BIG DATA EN LA ESTADÍSTICA PÚBLICA: RETOS ANTE LOS PRIMEROS PASOS

DAVID SALGADO
Instituto Nacional de Estadística

La digitalización de la economía y de la actividad humana en general ha traído en los últimos años, entre otras consecuencias, la generación y el almacenamiento de cantidades ingentes de datos en sistemas digitales (véase p. ej. The Economist (2010)). Sin duda, este fenómeno abre una ventana a nuevas fuentes de información con múltiples aplicaciones.

Es fácil argumentar el potencial de estos datos en el sector privado, no ya en los propios departamentos de *Business Intelligence* para una optimización de las operaciones de las compañías, sino como una oportunidad de nuevas líneas de negocio donde el dato generado por el servicio ofertado al cliente por la empresa, debidamente anonimizado y protegido, puede reusarse para análisis socioeconómicos de diversos aspectos de la sociedad con un creciente valor.

De hecho, así está explícitamente reconocido por la reciente Comunicación de la Comisión Europea titulada «*Building a European Data Economy*» (European Commission, 2017), donde se reconoce el creciente valor de los datos digitales en la economía europea y se apunta en el medio plazo a preparar iniciativas legales para propiciar un marco de cooperación basado en el libre flujo de datos en la Unión Europea, así como su acceso y re-uso bajo las garantías normativas de la protección de la información.

Por tanto, la reutilización de estos datos por el sector público y, en particular, para la elaboración de estadísticas oficiales aparece como una etapa natural en esta generación y acumulación de información. De hecho, el principio 5 de los Principios Fundamentales de las Estadísticas Oficiales de Naciones Unidas (ONU, 2014) claramente indica que «los datos para fines estadísticos pueden obtenerse de todo tipo de fuentes», indicando que «los organismos de estadística han de seleccionar la fuente con respecto a la calidad, la oportunidad, el coste y la carga que impondrá a los encuestados». Aunque recientemente adoptados por la Asamblea General de las Naciones Unidas (ONU, 2014), estos principios no consideran explícitamente la situación presente que los *Big Data* han traído. Por ello, recientemente el *Global Working Group on Big Data* de Naciones Unidas ha redactado sus *Recommendations for Access to Data from Private Organizations for Official Statistics (United Nations Global Working Group on Big Data*, 2016), reconociendo que para el propósito de producir estadísticas oficiales de calidad y relevantes, es preciso recabar datos de organizaciones privadas como inputs para su producción.

A priori es natural esperar que tales cantidades ingentes de información sean de utilidad en algún grado para la elaboración de las estadísticas oficiales, ya sea

como fuente primaria o secundaria de datos o como información auxiliar en el proceso de producción estadística. Sin embargo, el reto no es sencillo. El proceso de producción estadística oficial es un sistema complejo y los *Big Data* traen consigo cambios notables con un previsible impacto significativo desde diversos puntos de vista.

Desde nuestra experiencia (Salgado *et al*., 2016, 2017), la definición misma de *Big Data* para la producción de estadísticas oficiales debe ser adecuadamente perfilada para no perder matices importantes para la Estadística Pública. La mayor parte de estas nuevas fuentes de información precisan resolver diversas cuestiones, en muchos casos fuertemente entrelazadas entre sí, que van desde el acceso institucional a los datos, pasando por la metodología estadística que sustenta la inferencia de los datos recogidos respecto de la población de análisis (ya sea de personas, empresas, establecimientos,…) y por la tecnología necesaria hasta la calidad de las estadísticas producidas, aspectos todos ellos que van a ser desarrollados en los distintos epígrafes de este artículo. Algunos de estos retos se enmarcan en el proceso de modernización e industrialización de la producción estadística oficial, un proceso puesto en marcha en la última década aproximadamente que va más allá del uso de estas nuevas fuentes de datos.

DEFINICIÓN REVISADA DE *BIG DATA* PARA LA ESTADÍSTICA PÚBLICA

Es sobradamente conocida la caracterización del concepto de *Big Data* mediante la confluencia de las 3Vs (Laney, 2001): volumen, velocidad y variedad; si bien posteriormente más Vs han sido añadidas (Normandeau, 2013). Muchas nuevas fuentes de datos digitales de interés para la Estadística Pública no satisfacen estos criterios, como pueden ser los datos provenientes de la actividad de los terminales de telefonía móvil en las redes celulares de telecomunicaciones. Ciertamente, el volumen y la velocidad de generación son notablemente altos, pero son datos claramente estructurados sin mucha variedad.

En conexión con la producción estadística oficial, sin embargo, nosotros defendemos que existen características más relevantes para su uso en la Estadística Pública, en particular, en comparación con los datos de encuestas y de registros administrativos.

Por ello, en el ecosistema de datos de la Estadística Oficial proponemos parametrizar las fuentes de datos en términos de las siguientes características (Salgado *et al.*, 2017):

1. Los datos no contienen información sobre el proveedor de datos (informante), sino sobre terceras personas, ya sean físicas o jurídicas.

2. Los datos desempeñan un papel central en el proceso de negocio del proveedor de datos (informante).

3. Los datos poseen al menos una de las tres Vs de la definición clásica de *Big Data* (volumen, velocidad, variedad).

Estos rasgos permiten caracterizar actualmente las tres fuentes de datos de las Estadísticas Oficiales, a saber, las encuestas, los registros administrativos y los *Big Data*. Con breves ejemplos se aprecia mejor la relevancia de estas características. En las encuestas coyunturales tradicionales, por ejemplo, al sector servicios, se solicita a las empresas informar mensualmente sobre su personal contratado. Complementariamente, en algunas otras operaciones estadísticas, se trabaja con las cuentas de cotización a la Seguridad Social de las empresas, que provienen del correspondiente registro administrativo generado y gestionado por el Ministerio de Empleo y Seguridad Social. Por último, es posible rastrear y descargar datos de empleo de portales específicos y *webs* de empresas en Internet.

Adviértase cómo, en el caso de las encuestas, los datos no cumplen ninguna de las características anteriores: las empresas proporcionan sus propios datos de empleo contratado, no de otras empresas; estas cifras no desempeñan un papel esencial en su proceso de producción y, en general, se trata de un volumen muy bajo de datos que se genera (con esfuerzo) a petición expresa de la oficina pública de estadística (nunca por debajo del mes) y tienen una estructura muy concreta derivada del cuestionario diseñado y administrado por la oficina estadística.

En el caso del registro administrativo se cumplen las características 1 y 2, pero no la característica 3. Evidentemente los datos contenidos en el registro administrativo se refieren a terceras personas, no al propio Ministerio de Empleo y Seguridad Social (característica 1). Además, la generación y mantenimiento de este registro constituyen una de las funciones del Ministerio (característica 2). Por último, no se observa ninguna de las tres Vs clásicas que definen una fuente *Big Data*. Es posible que en el futuro ciertos registros administrativos sí aumenten en volumen y velocidad de alimentación y, por tanto, sea aconsejable introducir una nueva característica sobre la titularidad pública o privada de los datos. En la actualidad, no conocemos ningún registro administrativo con uso en la Estadística Pública que pueda considerarse un ejemplo de *Big Data*.

En el caso de portales específicos de ofertas de empleo, es evidente que las características 1 y 2 se cumplen: estos datos sobre terceros agentes (las empresas ofertantes) son el elemento central del negocio de los portales. Ambas características son la consecuencia directa de la digitalización de la economía por la que progresivamente los datos desempeñan un papel cada vez más central en los procesos de negocio, siendo, por tanto, claves en la estrategia de negocio de los agentes económicos. Además, dependiendo de las características propias de cada portal, bien el volumen, la velocidad o la variedad de los datos caracterizan la información. Compartir estos datos em-

pieza a ser equivalente a compartir buena parte de la materia prima de muchas actividades económicas.

En general, las nuevas fuentes de *Big Data* cumplen las tres características anteriores. Y todas ellas tienen consecuencias directas sobre el uso de estas fuentes para la elaboración de estadísticas oficiales, como mostraremos en las secciones siguientes.

ACCESO INSTITUCIONAL

El primer reto para la incorporación de los *Big Data* al proceso de producción estadística oficial es acceder a los datos. Este acceso debe producirse, no ya en condiciones provisionales para investigar el uso de esta fuente en la producción de estadísticas oficiales, sino especialmente en las condiciones estándares de producción que permitan el aseguramiento de la provisión de datos a medio y largo plazo.

Dada la naturaleza diversa de las fuentes *Big Data* es imposible ofrecer una descripción detallada de las circunstancias que dificultan el acceso a cada una de ellas, pero con carácter algo más genérico sí pueden describirse distintos factores.

Aspectos legales

La provisión de datos en la Estadística Pública está asegurada mediante un marco legal tanto nacional como internacional. En el ámbito nacional, el artículo 10.1 de la Ley 12/89 de la Función Estadística Pública reconoce el derecho de las autoridades estadísticas a solicitar tal provisión de datos. Además, el artículo 10.2 de este misma Ley establece la obligación de toda persona legal, física o jurídica, pública o privada, de proporcionar la información solicitada. En el ámbito europeo el artículo 285 del texto consolidado del Tratado Constitutivo de la Comunidad Europea (Comunidad Europea, 2002) otorga a los servicios estadísticos este mismo papel.

Sin embargo, al intentar acceder a alguna fuente *Big Data*, surgen cuestiones de carácter legal. ¿Respaldan las leyes el acceso institucional por parte de una oficina de estadística a este tipo de fuentes de información, en especial, tras la consideración de las características 1 y 2 propuestas en su definición revisada? En lo que sigue, nos centraremos en el ámbito nacional del Sistema Estadístico Nacional español, si bien la situación es similar en el resto del Sistema Estadístico Europeo con las debidas transposiciones en las regulaciones legales nacionales. No haremos consideraciones sobre la legislación autonómica al respecto.

Básicamente, en cualquier fuente *Big Data*, la cuestión legal se caracteriza por la confluencia de tres ámbitos legales. En primer lugar, la Ley 12/89 de la Función Estadística Pública confiere a la Administración Estadística del Estado amplia potestad para «solicitar datos de todas las personas físicas y jurídicas, nacionales y extranjeras, residentes en España», así como de «todas las instituciones y entidades públicas de la Administración del Estado, las Comunidades Autónomas y las Corporaciones Locales». En la producción estadística oficial tradicional, esta ley ampara al Instituto Nacional de Estadística y a otros servicios estadísticos del Estado para requerir datos incluso bajo riesgo de sanción administrativa en caso de omisión. Las características 1 y 2 de los *Big Data* a veces son aducidas para hacer una interpretación restrictiva de la norma. Organizaciones internacionales como la División Estadística de Naciones Unidas (United Nations Global Working Group on Big Data, 2016) o Eurostat (Eurostat, 2015) incluyen en sus estrategias para *Big Data* una revisión de los principios y recomendaciones para el acceso a estas nuevas fuentes de datos. Informes internos del *Steering Group* de la *Task Force on Big Data* del Sistema Estadístico Europeo (ESS Task Force on Big Data, 2017) concluyen que las legislaciones estadísticas nacionales y europea confieren suficiente potestad legal a las oficinas productoras para requerir el acceso a estos datos. Como veremos, otra cuestión diferente es la complejidad asociada.

En segundo lugar, puesto que la mayoría de estos datos son de titularidad privada y dentro de un sector de actividad económica concreto (salud, telecomunicaciones, …), en muchas ocasiones existe una legislación específica del sector en cuestión (sobre todo dirigida a salvaguardar los derechos del ciudadano) que impide el acceso a estos registros de datos (piénsese, por ejemplo, en los datos de telefonía móvil o en los historiales clínicos en hospitales). En el caso de los servicios estadísticos, el aparente conflicto suele resolverse mediante la intervención de la correspondiente Agencia de Protección de Datos.

En tercer lugar, no solo en España, sino en cualquier país del ámbito del Sistema Estadístico Europeo, existe una norma estricta relativa a la protección de datos de carácter personal, norma supervisada por una Agencia de Protección de Datos nacional. En el caso español, esta norma está dada por la Ley 15/99 Orgánica de Protección de Datos de Carácter Personal, supervisada por la Agencia Española de Protección de Datos (AEPD, 2017). Por la experiencia acumulada en los proyectos de investigación en marcha, parece claro que esta Agencia es, y será, fundamental en el acceso institucional a los datos velando y garantizando la privacidad de los datos de las personas. En países de nuestro entorno con acceso a algunas fuentes *Big Data* de carácter personal (no obstante, en fase de investigación, aún no en producción), el papel de esta agencia ha sido determinante para resolver puntualmente cuestiones legales de acceso garantizando la privacidad. Es pertinente, no obstante, recordar que la propia Ley 12/89 establece la obligación de los servicios estadísticos del Estado de guardar el secreto estadístico, no revelando ni compartiendo en ningún caso la identidad de las personas legales, ya sean físicas o jurídicas, que proporcionan los datos.

Características de los datos

En el sistema de producción estadística oficial tradicional, primero se deciden los agregados o índices que desean conocerse (índice de precios al consu-

mo, total de personas paradas, etc.), a continuación se diseña la operación estadística (desde la recogida hasta la difusión de resultados pasando por el procesamiento de los datos) y se ejecuta, recabando datos sobre una muestra de unidades estadísticas seleccionada mediante algún método estadístico apropiado. De hecho, el productor de estadísticas oficiales debe informar a la persona o empresa requerida que ha sido seleccionada en una muestra, que tiene la obligación de responder, que los servicios estadísticos del Estado tienen la obligación de guardar el secreto estadístico y que sus datos recabados en esta ocasión se emplearán con fines estadísticos exclusivamente para la operación estadística en cuestión.

La situación con los *Big Data* es diferente. Los datos ya han sido generados antes de saber incluso qué tipo de información estadística puede extraerse de ellos. De hecho, han sido generados para otros propósitos diferentes a la producción estadística oficial. La obligación de cederlos queda a las consideraciones legales anteriores. Finalmente, estos datos, por razones evidentes, pueden ser empleados para más de una estadística oficial (por ejemplo, los datos de telefonía móvil podrían emplearse para estadísticas de turismo, pero también para estadísticas de movilidad humana).

Además, al haber sido generados para otros fines, no están preparados para su procesamiento estadístico (carecen de metadatos asociados relativos al propósito estadístico). Por ello, la identificación de qué datos en concreto del sistema de información digital del proveedor son de interés para la producción estadística oficial no es una cuestión elemental (de hecho, una buena parte de los proyectos de investigación en marcha tienen como objetivo parcial identificar qué datos permitirían la elaboración de estadísticas oficiales).

Condiciones de acceso

Como se ha comentado, los *Big Data* presentan dos características muy relevantes: (i) no dan información sobre el proveedor de los datos, sino sobre terceros (clientes, subscriptores, usuarios,…) y (ii) son el medio (o una parte importante de él) por el que las empresas propietarias realizan su actividad económica y obtienen su beneficio. Esto conlleva evidentes dificultades más allá de las cuestiones legales.

Por un lado, la protección de su negocio corre paralela a la protección de sus datos, que, en muchos casos, son datos privados de sus clientes. La cesión de esta información, incluso a las oficinas públicas de estadísticas acostumbradas a recoger y procesar datos confidenciales, es percibida como un riesgo muy alto para su propio negocio no solo por la potencial pérdida de control sobre los datos sino por la imagen corporativa y la percepción social de desprotección de la información.

Por otro lado, el volumen de datos es considerablemente mayor que la información recogida a través de los cuestionarios (físicos o electrónicos) tradicionales. La extracción de estos datos de los sistemas de información corporativos puede requerir procesos más o menos complejos que, además de la potencial interferencia con las operaciones de negocio de la compañía, implican costes de diversa naturaleza (*hardware*, *software*, personal técnico altamente cualificado, …). La Estadística Oficial tiene como principio universal no pagar por los datos para fines estadísticos oficiales (United Nations Global Working Group on Big Data, 2016) (aunque los procedimientos de recogida suponen una buena parte del presupuesto de las oficinas de estadística). La cuestión de estos costes aún no está resuelta y, de hecho, está explícitamente mencionada en las *Recommendations for Access to Data from Private Organizations for Official Statistics* de Naciones Unidas (United Nations Global Working Group on Big Data, 2016).

Estrategia: principales líneas

Aunque el reto del acceso institucional a las fuentes *Big Data* se mantiene como uno de los escollos principales para su uso en la producción estadística oficial, las principales organizaciones estadísticas, ya sean productoras o usuarias, como Naciones Unidades, OCDE, Eurostat, así como el resto de socios del Sistema Estadístico Europeo, reconocen en las colaboraciones público-privadas (*public-private partnerships*) la mejor opción para conseguir este acceso institucional (Klein y Verhulst, 2017; Salgado *et al.*, 2017).

Aparte de los potenciales cambios legales necesarios en algunos países (no en el caso de España), estas colaboraciones tienen por objeto, no ya garantizar la provisión de datos sostenida en el tiempo y proteger jurídicamente a las compañías propietarias que los ceden para un uso no originalmente planificado (estadísticas oficiales), sino también buscar una relación simbiótica mutuamente beneficiosa basada en la elaboración conjunta de productos estadísticos con altos estándares de calidad. Todo parece apuntar a la integración parcial en una fase muy temprana del proceso de producción de las compañías generadoras y acumuladoras de datos. Esto, sin duda, supondrá un cambio de cultura y de filosofía de trabajo en las oficinas de estadística.

METODOLOGÍA ESTADÍSTICA: CONCEPTOS IMPORTANTES Y ERRORES COMUNES

El siguiente reto que los *Big Data* presentan para su uso en la producción estadística oficial es el conjunto de métodos estadísticos necesarios para su procesamiento y, en especial, para realizar las inferencias respecto de las poblaciones de interés (humanas, de empresas, de establecimientos, ...).

En todo ejercicio de estimación estadística existen ciertos conceptos esenciales que deben ser entendidos correctamente para poder evaluar la calidad de las estimaciones. Comentamos muy brevemente algunos de estos conceptos, así como errores comunes que suelen aparecer cuando se emplean los *Big Data* para inferir valores de una población de análisis.

Población, muestra e inferencia

De modo genérico, un ejercicio de estimación en la Estadística Pública consiste en proporcionar un valor numérico y su precisión para una propiedad de una población de análisis a partir de los datos recogidos de una muestra de tal población. Por ejemplo, al estudiar el número de personas paradas de la población activa se estima, en la población de personas mayores de 15 años (población de análisis), el número total de personas paradas (propiedad de la población) a partir de los datos recogidos de una muestra de esta población.

En abstracto, con suficiente generalidad práctica, si $Y = \sum_{k \in U} y_k$ denota la propiedad expresada por la variable *Y* de la población *U* a estudiar, la estimación se proporciona a través de los valores y_k de las unidades *k* de la muestra *s* mediante un estimador $\hat{Y}_s = \sum_{k \in s} \omega_{ks} y_k$, donde los coeficientes ω_{ks} se denominan pesos de muestreo y dependen de cómo se ha seleccionado la muestra *s*.

Es evidente que para cada muestra posiblemente seleccionada el valor de la estimación será diferente. Un estimador bien construido y con calidad es aquel que arroja todas estas estimaciones iguales en promedio al valor de la población que queremos estimar (matemáticamente $\mathbb{E}\hat{Y}_s = Y$). Además, las variaciones entre estas potenciales estimaciones deben ser muy pequeñas en torno al valor promedio *Y*. Gran parte del trabajo en una oficina de estadística está encaminado a cumplir estas dos propiedades.

Representatividad de la muestra

A menudo se plantea la calidad de las estimaciones como una cuestión de representatividad de la muestra respecto a la población de análisis. No obstante, debemos aclarar que el concepto de representatividad de la muestra es peligrosamente escurridizo (Kruskal and Mosteller, 1979a, 1979b, 1979c, 1980). Una definición inequívoca (que no hace referencia al mecanismo de respuesta) es la que afirma que una muestra *s* es representativa de una población *U* respecto de una variable *Y* si las distribuciones de valores de esta variable en la muestra y en la población son muy similares (Bethlehem, 2009). Si bien esta definición matemática es rigurosa, sin embargo, es de poca utilidad para los efectos habituales en la Estadística Pública, al menos por dos motivos: (i) la distribución de los valores en la población de las variables de interés nunca es conocida y (ii) aun teniendo una muestra representativa para un conjunto de variables $Y_1, \ldots, Y_p$, al estudiar una nueva variable puede suceder que la muestra deje de ser representativa, al menos, para esta nueva variable. Existen variantes del concepto de representatividad que involucran el mecanismo de respuesta, pero esto sólo ilustra la complejidad del concepto (véase p.ej. Schouten, Cobben, y Bethlehem (2009)).

El concepto de representatividad, no obstante, recoge cualitativamente cierta medida de calidad del proceso de selección de la muestra, que se realiza sobre un listado completo de todas las unidades de la población, ya sea de personas, empresas, establecimientos, etc. Estos listados son generados y mantenidos por las oficinas de estadística pública. La selección se realiza de modo probabilístico de acuerdo con una metodología internacional establecida desde los años 1930-40, asegurando que cada unidad de la población tiene una probabilidad no nula de ser aleatoriamente escogida. Esto asegura que las dos propiedades de construcción de estimadores mencionadas anteriormente puedan ser satisfechas.

Por todo ello, desaconsejamos considerar el concepto de representatividad de la muestra, sino perseguir las propiedades anteriores necesarias de los estimadores para obtener estimaciones de calidad.

Tamaño de la muestra

Con los *Big Data* es tremendamente común escuchar el argumento del tamaño muestral: teniendo tal cantidad de datos, las estimaciones son (casi automáticamente) más precisas. A veces, el argumento va más allá descalificando la Estadística como herramienta de análisis pregonando el uso exclusivo de técnicas de inteligencia artificial y aprendizaje automático.

Ya en 1949, Yates (1949) advertía que una encuesta por muestreo probabilístico es más precisa que un censo (1). La justificación de esta afirmación proviene de los llamados *errores de medida*. Al determinar los valores y_k de la variable de interés el proceso de determinación es necesariamente imperfecto y, por tanto, aparecen errores $y_k = y_k^{(0)} + \epsilon_k$, donde $y_k^{(0)}$ denota el valor verdadero que quiere determinarse. Si estos errores no se corrigen, el estimador en realidad sería de la forma $\hat{Y}_s = \sum_{k \in s} \omega_k y_k^{(0)} + \sum_{k \in s} \omega_k \epsilon_k$, que implicaría que, en promedio, las estimaciones no son iguales a la cantidad investigada. Por tanto, no se cumpliría la primera propiedad requerida más arriba para alcanzar estimaciones de calidad. Si el error de medida no es completamente aleatorio, cuanto mayor sea el tamaño de la muestra, mayor será el error en la estimación.

En las oficinas de estadística se desarrolla toda una fase del proceso de producción encaminada a detectar estos errores y corregirlos. Es evidente que, dados los costes de producción, es mejor corregir los errores en una muestra (p. ej. de 10.000 empresas) que de un censo (p. ej. de 1.000.000 empresas). A menudo se aduce que los datos digitales no contienen errores (la V de veracidad), pero, si bien esto debe comprobarse (2), en la práctica la adecuación de las variables digitales a las definiciones de variables estadísticas como consecuencia de su generación sin un sistema de metadatos estadísticos subyacente surge como una nueva fuente de errores de medida. Por tanto, nuevamente cuanto mayor sea el tamaño de la muestra, en realidad, mayor riesgo de cometer este error de estimación surge.

Sin embargo, efectivamente cuanto mayor es el tamaño muestral, menor es la variabilidad de las estimaciones, pero si el error anterior existe, entonces esta-

ríamos proporcionando estimaciones con un notable error de medida (por el efecto aditivo) pero con poca variabilidad entre todas las posibles estimaciones, esto es, una estimación errónea con casi toda seguridad.

Muestreo probabilístico, modelos estadísticos y aprendizaje automático

¿Quiere decir esto que deban mantenerse a toda costa los métodos tradicionales de inferencia en la Estadística Oficial con las nuevas fuentes de datos? Esto es sencillamente imposible.

En primer lugar, obsérvese que los datos digitales proporcionados por estas fuentes no provienen de una selección de una muestra probabilística diseñada por el estadístico oficial, sino que son generados por mecanismos aleatorios desconocidos. Por tanto, todo el esquema de inferencia anterior basado en muestras probabilísticas es inútil.

En segundo lugar, en la actualidad el acceso a estos datos digitales se plantea de modo completamente anonimizado, siendo la identificabilidad una de las características fundamentales de tales métodos tradicionales (Cassel *et al.*, 1977). Por tanto, nuevamente la inferencia al uso no puede aplicarse.

No obstante, en algunas circunstancias (p. ej. en la estimación en pequeños dominios), la Estadística Pública dispone de métodos alternativos basados en la modelización de los valores de las variables de la población, permitiendo superar las circunstancias anteriores. Estos métodos sólo se emplean por algunas oficinas de estadística y en determinadas circunstancias. ¿Por qué no aplicarlos a las nuevas fuentes *Big Data*?

El uso de modelos estadísticos será, sin duda, necesario, aunque aún está por llegar una investigación exhaustiva al respecto para cada fuente *Big Data*. Pero ya son conocidas las implicaciones que tiene este cambio de paradigma (Smith, 1976, 1994). El muestreo probabilístico fue escogido como la solución estándar al problema de la inferencia en la Estadística Pública porque libera al estadístico oficial de realizar hipótesis *a priori* sobre los valores de las variables. En los modelos estadísticos, esto, sin embargo, no es así, corriendo el grave riesgo de producir malas estimaciones si las hipótesis no son válidas. Un ejemplo muy conocido con una fuente *Big Data* se encuentra en la estimación de la prevalencia de la gripe en EE.UU. realizada por Google a partir exclusivamente de las búsquedas de términos relacionados en su buscador de Internet (Butler, 2013). Son las llamadas *Google Flu Trends*. Las estimaciones se realizaron mediante técnicas de modelización estadística y, posteriormente, se compararon con las cifras oficiales recogidas de los hospitales (Olson *et al.*, 2013). Cuando las hipótesis a priori dejaron de ser válidas, las estimaciones resultantes perdieron su calidad. La Estadística Pública ha primado siempre el principio de independencia de su trabajo (de hecho, es el primer principio del Código de Buenas Prácticas europeo (European Statistical System, 2011)).

Las técnicas analíticas asociadas a los *Big Data*, sin embargo, incluso parecen proporcionar cierta solución para este problema. Las técnicas de aprendizaje automático permiten construir modelos estadísticos y actualizarlos conforme a los valores de las variables que se vayan procesando. Aquí la cuestión se torna muy sutil, recordando el debate entre racionalismo y empiricismo (Starmans, 2016).

Ilustrémoslo con un ejemplo muy conocido en la filosofía de la ciencia (Kuhn, 1957): el cambio de paradigma científico entre la cosmología de Ptolomeo y la de Copérnico. En la Antigüedad, el sistema ptolemaico permitía calcular el movimiento de cualquier objeto celeste a través de la introducción de más elementos de cómputos (más epiciclos, deferentes, ...). Se disponía de todo un sistema de computación que tan sólo necesita un *input* adecuado para producir los resultados precisos. La llegada del nuevo sistema copernicano introducía un elemento explicativo (la ley de la gravitación de Newton) que permitía reproducir estos resultados, aunque comprendiendo la naturaleza de los fenómenos.

Salvando ciertas distancias, las técnicas de aprendizaje automático son similares al sistema ptolemaico, pues proporcionan un método de cómputo que se adecúa a los datos de entrada produciendo bien uno u otro resultado, sin mayor explicación de los fenómenos subyacentes que relacionan las variables involucradas. Los modelos estadísticos, sin embargo, al igual que el sistema copernicano, arrojan cierta luz sobre la naturaleza de los fenómenos y de las variables involucradas, produciendo asimismo los resultados adecuados siempre y cuando las hipótesis de partida (la explicación del fenómeno) sean válidas.

En los ejercicios de análisis que utilizan *Big Data*, en especial en la investigación sociológica y económica, parece natural afirmar que las técnicas de aprendizaje automático no son completamente satisfactorias. Sin embargo, la Estadística Pública persigue medir ciertas cantidades independientemente de su explicación o de su naturaleza. Por tanto, ¿cabe el uso del aprendizaje automático siempre y cuando las estimaciones sean de calidad? Esto forma parte de la necesidad de analizar el uso de *Big Data* para la producción estadística oficial.

TECNOLOGÍA: EL ENTORNO DE COMPUTACIÓN

Sin duda, el aspecto más visible de los *Big Data* es el aspecto tecnológico. Las oficinas productoras de estadística poseen unidades especializadas en el tratamiento y procesamiento informático de los datos, pero los *Big Data* requieren una actualización de esta infraestructura. Básicamente el origen de estas nuevas necesidades del marco tecnológico proviene del hecho que los *Big Data* no caben en la memoria de una única computadora. El almacenamiento y procesamiento debe realizarse necesariamente en un *cluster*.

La nueva infraestructura informática estará fuertemente condicionada por dos factores. Por un lado, dependerá de las necesidades concretas de almacenamiento

y procesamiento de datos, lo que, en última instancia, dependerá del modelo de negocio acordado con los proveedores de datos. Por ejemplo, si en la colaboración con las corporaciones privadas generadoras y poseedoras de los datos se acuerda realizar un pre-procesamiento y agregación *in-situ* dentro de los sistemas de información de las compañías para disminuir la exposición de los datos al exterior, es posible que el volumen de información dentro de las oficinas de estadísticas no sea crítico. De igual modo, la periodicidad con la que se acuerde acceder a estos datos condicionará el tipo de herramientas ya sean para procesar en tiempo real o en diferido. Del mismo modo, si los datos se estructuran antes de llegar a los sistemas informáticos de las oficinas de estadística, las necesidades serán obviamente diferentes. Todo ello además deberá ser analizado fuente a fuente y proveedor a proveedor.

Por otro lado, la infraestructura informática dependerá también de la estrategia de modernización e industrialización del proceso de producción estadística que lleva en marcha varios años en la comunidad internacional (UNECE, 2017). Tanto la arquitectura del *hardware* como las herramientas de *software* deberán ser revisadas dentro de esta estrategia de industrialización.

Es prácticamente imposible concretar qué tipo de arquitectura y qué herramientas son idóneas sin conocer los detalles de las necesidades de almacenamiento y procesamiento. En cualquier caso, sí existen ya opciones que permiten una industrialización a gran escala del entorno de computación de una oficina estadística. El empleo de *clusters* para la computación en paralelo independientemente del uso de fuentes *Big Data* constituye una modernización importante de los procesos de cómputo (piénsese por ejemplo en el procesamiento de datos censales). Esto conlleva retos, no solo desde el punto de vista de la arquitectura del *hardware,* sino de la programación de algoritmos paralelizados. Por supuesto, no cabe otra opción que estas infraestructuras sean centralizadas, dando servicio a todas las operaciones estadísticas producidas en una oficina, de tal modo que el almacenamiento y procesamiento se realicen siempre de modo centralizado. Herramientas como la virtualización (Wikipedia, 2017a) pueden permitir dar soluciones adecuadas a cada circunstancia (por ejemplo, habrá operaciones que por su volumen de datos no requieran el uso de un *cluster* ni de paralelismo alguno). Todo ello conduce de algún modo al concepto de *cloud computing* (Wikipedia, 2017b), nube que, en el caso de la producción estadística oficial, debe ser de acceso restringido y con las mismas salvaguardas de seguridad informática que en los sistemas tradicionales.

Si la producción de estadísticas se asemeja a la escritura de una novela, el cambio necesario en la infraestructura equivale a pasar de la caligrafía manual al procesamiento de textos con un ordenador. Esta modernización debe ser además acometida manteniendo la producción estadística cotidiana comprometida en los calendarios de difusión de las operaciones estadísticas.

LA CALIDAD DE LA PRODUCCIÓN ESTADÍSTICA OFICIAL

La calidad de las estadísticas oficiales ha sido siempre un elemento central de la producción estadística oficial. Ya llevó más de 30 años (desde finales del siglo XIX hasta finales de la década de los 30) convencerse de que las técnicas de muestreo probabilístico permiten formalmente obtener resultados más precisos que un censo (3). El control de la acuracidad, esto es, tanto del sesgo como de la varianza, fueron el objetivo central del diseño de todas las operaciones estadísticas. La prioridad era controlar el llamado *total survey error*, esto es, el conjunto de errores tanto de muestreo como ajenos al muestreo, dando lugar al *total survey error paradigm* (Biemer, 2010).

Posteriormente, en la década de los 80, al igual que en muchas otras industrias, el concepto de calidad sufrió una revolución. En la Estadística Oficial surgieron las dimensiones de la calidad. Ya no sólo debe garantizarse la acuracidad, sino también la relevancia, la puntualidad (4) y oportunidad, la coherencia y comparabilidad y otros factores. El paradigma *total survey error* quedó así integrado en el *total survey quality framework* (Biemer, 2010).

Finalmente, en el ámbito del Sistema Estadístico Europeo desde hace unos años el llamado Código de Buenas Prácticas (CBP) de las Estadística Europeas (European Statistical System, 2011) constituye la base para la garantía de la calidad en el proceso de producción. El CBP está formado por 15 principios divididos en 3 bloques, que son (i) entorno institucional (principios 1 a 6), (ii) procesos estadísticos (principios 7 a 10) y (iii) producción estadística (principios 11 a 15). Cada uno de estos principios da origen a una serie de indicadores que calibran diversos aspectos relacionados con cada principio.

Así, por ejemplo, el principio 8 establece que «las estadísticas de calidad se apoyan en procedimientos estadísticos adecuados, aplicados desde la recogida de los datos hasta la validación de los mismos». Este principio posee nueve indicadores asociados, los tres últimos son:

> 8.7: Las autoridades estadísticas participan en el diseño de los datos administrativos para adecuarlos en mayor medida a los fines estadísticos.
>
> 8.8: Se establecen acuerdos con los propietarios de los datos administrativos, en los que se establece el compromiso común para el uso de dichos datos con fines estadísticos.
>
> 8.9: Las autoridades estadísticas colaboran con los propietarios de los datos administrativos para garantizar la calidad de los datos.

Estos indicadores hacen referencia explícita a una de las fuentes de datos cada vez más importante: los registros administrativos. Se observa cómo los principios persiguen calibrar los diversos detalles relacionados con cada aspecto de la producción.

En esta misma línea, la implementación sistemática del CBP tiene su principal desarrollo en el Marco de Garantía de la Calidad del Sistema Estadístico Europeo (European Statistical System, 2012). Se trata de un documento que identifica diversas actividades, métodos y herramientas para esta implementación. La primera versión, de agosto de 2011, cubría los principios 4 y 7 a 15. Tras la actualización del CBP en septiembre de 2011, se revisó y adoptó consiguientemente el marco de calidad vigente, que fue aprobado por el Grupo de Trabajo de Calidad de las Estadísticas en noviembre de 2012. Esta versión no contiene consideraciones respecto al uso de *Big Data* en la producción de estadísticas oficiales.

El reto es evidente. ¿Cómo deberán cambiar tanto el CBP como el marco de garantía para asegurar la calidad en la producción con *Big Data*? Previsiblemente muchos de los principios se verán afectados por el uso de *Big Data* debido a los retos expuestos anteriormente, esto es, al acceso institucional, al uso de nueva metodología estadística y a la necesidad de nuevas tecnologías informáticas. Como ejemplo, los indicadores 8.7, 8.8 y 8.9 expuestos anteriormente necesitarán claramente ser completados con sus equivalentes para las fuentes *Big Data* -- si así finalmente se concluye que es posible y conveniente o necesario.

Obsérvese cómo paralelamente a la adaptación y revisión del CBP y del marco de calidad es preciso resolver los tres retos expuestos anteriormente y analizar detalladamente la interrelación existente entre ellos.

CONCLUSIONES

La Estadística Pública se encuentra en un proceso de modernización e industrialización en el que la incorporación de los *Big Data* al proceso de producción aparece como uno de los elementos más visibles.

Antes de su uso generalizado para la producción de estadísticas oficiales, existen retos notables que deben superarse para garantizar la calidad de las estadísticas que se nutren de estas fuentes de datos. Los retos principales son el acceso institucional a los datos, la nueva metodología estadística para el tratamiento de estos datos y los cambios en la infraestructura tecnológica. Todo ello conlleva la necesaria renovación del perfil profesional del estadístico oficial que va desde la formación académica universitaria integrando disciplinas como la Estadística y la Informática hasta planes de formación continua en las oficinas de estadística.

Si la modernización y la industrialización son ya una necesidad ante el agotamiento del modelo de producción tradicional, la inclusión de los *Big Data* en el proceso de producción es una necesidad ante el riesgo de que las oficinas de estadística pierdan relevancia en la producción de estadísticas. Esto supone un cambio de cultura en la Estadística Pública que requiere un refuerzo de su colaboración con los sectores privado y académico y, seguramente, el diseño coordinado de una estrategia de acceso y reaprovechamiento de estos datos por partes de las Administraciones Públicas.

NOTAS

[1] Esta afirmación debe ser matizada en relación con funciones de coste.

[2] En la investigación con datos de identificación automática de navíos (datos AIS), nuestros colegas han podido comprobar cómo varios cientos de navíos en el Mar Mediterráneo son detectados por el sistema en el Desierto del Sáhara (Consten, 2017).

[3] Como hemos apuntado, esta afirmación se encuentra ya en el libro de 1949 de uno de los padres del muestreo probabilístico, F. Yates, *Sampling methods for censuses and surveys*, y se justifica por las dificultades que surgen de los llamados errores ajenos al muestreo (Lessler y Kalsbeek, 1992), esto es, errores no relacionados con el procedimiento de selección de la muestra. Véase la sección 4.3.

[4] *Timeliness.*

BIBLIOGRAFÍA

AEPD (2017). *Agencia Española de Protección de Datos*. https://www.agpd.es.

BETHLEHEM, J. (2009). *Applied Survey Methods - A Statistical Perspective.* Amsterdam: Wiley.

BIEMER, P. (2010). «Total survey error: design, implementation, and evaluation». *Public Opinion Quaterly, 74*, 817-848.

BUTLER, D. (2013). «When Google got flu wrong». *Nature, 494*, 155-156.

CASSEL, C.-M., SÄRNDAL, C., and WRETMAN, J. (1977). *Foundation of inference in survey sampling.* New York: Wiley.

COMUNIDAD EUROPEA (2002). Versión consolidada del Tratado Constitutivo de la Comunidad Europea. *Diario Oficial de las Comunidades Europeas C325/33-184*.

CONSTEN, A. (2017). ESSnet Pilot on AIS Data. Dissemination Workshop of the ESSnet on Big Data. Sofia, 23-24 February, 2017. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/77/Presentation_ WP4_20170223-24_Sofia_AIS_data.pdf.

ESS TASK FORCE ON BIG DATA (2017). *Results from the analysis project on legal issues related to the use of Big Data.* Doc.DDG.TF.BD 2017 04 27-28-4-legal issues. *Internal document.*

EUROPEAN COMMISSION (2017). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the Mid-Term Review on the implementation of the Digital Market Strategy. A connected Digital Single Market for All. *COM(2017) 228 final*.

EUROPEAN STATISTICAL SYSTEM (2011). European Statistics Code of Practice. http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15.

EUROPEAN STATISTICAL SYSTEM (2012). ESS Quality Assurance Framework v1.2. http://ec.europa.eu/eurostat/ documents/64157/4392716/qaf_2012-en.pdf/8bcff303-68da-43d9-aa7d-325a5bf7fb42.

EUROSTAT (2015). *Big Data*. https://ec.europa.eu/eurostat/cros/content/big-data_en.

KLEIN, T. and VERHULST, S. (2017). «Access to new data sources for statistics: business models and incentives for the corporate sector». *OECD Statistics Working Papers* (Working Paper No. 82), 1-38.

KRUSKAL, W. and MOSTELLER, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review, 47*, 13-24.

KRUSKAL, W. and MOSTELLER, F. (1979b). «Representative sampling, II: scientific literature, excluding statistics». *International Statistical Review, 47*, 111-127.

KRUSKAL, W. and MOSTELLER, F. (1979c). «Representative sampling, III: the current statistical literature». *International Statistical Review, 47*, 245-265.

KRUSKAL, W. and MOSTELLER, F. (1980). Representative sampling, IV: the history of the concept in Statistics, 1895-1939». *International Statistical Review, 48*, 169-195.

KUHN, T. (1957). *The Copernican revolution.* Boston: Harvard University Press.

LANEY, D. (2001). 3D Data management: controlling data volume, velocity y variety. META Group. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-y-Variety.pdf

LESSLER, J. and KALSBEEK, W. (1992). *Nonsampling error in surveys.* New York: Wiley.

NORMANDEAU, K. (2013). Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. *insideBigData*, 12 September, 2013. http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.

OLSON, D., KONTY, K., PALADINI, M., VIBOUD, C., and SIMONSEN, L. (2013). «Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales». *PLOS Computational Biology, 9*, 156-170.

ONU (2014). *Principios Fundamentales de las Estadísticas Oficiales de Naciones Unidas.* New York: Naciones Unidas. https://unstats.un.org/unsd/dnss/gp/FP-New-S.pdf.

SALGADO, D., ALEXYRU, C., DEBUSSCHERE, M., DUPONT, F., PIELA, P., and RADINI, R. (2016). *Current status of access to mobile phone data in the ESS.* ESSnet on Big Data WP5 Deliverable 1.1. https://webgate.ec.europa.eu/fpfis/ mwikis/essnetbigdata/images/6/65/WP5_Deliverable_1.1.pdf.

SALGADO, D., ALEXYRU, C., OANCEA, B., DEBUSSCHERE, M., DUPONT, F., PIELA, P., WILLIAMS, S. (2017). *Guidelines for the access to mobile phone data within the ESS.* ESSnet on Big Data WP5 Deliverable 1.2. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/ 6/65/WP5.Deliverable_1.2.pdf

SCHOUTEN, B., COBBEN, F., and BETHLEHEM, J. (2009). «Indicators for the representativeness of survey response». *Survey methodology, 35*(1), 101-113.

SMITH, T. (1976). «The foundations of survey sampling: a review». *Journal of the Royal Statistical Society A, 139*, 183-204.

SMITH, T. (1994). «Sample surveys 1975-1990: An age of reconciliation?» *International Statistical Review, 62*, 5-19.

STARMANS, R. (2016). «The advent of data science: some considerations on the unreasonable effectiveness of data». En P. Bühlmann, P. Drineas, M. Kane, y M. van der Laan, *Hybook of Big Data* (págs. 3-20). Amsterdam: Chapman y Hall/CRC Press.

THE ECONOMIST. (2010). The data deluge. *http://www.economist.com/node/15579717*.

UNECE (2011). *Modernization of official statistics*. https://www.unece.org/stats/mos.html.

UNITED NATIONS GLOBAL WORKING GROUP ON BIG DATA (2016). *Recommendations for access to data from private organizations for Official Statistics.* Dublin: United Nations.

WIKIPEDIA (2017A). *Virtualization*. https://en.wikipedia.org/wiki/Virtualization.

WIKIPEDIA (2017B). *Cloud Computing.* https://en.wikipedia.org/wiki/Cloud_computing.

YATES, F. (1949). *Sampling methods for censuses and surveys.* London: Charles Griffins.