabajo se usa el término aprendizaje en sentido amplio. Estrictamente hablando se debe distinguir entre la aproximación basada en el aprendizaje y la basada en el diseño. En la visión basada en el aprendizaje los datos son necesarios para derivar la función objetivo mientras que en la basada en el diseño se postula una función objetivo con anterioridad a disponer de los datos.

con los datos utilizados normalmente en econometría así como la diferencia en perspectivas: mientras la econometría transitaba hacia técnicas que permitieran identificar y explicar los efectos causales, el ML solo está interesado en producir modelos predictivos y, por tanto, correlaciones. En econometría siempre tenemos un componente aleatorio que permite obtener bandas de confianza y realizar inferencias. Este no es el caso, en general, de las técnicas de ML. Estas diferencias no son complejas para un estudiante que ha superado un primer curso de econometría[2]. Los problemas, como ya sucedió en el pasado cuando otras técnicas estadísticas saltaron al trabajo con datos económicos, empiezan por la nomenclatura. Los libros de texto sobre machine learning hablan de *features* en lugar de variables explicativas, *labels* en lugar de variables endógenas, de muestras de entrenamiento y testeo en lugar de simplemente referirse a una muestra, etc. De la misma manera hablan obsesivamente de *sobreajuste* y el uso de *cross validation* sin referirse la mayor parte de las veces a R^2 o el R^2 ajustado. Finalmente se habla de LASSO, concepto que un estudiante de econometría de hace 30 años habría identificado como una generalización de la regresión *ridge* pero al que la mayoría de los estudiantes más recientes no han sido expuestos.

A estas diferencias terminológicas se unen profundas diferencias conceptuales. Un primer punto son los procedimientos de aprendizaje no supervisado que chocan con la visión tradicional de la econometría que trata con procedimientos supervisados. La teoría de la generalización y sus implicaciones respecto a la existencia de muestra de entrenamiento y muestras de testeo supone otra novedad conceptual. El estudiante de econometría está acostumbrado a la descomposición del error cuadrático medio de un estimador entre sesgo al cuadrado y varianza. Cuando hablas de los componentes del error esperado en la muestra de testeo tiene tres componentes (sesgo cuadrado, varianza y ruido) la influencia de la visión econométrica tradicional dificulta la comprensión del *trade-off* entre varianza y sesgo derivado de la complejidad creciente del algoritmo. Es más, cuesta explicar que el efecto no tiene relación con N (el tamaño de la muestra) sino con K (la complejidad del algoritmo). Si el sesgo es elevado hay que usar modelos más complejos (kernelizar), añadir features o utilizar *boosting* para algoritmos que producen resultados con sesgo. Si el problema es el contrario (varianza elevada o *sobreajuste*) la opción es añadir más datos, reducir la complejidad del modelo, por ejemplo, usando técnicas de regularización, *bagging*, etc. Otros aspectos difíciles de transmitir son la factibilidad del aprendizaje o, en la teoría de la generalización, la importancia de la dimensión de Vapnik por contraposición al simple número de parámetros a estimar.

3. CALIFICACIONES CREDITICIAS, RECOMENDACIONES E INGENIERÍA INVERSA

Montalvo y Reynal-Querol (2020) analizan la influencia del género del gestor encargado de la decisión sobre la concesión de un crédito sobre la posterior probabilidad de impago. Se dispone de información sobre más de un millón de solicitudes de créditos con las

[2] Pongamos un estudiante que entiende con claridad el funcionamiento del modelo lineal, incluida la predicción, y que sabe trabajar con variables dicotómicas tanto explicativas como endógenas.

características de la solicitud, destino, situación financiera del solicitante, colateral, etcétera. así como la calificación crediticia y la recomendación generada por el sistema experto. Los resultados muestran que las mujeres concedieron los créditos con recomendación negativa muchas menos veces que los hombres durante el periodo 2000-2008 y que este mayor seguimiento de la recomendación produjo una proporción menor de impagos cuando comenzó la crisis financiera. Para profundizar en este resultado nos planteamos realizar un análisis de discontinuidad en torno a la puntuación que marcaba la recomendación de no concesión. Sin embargo, la situación no era sencilla, puesto que la visualización de la relación entre *score* y recomendación mostraba con claridad que no existían unos límites claramente definidos en función de la puntuación. La figura 1 muestra un ejemplo con los resultados de la tabla referida a créditos hipotecarios para no clientes[3]. El eje de las abscisas representa el momento de concesión del crédito y el de ordenadas la puntuación asignada. La escala de grises representa el año de concesión. El primer panel contiene los créditos que el sistema de recomendación considera de riesgo bajo o muy bajo (VLR y LR), el segundo los créditos de riesgo neutral (NR), el tercero los de riesgo alto (HR) y el cuarto los de riesgo muy alto (VHR). Como se puede comprobar las puntuaciones para diferentes tablas se solapan a pesar de que las recomendaciones de riesgo bajo o neutro suponían concesión, mientras que las de riesgo alto o muy alto implicaban que, por lo general, debían rechazarse.

De esta forma es evidente que no existe una relación lineal entre puntuaciones y recomendaciones que permita construir un diseño de discontinuidad. Además, se comprueba que las tablas cambian en el tiempo, lo que supone una dificultad adicional.

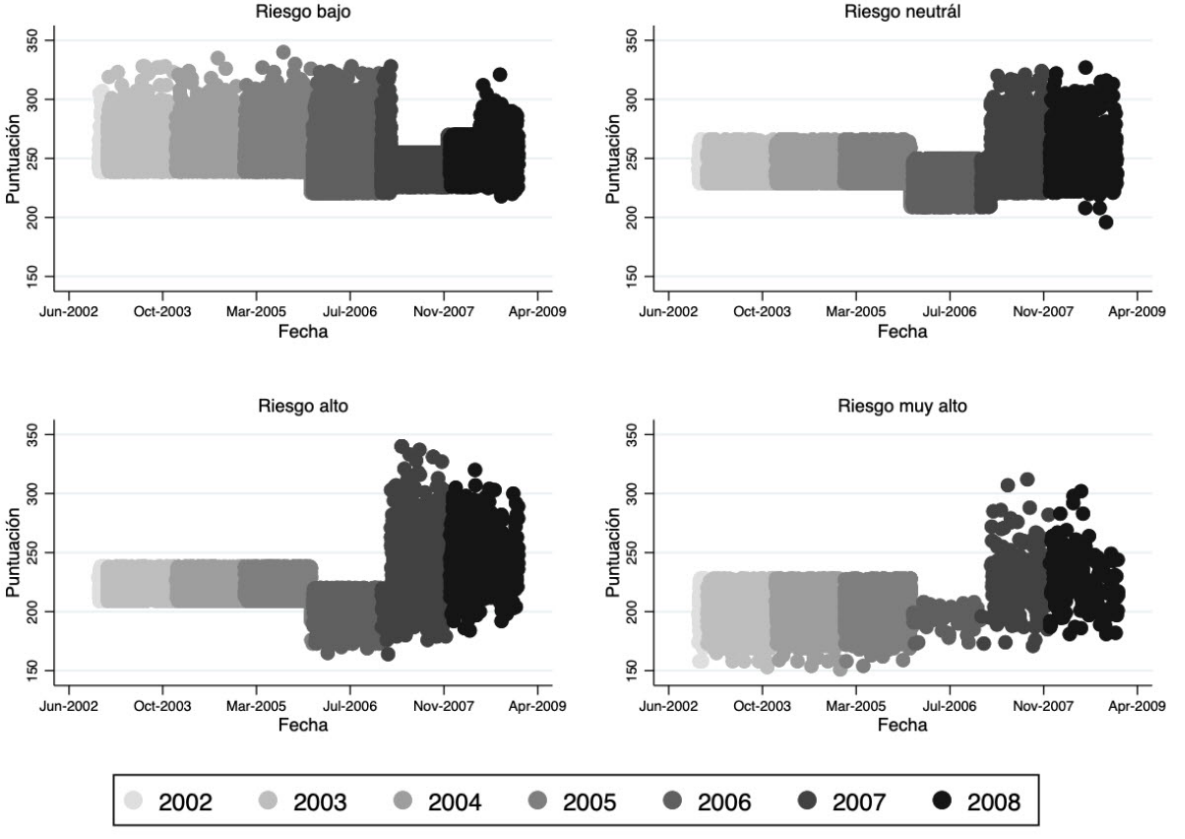
Como los datos se referían a un periodo bastante antiguo y la entidad que había proporcionado la información había desaparecido fue imposible encontrar los algoritmos que habían generado las recomendaciones[4]. De esta forma, la única opción viable fue utilizar ingeniería inversa: descubrir los algoritmos a partir de las recomendaciones y el resto de variables disponibles. Las variables consideradas fueron el *scoring,* la relación préstamo/valor *(loan to value, LTV)*, la relación préstamo/renta *(loan to income, LTI),* el tamaño del préstamo, la extensión temporal, el crédito sobre el patrimonio, la edad, el número de años como cliente, el balance medio en los últimos 6 meses, la ratio entre el balance medio a 6 y 12 meses, el tipo de contrato y su duración, el destino del crédito y el estado civil, entre otras. En principio se probaron diversos algoritmos, pero el que generalizaba mejor era un simple árbol de decisión. De hecho, no hizo falta ningún tipo de *bagging* o versiones tipo *random forest.* El grado de impureza de las ramas finales era 0 y así sucedía en todas las tablas. Por ejemplo, la figura 2 muestra el ajuste del árbol de decisión para las recomendaciones de los créditos hipotecarios a no clientes donde las predicciones de todos los tipos de recomendaciones de la última hoja se corresponden con precisión con las recomendaciones efectivamente observadas. Las únicas variables que tienen capacidad predictiva son la puntuación y el LTI.

[3] El banco mantenía 13 tablas de puntuación en función del producto (crédito hipotecario, consumo, subrogación hipotecaria, etc.) y de la relación del solicitante con el banco (cliente, no cliente).

[4] Se intentó contactar con los antiguos gestores de riesgos, pero obviamente no recordaban con precisión los algoritmos utilizados para cada producto y no tenían los manuales que explicaban los procedimientos. También contactamos con la empresa que colaboró en la implementación de los algoritmos que tampoco nos pudo dar información.

Figura 1.

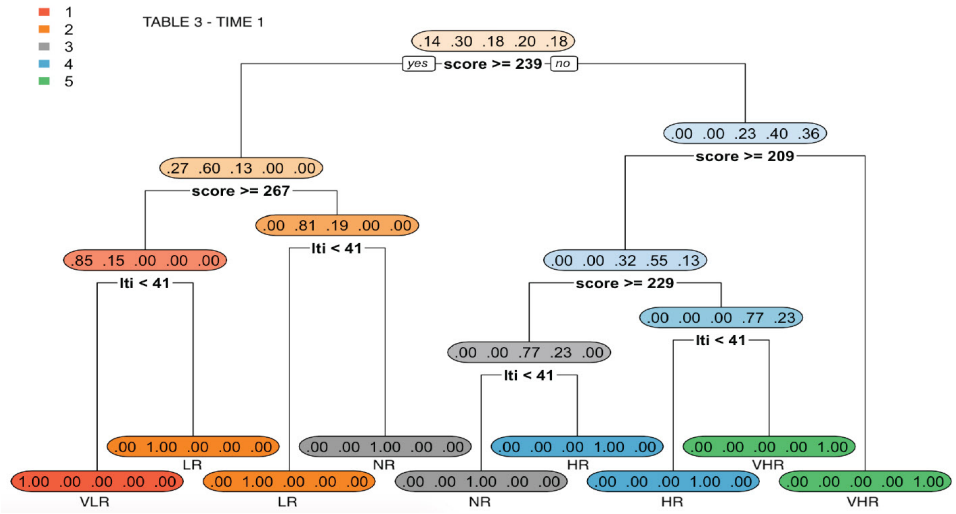
Relación entre puntuaciones y recomendaciones para créditos hipotecarios a no clientes



Fuente: Montalvo y Reynal-Querol (2020).

Figura 2.

Árbol de decisión para la tabla 3 (primer periodo)



Fuente: Elaboración propia.

Este ejercicio muestra un ejemplo claro de utilización adecuada y efectiva de técnicas de ML en una situación donde la única solución es la reingeniería del algoritmo que generó los datos originales.

4. *BIG DATA,* COVID-19 Y ECONOMÍA EN TIEMPO REAL

La propagación de COVID-19 ha afectado enormemente la actividad económica en todo el mundo. Los gobiernos han tomado muchas medidas para responder a la pandemia. Existe un alto grado de incertidumbre sobre el efecto de las respuestas de política y la idoneidad del monto total de apoyo público a la economía. Sin embargo, la mayoría de los indicadores oficiales y las estadísticas macroeconómicas tienen una frecuencia baja y se producen con grandes retrasos. Este es un desafío importante para los responsables de la formulación de políticas en sus esfuerzos por adaptar sus respuestas para aplanar la curva de recesión (Gourinchas, 2020) después de aplanar la curva de infección.

La necesidad de tener información actualizada con alta frecuencia sobre las principales macromagnitudes y el impacto de las medidas de política económica ha generado diversas iniciativas internacionales para rastrear en tiempo real (o con indicadores de alta frecuencia) la evolución de la actividad económica. Por ejemplo, Cicala (2020) utiliza el uso de electricidad de la red europea para representar la evolución de la actividad económica, ya que el consumo de electricidad está altamente correlacionado con su uso empresarial. Bick y Blandin

(2020) recurren a una encuesta en tiempo real *(RPS),* siguiendo la estructura de *CPS (Current Population Survey)* para construir estimaciones de alta frecuencia de empleo, horas trabajadas e ingresos. Chen *et al.* (2020) muestran que estas medidas de alta frecuencia de consumo de energía y horas trabajadas están fuertemente correlacionadas con los indicadores de movilidad del *Informe de Movilidad de la Comunidad de Google.* Chetty *et al.* (2020) han creado un rastreador económico para medir la actividad económica con alta frecuencia en los EE. UU. El objetivo es utilizar información de empresas privadas para construir agregados que tengan utilidad para el seguimiento de la economía en tiempo real y la medición del impacto de las medidas de política económica. La ventaja de esta información es su inmediatez, la disponibilidad con muy alta frecuencia (incluso diaria) y la posibilidad de agregarla en función de muy diferentes criterios. Chetty *et al.* (2020) utilizan información económica anónima de empresas privadas para medir el gasto de los consumidores (transacciones basadas en tarjetas de Affinity Solutions); la tasa de cierre de pequeñas empresas (negocios que realizan transacciones en un día determinado desde Womply); el tiempo dedicado al trabajo (datos GPS proporcionados por Google); o las horas trabajadas en pequeñas empresas (proporcionadas por Homebase).

En los últimos años, la investigación económica ha comenzado también a aprovechar la ingente cantidad de datos disponibles en las aplicaciones de gestión de finanzas personales o en las cuentas bancarias de las entidades financieras, para analizar diversas hipótesis. Gelman *et al.* (2014) utilizan datos diarios de un agregador financiero para estudiar la hipótesis del ingreso permanente (PIH), encontrando un exceso de sensibilidad del consumo al ingreso. Gelman *et al.* (2014) encuentran que, alrededor del período de recepción del salario y los pagos de la Seguridad Social, hay un significativo aumento del consumo. Este resultado se produce tanto en el gasto total, como en el gasto no recurrente y el gasto no esencial (como, por ejemplo, comida rápida y cafeterías). Olafsson y Page (2018) estudian también la PIH utilizando datos del agregador de *software* financiero de Islandia Meniga. Como Gelman *et al.* (2014) encuentran una respuesta significativa del consumo en el día de pago que es consistente en todas las categorías de ingresos y gastos. Baker (2018) también utiliza datos de un sitio web de finanzas personales en línea y encuentra que la heterogeneidad en la elasticidad del consumo se puede explicar por el crédito y la liquidez.

La disponibilidad de información económica muy granular y casi en tiempo real permite analizar el impacto de las políticas públicas con muy pequeños retardos respecto al momento de su ejecución. Por ejemplo, diversas investigaciones ya han comenzado a utilizar datos de alta frecuencia para analizar el impacto de los paquetes de estímulo económico para mitigar el efecto de la epidemia de COVID-19 en la actividad económica. Dos ejemplos para el caso de Estados Unidos son el efecto sobre el empleo agregado del Programa de Protección de Pagos (PPP) para pymes o el efecto sobre el consumo de los cheques de estímulo enviados por la Administración Federal. Autor *et al.* (2020) muestran que el empleo en las empresas elegibles aumentó entre el 2 y el 4,5 %, aunque los resultados son preliminares. Baker *et al.* (2020), utilizando datos de agregación financiera y aplicaciones de servicios, muestran que la propensión marginal a consumir a partir de estos cheques se sitúa entre el 0,26 y el 0,36 dependiendo de que la muestra tenga los pesos originales o se posestratifique para representar a la población de Estados Unidos. Cox *et al.* (2020) utilizan información de cuentas bancarias

anónimas de varios millones de clientes de JP Morgan Chase para estudiar el efecto heterogéneo de la pandemia de COVID-19 en el gasto y el ahorro. Sheridan *et al.* (2020) utilizan datos de un gran banco escandinavo para mostrar que las leyes de distanciamiento social tuvieron un pequeño impacto en la actividad económica. Sheridan *et al.* (2020) comparan los datos de Suecia, que no impuso un confinamiento estricto, con los datos de Dinamarca, que sí cerró la economía. La diferencia como resultado del cierre son 4 puntos porcentuales adicionales de reducción del gasto en Dinamarca. Este resultado concuerda con los hallazgos en Chetty *et al.* (2020) que muestran que la fuerte reducción en el gasto de los consumidores, el gasto de las pequeñas empresas y el tiempo dedicado al trabajo en la economía de EE. UU. comenzó semanas antes de la orden de quedarse en casa y el cierre del negocio no esencial y no se recuperó inmediatamente después del levantamiento de la orden de permanecer en casa.

Desde el comienzo de la pandemia de COVID-19, un número creciente de investigaciones han utilizado datos de agregadores financieros, TPV y tarjetas de crédito, y cuentas bancarias para analizar con datos semanales o diarios, la evolución del gasto total, por tipologías y por nivel de renta de los individuos. Además de los ejemplos comentados de Estados Unidos o Dinamarca también existen trabajos para el caso de Francia (Bounie *et al.,* 2020), Reino Unido (Hacioglu, Kanzig y Surico, 2020), Japón (Kubota, Koichiro y Toyama, 2020), Países Bajos (Golec *et al.,* 2020), China (Chen, Quian y Wen, 2020) o Portugal (Dos Santos, Carvalho y Peralta, 2020).

En España, Carvalho *et al.* (2020) utilizan las transacciones con tarjeta de crédito y TPV de BBVA (1.300 millones de transacciones) para medir la evolución del consumo por categorías. CaixaBank Research publica notas semanales que analizan el uso de tarjetas de crédito en TPV, transacciones *online* y retiradas de efectivo para estimar el impacto de las medidas de confinamiento en el consumo. Este tipo de información también es explotada habitualmente por otros bancos como Abanca (Observatorio Abanca por IESIDE) o el Banc Sabadell (Pulso)[5].

Para el caso de España, García-Montalvo y Reynal-Querol (2020) utilizan datos de 300 millones de transacciones en cuentas bancarias vinculadas por el agregador financiero Fintonic, una de las Fintech de gestión de finanzas personales más grandes de España, para analizar el efecto distributivo de la pandemia en el gasto. En este trabajo se analizan tres dimensiones de la heterogeneidad: ingreso, saldo de la cuenta y edad. García-Montalvo y Reynal-Querol (2020) muestran que, a diferencia de los resultados en otros países como EE. UU. (Cox *et al.,* 2020; Chetty *et al.,* 2020) y Reino Unido (Hacioglu, Känzig y Surico, 2020), no se encuentran diferencias significativas en la evolución del gasto agregado por nivel de ingresos: la caída y la recuperación del gasto es similar para todos los cuartiles de la distribución de la renta. Sin embargo, sí aparece alguna diferencia en la evolución del gasto por edad y por el saldo medio de las cuentas bancarias.

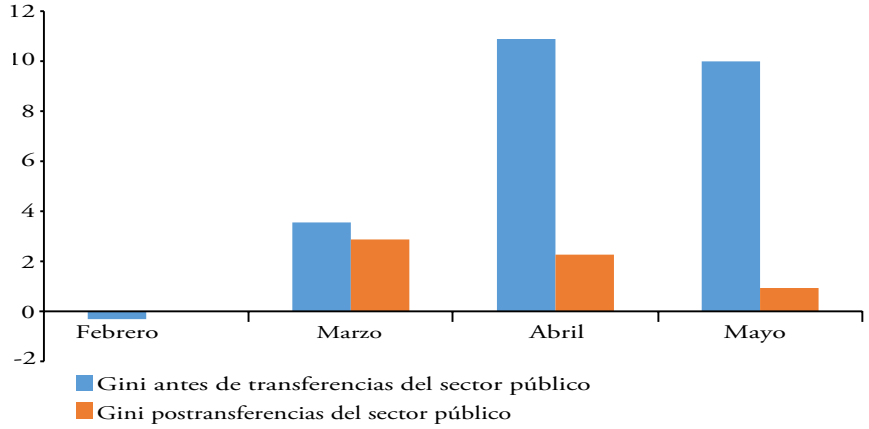
Una de las consecuencias habituales de las pandemias es el aumento de la desigualdad (Wade, 2020). En un trabajo muy diferente a los anteriores, Aspachs *et al.* (2020) describen

[5] El Banco de España también ha utilizado POS para rastrear el gasto durante la pandemia (Gonzalez, Urtasun y Perez, 2020).

una metodología para calcular la evolución de la desigualdad de los salarios utilizando datos de las cuentas bancarias. A diferencia de las muestras poco representativas que suelen ser habituales en los estudios que utilizan este tipo de datos[6], Aspachs *et al.* (2020) muestran que la distribución de los salarios de las nóminas depositadas en CaixaBank es prácticamente idéntica a la distribución de salarios de la *Encuesta de Estructura Salarial* del INE actualizada a 2020[7]. A partir de una base de datos de más de 3 millones de observaciones mensuales, Aspachs *et al.* (2020) calculan los movimientos mensuales en la distribución de los salarios (antes y después de considerar las políticas públicas) así como diferentes indicadores de desigualdad. La figura 3 muestra el enorme incremento del índice de Gini asociado al comienzo de la pandemia (más de 11 puntos) y la capacidad de los ERTE y la extensión del seguro de desempleo para moderar este incremento, aunque sin llegar a compensarlo totalmente. La granularidad de los datos permite analizar la evolución de la desigualdad por género, edad, nacionalidad, localización geográfica, etcétera.

Figura 3.

Cambio en la desigualdad salarial (Gini) en España entre febrero y mayo de 2020



Fuente: Aspachs *et al.* (2020).

5. PREDICIENDO PRECIOS DE INMUEBLES A NIVEL DE CÓDIGO POSTAL

Los cursos sobre Machine Learning suelen utilizar la predicción de precios residenciales como una de las aplicaciones más frecuentes. Supongamos que quisiéramos predecir los precios de distintos tipos de inmuebles (residenciales, comerciales, oficinas y suelo) en España a nivel de código postal a largo plazo. Contamos con información sobre los precios de

[6] Especialmente si se usan datos de agregadores financieros, como por ejemplo Gelman *et al.* (2014) o Baker (2018).

[7] Aspachs *et al.* (2020) también comprueban que la distribución de las variables demográficas (género y edad) son prácticamente idénticas a las observadas en las encuestas para los trabajadores por cuenta ajena.

compraventa trimestrales durante varios años en aquellos códigos donde se han producido transacciones así como de los precios de oferta y el número de transacciones en cada código postal. García-Montalvo *et al.* (2018) proponen un procedimiento basado en estimar la elasticidad respecto al crecimiento de la renta en cada código postal para cada tipo de inmueble[8]. Las predicciones pretendían utilizar la metodología más adecuada al problema sin perder de vista que la incertidumbre asociada a un ejercicio de esta naturaleza es enorme. García-Montalvo *et al.* (2018) aplican un procedimiento para hacer predicciones del PIB y los precios de la vivienda a largo plazo usando una combinación de VARs e inferencia a baja frecuencia (útil para predicciones a muy largo plazo). Las proyecciones de los precios agregados de la vivienda se supone que inciden en los cambios de los precios a nivel de código postal por analogía con un modelo financiero de valoración basado en un único factor.

Para la predicción a nivel de código postal se utiliza un modelo jerárquico (HML) que, posteriormente, permite agregar las predicciones a nivel municipal, provincia, regional e incluso nacional. Además de los precios agregados se incluye un factor idiosincrático asociado a la presión de la demanda en cada código postal. Se utilizan tres indicadores alternativos para cada código postal: la proporción de unidades en venta que disminuyeron su precio entre dos periodos consecutivos; la proporción de unidades que se mantienen de un periodo al siguiente; y la proporción de transacciones sobre las unidades ofertadas en los portales en el trimestre anterior.

En el modelo de precios residenciales tanto los parámetros del efecto del precio agregado como de la *proxy* de la presión de la demanda cambian para cada código postal. Esto implica que solo el modelo de precios residenciales supone estimar miles de parámetros.

5.1. Predicción agregada a largo plazo

El modelo básico es un vector autorregresivo bayesiano *(BVAR)* con variables $i \in 1:I$, periodos $i \in 1:T$ y retardos $l \in 1:L$)

$$y_t = \alpha + \sum_{l=1}^{L} \mathrm{B}_l y_{t-l} + \varepsilon_t, \quad \varepsilon_t \,|\, \Sigma_\varepsilon \sim \left(0, \Sigma_\varepsilon\right) \qquad [1]$$

donde B_l son $I \times I$ matrices de coeficientes Σ_ε y $I \times I$ es una matriz positiva definida con la covarianzas de los residuos[9]. Las variables incluidas en el vector son: PIB real, formación bruta de capital, consumo privado, tasa de desempleo, EONIA *(overnight),* IBEX 35 e índice de precios de la vivienda (IPV) del INE.

[8] El análisis incluye también predicciones para el número de transacciones y el precio de los alquileres aunque, en esta sección solo nos centraremos en el apartado de los precios de los inmuebles residenciales.

[9] Ver García-Montalvo *et al.* (2018) para los detalles sobre como se seleccionan las variables y los retardos a incluir en la especificación final usando un procedimiento de maximización de la distribución *a posteriori* (MAP).

El modelo anterior es útil para obtener predicciones de corto plazo, pero es menos apropiado para aprender la distribución estacionaria. Müller y Watson (2019) proponen un procedimiento para realizar predicciones a largo plazo del panel de procesos como una suma de un proceso Gausiano estacionario y un proceso de ruido idiosincrático, para luego estimar los momentos estacionarios del proceso Gausiano. Esta aproximación es efectivamente una descomposición espectral de un proceso Gausiano utilizando funciones trigonométricas.

La idea es proyectar las series temporales sobre un conjunto de basis functions $\psi_k(t) = \cos(\pi j\,(t\text{-}0{,}5)/T)$ considerando el modelo

$$y_t = \alpha + \mathrm{B}\psi_t + \varepsilon_t, \quad \psi_t = \sqrt{2}[\psi_1(t)\ \ldots\ \psi_K(t)]^\top, \quad \varepsilon_t \,|\, \Sigma_\varepsilon \sim (0, \Sigma_\varepsilon). \qquad [2]$$

Utilizamos $K = 12$ bases y la tasa de crecimiento anualizada entre trimestres, dado que las K bases capturan periodicidades mayores de $2T/K$ (Müller y Watson, 2019). Este modelo no pretende predecir la evolución temporal sino filtrar los componentes de baja frecuencia $\mathrm{B}\psi_t$ y usarlos en la estimación de las características estacionarias. Para simplificar el análisis, fijamos la prior conjugada en las columnas β_j de la matriz de correlación B:

$$\beta_j \,|\, \Sigma_\varepsilon \sim (0, g\Sigma_\varepsilon) \qquad [3]$$

Para predecir el proceso estacionario, $\tilde{\boldsymbol{y}}_t$, García-Montalvo *et al.* (2018) usan un modelo de error de dos componentes. El primer término, el error η_t, proviene de la variación a baja frecuencia explicada por las *basis functions,* mientras el segundo término de error ε_t representa la variación residual no estructurada.

$$\tilde{\boldsymbol{y}}_t = \alpha + \eta_t + \varepsilon_t, \qquad \varepsilon_t |\, \Sigma_\varepsilon \sim (0, \Sigma_\varepsilon) \qquad [4]$$

Podemos inferir la covarianza de η_t de $\mathrm{B}\psi_t$ utilizando las siguientes expresiones:

$$\eta_t | \mathrm{B} \sim (0, \Sigma_\eta(\mathrm{B})), \qquad \Sigma_\eta(\mathrm{B}) = T^{-1}\mathrm{B}\left(\sum_t (\psi_t)\right)\mathrm{B}^\top \approx T^{-1}\mathrm{B}\Psi\Psi^\top\mathrm{B}^\top. \qquad [5]$$

Por tanto, la distribución a largo plazo de $\tilde{\boldsymbol{y}}_t$ es

$$\tilde{\boldsymbol{y}}_t | (\alpha, \mathrm{B}, \Sigma_\varepsilon) \sim (\alpha, \Sigma_\varepsilon + T^{-1}\,\mathrm{B}\Psi\Psi^\top\mathrm{B}^\top) \qquad [6]$$

García Montalvo *et al.* (2018) explican cómo formular una versión Bayesiana de este modelo para combinar posteriormente con el BVAR. Esta combinación depende de los supuestos que se adopten sobre el panel VAR. Para evitar tener que hacer supuestos fuertes se puede adoptar una aproximación *ad-hoc* que directamente interpola la predicciones de corto y largo plazo. Específicamente se puede definir una combinación convexa de la distribución predictiva $f(\tilde{\boldsymbol{y}}_{t+s}|\boldsymbol{y}_t)$ dadas las observaciones en t y una distribución estacionaria $f(\tilde{\boldsymbol{y}}_{t+s})$:

$$g(\tilde{\boldsymbol{y}}_{t+s}|y_t) = e^{-\delta s} f(\tilde{\boldsymbol{y}}_{t+s}|y_t) + (1-e^{-\delta s}) f(\tilde{\boldsymbol{y}}_{t+s}) \quad [7]$$

donde el parámetro $\delta > 0$ determina la intensidad de la dinámica transicional. En este ejercicio se utiliza un $\delta = 0{,}1$ lo que implica que la dinámica de largo plazo empieza a tener relevancia a partir de aproximadamente el segundo periodo.

5.2. Modelo multinivel jerárquico

Antes de plantarse la estimación del modelo HML es necesario resolver la *sparcity* de información a nivel de código postal. Para producir un panel denso de tasas de crecimiento de precios de vivienda se utilizan dos procedimientos para agrupar observaciones de códigos postales: en primer lugar se usa un algoritmo basado en la proximidad de los códigos postales y un número mínimo de observaciones por clúster. En segundo lugar se utiliza una aproximación más formal que usa todas las variables disponibles a nivel de código postal (distancia entre códigos postales de una misma provincia, tensión del mercado, volumen de transacciones, etc.). Con estas variables se construyen 50 random forest (RF) de profundidad máxima 3 (para evitar sobreajuste) y se entrenan estos RF para predecir el precio por metro cuadrado. Se usan las 20 variables más importantes en media (entre los 50 RF) para entrenar un algoritmo K-means con el número de clústeres siendo función del número de transacciones en cada provincia.

Una vez clusterizados los códigos postales se estima un modelo de regresión lineal jerárquico,

$$y_{zt} = \boldsymbol{\beta}_z^{\top} \boldsymbol{x}_{zt} + \epsilon_{zt}, \qquad \epsilon_{zt} \sim (0, \tau^2) \quad [8]$$

donde y_{zt} es la respuesta en el código postal z en el periodo t, β_z es un J-vector de coeficientes pertenecientes al código postal z y x_{zt} es un vector de factores en el tiempo t. Dado que existe un gran número de parámetros, lo que podría producir sobreajuste y disminuir la predicción en la muestra de testeo, especialmente para códigos postales con pocas transacciones, se mezcla información de varios códigos postales via *priors* jerárquicas:

$$\beta_z \sim (\mu, \Sigma) \quad [9]$$

El modelo se completa con las siguientes *priors* para los *nuisance parameters:*

$$\tau^2 \sim (0{,}01, 0{,}01), \qquad p(\mu) \propto 1, \qquad \Sigma \sim (J, J\Pi) \quad [10]$$

En particular, la modelización de los precios residenciales incluye la variable comentada anteriormente (tensión del mercado), generando la siguiente especificación,

$$\Delta \log p_{zt} = \alpha_z + \beta_z \Delta \log h_t + \gamma_z s_z + \varepsilon_{zt} \quad [11]$$

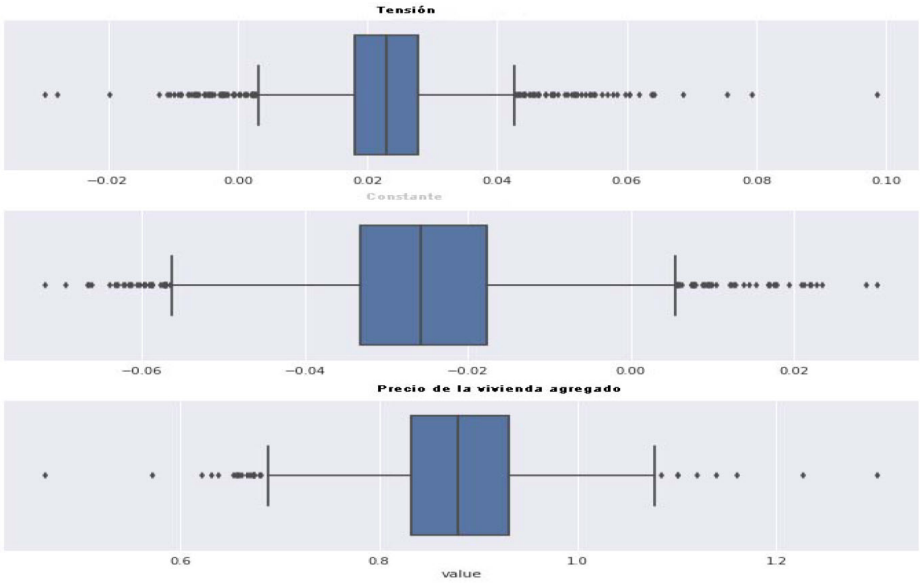
donde p_{zt} es el precio medio del metro cuadrado en el código postal z en t, h_t corresponde al índice IPV del INE en t, y s_z es la media de la tensión del mercado en cada código postal z[10]. Agregando los datos de los códigos postales se comprueba que el supuesto de linealidad respecto a la evolución del precio de la vivienda agregado es apropiado en este caso (ver García-Montalvo *et al.,* 2018).

Los resultados de la distribución *a posteriori* de los coeficientes, aparecen en el *box plot* de la figura 4.

La figura 5 muestra la distribución de los residuos versus los valores ajustados donde no parece haber desviaciones claras del supuesto de linealidad.

Figura 4.

Distribución *a posteriori* de los parámetros

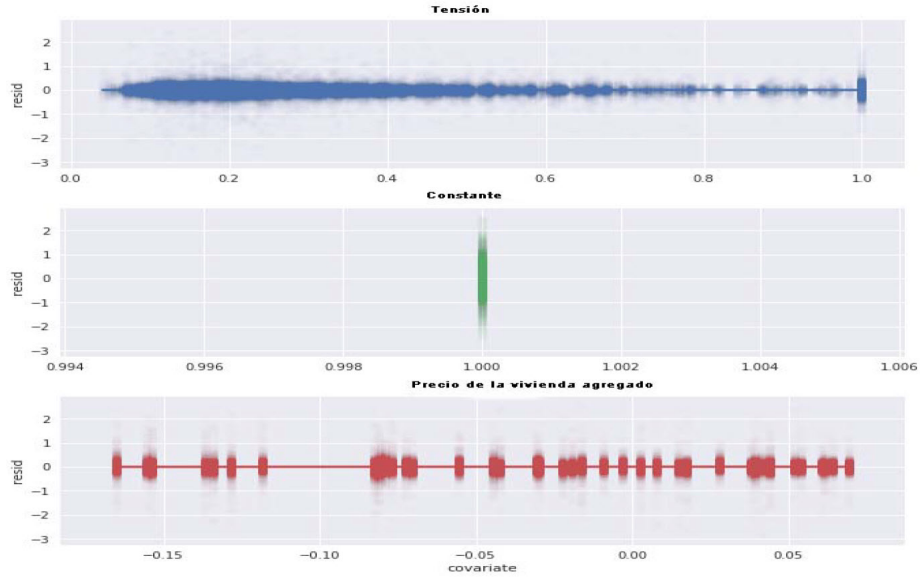


Fuente: Elaboración propia.

[10] La estimación del modelo utiliza un *Gibbs sampler* que proporciona resultados satisfactorios con muestras efectivas de unas 10.000 iteraciones. También se plantean tres escenarios en función de las bandas de confianza del 90 % alrededor de la predicción del crecimiento del precio de la vivienda agregado.

Figura 5.

Residuos



Fuente: Elaboración propia.

6. OTRAS APLICACIONES: LUCES Y ELECCIONES

Otras dos áreas donde se están multiplicando las aplicaciones son la utilización de imágenes de satélite como indicadores para el nivel de desarrollo o pobreza por áreas pequeñas y la predicción de resultados electorales.

6.1. Imágenes de satélite

La disponibilidad de gran cantidad de datos obtenida a partir de sensores e imágenes de satélites ha abierto también una gran ventana de oportunidad para utilizar estos datos en la medida del nivel de desarrollo, pobreza o desigualdad. Los primeros trabajos que utilizan la luz nocturna captada por satélites para medir el nivel de desarrollo o el crecimiento aparecieron unos años después de que se publicaran los primeros estudios de este tipo en las revistas sobre sensores e ingeniería de señales. Chen y Nordhaus (2011) y Henderson, Storeygard y Weil (2012) utilizan modelos muy similares. En Henderson, Storeygard y Weil (2012) la señal de la luz nocturna se utiliza como medida de actividad económica en un modelo de error de medida clásico donde tanto el crecimiento de la intensidad de la luz como el crecimiento del PIB están relacionados con el crecimiento de la renta verdadera no observable. A partir

de este modelo se derivan los pesos óptimos para minimizar el error de medida. Utilizando este procedimiento, Henderson, Storeygard y Weil (2012) muestran que la intensidad de la luz por km^2 tiene capacidad predictiva sobre la tasa de crecimiento real del PIB. Utilizando unos supuestos similares, Pinkovskiy y Sala-i-Martin (2016) rechazan la hipótesis de que el peso óptimo de las encuestas utilizadas para aproximar la renta per cápita sea superior al de las cuentas nacionales. Asimismo, muestran que las encuestas son peores que la contabilidad nacional para aproximar los resultados que importan para la población con menor nivel de renta en países en vías de desarrollo[11]. García-Montalvo y Reynal-Querol (2020) utilizan la luz nocturna para medir índices de desigualdad con el objetivo de encontrar un procedimiento que pueda permitir calcular índices de Gini en zonas donde, por su nivel de desarrollo o por la existencia de conflictos, no existen datos de encuestas ni contabilidad nacional que permitan hacer un cálculo mínimamente fiable del nivel de desigualdad. El procedimiento consiste en dividir la extensión del área deseada en pequeños píxeles y dividir la intensidad de la luz por el número de habitantes después de alienar las tramas. Este procedimiento genera suficientes observaciones para, con posterioridad, realizar un *ranking* de luz media por habitante de cada pixel. Por ejemplo, a nivel de país, la media de píxeles de los países es de 4,4 millones. Se aplican diferentes correcciones a los datos de luz nocturna para evitar los problemas del *top-coding* y de baja intensidad. Una vez calculados los cocientes y ordenados se calcula el índice de Gini. Para la validación cruzada de este procedimiento se utilizan los datos de países disponibles a partir de obtener un factor común a los estimadores disponibles en las diferentes bases de datos existentes. A partir de la correlación con dicho factor se calculan los pesos del Gini medido a partir de la luz.

6.2. Predicción de elecciones con nuevos partidos políticos

García-Montalvo, Papaspiliopoulos y Strumpf-Fetizon (2019) proponen una nueva metodología para predecir los resultados de elecciones en presencia de nuevos partidos políticos sin historial electoral, o con un historial muy limitado. Esta situación ha sido muy común en los países europeos desde 2008 con el avance de nuevos partidos de orientación populista. El procedimiento combina un modelo fundamental con todas las encuestas disponibles a través de un mecanismo bayesiano de síntesis de evidencia.

El modelo fundamental predice, a nivel nacional, la proporción de votos y escaños utilizando microdatos de las encuestas electorales del CIS así como datos demográficos. El modelo para las encuestas frecuentes tiene en cuenta no solo su nivel de incertidumbre sino también la calidad de la empresa/institución que produce los resultados de cada encuesta. El procedimiento utiliza todas las encuestas disponibles durante todas las elecciones que se producen con anterioridad al periodo para el cual se busca la generalización. La especificación incluye un sesgo invariante en el tiempo para cada empresa/institución que realiza encuestas (*house effect*), un sesgo invariante para cada empresa que se estima para cada elección sepa-

[11] Los análisis recientes no solo utilizan la luz nocturna. Por ejemplo Jean *et al.* (2016) complementan la luz nocturna con imágenes diarias sobre el tipo de tejados de las viviendas, por ejemplo, para complementar la información en el cálculo de la pobreza por áreas pequeñas.

radamente (*election effect*), un tendencia lineal con un coeficiente que puede cambiar para distintas elecciones pero es común a todas las empresas demoscópicas; y, finalmente, un error idiosincrático que es específico de cada encuesta debido a diferencias en la metodología utilizada y la variabilidad muestral. La hibridación de ambos modelos se produce a partir de sus probabilidades condicionadas. García-Montalvo, Papaspiliopoulos y Strumpf-Fetizon (2019) aplican este procedimiento a las elecciones españolas en 2015 y muestran que el procedimiento genera mejores predicciones, especialmente en términos de escaños, que especificaciones alternativas.

7. CONCLUSIONES: LECCIONES APRENDIDAS

Las técnicas de ML y la creciente disponibilidad de grandes bases de datos generan enormes posibilidades para el avance de la ciencia económica. Permiten analizar el impacto de las políticas económicas prácticamente en tiempo real y avanzar en modelizaciones cada vez más complejas apoyadas en millones de datos. Pero no son ni mucho menos una panacea. El *hype* actual del big data en economía tiene similitudes con lo sucedido a mitad de los años noventa con la revolución de los experimentos randomizados. La extensión de esta metodología llevó a un desprecio creciente de otros procedimientos estadísticos. Sin embargo, las técnicas empíricas dependen crucialmente del tipo de aplicación, la cantidad de datos disponibles y la calidad de los mismos. Además las técnicas no son excluyentes: por ejemplo, en la actualidad existen diversas modelizaciones que permiten combinar resultados utilizando big data o, en general, datos observacionales, y datos de experimentos[12].

En algunas ocasiones las técnicas de ML son esenciales como en el caso de la ingeniería reversa comentado en la tercera sección puesto que los datos han sido generados por algún tipo de algoritmo habitual en ML. También es importante cuando se intenta obtener un nuevo indicador que, bien por lo caro que puede resultar realizar encuestas o lo difícil que puede ser hacerlo para áreas pequeñas, es mejor aproximar utilizando otras fuentes de datos. Este es el caso del uso de las imágenes de satélite para aproximar el nivel de desarrollo o la desigualdad para áreas pequeñas o no delimitadas por fronteras administrativas. En el seguimiento de la economía en tiempo real, la utilización de datos administrativos es la mejor opción para avanzar por mucho que se puedan realizar encuestas por Internet con elevada periodicidad.

Sin embargo, en otros casos se plantean ejercicios que los datos disponibles difícilmente podrán responder. Este sería el caso de la predicción del crecimiento de los precios de la vivienda por código postal. Es cierto que la predicción es mejor que otros procedimientos alternativos pero, en cualquier caso, el nivel de error en la generalización es elevado. Por tanto, si el objetivo es mejorar otros procedimientos, o buscar una metodología con una buena justificación, es un ejercicio razonable pero, si el objetivo es tener una buena predicción, la conclusión es diferente.

[12] Imbens, Chetty y Athley (2020) presentan una posibilidad de realizar dicha combinación utilizando un supuesto de *latent unconfoundedness* para los datos observacionales.

Una de las características más importantes de las técnicas de ML es la insistencia en la comparación de diferentes hipótesis (especificaciones, modelos o técnicas) para determinar cuál es la que tiene mejores propiedades predictivas. Desafortunadamente en muchas aplicaciones económicas este ejercicio no se realiza y se toman las técnicas más complejas como las más apropiadas, sin tener en cuenta la importancia que el campo del ML otorga al problema del sobreajuste. Otro problema importante es la escasa tendencia de los economistas a mirar y limpiar los datos antes de realizar ningún análisis. Teniendo en cuenta la baja señal sobre ruido de muchas de las bases de datos masivas este defecto disciplinario se hace particularmente relevante. A pesar de mi insistencia resulta casi imposible que un ayudante de investigación mire los datos antes de empezar a hacer análisis. No es infrecuente encontrar que ni siquiera la carga de los datos se produjo de la forma esperada y ya se han realizado multitud de análisis.

Otra característica relevante del big data es la reutilización de datos creados con otros fines, y la fusión de estas bases de datos con el objetivo de realizar investigación. La posibilidad de fusionar bases de datos reutilizadas permite investigar temas que, de otra forma, no podrían ser tratados. García-Montalvo y Raya (2012) fusionan datos de diversas fuentes para conseguir una base de datos que contiene, para cada inmueble residencial, el precio de venta de mercado, el precio registral y el valor de tasación. De esta forma comprueban que a un crédito valor del 83 % le correspondían, durante la gran expansión inmobiliaria de los años 2000, un crédito sobre precio de compra medio en el entorno del 110 %. Esto, lógicamente, tiene implicaciones significativas en términos de política macroprudencial (García-Montalvo y Raya, 2018). Estos mismos datos permiten calcular la proporción de operaciones que implicaron algún tipo de pago opaco a la Hacienda Pública (García-Montalvo, Piolatto y Raya, 2020).

Asimismo, las técnicas de ML insisten en la capacidad predictiva de los procedimientos lo que implica que no importa si el método es una caja negra mientras el error de generalización sea pequeño. Sin embargo, como economistas nos interesa entender los motivos que explican los resultados empíricos que obtenemos más allá de la simple predicción. Esta es una área de investigación que será cada vez más importante en el futuro del data science, pues los sesgos de los algoritmos de aprendizaje en aplicaciones reales requieren cada vez más de auditorías de los procedimientos para comprender su origen e intentar resolverlo. Este es el caso de la discriminación en los sistemas de puntuación crediticia aunque se evite considerar la variable que describe las diferencias entre los grupos (por ejemplo, afroamericanos, inmigrantes o mujeres).

En el caso español la utilización de datos administrativos continúa en un claro retraso frente a otros países, especialmente los escandinavos. De hecho existe una clara animosidad hacia la investigación basada en big data como demostró la enorme polémica que generó el estudio del INE sobre movilidad utilizando el posicionamiento de teléfonos móviles. A pesar de que los datos se iban a agregar geográficamente y estaban anonimizados la opinión pública puso el grito en el cielo. Lo sorprendente es que la mayoría de los ciudadanos estén dispuestos a ceder todos sus datos por bajarse una aplicación de linterna al teléfono móvil y se escandalicen por la utilización de datos anonimizados para la consecución de un objetivo socialmente deseable a partir de un estudio de una institución pública. Es

necesaria una estrategia común para explicar la importancia de este tipo de estudios y todas las salvaguardias sobre privacidad que se aplican. Por ejemplo la AIReF ha realizado varias reuniones de expertos sobre estos temas y ha presentado un documento de opinión, AIReF (2020), con una propuesta para una estrategia nacional para avanzar en la disponibilidad de datos administrativos para la investigación sin ánimo de lucro que redunde en beneficio del interés público. En este sentido el anuncio del INE, la Agencia Tributaria, la Seguridad Social y el Banco de España de comenzar a trabajar conjuntamente en un sistema de acceso a sus bases de datos con fines científicos es una excelente noticia.

Finalmente, la creciente utilización de técnicas de ML en economía y finanzas ha multiplicado el número de revisiones de la literatura sobre técnicas de ML escritos por económetras. Desgraciadamente la gran mayoría son simplemente malos resúmenes de libros de texto clásicos en el ámbito del ML. Seguramente, como ya sucediera con cierta frecuencia en el pasado, los economistas acabaremos redescubriendo muchos de los resultados clásicos de las técnicas de ML.

En resumen, las técnicas de aprendizaje automático requieren de una serie de condiciones para que dicho aprendizaje sea factible. Esta situación no está garantizada para cualquier aplicación o conjunto de datos disponible por muy grande que sea el mismo.

Referencias

AIReF (2020). Opinión para una estrategia de acceso a datos administrativos. *Opinión,* 1/20.

Aspachs, O., Durante, R., Graziano, A., Mestres, J., García-Montalvo, J. y Reynal-Querol, M. (2021). Real-time inequality and the welfare state in motion: evidence from COVID-19 in Spain.

Athey, S., Chetty, R. e Imbens, G. (2020). *Using Experiments to Correct for Selection in Observational Studies.* Mimeo.

Autor, D., Cho, D., Crane, L., Goldar, M., Lutz, B., Montes, J., Peterman, W., Ratner, D., Villar, D. y Yildirmaz, A. (2020). An evaluation of the Paycheck Protection Program using administrative payroll microdata. *MIT Working Paper.*

Baker, S. (2018). Debt and the response of household income to shocks: validation and application of linked financial account data. *Journal of Political Economy,* 126(4), pp. 1504-1556.

Baker, S., Farrokhnia, R., Meyer, S., Pagel, M. y Yannelis, C. (2020). Income, liquidity, and the consumption response to the 2020 economic stimulus payments. De próxima aparición en *Review of Asset Pricing Studies.*

Bick, A. y Blandin, A. (2020). *Real time labor market estimates during the 2020 coronavirus outbreak.* Mimeo, Arizona State University.

Bounie, D., Camara, Y., Fize, E., Galbraith, J., Landais, C., Pazem, T. y Savatier, B. (2020). *Consumption dynamics and COVID fiscal stimulus strategy: evidence from France.* Mimeo.

Carvalho, V. M., Hansen, S., Ortiz, A. Garcia, J. R., Rodrigo, T., Rodriguez Mora, S. y Ruiz de Aguirre, J. (2020). Tracking the Covid-19 crisis with high-resolution transaction data. *CEPR Discussion Paper,* 14642

Chen, S., Igan, D. Pierri, N. y Presbitero, A. F. (2020). Tracking the Economic Impact of COVID-19 and Mitigation Policies in Europe and the United States. *IMF WP/20/125.*

Chen, X. y Nordhaus, W. D. (2011). Using Luminosity Data as a Proxy for Economic Statistics. *Proceedings of the National Academy of Sciences,* 108(21), pp. 8589-8594.

Chen, H., Quian, W. y Wen, Q. (2020). The impact of the COVID-19 pandemic on consumption: learning from high frequency transaction data. Mimeo.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. y el equipo The Oppotunity Insights (2020). How did Covid-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data. *NBER Working Paper,* 27.431.

Cicala, S. (2020). *Early Economic Impacts of COVID-19 in Europe: A View from the Grid.* Mimeo.

Cox, N., Ganong, P., Noel, P., Vara, J., Wong, A., Farrell, D. y Greig, F. (2020). Initial impact of the pandemic on consumer behavior: evidence from linked income, spending, and savings data. De próxima aparición en *Brookings Papers on Economic Activity.*

Dos Santos, J., Carvalho, B. y Peralta, S. (2020). *What and how did people buy during the Great lockdown? Evidence from electronic payments.* Mimeo.

García-Montalvo, J. (2014). El impacto del Big Data en los servicios financieros. *Papeles de Economía Española,* 114, pp. 43-59.

García-Montalvo, J., Muñoz, J. C. y Reynal-Querol, M. (2020). *Measuring inequality from above.* Mimeo.

García-Montalvo, J., Papaspiliopoulos, P., Rossell, D. O. y Strumpf-Fetizon, T. (2018). *Predicting real estate prices at zip code.* Mimeo.

García-Montalvo, J., Papaspiliopoulos, P. y Strumpf-Fetizon, T. (2019). Forecasting electoral outcomes with new parties' competition. *European Journal of Political Economy,* 59, pp. 52-70.

García-Montalvo, J., Piolatto, A. y Raya, J. M. (2020). Transaction tax evasion in the housing market. *Regional Science and Urban Economics,* 81, pp. 1-17.

García-Montalvo, J. y Raya, J. M. (2012). What is the right price of Spanish residential real estate? *Spanish Economic and Financial Outlook,* 1(3), September, pp. 22-29.

García-Montalvo, J. y Raya, J. M. (2018). Constraints on LTV as a macroprudential tool: a precautionary tale. *Oxford Economic Papers,* 70(3), pp. 821-845.

García-Montalvo, J. y Reynal-Querol, M. (2020). Distributional Effects of COVID-19 on Spending: A First Look at the Evidence from Spain. *Barcelona GSE Working Paper,* 1.201 September, 2020.

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

Golec, P., Kapetanios, G., Neuteboom, N., Ritsema, F. y Ventouri, V. (2020). *Consumtion during the COVID-19 pandemic: loackdown or fear? Evidence from transaction data for the Netherland.* Mimeo.

González, J., Urtasum, A. y Perez, M. (2020). Evolución del consumo en España durante la vigencia del estado de alarma: un análisis a paritr del gasto con tarjetas de pago. *Articulos Analiticos,* 3/2020.

Gourinchas, P. O. (2020). Flattening the pandemic and recession curves. *Mitigating the COVID Economic Crisis: Act Fast and Do Whatever, pp.* 31-40.

Hacioglu, S., Kanzig, D. y Surico, P. (2020). The distributional impact of the pandemic. *CEPR Discussion Paper,* 15.101.

Henderson, J., Storeygard, A. y Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review,* 102, pp. 994-1028.

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

Kobota, S., Koichiro, O. y Toyama, Y. (2020). Consumption responses to COVID-19 payments: evidence from a natural experiment and bank account data. *Covid Economics,* 62, pp. 90-123.

Müller, U. K. y Watson, M. W. (2019). Low-Frequency Analysis of Economic Time Series. De próxima aparición en S. Durlauf, L. P. Hansen, J. J. Heckman y R. Matzkin (eds.), *Handbook of Economics,* Vol. 7. North Holland.

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

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

Sheridan, A., Andersen, A., Hansen, E. y Johannesen, N. (2020). Social distancing laws cause only small losses of economic activity during the COVID-19 pandemic in Scandinavia, *Proceeding of the National Academy of Sciences,* August, 1-6.

Wade, L. (2020). An unequal blow. *Science, 368,* 6492, pp. 700-703.

Xiong, T., Bao, Y., Hu, Z. y Zhang, L. (2015). A combination method for interval forecasting of agricultural commodity futures prices. *Knowledge-Based Systems,* 77, pp. 92-102.