CUADERNOS ECONÓMICOS ICE INFORMACIÓN COMERCIAL ESPAÑOLA

La gestión de los datos administrativos en España: diagnóstico y retos de futuro*

The Management of Administrative Data in Spain: Diagnosis and Challenges

Miguel Almunia
CUNEF Universidad

Pedro Rey-Biel
ESADE, Universidad Ramón Llull

Resumen

Los países de nuestro entorno han dado pasos importantes para gestionar y facilitar el acceso a la ingente cantidad de datos que generan y recopilan las Administraciones públicas, lo que supone una oportunidad para mejorar la evaluación de las políticas públicas, promover que el diseño de nuevas políticas se base en la evidencia disponible y expandir el conocimiento científico utilizando los registros administrativos. Sin embargo, a pesar de algunas iniciativas aisladas recientes, España se está quedando rezagada. A partir del análisis comparado de la experiencia en otros países avanzados y destacando las ventajas e inconvenientes de las distintas opciones que tenemos, planteamos propuestas de mejora en el acceso a datos administrativos en España con el fin de aprovechar el potencial de los datos existentes para impulsar las políticas basadas en la evidencia científica.

Palabras clave: *datos administrativos, acceso, investigación, políticas basadas en la evidencia.*
Códigos JEL: *B40, C81, C82, I38.*

Abstract

The neighboring countries have taken important steps to manage and facilitate access to the vast amount of data generated and compiled by public administrations, which represents an opportunity to improve the evaluation of public policies, promote the design of new policies building on the available evidence and expand scientific knowledge using administrative records. However, despite some recent isolated initiatives, Spain is lagging behind. Based on the comparative analysis of the experience in other advanced countries and highlighting the advantages and disadvantages of the different options that we have, we make proposals to improve the access to administrative data in order to take advantage of the potential of existing data to promote policies based on scientific evidence in Spain.

Keywords: *administrative data, access, research, evidence-based policies.*

* Una versión anterior de este artículo se publicó como Policy Brief de ESADE EcPol bajo el título «Por un cambio de cultura en la gestión de datos en España: una propuesta de reforma», disponible en: https://dobetter.esade.edu/es/informe-gestion-datos-covid

DOI: https://doi.org/10.32796/cice.2021.102.7313
Recibido: Septiembre 2021 · Aceptado: Octubre 2021

1. Introducción

La recolección y procesamiento de datos son actividades esenciales que los Estados deben realizar para cumplir muchas de sus funciones básicas. Por poner algunos ejemplos recientemente relevantes, las Administraciones públicas necesitan registros de la actividad económica de los contribuyentes para recaudar impuestos, registros de afiliación a la seguridad social para gestionar los sistemas de previsión social, o información sobre los contagiados (y sus contactos), ingresados y fallecidos para afrontar una pandemia. No es casualidad que el término «estadística» provenga originalmente del latín *statisticum*, «relacionado con el Estado».

Los enormes avances en las tecnologías de la información y la comunicación (TIC) permiten el almacenamiento de una cantidad cada vez mayor de datos a un coste cada vez más bajo. Esto ha llevado a los gobiernos de muchos países a mantener grandes registros administrativos sobre tributación, salud, educación y programas sociales. A su vez, la mejora de nuestra capacidad de procesamiento y análisis de datos supone una oportunidad de expandir el conocimiento utilizando los registros administrativos. La posibilidad de realizar evaluaciones de las políticas públicas y diseñar nuevas políticas basadas en la evidencia existente depende crucialmente de la disponibilidad y el aprovechamiento de los datos.

La administración realiza una evaluación interna de seguimiento de sus actividades mediante el uso de indicadores, que requieren del uso de datos administrativos, como por ejemplo en las políticas activas de empleo. Sin embargo, la administración no puede, ni debe, tener el monopolio del análisis de los datos que posee. No puede, en primer lugar, porque no es su función principal. La prioridad de la administración es gestionar las políticas públicas, no realizar un análisis profundo de los datos de los que dispone. En segundo lugar, porque los recursos públicos disponibles para este fin son reducidos. Aunque la administración cuenta con algunos organismos especializados en producir información estadística, como el Instituto Nacional de Estadística (INE) o el Centro de Investigaciones Sociológicas (CIS), los recursos económicos de los que disponen para evaluar políticas públicas son limitados. Esta limitada capacidad de análisis de los datos administrativos supone un despilfarro de los recursos destinados a recolectarlos y una ineficiencia respecto al enorme potencial que el uso de esos datos podría tener para el desarrollo y la evaluación de las políticas públicas. Además, no es deseable que la administración tenga el monopolio del análisis de estos datos, porque podrían surgir conflictos de interés al hacerlo. La creación de una entidad independiente como la Autoridad Independiente de Responsabilidad Fiscal (AIReF) ha significado un paso hacia delante en este sentido, que esperamos sea aún mayor con la reciente ampliación de su mandato actual, inicialmente restringido únicamente a la evaluación de políticas fiscales.

Permitir que la comunidad científica tenga acceso, de manera controlada, a registros administrativos resultaría en un enorme incremento de nuestra capacidad para analizarlos. Expandiría nuestro conocimiento sobre la efectividad de las políticas públicas, aumentaría la transparencia de la actuación de las administraciones y, por

lo tanto, profundizaría la calidad de nuestra democracia (Arellano, 2018). Los incentivos están alineados: la administración se beneficiaría porque su capacidad de analizar los datos es limitada, mientras que la comunidad investigadora tendría la oportunidad de explotar esta ingente fuente de información para realizar investigación académica de calidad. Los nuevos conocimientos obtenidos se podrían utilizar para diseñar políticas públicas basadas en la evidencia, algo que desafortunadamente es poco habitual en nuestro país.

Los datos administrativos presentan varias ventajas respecto a los datos provenientes de otras fuentes, como las encuestas, para la evaluación de las políticas públicas. Al cubrir a toda la población, se consigue un tamaño muestral mucho mayor, corrigiendo posibles sesgos de selección que a menudo afectan a las encuestas. Los datos administrativos tienen una tasa de no respuesta mucho menor, lo cual permite disponer de bases longitudinales (panel) más extensas en el tiempo y con menor desgaste *(«attrition»)*. Además, tienen menos errores de medición, aunque estos nunca se eliminan por completo. Por último, los registros administrativos se pueden cruzar con otras fuentes de datos utilizando identificadores individuales o de hogar. Esto es esencial para poder realizar análisis más profundos sobre causalidad y evaluación de políticas públicas.

Los datos administrativos también presentan algunas limitaciones que se deben tener en cuenta (Slemrod, 2016). En primer lugar, los datos se recogen atendiendo a objetivos de gestión, lo que afecta al formato, estructura y alcance de los mismos. Por ello, el número de variables disponibles en cada registro suele ser limitado. Por ejemplo, los datos fiscales no suelen tener información socioeconómica de los contribuyentes, más allá de la edad y el sexo. Esta es otra de las razones por las que la posibilidad de cruzar datos de distintos registros es esencial para explotar su verdadero potencial. Además, al haber sido recogidos con el fin principal de facilitar la gestión, la construcción de bases de datos para investigación a partir de registros administrativos supondrá un importante esfuerzo adicional. Idealmente, este trabajo lo deberían realizar conjuntamente funcionarios públicos e investigadores, dado que los primeros tienen el mejor conocimiento de los datos y los segundos tienen mayor experiencia analizando datos para investigación. Por último, el volumen de las bases de datos administrativos es grande, a veces difícil de manejar. Sin embargo, esta limitación es superable dado el rápido avance en las tecnologías de almacenamiento en la nube y el desarrollo de nuevas herramientas de análisis como el *machine learning* o la inteligencia artificial.

Un factor importante por considerar es la protección de la confidencialidad e integridad de los datos administraivos con información sensible sobre los ciudadanos. Para ello, deben utilizarse métodos de anonimización de los datos, eliminando identificadores individuales y garantizando que no se pueda trazar la identidad de ningún individuo concreto usando otras variables. Igualmente, deben crearse protocolos que garanticen que el acceso a los datos sea seguro desde este punto de vista, como han hecho otros países. Es importante resaltar que el anonimato no es incompatible con la creación de códigos individuales, que permitan identificar y emparejar variables

procedentes de distintas bases de datos pero pertenecientes a la misma unidad de observación (por ejemplo ciudadano, empresa o institución).

La mayoría de los países de nuestro entorno han creado instituciones y sistemas de acceso a los datos administrativos que protegen la confidencialidad. Algunos de estos sistemas se describen en detalle en la sección 3 de este artículo. Para diseñar sistemas similares en España no tenemos que reinventar la rueda, pero sí elegir entre las diversas opciones que se analizan en la sección 2 e inspirarnos en los modelos que han creado estos países y adaptarlos a nuestra estructura administrativa. La reciente Directiva UE 2019/1024 de 20 de junio de 2019, relativa a los datos abiertos y la reutilización de la información del sector público da un claro soporte legal al respecto, puesto que prevé que «cualquier documento o conjunto de datos conservado por organismos públicos (incluidas empresas públicas y organismos de investigación financiados públicamente) sea puesto a disposición general del público, para fines comerciales o no comerciales, y en formatos abiertos, legibles por máquina, accesibles, fáciles de localizar y reutilizables, junto con sus metadatos». Esta Directiva debería haber sido objeto de transposición a la legislación nacional de los países miembros antes del 17 de julio de 2021, sin embargo la web de Transparencia del Gobierno apunta que «la actividad lleva cierto retraso y puede no llegar a trasponerse en el plazo previsto».

Afortunadamente, no partimos de cero. En España existen ya algunas bases de datos administrativos disponibles para la investigación, como la Muestra Continua de Vidas Laborales (MCVL) o el Panel de Declarantes del IRPF, que se describen en la sección 4. Además, existen casos aislados de investigadores que han conseguido obtener el acceso a ciertos datos, la mayor parte de las veces gracias a contactos personales y a eternas gestiones, no regidas con criterios objetivos sobre el interés público de las investigaciones que proponen. No existen, sin embargo, protocolos generalizados que regulen la colaboración y la cesión de datos entre las administraciones y la comunidad científica. Es obvio que en un país con un alto grado de descentralización, y con tantos niveles de administración como el nuestro, no es sencillo acordar y escoger entre las diversas formas de hacerlo. Por ello, precisamente, es necesario un debate sosegado, que se enriquezca con la experiencia ganada de otros países de nuestro entorno, con realidades políticas muy diversas.

La disponibilidad de mejores datos también permitiría diseñar medidas políticas bajo criterios que permitieran su evaluación, creando una colaboración activa, y no sólo pasiva, entre investigadores y administraciones. En este sentido, fuera de nuestras fronteras destacan iniciativas como el *Behavioral Insights Team* o las agencias del tipo *What Works (¿Qué Funciona?)* del Reino Unido. La creación de agencias que colaboren con investigadores para la realización y evaluación de políticas basadas en la evidencia puede ser uno de los muchos pasos necesarios hacia la modernización de la administración pública en España. Afortunadamente, muchos de estos pasos (disponibilidad de datos públicos, creación de una agencia de evaluación, apertura a la colaboración científica para el desarrollo de intervenciones que permitan basar

las políticas en la evidencia), están alineados, crean sinergias y pueden marcar un verdadero cambio en la forma de tomar decisiones políticas en nuestro país. Una señal esperanzadora en este sentido es que el Real Decreto-Ley por el que se aprobó el Ingreso Mínimo Vital incluye en su artículo 30 el compromiso de evaluar los efectos de esta política en concreto. Esperemos que esta evaluación no solo se lleve a cabo bajo los estándares metodológicos más exigentes, sino que siente el precedente para la evaluación de muchas otras políticas públicas.

2. Marco conceptual

Existen diversos modelos para facilitar el acceso público a datos administrativos, cada uno con sus ventajas y sus inconvenientes. Dichos modelos difieren fundamentalmente respecto a qué instituciones custodian los datos, cómo y a quién se facilita cada tipo de datos y en qué condiciones pueden obtenerse y tratarse. Naturalmente, la estructura administrativa de cada país es determinante para la elección de un modelo u otro. A continuación describimos las principales decisiones que se deben tomar en el diseño de un sistema de acceso a datos administrativos para investigación, destacando las ventajas e inconvenientes de las diferentes opciones. En la siguiente sección, describiremos algunos modelos concretos de países de nuestro entorno.

2.1. ¿Qué institución debe custodiar los datos?

En todos los países existen múltiples organismos públicos que recaban información de los ciudadanos y mantienen registros de datos. Una de las decisiones principales cuando se diseña un sistema de acceso a datos es qué institución (o instituciones) será la encargada de custodiarlos. Una opción es adoptar un modelo centralizado en el que solo una institución se encargue de esta tarea. Habitualmente este papel recae en el instituto nacional de estadística, que opera como un «tercero seguro»: recibe datos de diversas entidades gubernamentales –que no necesariamente quieren compartir datos entre sí– y se encarga de combinar los datos (usando identificadores individuales) y modificar su formato para que sean útiles para los distintos usuarios, como pueden ser los investigadores. Otra opción es adoptar un modelo descentralizado en el que cada institución generadora de datos se ocupa de la custodia y cesión de estos a los investigadores. También pueden existir modelos intermedios en los que varias instituciones se agrupen en consorcios para proveer acceso a sus datos, funcionando una de ellas como tercero seguro, y otras tengan un sistema de acceso aparte.

El modelo centralizado tiene dos ventajas principales. En primer lugar, la división de tareas entre instituciones que solamente proveen datos y un tercero seguro –que los custodia y cede– puede llevar a una especialización y ganancias de eficiencia.

Los aspectos legales de la cesión de datos a investigadores, así como la logística de los proyectos de investigación, requieren un proceso largo de aprendizaje y en estas actividades hay rendimientos de escala. En segundo lugar, un sistema centralizado proporciona una mayor transparencia al proceso desde la perspectiva de los investigadores, que saben inmediatamente a qué institución dirigirse si necesitan datos de un país concreto.

Por otro lado, los modelos descentralizados también tienen algunas ventajas. En primer lugar, determinadas instituciones pueden tener requisitos particulares de confidencialidad y por ello prefieren tener un control más directo sobre el acceso a sus propios datos. En segundo lugar, en países con una estructura gubernamental muy descentralizada, un sistema que replique esa estructura se puede ajustar mejor al diseño institucional.

2.2. ¿Cuáles deben ser los criterios de acceso a los datos?

La necesidad de proteger la confidencialidad de los datos administrativos merece una consideración especial a la hora de definir quiénes pueden tener acceso a ellos. Por otro lado, para asegurar que su uso tenga el máximo impacto posible, es deseable que se priorice el acceso de investigadores con proyectos bien definidos que tengan interés desde el punto de vista académico y/o desde la perspectiva de la evaluación de políticas públicas.

Para que el sistema funcione correctamente es importante que los criterios de acceso a los datos sean uniformes y transparentes, al contrario de lo que sucede actualmente en España. Cualquier investigador que desee acceder a los datos debería presentar una solicitud explicando el objeto de su estudio y especificando los registros a los que necesitaría acceder. Estas solicitudes deberían ser evaluadas por un comité científico en función de sus méritos científicos y el interés de la propuesta. Además, debería ser evaluada por un comité de productores de datos (representantes de las Administraciones públicas) para dictaminar la utilidad del estudio para el sector público y la viabilidad de ofrecer los datos requeridos en un formato apto para la investigación. Si estas evaluaciones tienen un resultado positivo, el investigador deberá firmar un contrato comprometiéndose a cumplir las reglas de uso y, especialmente, proteger la confidencialidad de los datos. Este contrato debe incluir la posibilidad de imponer sanciones al investigador si se incumplen las normas, con un objetivo disuasorio. Hay que tener en cuenta que cualquier tipo de infracción por parte de un investigador tendría también un efecto reputacional enorme, porque el resto de la comunidad científica podría ver en riesgo el acceso a los datos por la irresponsabilidad de una sola persona.

2.3. ¿El acceso a los datos debe ser de pago o gratuito?

Existen algunos argumentos para que el acceso a datos administrativos sea gratuito. En primer lugar, se trata de datos públicos que pertenecen al conjunto de la ciudadanía, y se podría esgrimir el principio de transparencia de los datos públicos para que se ofrezcan de manera gratuita. Además, la gratuidad evitaría cualquier tipo de discriminación del acceso entre investigadores con pedigrí, o que trabajan en instituciones con mayores recursos, e investigadores con menor experiencia o de instituciones más modestas.

Sin embargo, no se puede ignorar que la implementación de un buen sistema de acceso a datos como los que se describen en este documento tiene un coste que puede ser sustancial, ya que incluiría los salarios del personal que contrate el sector público (administrativos, expertos de sistemas informáticos, expertos en manejo de bases de datos, etc.) que sería adicional a las plantillas ya existentes. También habría que contabilizar el coste de los equipos informáticos y el software necesario para el funcionamiento del sistema. Aunque algunos de estos costes son fijos y solo se tienen que incurrir una vez, otros son variables y dependen directamente del número de proyectos de investigación que se desarrollen al mismo tiempo.

Por tanto, consideramos que sería razonable que se cobre algún tipo de tasa a los usuarios de este servicio que ayude a financiarlo. Una posible ventaja de este sistema es que los investigadores necesitarán obtener recursos para sus investigaciones en convocatorias competitivas, lo cual facilitará el filtrado de buenos proyectos y generará una selección positiva. Para que estas tasas no supongan una barrera a la entrada, se podría proporcionar acceso gratuito a muestras pequeñas de las distintas bases de datos para que los investigadores puedan explorarlos antes de tener que pagar las tasas. En este modelo, las tasas se cobrarían para el acceso a los datos completos, y especialmente cuando el uso de los datos suponga un mayor coste para la agencia proveedora del servicio. Por ejemplo, en los casos en que sea necesario hacer nuevos cruces de bases de datos o se requiera una gran cantidad de recursos computacionales. En la sección 3 se mencionan varios ejemplos de países en los que se cobran tasas a los investigadores por el uso de los datos.

2.4. ¿Deberían revisar los resultados de las investigaciones los organismos proveedores de datos?

Esta es una pregunta delicada porque afecta a la credibilidad de cualquier investigación que se realice con datos administrativos, así como a los incentivos que puedan tener los diferentes organismos para ceder sus datos. Por un lado, es razonable que los proveedores de datos tengan un interés en los resultados obtenidos y, además, quieran tener la posibilidad de revisar los resultados para comprobar que son factualmente correctos y que en ningún caso se viola la confidencialidad. Sin embargo, es fundamental que estos organismos se comprometan de antemano a

permitir la publicación de los estudios independientemente de si están de acuerdo o no con las conclusiones. En este sentido, un sistema de pre-registro de los estudios que se están llevando a cabo, de forma que posteriormente haya que reportar los resultados, podría contribuir a dar credibilidad al sistema y a evitar incentivos perversos[1].

2.5. ¿El acceso a los datos debería ser en salas seguras o a través de conexión remota?

En muchos casos, las instituciones proveedoras de datos prefieren que el acceso se realice exclusivamente en salas seguras de la propia institución, porque se considera que es la única manera de garantizar la seguridad del acceso y la confidencialidad.

Sin embargo, este modelo tiene muchas desventajas. En primer lugar, discrimina a investigadores no basados en la misma ciudad (a menudo la capital). En segundo lugar, en la práctica genera enormes pérdidas de eficiencia para los investigadores. Esto sucede porque los códigos que se utilizan para analizar grandes bases de datos a menudo tardan horas en ejecutarse y durante estos periodos de inactividad los investigadores no pueden conectarse a internet o trabajar en otras tareas, porque estas actividades no están permitidas en las salas seguras. Además, se debe tener en cuenta que los investigadores a menudo también tienen obligaciones docentes (y de gestión), lo que les impide pasar días completos en la sala segura trabajando en un único proyecto. Esto hace que los proyectos de investigación progresen a una velocidad mucho más lenta en los sistemas de acceso presencial a los datos.

Varios de los modelos que se presentan en la sección 3 demuestran que el acceso remoto a través de una conexión segura elimina estos costes de eficiencia a cambio de un incremento muy limitado en la probabilidad de que se quiebre la confidencialidad de los datos (Card *et al*., 2010; Almunia *et al*., 2019). Se trata, por lo tanto, de un aspecto que se debe tener muy en cuenta. En la sección 4 se menciona cómo algunas instituciones como el Banco de España ya se están moviendo en esta dirección.

3. Modelos de acceso a datos administrativos con fines de investigación

En esta sección se describen las características principales de tres modelos existentes de acceso a datos administrativos, incluyendo una descripción de algunos de los estudios más relevantes que se han realizado con cada modelo.

[1] Este sistema de pre-registro de estudios se ha instaurado con éxito en la investigación económica a través del *American Economic Association's registry for randomized controlled trials*, que se puede consultar en https://www.socialscienceregistry.org/

3.1. El modelo nórdico

Los países nórdicos (Dinamarca, Finlandia, Noruega y Suecia)[2] han sido pioneros en la creación de mecanismos para que la comunidad científica pueda realizar investigaciones utilizando sus datos administrativos de forma segura. En estos cuatro países, el organismo encargado de recopilar datos de diferentes registros administrativos es el instituto nacional de estadística respectivo. Las páginas web de estos institutos de estadística tienen un diseño moderno y accesible, y todas ellas incluyen una sección en la que se explica cómo se puede acceder a sus microdatos para fines de investigación.

Estos repositorios incluyen microdatos anonimizados sobre población (censos), empleo (historiales de cotización a la seguridad social), empresas (registros mercantiles), impuestos (renta, sociedades, IVA y especiales), sanidad (historiales hospitalarios), educación (resultados académicos), vivienda (precios de venta y alquiler), justicia (procesos judiciales), medio ambiente (datos históricos de clima y de partículas contaminantes) y comercio internacional (transacciones de importación y exportación).

Los criterios de acceso de investigadores a los datos varían entre países. En el caso de Dinamarca y Noruega es necesario tener una afiliación a una de las instituciones de investigación del país que esté aprobada por el instituto de estadística, mientras que los datos de Finlandia y Suecia se pueden obtener teniendo una afiliación con una institución dentro del Espacio Económico Europeo.

En todos los casos, los investigadores deben realizar una solicitud detallada en la que explican el diseño de su estudio y las bases de datos específicas a las que necesitan acceder para su investigación. En algunos casos los datos están disponibles en el formato deseado, pero en otros se requiere un trabajo por parte del instituto de estadística para vincular múltiples bases de datos. Existe también la posibilidad de que los investigadores aporten datos externos y el instituto de estadística los cruce con sus propios datos utilizando identificadores individuales, para crear una nueva base de datos anonimizada.

El acceso a los datos se realiza en todos los casos de forma remota por internet. Es decir, los investigadores reciben unas claves de acceso para conectarse a un servidor remoto donde están almacenados los datos y está instalado el software necesario para su análisis. Cada investigador tiene acceso exclusivamente a los datos de su estudio y por tanto no comparte el espacio con otros investigadores. El acceso remoto no permite la transferencia directa de archivos al ordenador de los investigadores. Cuando estos quieren extraer sus resultados, los archivos son revisados por personal del instituto de estadística para garantizar que se cumplen los requisitos de confidencialidad.

Los institutos de estadística cobran tasas por el uso de sus datos a los investigadores. Estas tasas incluyen un cargo por cada hora de trabajo que suponga la preparación de las bases de datos (alrededor de 100-150 euros por hora de trabajo) y también el coste de mantenimiento del servidor remoto (entre 150 y 300 euros por mes según el

[2] Habitualmente se incluye a Islandia en este grupo de países, pero no lo hemos incluido en nuestro análisis dado su reducido peso en términos de población.

volumen de datos almacenado), entre otros. Estas tasas garantizan que el sistema es sostenible, dado que asegura que se pueden financiar los costes, tanto de recursos humanos como de sistemas informáticos, aunque aumente el número de proyectos de investigación.

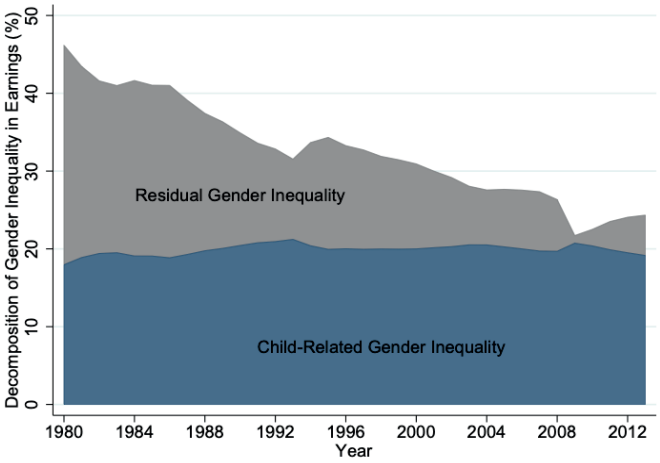
Ejemplos del Modelo Nórdico: Los efectos de distintos incentivos sobre el ahorro y los efectos de los hijos sobre la brecha salarial entre hombres y mujeres

a) Chetty *et al*. (2014a) estudia las decisiones de ahorro de los trabajadores en Dinamarca. Los investigadores combinan datos individuales de múltiples registros administrativos: declaraciones del impuesto sobre la renta, censo de población, una base de datos integrada de relaciones laborales, saldos de cuentas bancarias y aportaciones anuales a planes de pensiones (individuales y de empresa). Esta impresionante combinación de datos les permite observar todas las decisiones de ahorro del conjunto de la población danesa en el periodo 1995-2009. Los autores encuentran que los incentivos fiscales al ahorro previsional tienen un efecto moderado sobre el ahorro total (incluyendo planes de pensiones y otros instrumentos financieros). Sin embargo, estiman que un 85 % de los daneses son ahorradores «pasivos», que aceptan las contribuciones ofrecidas por sus empleadores por defecto, y solo un 15 % son ahorradores «activos», que reajustan su uso de distintos instrumentos de ahorro para aprovechar al máximo los incentivos fiscales, sin cambiar su ahorro total. Estos resultados tienen unas implicaciones claras para el diseño de los incentivos al ahorro previsional: es más efectivo regular que los planes de pensiones de empresa tengan, por defecto, un porcentaje alto de contribución que ofrecer un incentivo fiscal a todas las contribuciones a planes de pensiones.

b) Otro estudio que ha tenido un gran impacto es el de Kleven *et al*. (2019), que estudia, también para Dinamarca, la penalización laboral que sufren las mujeres cuando tienen su primer hijo. Los autores combinan datos de impuestos sobre la renta, censos de población, relaciones laborales y registros de nacimiento, que les permiten vincular a padres y madres con sus hijos. El principal hallazgo es que, con el nacimiento del primer hijo, la renta anual de la madre cae un 20 %, mientras que la del padre se mantiene constante. Esta brecha se mantiene inalterada durante los siguientes 10 años, y la misma tendencia se observa en las horas trabajadas, la tasa de empleo y en el salario por hora. Como se ve en la Figura 1, aunque la brecha salarial está decreciendo, el nacimiento del primer hijo explica cada vez un porcentaje mayor de la brecha en Dinamarca. Este estudio, junto con otros posteriores que se han hecho para otros países (incluyendo uno para España: De Quinto *et al*., 2020), indican que para cerrar la brecha salarial entre hombres y mujeres es fundamental diseñar políticas públicas que lleven a un reparto más equitativo de la crianza de los hijos entre padres y madres.

FIGURA 1
DESIGUALDAD DE GÉNERO EN RENTA DEBIDA A LOS HIJOS

A: Child-Related Inequality vs Non-Child Inequality



FUENTE: Kleven *et al*. (2019).

3.2. El modelo continental

Otros países de Europa continental han adoptado modelos diferentes al nórdico, pero que también están dando como resultado una fructífera colaboración entre la administración pública y la comunidad investigadora. La principal diferencia con el modelo nórdico es que no se construye alrededor del instituto nacional de estadística, sino que se han creado instituciones dedicadas específicamente a la tarea de recopilar y preparar los datos para su uso en investigación, además de la gestión de las solicitudes y los proyectos.

Además de los sistemas de acceso a datos en cada país, que describimos a continuación, en 2017 se creó INEXDA (International Network for Exchanging Experience on Statistical Handling of Granular Data), una red de bancos centrales (principalmente europeos, entre ellos el Banco de España) e institutos estadísticos. Los objetivos de este grupo son compartir experiencias en el manejo estadístico de datos granulares (también llamados microdatos) y homogeneizar los datos que gestionan los bancos centrales, de forma que las estadísticas producidas en distintos países sean comparables entre sí. El objetivo a largo plazo es que los datos estén disponibles para el análisis, el diseño de políticas públicas y la investigación. Aunque todavía no existe una lista completa de los datos que estarán disponibles, se pueden consultar las conclusiones del grupo de trabajo que ha definido los protocolos de acceso a los datos, las normas para combinar distintos registros y la información que se recolectará de los equipos encargados de cada proyecto de investigación.

3.2.1. Francia

En el caso de Francia, en el año 2010 se creó un centro de acceso seguro a datos (CASD) que actúa como tercero seguro y dispone de registros administrativos de impuestos, trabajo, empresas, finanzas, justicia, medio ambiente, agricultura y sanidad, todos ellos provenientes de los diferentes ministerios y del instituto de estadística francés (INSEE). El CASD está constituido como un «grupo de interés público» que reúne a varias instituciones: INSEE, GENES (las *grandes écoles* nationales), CNRS (centro nacional de investigación) y dos universidades: École Polytechnique y HEC Paris. Según se indica en su página web: «El objetivo principal de este grupo, de carácter industrial y comercial, es organizar e implementar servicios de acceso seguro a datos confidenciales para investigación, estudio, evaluación o innovación sin ánimo de lucro».

Una característica particular del CASD es que ha creado su propia tecnología, la *SD-box*, para permitir el acceso remoto a los datos por parte de investigadores externos. Se trata de un aparato (similar a un descodificador) que, conectado a un monitor y a una red de internet, permite a los investigadores tener acceso remoto a los datos de su proyecto. Al igual que el modelo nórdico, los investigadores interesados tienen que enviar una solicitud detallando el objetivo de su estudio y las bases de datos requeridas, y el servicio tiene un coste que depende de varios factores. En este caso, los investigadores con proyectos aprobados deben viajar a París para recibir un breve curso sobre las normas para proteger la confidencialidad de los datos.

Para su funcionamiento, el CASD se apoya en tres comités: un comité científico, un comité de productores de datos y un comité de seguimiento de la política de seguridad de los sistemas de información. El primero, formado por 16 investigadores, asesora a la dirección del CASD en materias de prospectiva, innovación, ética y estrategia científica. El segundo comité asesora a la dirección en asuntos relacionados con las condiciones de acceso a los datos, la documentación, el archivo y la difusión de información. El tercero asesora sobre temas relacionados con la protección de la confidencialidad e integridad de los datos. Una lista completa (e impresionante) de todos los estudios que se están realizando con datos del CASD está disponible.[3]

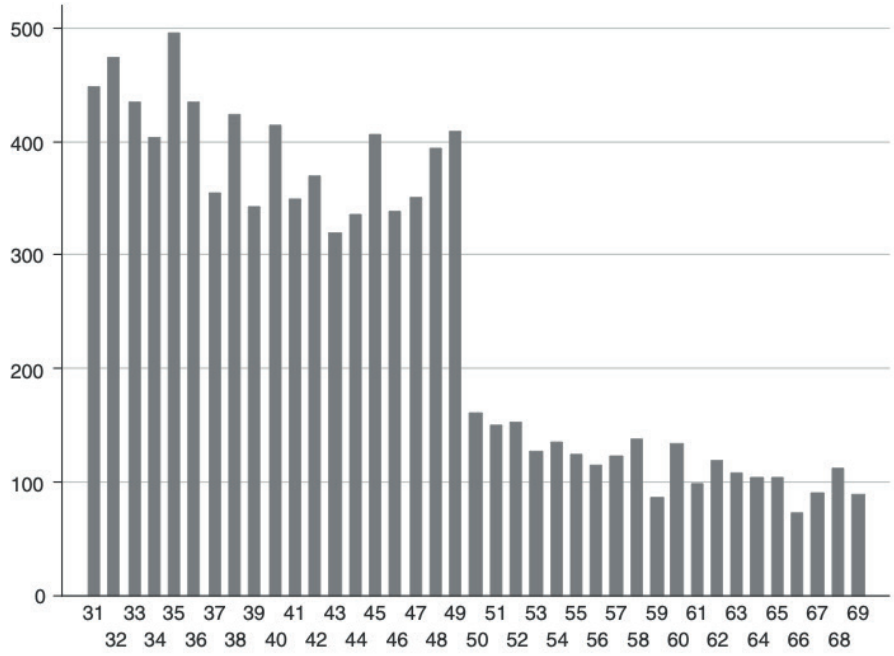
Uno de los grandes beneficiarios de la creación del CASD ha sido el Institut de Politiques Publiques, un think tank creado por la Paris School of Economics y las Grandes Écoles para promover la evaluación cuantitativa de las políticas públicas utilizando métodos de vanguardia. Este instituto ha tenido un notable impacto sobre el debate público en Francia, beneficiando al conjunto de la sociedad con sus evaluaciones independientes.

[3] https://www.casd.eu/en/projects-hosted-at-casd/

Ejemplo del Modelo Continental Francés. Evaluación de políticas públicas en Francia gracias a la creación del CASD en 2010

a) Garicano *et al*. (2016) analiza los efectos de las regulaciones laborales sobre el tamaño empresarial. Para ello utilizan datos sobre la población de empresas manufactureras en Francia, mostrando que hay una notable acumulación de empresas justo por debajo del umbral de 50 trabajadores, a partir del cual se aplican una serie de regulaciones que hacen el coste de los empleados mayor para las empresas. Los autores estiman que estas regulaciones reducen la competitividad global de las empresas francesas, al incentivar la actividad de las más pequeñas, que son menos productivas en promedio.

FIGURA 2
NÚMERO DE EMPRESAS POR TAMAÑO DEL EMPLEO EN FRANCIA



FUENTE: Garicano *et al*. (2016).

3.2.2. Alemania

La Agencia Federal de Empleo creó en el año 2004 un Centro de Datos para Investigación *(Forschungsdatenzentrum*, FDZ) que ofrece varias bases de datos provenientes de registros administrativos de la seguridad social alemana. Entre ellas

está la base integrada de historiales laborales (Integrated Employment Biographies, IEB), que contiene información sobre el censo de relaciones laborales del sector privado, incluyendo la remuneración y los días trabajados en cada empleo año a año, información sobre el nivel educativo y la ocupación del trabajador, el sector de la empresa, el régimen de dedicación (tiempo parcial o completo) y un identificador de establecimiento[4].

Existen tres modalidades de acceso a estos registros administrativos: presencial, remoto y a través de bases de datos reducidas para uso científico. Para facilitar el acceso en modo presencial a investigadores de todo el mundo, el FDZ ha llegado a acuerdos con diversas instituciones académicas y de investigación para establecer salas seguras de investigación en universidades de varios países como Estados Unidos (Michigan, Berkeley, Harvard, Princeton, Cornell, UCLA), Reino Unido (UCL, Essex), Canadá (Vancouver School of Economics), Francia (CASD) e incluso España (Universidad Carlos III de Madrid y Banco de España). El acceso a estos datos está abierto a investigadores de cualquier país del mundo y no tiene coste. Bajo estos acuerdos se están produciendo cientos de estudios anuales que cruzan la información proveniente de los distintos registros administrativos bajo condiciones de anonimato.

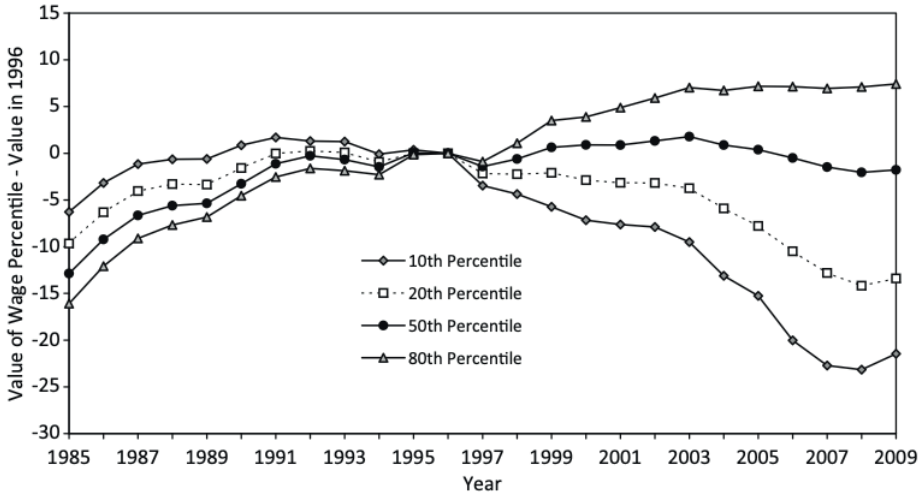
Ejemplos del Modelo Continental Alemán. Evaluación de la desigualdad salarial y de la carga del impuesto de sociedades entre empresarios y trabajadores.

a) Card *et al*. (2013) utilizan la base IEB para estudiar la desigualdad salarial. Los autores explotan la riqueza de los datos vinculados para estimar un modelo con efectos fijos de trabajador y establecimiento. Los resultados muestran que, a lo largo del periodo 1985-2009, se incrementaron tanto la dispersión en la productividad de los trabajadores como la dispersión en los salarios ofrecidos por las empresas (para un cierto nivel de productividad). En paralelo a estos procesos, creció la probabilidad de que los trabajadores más productivos se asociasen a los establecimientos con mayores salarios, dando como resultado un notable incremento en la desigualdad salarial en Alemania en este periodo. Esta investigación difícilmente hubiera podido realizarse con los datos de una encuesta o con una pequeña muestra de los datos administrativos, porque la identificación de los parámetros del modelo depende de los trabajadores que cambian de empresa: si ningún trabajador cambia de empresa, no se puede distinguir el efecto del trabajador y del establecimiento sobre el salario.

[4] Una empresa puede tener varios establecimientos.

FIGURA 3

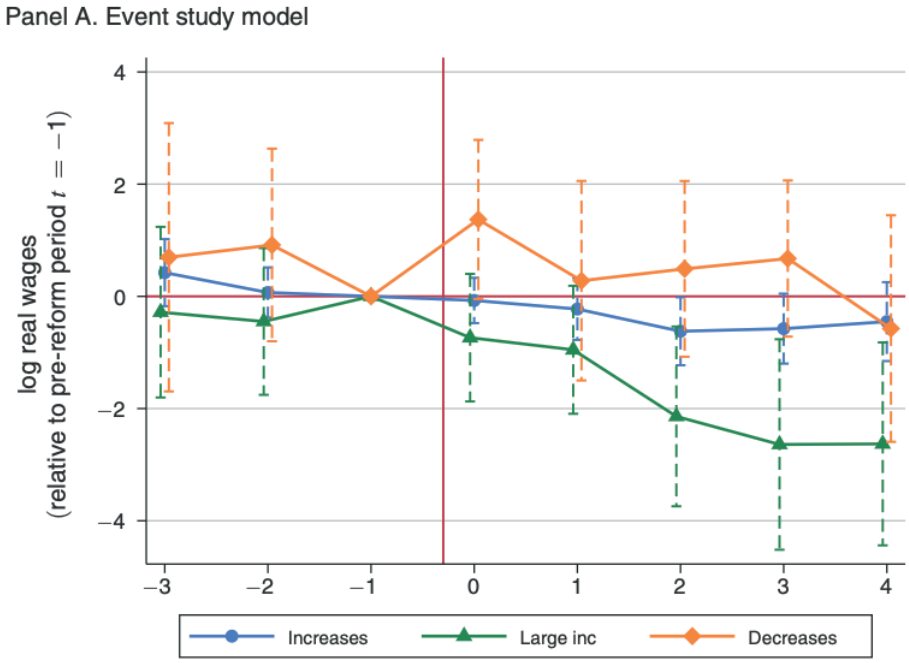
TENDENCIAS EN PORCENTAJES DEL LOGARITMO DE LOS SALARIOS DIARIOS PARA HOMBRES DE ALEMANIA DEL ESTE



FUENTE: Card *et al*. (2013).

b) Fuest *et al*. (2018) responden a una pregunta clásica: ¿quién soporta la carga de los impuestos sobre beneficios empresariales, el empresario o los trabajadores? La teoría nos dice que el sujeto que tiene la obligación legal de pagar un impuesto (en este caso, el empresario) no es necesariamente el que soporta toda la carga del impuesto, ya que puede ajustar su comportamiento de forma que parte de la carga se traslade a otro agente (en este caso, los trabajadores, a través de menores salarios). Los autores analizan miles de reformas de este impuesto local sobre los beneficios empresariales en Alemania para estimar hasta qué punto se traslada la carga impositiva a los trabajadores. Para ello, combinan datos a nivel municipal con el panel de empresas y trabajadores ofrecido por la seguridad social alemana, que incluye información sobre salarios a nivel individual. El principal resultado es que, cuando se produce una subida del impuesto sobre beneficios en un municipio, los trabajadores soportan el 51 % del coste de dicha reforma. Es importante destacar que no necesariamente bajan los salarios en términos nominales, sino que suben menos de lo que lo hacen en otros municipios comparables en los que no cambió la escala del impuesto.

FIGURA 4
LOGARITMO DE LOS SALARIOS REALES RELATIVOS ANTES DE LA REFORMA



FUENTE: Fuest *et al*. (2018)

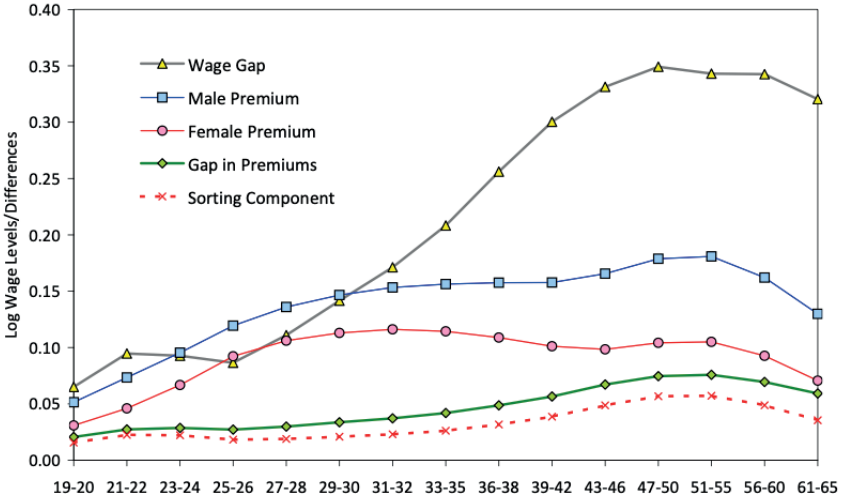
3.2.3. Portugal

Portugal ha desarrollado en los últimos años varios mecanismos de acceso a datos administrativos. Por un lado, en 2014 se formalizó la creación de un consorcio entre el *Instituto Nacional de Estatística* (homólogo del INE), la *Fundaçao para a Ciência e Tecnologia* (similar al CSIC) y la *Direção Geral das Estatística da Educação e Ciência*. El objetivo de este consorcio es proveer de acceso a los microdatos del INE a científicos afiliados a centros de investigación portugueses, incluyendo a estudiantes de doctorado y máster (en cuyo caso la solicitud debe ser refrendada por sus directores académicos). Los datos disponibles incluyen registros administrativos de empresas, relaciones laborales y comercio internacional, así como una gran variedad de encuestas con información demográfica, de salud y educación, entre otros. El acceso a los microdatos se debe realizar en las instalaciones del INE portugués o en centros seguros aprobados por el mismo.

Ejemplo del Modelo Continental Portugués. Las causas de la brecha salarial

a) Una de las bases de datos administrativos que más atención ha recibido de los investigadores son los *Quadros de Pessoal,* que recogen información detallada de todas las relaciones laborales en el sector privado (incluyendo duración de los contratos y remuneraciones) en Portugal durante varias décadas. Utilizando esta base de datos, Card *et al*. (2016) analizan la contribución de las empresas a la brecha salarial entre hombres y mujeres en Portugal. Gracias a la riqueza de estos datos, pueden distinguir dos efectos: por un lado, las mujeres tienen un menor poder de negociación salarial; por otro lado, las mujeres tienden a trabajar en empresas que pagan menores salarios. Estos dos efectos explican un 20 % de la brecha salarial total entre hombres y mujeres en Portugal.

FIGURA 5
EVOLUCIÓN DE LA BRECHA SALARIAL Y SUS COMPONENTES DURANTE EL CICLO VITAL



FUENTE: Card *et al*. (2016).

3.3. El modelo anglosajón

En los países anglosajones también ha habido un enorme interés recientemente por el uso de datos administrativos para la investigación. El modelo de acceso es generalmente más descentralizado que en los países nórdicos o de Europa continental, pues distintos organismos gubernamentales diseñan sus propios sistemas y protocolos de acceso a datos.

3.3.1. Estados Unidos

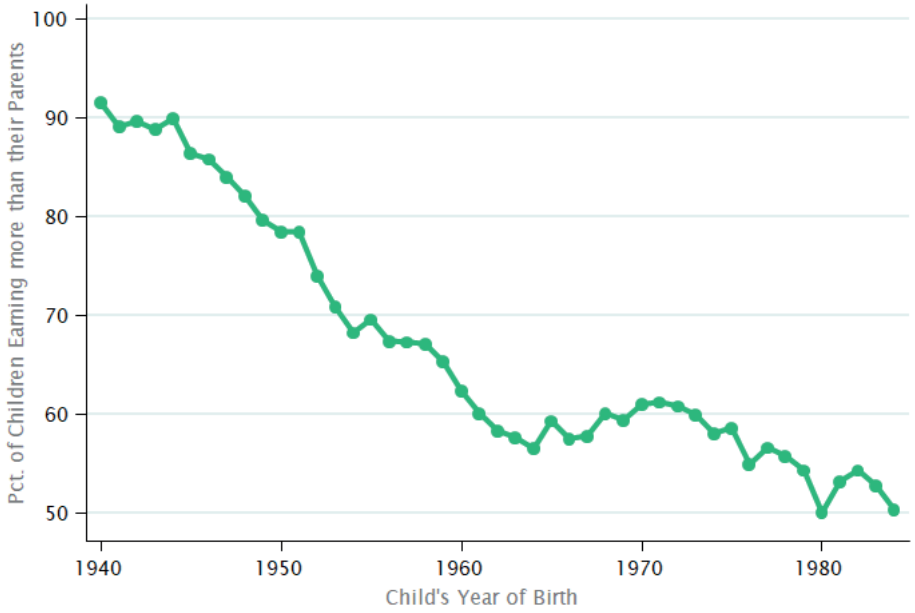
En Estados Unidos, un grupo de economistas de las universidades de Harvard y Berkeley escribió en septiembre de 2010 una carta abierta a la National Science Foundation (NSF) para pedir que se establecieran sistemas de acceso a datos administrativos (Card *et al*., 2010). En esta carta, argumentaban que se estaba realizando una gran cantidad de evaluaciones de políticas públicas en países europeos por la mayor disponibilidad de datos administrativos, y que Estados Unidos se estaba quedando atrás en este aspecto, perdiendo la posición de país líder en investigación en ciencias sociales. A raíz de esa iniciativa, instituciones como el Internal Revenue Service (la agencia tributaria estadounidense) pusieron en marcha sistemas de acceso a microdatos fiscales para investigación, dando lugar a una explosión de estudios sobre los efectos de las políticas públicas en Estados Unidos.

Ejemplo del Modelo Anglosajón en Estados Unidos. Nueva evidencia sobre movilidad intergeneracional usando los microdatos tributarios

a) Uno de los trabajos resultantes analiza los cambios en la movilidad intergeneracional en Estados Unidos (Chetty *et al*. 2014). Utilizando datos de declaraciones del impuesto sobre la renta, observan la renta de todos los estadounidenses nacidos en 1980-82 cuando tienen aproximadamente 30 años (en 2011-12) y vinculan esta información con la renta declarada por sus padres en el periodo 1996-2000, cuando los hijos eran adolescentes. Esto les permite estimar la relación entre la posición relativa de la renta de los padres y la de los hijos cuando son adultos. Los autores encuentran una gran variación en la movilidad intergeneracional entre distintas ciudades: la probabilidad de que un niño nacido en una familia que está en el 20 % más pobre de EE.UU. pase a estar en el 20 % más rico del país en su edad adulta es del 12,9 % en San José (en el *Silicon Valley* californiano) y solo del 4,4 % en Charlotte (en el estado sureño de Carolina del Norte). Comparando estos resultados con estudios de otros países, la movilidad intergeneracional en San José es similar a la de Dinamarca, pero la de Charlotte y otras ciudades del sureste es inferior a la de todos los demás países avanzados para los que existen datos. Los datos agregados de movilidad por ciudades se pueden consultar online. Estos resultados sugieren que EE.UU. ya no es, en gran parte de su territorio, una «tierra de oportunidades» y que la idea del sueño americano cada vez choca más con la realidad de una sociedad desigual en la que el éxito de una persona depende más de quiénes son sus padres que de su esfuerzo y mérito personal. Adicionalmente, el porcentaje de jóvenes que tiene mayor renta que sus padres al llegar a la edad adulta no ha parado de bajar (véase Figura 6). Este estudio, y otros relacionados que también hacen uso de datos tributarios, ha tenido un enorme impacto en el debate público sobre la desigualdad económica y racial en EE.UU.

FIGURA 6

PORCENTAJE DE HIJOS CON GANANCIAS MAYORES QUE SUS PADRES, POR AÑO DE NACIMIENTO



FUENTE: www.opportunityinsights.org/nationaltrends

3.3.2. Reino Unido

En el caso del Reino Unido, la agencia tributaria (Her Majesty's Revenue and Customs, HMRC) también ha creado un mecanismo para permitir el acceso de investigadores a datos fiscales y aduaneros, llamado HMRC DataLab. Al igual que en otros países, para obtener acceso a los datos se completa una solicitud en la que se explican los objetivos del estudio y las bases de datos necesarias. Para algunos proyectos se han combinado datos fiscales con otros datos externos, como por ejemplo los provenientes de registros mercantiles. El acceso a los datos solo se puede realizar presencialmente en una de las sedes de HMRC en Londres, lo cual limita parcialmente el acceso a investigadores residentes en otras ciudades del país. Además, el servicio no está sujeto al pago de ninguna tasa, una de las razones por las que no se ha realizado la inversión en un sistema de acceso remoto. El estudio de Almunia *et al*. (2019) realiza una comparación detallada del sistema británico con el finlandés.

Además, desde 2017 se puede acceder a microdatos del instituto de estadística británico (Office of National Statistics, ONS). El acceso se realiza a través de una red de centros seguros distribuidos por todo el país, donde los investigadores deben acudir en persona. Dada la restricción al acceso remoto, la pandemia ha supuesto una limitación en el acceso a datos para diversos proyectos de investigación.

Ejemplos del Modelo Anglosajón en el Reino Unido: Estudios sobre el efecto de las políticas fiscales en Reino Unido gracias al HMRC DataLab

a) Este sistema de acceso ha permitido que se realizaran investigaciones de alto impacto, como un estudio sobre los efectos de los impuestos a las transacciones inmobiliarias sobre el mercado de la vivienda (Best & Kleven, 2018), otro sobre los salarios obtenidos por estudiantes graduados en distintas universidades británicas (Belfield *et al*., 2018), una estimación del tipo impositivo efectivo que pagan las multinacionales extranjeras en Reino Unido en comparación con empresas locales del mismo tamaño (Bilicka, 2019) o el impacto de los incentivos fiscales a las donaciones a ONG y fundaciones (Almunia *et al*., 2020).

4. El acceso actual a datos administrativos en España

El acceso a datos administrativos para investigación en España es limitado en comparación con los países mencionados en la sección anterior, aunque no partimos de cero. España comparte con otros países europeos una gran ventaja a la hora de combinar diferentes registros administrativos: la existencia de un identificador individual que coincide con el número del documento nacional de identidad (DNI). Aunque para la gran mayoría de estudios, dicho identificador ha de ser anonimizado mediante un código, el identificador permitiría, en teoría, seguir a un mismo ciudadano cruzando distintas bases de datos de forma que se pudieran estudiar los efectos de variables pertenecientes a distintas fuentes.

Una visión de conjunto de la actual situación en España muestra a una serie de instituciones dando pequeños pero importantes pasos hacia el acceso a los datos. Sin embargo, muchas de estas iniciativas son parciales y aisladas, por lo que se las puede considerar como el sustrato sobre el que esperemos que vaya fermentando una visión más global que supere las muchas limitaciones actuales para acceder a los relativos escasos datos disponibles. En España existen tres fuentes básicas de datos administrativos a las que los investigadores han podido tener acceso: la Muestra continua de vidas laborales, el Panel de empresas-trabajadores y el Panel de declarantes del IRPF.

La Muestra Continua de Vidas Laborales (MCVL), disponible desde 2004, es una muestra aleatoria que contiene información (anonimizada) sobre un 4 % de los cotizantes y beneficiarios de la Seguridad Social. Se trata de una base de datos longitudinal que permite seguir el historial laboral de estos individuos a lo largo del tiempo. Estos datos se han utilizado en un gran número de estudios, por ejemplo para analizar la evolución de la desigualdad salarial a lo largo del ciclo económico (Bonhomme & Hospido, 2017) o evaluar la empleabilidad de los parados de larga duración (Bentolila *et al*., 2017).

El Panel de Empresas-Trabajadores (PET) se presentó en 2019 y supone un avance importante respecto a la MCVL. El muestreo se hace a nivel de empresa, en lugar

de trabajador, de forma que los investigadores observan a todos los individuos que durante el periodo 2013-2016 han estado afiliados al Régimen General de la Seguridad Social a través de cada empresa seleccionada en la muestra. La muestra cubre un 3-5 % de las empresas pequeñas (entre 1 y 9 trabajadores) y un 15% de las más grandes (más de 500 trabajadores). La manera de construir la muestra permitirá a los investigadores analizar cuestiones como los flujos de creación y destrucción de empleo, la contribución de los salarios a la desigualdad de la renta, los efectos del uso (y abuso) de los contratos temporales y la segregación horizontal y vertical de los trabajadores.

El Instituto de Estudios Fiscales (IEF) ofrece acceso a microdatos del Impuesto sobre la Renta de las Personas Físicas (IRPF) para el periodo 1999-2015 en dos formatos: muestras aleatorias anuales (que incluyen un 10-15 % del total de declaraciones anuales) y un panel de declarantes (que incluye un 3 % de las declaraciones anuales, en media). Las muestras son especialmente útiles para hacer ejercicios de microsimulación, mientras que el panel es útil para analizar los efectos de reformas fiscales sobre el comportamiento de los contribuyentes, como se hace en un estudio reciente que estima la elasticidad de la renta respecto a cambios en el tipo impositivo marginal (Almunia & López-Rodriguez, 2019).

Estas tres bases de datos administrativos se distinguen de las descritas anteriormente para otros países en que se trata de muestras representativas, pero que incluyen sólo un porcentaje de la población total. Desde el punto de vista del análisis estadístico, asumiendo que el muestreo esté hecho correctamente, es posible inferir características generales de la población a partir de una muestra. Sin embargo, muchas preguntas de investigación sólo se pueden abordar con datos poblacionales, porque las muestras dejan de ser representativas para subgrupos muy específicos. Por ejemplo, un estudio reciente de Llaneras *et al*. (2020) ha tratado de replicar el análisis de movilidad intergeneracional realizado por Chetty *et al*. (2014) para Estados Unidos, pero las limitaciones de los datos disponibles hacen que este análisis no sea directamente comparable (como ha argumentado Polavieja, 2020). Esta limitación también afecta a estudios sobre la desigualdad de la renta y la riqueza, en los que es crucial tener información detallada sobre las personas que están dentro del top 1 % de la distribución.

Otra desventaja de estas bases de datos es que, en la actualidad, no se pueden cruzar con otros registros administrativos. Por lo tanto, la MCVL no permite realizar un análisis completo de los efectos de cambios en el salario mínimo sobre distintos tipos de empresas, según su tamaño, sector o localización geográficas. Tampoco podemos estudiar con datos administrativos el efecto de distintos niveles de educación sobre la renta, o los efectos de tener hijos sobre la desigualdad salarial entre los cónyuges. Es paradójico que los investigadores españoles puedan acceder al universo de historiales laborales de Alemania a través de las salas seguras establecidas por el FDZ alemán en la Universidad Carlos III de Madrid y el Banco de España, pero no tengan acceso a los datos equivalentes de la seguridad social española. No obstante, la reciente apertura de salas seguras en distintas ciudades del Ministerio de

Seguridad Social, supone un primer paso en la buena dirección. Aunque a fecha de hoy no es posible disponer del universo de historias laborales, y los investigadores deben solicitar extracciones concretas de los datos que necesitan, se espera que en breve se elimine esta traba administrativa. La web de la Seguridad Social no tiene ninguna documentación sobre los datos que están disponibles en las nuevas salas seguras. Esta información debería publicarse (en español y en inglés) para que los potenciales investigadores sepan a qué datos pueden acceder. Quizá incluso se podría ofrecer una base de datos «ficticios» que mostrara valores ficticios pero que diera cuenta de las variables disponibles, como de hecho ya hace el FDZ alemán, de forma que se facilitase el desarrollo de preguntas posibles de investigación y el desarrollo de códigos para el análisis, antes de acudir a las salas seguras presencialmente a realizar la petición expresa de datos.

Una limitación adicional para el análisis es la segmentación geográfica. Por ejemplo, y respecto a las bases de datos tributarios, no incluyen a las Comunidades Autónomas de Régimen Foral (Navarra y País Vasco), por lo que no se puede obtener una visión del conjunto del país ni se pueden analizar los flujos migratorios (de personas y empresas) desde y hacia estas dos Comunidades Autónomas como respuesta a diferencias en la regulación.

Por otro lado, el Banco de España ha lanzado recientemente una iniciativa, el BE-Lab, para dar acceso a investigadores externos a varias de sus bases de datos. Desde julio de 2019, se puede acceder a la Central de Balances, que compila la documentación aportada anualmente por las empresas no financieras al Registro Mercantil. Un aspecto positivo de esta iniciativa es que una vez se ha creado la infraestructura para el acceso seguro a datos (en las dependencias del Banco de España), se han seguido dando pasos para facilitar el acceso. En este sentido, destaca el hecho que desde mayo de 2021, se puede acceder a los datos del BELab de forma remota, lo cual supone un avance muy positivo en línea con el punto 2.5 de nuestro marco conceptual.

Por su parte, el INE proporciona acceso a través de su página web a microdatos de las encuestas que realiza. Estas encuestas tienen información valiosa, pero tienen las desventajas propias de realizarse con muestras parciales de la población. Además, los únicos microdatos administrativos que se ofrecen en esta página son los de nacimientos y defunciones. Estos últimos son útiles para estadística descriptiva, pero, al no poder cruzarse con ninguna otra fuente de datos, no permiten realizar un análisis más profundo de relaciones entre distintas variables.

Finalmente, a nivel de las Comunidades Autónomas hay algunas iniciativas prometedoras, como el DataResSS (*Data for Research in Social Sciences*), un consorcio de la Barcelona Graduate School of Economics y el Instituto de Estadística de Cataluña (Idescat). El objetivo de este organismo es poner datos administrativos de Cataluña que posee Idescat a disposición de investigadores de vanguardia, tanto españoles como extranjeros. El DataResSS tiene un comité académico que evalúa las propuestas de investigación en función de su calidad académica, contribución al diseño de políticas públicas, cuestiones éticas relacionadas con la confidencialidad, y su factibilidad y extensión temporal esperada. Uno de los aspectos positivos del

diseño de DataResSS es la cooperación entre un grupo de universidades de máximo nivel con el instituto de estadística de la Comunidad Autónoma.

Quizá el paso reciente más esperanzador ha sido la reciente declaración conjunta del Instituto Nacional de Estadística (INE), la Agencia Tributaria (AEAT), el Ministerio de la Seguridad Social y el Banco de España comprometiéndose a «comenzar a trabajar conjuntamente en el desarrollo de un sistema de acceso a sus bases de datos con fines científicos de interés público». Dicha declaración supone un importante paso para desarrollar una estrategia más global por parte precisamente de las instituciones que ya han sido las impulsoras a nivel individual del relativamente escaso acceso a datos administrativos en España. No obstante, el comunicado no concreta ni el método de acceso a los datos que se busca ni la infraestructura que se utilizará, ni la estrategia que se seguirá para incorporar nuevas bases de datos. Nos parece además particularmente importante que se produzca un compromiso expreso para dotar de recursos, tanto informáticos como humanos, a esta iniciativa, de forma que no se quede en una intención por colaborar, sino que llegue a hacerse realidad, e inspire a pasos aún más ambiciosos. Para conseguirlo creemos que serían muy útiles las recomendaciones incluidas en el reciente informe de la AIREF, institución también relevante e implicada desde 2017 en la evaluación de políticas públicas y desde su formación en una estrategia de apertura de los datos administrativos.

5. ¿Cómo diseñar un sistema de acceso a datos administrativos para investigación en España?

En las secciones anteriores hemos explicado las distintas opciones que existen y hemos resumido la experiencia de países de nuestro entorno para demostrar que es posible ofrecer acceso a datos administrativos para investigación de forma segura y con importantes resultados en cuanto a la cantidad y calidad de la investigación sobre políticas públicas. A continuación, presentamos algunas reflexiones específicas sobre la mejor manera, a nuestro modo de ver, de adaptar estas experiencias a las particularidades de las administraciones públicas en España.

Nuestras reflexiones se centran en los cuatro aspectos clave que discutimos en la sección 2: ¿Qué institución, o instituciones, deben cumplir el papel de custodio de los datos y de «tercero seguro»? ¿Cuáles deberían ser los criterios de acceso a los datos? ¿Cómo se garantizará la confidencialidad de los datos? ¿Cómo se debería realizar el acceso a los datos?

5.1. ¿Qué institución debería jugar el papel de depositaria de los datos y tercero seguro?

Una posible ventaja de adoptar el modelo nórdico, donde el instituto de estadística centraliza el acceso a todos los registros administrativos, es que nuestro INE ya

dispone de un respaldo legal para recopilar y custodiar registros administrativos. Sin embargo, es importante ser conscientes de que el INE ya tiene asignadas una gran cantidad de funciones, como son la elaboración de la Contabilidad Nacional, el Censo de Población y numerosas encuestas. Por lo tanto, si se quiere asignar esta nueva tarea al INE, será indispensable dotarlo de los recursos humanos y técnicos suficientes para llevarla a cabo.

Una alternativa viable es seguir el modelo europeo continental, basado en la creación de consorcios de varias instituciones públicas con algunas entidades de investigación. En el caso de España, este tipo de consorcio podría incluir a las instituciones que ya han expresado formalmente su voluntad de colaborar: el INE, la Agencia Tributaria, la Seguridad Social, el Banco de España y la AIReF. Sería también deseable que las Comunidades Autónomas pudieran formar parte de este consorcio, quizá a través de sus propios institutos de estadística. De hecho, no es necesario que se forme un único consorcio. Podría comenzarse a trabajar con acuerdos puntuales para compartir ciertas bases de datos aprovechando que algunas instituciones, como el Banco de España, el INE o la AIReF, ya tienen creada una infraestructura de centros seguros. Paulatinamente se podrían ir incorporando otras instituciones una vez se compruebe que el sistema funciona adecuadamente y se comiencen a observar los resultados positivos que se obtienen de la evaluación de las políticas públicas.

5.2. Criterios de acceso a los datos

Los criterios de acceso a los datos deben ser uniformes y transparentes, al contrario de lo que sucede actualmente. Cualquier investigador que desee acceder a los datos debería presentar una solicitud explicando el objeto de su estudio y especificando los registros a los que necesitaría acceder. Esto implica que no se restrinja el acceso solamente a investigadores asociados a unas pocas instituciones de reconocido prestigio o a investigadores con un historial de publicaciones. El acceso debería estar abierto a cualquier investigador con una buena idea y, especialmente, a investigadores jóvenes (incluso estudiantes de doctorado, como argumentan Card *et al*. (2010) en el contexto de Estados Unidos). En este último caso se puede requerir el apoyo del director de tesis como garantía de compromiso con la protección de la confidencialidad y la calidad de la investigación.

Idealmente, deberían existir dos comités que revisen las solicitudes de acceso a los datos: el primero estaría formado por investigadores de reconocido prestigio que evalúe el interés de cada proyecto desde el punto de vista académico. El segundo estaría formado por personas de la institución generadora de los datos (o del «tercero seguro») que evalúen la viabilidad del proyecto desde el punto de vista logístico. Por ejemplo, si es posible combinar las bases de datos requeridas o acceder a ciertos datos que tengan una especial protección por su confidencialidad. Estos comités deberían reunirse con suficiente frecuencia, por ejemplo mensual o trimestralmente, para evaluar las solicitudes de acceso a los datos sin causar un retraso excesivo a los

proyectos de investigación. En cualquier caso, el proceso de solicitud debería ser sencillo: los procesos de investigación no siempre siguen procesos lineales, por lo que como punto de partida no se debería exigir más que una pregunta o hipótesis de investigación bien definida y una breve descripción del contexto en el que se plantea esa pregunta. Un formulario de pocas páginas es suficiente para recoger la información necesaria para la inmensa mayoría de proyectos de investigación. Un proceso demasiado extenso y complicado sería contraproducente ya que funcionaría como barrera a la entrada.

Por último, es crucial tener en cuenta la capacidad del sistema para absorber un cierto número de proyectos de investigación simultáneamente. Las limitaciones de capacidad pueden surgir por muchos factores: carga de trabajo en los comités de evaluación de proyectos, capacidad de las salas seguras (en caso de acceso presencial) o de los servidores (en caso de acceso remoto), retrasos en el proceso de revisión de nuevos resultados, etc.

5.3. Modo de acceso a los datos

El objetivo a medio plazo debe ser que exista la posibilidad de acceder a los datos de manera remota a través de internet, como se hace en varios países de nuestro entorno. Dicho esto, la secuencia habitual que han seguido otros países es desarrollar primero una infraestructura de centros seguros, en los que se maximiza el control de la actividad de los investigadores, que deben acudir presencialmente. Por ello, este suele ser el modo de acceso preferido de instituciones que ceden sus datos por primera vez. Sin embargo, la experiencia internacional indica que, una vez los sistemas están en marcha y funcionan adecuadamente, se suele producir un aumento de la demanda de nuevos proyectos que en poco tiempo supera la capacidad de estos centros seguros presenciales. Además, los centros seguros presenciales generan desigualdad territorial, pues el acceso es mucho más sencillo para los investigadores que residen en la misma ciudad (habitualmente la capital del país) que para los que están en otros lugares. Por todas estas razones, todos los modelos descritos anteriormente han terminado convergiendo hacia el acceso remoto a través de internet, haciendo uso de las tecnologías existentes. Como ha demostrado la experiencia reciente del INE, que cuenta con una sala segura para fines investigadores a la que desde hace poco se puede acceder de forma remota, es posible ofrecer acceso remoto y tener un control sobre la actividad de los investigadores muy similar al de los centros seguros presenciales. Por último, la opción del acceso remoto puede fomentar que accedan a los datos también investigadores ubicados fuera de España, expandiendo el grupo de investigadores potenciales que harán uso de los datos. En definitiva, la inversión a corto plazo en una tecnología segura de acceso remoto puede reducir enormemente los costes administrativos a largo plazo.

5.4. Cómo garantizar la confidencialidad

Este aspecto, que en muchas ocasiones se cita como una barrera insalvable, se ha solventado en todos los países de nuestro entorno cuyos sistemas hemos descrito anteriormente. ¿Existe alguna diferencia radical entre el sistema legal español y el de otros países europeos que impida el uso de datos administrativos con fines de investigación? La respuesta es no, por la sencilla razón de que si los datos están correctamente anonimizados y existe un control sobre el uso que se hace de ellos, en ningún momento se produce una quiebra de la confidencialidad. Tanto la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, como la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, son explícitas en este punto. La Directiva UE 2019/1024 de 20 de junio de 2019 aplica por igual a todos los países miembros de la Unión Europea e indica claramente que cualquier documento conservado por organismos públicos sea puesto a disposición general del público, por lo que, con las debidas precauciones, el impulso legislativo desde Europa hacia la accesibilidad de los datos es tajante.

El paso más básico en el proceso de anonimización de los datos consiste en sustituir los identificadores individuales (por ejemplo, DNI de personas o CIF de entidades jurídicas) por otros valores utilizando un algoritmo al que solo tenga acceso el depositario de los datos, pero nunca el investigador. En ocasiones, si las bases de datos administrativos son muy detalladas, puede ser necesario ir más allá y eliminar o distorsionar otra información (ubicación, edad, género, renta) que pueda permitir la identificación de un individuo o empresa concretos. Los detalles de estas técnicas van más allá del alcance de este artículo, pero es importante resaltar que se ponen en práctica en todos los países mencionados anteriormente y por lo tanto no hace falta inventar nuevos métodos.

La práctica habitual en los centros de acceso a datos administrativos es que los investigadores no pueden extraer ningún resultado hasta que no sea revisado por personal del centro donde los haya consultado. Este personal se cerciorará de que no se viola la confidencialidad, por ejemplo, poniendo como requisito que haya un mínimo de unidades (personas, empresas) en cada celda de las tablas de resultados.

6. Conclusiones

La modernización de la administración pública para poder implementar y evaluar políticas basadas en la evidencia pasa necesariamente por mejorar el acceso a los datos públicos y de diversos agentes facilitando la creación de conocimiento crucial tanto para el diseño y la evaluación de políticas como para la comunidad investigadora.

España se está quedando atrás respecto a otros países de nuestro entorno, que han creado agencias públicas de acceso a datos administrativos siguiendo distintos modelos que se adaptan a las características propias de su administración. El éxito de estas agencias nos permite usar como ejemplo los mejores modelos de cada una de

las importantes decisiones políticas que deben tomarse (centro seguro, confidencialidad, modo de acceso…) que mejor se adecúen a nuestra realidad territorial.

El cambio de cultura necesario para que nuestra muy fragmentada administración comparta y coordine sus datos supone un reto importante. Sin embargo, casos recientes muy notorios como la caótica gestión de la información durante la pandemia de COVID-19 abren una ventana de oportunidad para conseguir un consenso de todos los sectores de la sociedad sobre que este cambio de cultura es imprescindible.

Existen cada vez más casos aislados de éxito en nuestro país, e incluso iniciativas conjuntas de varias instituciones, sobre el potencial que abre el acceso público a los datos administrativos. Estos casos muestran que en España ni partimos de cero ni es inviable conseguirlo. La articulación de un Plan Nacional para impulsar la creación de una agencia de datos puede ir dando pequeños pasos, utilizando aquellos datos y administraciones menos sensibles, que afiancen el gran cambio necesario.

Referencias bibliográficas

Almunia, M., Harju, J., Kotakorpi, K., Tukiainen, J., & Verho, J. (2019). Expanding access to administrative data: the case of tax authorities in Finland and the UK. *International Tax and Public Finance*, *26*(3), 661-676. https://link.springer.com/article/10.1007/s10797-018-9525-0

Almunia, M., Guceri, I., Lockwood, B., & Scharf, K. (2020). More giving or more givers? The effects of tax incentives on charitable donations in the UK. *Journal of Public Economics*, *183*, 104114. https://www.sciencedirect.com/science/article/abs/pii/S0047272719301768

Arellano, Manuel (2018, 20-21 de agosto). *El acceso a los microdatos administrativos públicos: la nueva frontera de la investigación económica y social* [sesión de conferencia]. XVIII Aula de Verano Ortega y Gasset, Universidad Internacional Menéndez Pelayo (UIMP). http://www.cemfi.es/~arellano/arellano-presentacion-uimp-2018.pdf

Belfield, C., Britton J., Buscha F., Dearden L., Dickson M., Van Der Erve L., Sibieta L., Vignoles A., Walker I., & Yu Zhu (2018). *The relative labour market returns to different degrees*. Institute of Fiscal Studies. https://www.ifs.org.uk/publications/13036

Bentolila, S., García-Pérez, J. I., & Jansen, M. (2017). Are the Spanish long-term unemployed unemployable? *SERIEs*, *8*(1), 1-41. https://link.springer.com/article/10.1007/s13209-017-0155-z

Best, M. C., & Kleven, H. J. (2018). Housing market responses to transaction taxes: Evidence from notches and stimulus in the UK. *The Review of Economic Studies*, *85*(1), 157-193. https://www.henrikkleven.com/uploads/3/7/3/1/37310663/bestkleven_landnotches_sep2016.pdf

Bilicka, K. A. (2019). Comparing UK tax returns of foreign multinationals to matched domestic firms. *American Economic Review*, *109*(8), 2921-53. https://www.aeaweb.org/articles?id=10.1257/aer.20180496

Card, D., Cardoso, A. R., & Kline, P. (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, *131*(2), 633-686. https://doi.org/10.1093/qje/qjv038

Card, D., Heining, J., & Kline, P. (2013). Workplace heterogeneity and the rise of West German wage inequality. *The Quarterly journal of economics*, *128*(3), 967-1015. https://doi:10.1093/qje/qjt006

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*. https://eml.berkeley.edu/~saez/card-chetty-feldstein-saezNSF10dataaccess.pdf

Chetty, R., Friedman, J. N., Leth-Petersen, S., Nielsen, T. H., & Olsen, T. (2014). Active vs. passive decisions and crowd-out in retirement savings accounts: Evidence from Denmark. *The Quarterly Journal of Economics*, *129*(3), 1141-1219. http://www.rajchetty.com/chettyfiles/ret_savings.htm

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553-1623.

https://eml.berkeley.edu/~saez/chetty-friedman-kline-saezQJE14mobility.pdf

De Quinto, A., Hospido L., & Sanz C. (2020). *The Child Penalty in Spain*. Documento Ocasional 2017, Banco de España. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/20/Files/do2017e.pdf

Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo de 20 de junio de 2019 relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida). *Diario Oficial de la Unión Europea*. http://data.europa.eu/eli/dir/2019/1024/oj

Fuest, C., Peichl, A., & Siegloch, S. (2018). Do higher corporate taxes reduce wages? Micro evidence from Germany. *American Economic Review*, *108*(2), 393-418. https://www.aeaweb.org/articles?id=10.1257/aer.20130570

Garicano, L., Lelarge, C., & Van Reenen, J. (2016). Firm size distortions and the productivity distribution: Evidence from France. *American Economic Review*, *106*(11), 3439-79. https://www.aeaweb.org/articles/pdf/doi/10.1257/aer.20130232

Kleven, H., Landais, C., & Søgaard, J. E. (2019). Children and gender inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, *11*(4), 181-209. https://www.aeaweb.org/articles?id=10.1257/app.20180010

Kleven, H., Landais, C., Posch, J., Steinhauer, A., & Zweimüller, J. (2019). Child Penalties across Countries: Evidence and Explanations. *AEA Papers and Proceedings*, *109*, 122-26. https://www.aeaweb.org/articles?id=10.1257/pandp.20191078

Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público. *Boletín Oficial del Estado*, nº 276, de 17 de noviembre de 2007. https://www.boe.es/eli/es/l/2007/11/16/37

Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno. *Boletín Oficial del Estado*, nº 295, de 10 de diciembre de 2014. https://www.boe.es/eli/es/l/2013/12/09/19/con

Llaneras, K., Medina O., & Costas E. (2020). *Atlas de Oportunidades*, proyecto conjunto de la Fundación COTEC y la Fundación Felipe González. https://www.cotec.es/fundacionfelipegonzalez/oportunidades/proyecto/

Polavieja, J. G. (2020). Grandes Datos, Grandes Sesgos, Grandes Errores: Sobre el Atlas de Oportunidades. *Revista Internacional de Sociología*, *78*(3), 166. https://www.javierpolavieja.com/sobre-el-atlas-de-oportunidades

Portal de transparencia. Gobierno Abierto. Compromiso 2. *Incorporación de Directiva (UE) 2019/1024*. Consultado el 16 de octubre de 2021. https://transparencia.gob.es/transparencia/transparencia_Home/index/Gobierno-abierto/seguimientoIVPlanGA/seguimiento_C2/2-3-1-IncorporacionDirectiva.html

Real Decreto-ley 20/2020, de 29 de mayo, por el que se establece el ingreso mínimo vital. *Boletín Oficial del Estado*, n.º 154, 1 de junio de 2020. https://www.boe.es/eli/es/rdl/2020/05/29/20/con

Slemrod, J. (2016). Caveats to the research use of tax-return administrative data. *National Tax Journal*, *69*(4), 1003-1020. http://www.ntanet.org/NTJ/69/4/ntj-v69n04p1003-1020-caveats-research-use-tax-return-data.pdf?v=%CE%B1&r=19426893815398216