

Big Data y la Estadística Oficial: retos

David Salgado

Dpto. Metodología y Desarrollo de la Producción. INE

LA HUELLA DIGITAL

La expresión *Big Data* lleva de moda cierto tiempo induciendo con este adjetivo a concentrarse en el volumen de información al que alude descuidando quizá su velocidad de generación y la variedad de su estructura¹.

El primero de los Principios Fundamentales de las Estadísticas Oficiales de Naciones Unidas² establece que “las estadísticas oficiales constituyen un elemento indispensable en el sistema de información de una sociedad democrática y proporcionan al gobierno, a la economía y al público datos acerca de la situación económica, demográfica, social y ambiental”.

En este sentido, el verdadero potencial de los *Big Data* para la Estadística Oficial, en nuestra opinión, reside en que constituyen una huella digital de la actividad humana. Esto es, un creciente número de actividades humanas dejan su rastro en sistemas de información digitales que, a priori, pueden ser empleados para generar información y conocimiento, entre otros ámbitos, a través de la producción estadística oficial.

Se trata de la huella digital, que no requiere de modo obligado de grandes volúmenes de datos. Existen métodos de recogida de datos actualmente que se basan en esta idea. La recogida automática de datos (*automatic data collection*) es un método que reduce la carga del informante extrayendo la información de modo automático de las bases

de datos del informante. En el INE, por ejemplo, en la Encuesta de Ocupación Hotelera los establecimientos hoteleros pueden optar por la instalación y configuración de una aplicación informática que se conecta a la base de datos de la recepción del establecimiento y genera un fichero XML que se sube a la web del INE, bien automáticamente, bien manualmente. La carga al informante es mínima. Al mismo tiempo que el detalle de la información proporcionada es muy alto³.

En el caso de las nuevas fuentes de información *Big Data*, el proceso se complica pues se trata habitualmente de volúmenes de datos enormes, generados a una gran velocidad y con una diversidad en su estructura muy alta. Ninguno de estos factores está presente en el mencionado método de recogida, pero la idea es la misma: una actividad humana (alojarse en un establecimiento hotelero) queda registrada en un sistema digital.

Desde hace unos años diversas organizaciones internacionales de nuestro entorno están esforzándose por incluir estas nuevas fuentes de información *Big Data* en el proceso de producción de las estadísticas públicas. Muy pocos casos han alcanzado el éxito.

El proceso de producción de estadísticas oficiales es un proceso de notable complejidad, pues involucra un gran número de componentes interconectados de modo múltiple e irregular entre sí y con la participación de diferentes unidades de producción. Actualmente las oficinas de estadística están esforzándose por analizar y adaptar estos procesos de producción a las nuevas fuentes de información.

A continuación, ofrecemos una visión de conjunto de los retos que deben superarse para alcanzar el objetivo de integrar los *Big Data* en la producción cotidiana de una oficina de estadística.

De modo genérico, podemos reconocer cinco grandes ámbitos de la producción donde se observan retos a superar para integrar los *Big Data* en los procesos cotidianos. Éstos son: el acceso institucional a los datos, la metodología estadística, el marco tecnológico, el marco de calidad y otros retos de carácter más transversal.

Un creciente número de actividades humanas dejan su rastro en sistemas de información digitales, que pueden ser empleados para generar información y conocimiento a través de la producción estadística oficial

ACCESO INSTITUCIONAL

Un requisito implícito para la producción de estadísticas oficiales es el acceso a los datos de modo sostenido en el tiempo. Tradicionalmente, este acceso se ha conseguido gracias a la legislación estadística, en particular a través de las leyes de la estadística pública tanto a nivel nacional como europeo, que reconocen el derecho de las oficinas de estadística a disponer de los datos y la obligación de los ciudadanos (ya sean personas físicas o jurídicas) a proporcionar tales datos en tiempo y forma.

La mayoría de los *Big Data* son generados, almacenados y procesados por corporaciones privadas como resultado de su actividad de negocio, siendo datos (i) con información sobre terceras personas (usualmente sus clientes), (ii) no preparados para la explotación estadística sin un procesamiento previo, y (iii) con un gran volumen y velocidad de generación.

El acceso institucional de modo sostenido presenta un reto desde diversos ángulos. En primer lugar, confluyen la legislación del sector de actividad de las corporaciones (que regula el uso de la información generada por la actividad económica), la legislación estadística (que otorga a la administración estadística pública derechos sobre los datos) y la legislación de protección de datos de

carácter personal (que de modo genérico protege la información sensible de los ciudadanos). El marco legal con las nuevas fuentes es indudablemente más complejo.

En segundo lugar, la estructura de estas corporaciones productoras de *Big Data* no es homogénea y es probable que, mientras algunas dispongan de un sistema de información adecuado para compartir esta información, otras puedan necesitar inversiones para ello. En algunas actividades económicas, el núcleo del negocio no es el tratamiento de estos datos, sino que estos se generan como consecuencia de otra actividad principal. Esto complica el igual tratamiento que todas deben recibir frente a la Administración Pública si todas ellas deben proporcionar tales datos.

En tercer lugar, la operatividad del proceso de acceso a los datos cuando el volumen, la velocidad de generación y la variedad de su estructura son tan altos se complica. Indudablemente no es lo mismo administrar una entrevista a través de un cuestionario en papel, electrónico o telefónico que acceder a volúmenes de información del orden de los terabytes. Si, además, estos datos han de ser combinados con datos oficiales para mejorar los procesos de estimación, la situación se vuelve aún más compleja.

Por último, en algunos casos es preciso disponer de una infraestructura informática para



Es imperante que el perfil profesional del estadístico oficial integre de manera unificada formación estadística e informática para satisfacer las demandas que los nuevos procesos de producción requerirán

extraer la información y preprocesarla para su explotación estadística posterior, así como personal cualificado para llevarlo a cabo. Las recientes Recomendaciones para el Acceso a Datos de Organizaciones Privadas para las Estadísticas Oficiales [4] (actualmente en elaboración), reconociendo el derecho de las oficinas de estadística a los datos de manera gratuita, observan de modo escrupuloso los costes y el esfuerzo requeridos por parte de las corporaciones privadas promoviendo el equilibrio entre ambas partes.

METODOLOGÍA ESTADÍSTICA

Existen tres aspectos fundamentales que la metodología estadística oficial debe resolver para el empleo de los *Big Data* en la producción.

En primer lugar, el estadístico oficial no tiene control sobre las variables que se recogen, como sucede cuando diseña un cuestionario con una estructura muy concreta. La relación entre las variables recogidas y las variables de interés estadístico necesitan un análisis sistemático y profundo, como sucede con la explotación estadística de los registros administrativos. La diferencia estriba ahora en la complejidad computacional añadida.

En segundo lugar, la Estadística Pública tradicional basa su metodología estadística en la inferencia basada en los diseños muestrales. Esto es, a partir de diversos registros administrativos se generan marcos de población (conjuntos de todas las unidades estadísticas del país, como personas, hogares, empresas, establecimientos,...). Sobre ellas se selecciona una muestra de acuerdo con un diseño muestral probabilístico. Se recogen los datos de las unidades estadísticas seleccionadas y se emplean para construir estimaciones a partir de tales diseños muestrales y estimadores insesgados cuya varianza se controla con información auxiliar⁵.

La ventaja del uso de los diseños muestrales re-

side en que la estimación de la cantidad de interés se realiza sin necesidad de efectuar ninguna hipótesis *a priori* sobre la distribución de los valores de las variables en la población de interés⁶. Está libre de criterios más o menos subjetivos del productor.

Este esquema no parece poder aplicarse con los *Big Data*. Por tanto, la cuestión de la inferencia (¿basada ahora en modelos estadísticos?, ¿con estadística bayesiana?) debe ser igualmente estudiada.

Por último, en relación directa con los puntos anteriores, los diseños muestrales permiten dar una medida de la acuracidad de los estimadores empleados para generar las estimaciones también libre de hipótesis *a priori* sobre la distribución de los valores en la población. Si la metodología cambia, en cualquiera de los casos debe permitir calcular estas medidas de acuracidad. Esto está íntimamente relacionado con el apartado siguiente sobre la calidad.

MARCO TECNOLÓGICO

Al hablar de *Big Data* es inevitable mencionar las tecnologías involucradas para su almacenamiento y procesamiento. El marco tecnológico necesario para tratar con este tipo de datos es claramente más complejo que los sistemas de información habituales de una oficina estadística.

Esto supone que la Estadística Oficial, en caso de que el modelo de negocio pase por procesar datos en las oficinas, necesita invertir en la infraestructura necesaria para almacenar y procesar estos volúmenes de datos.

En cualquiera de los casos, sin embargo, el estadístico oficial deberá tener las capacidades necesarias para saber implementar los procesos, ya sean nuevos o más tradicionales, en entornos de computación distribuida. El estadístico oficial deberá afrontar el diseño y ejecución de algoritmos en clústeres de ordenadores.

CALIDAD

El cuarto gran reto se centra en el marco de garantía de la calidad (*Quality Assurance Framework*), en el caso de las estadísticas europeas, en el ámbito del Sistema Estadístico Europeo (SEE)⁷.

Desde hace aproximadamente una década, los productores de estadísticas oficiales europeos se rigen por el llamado Código de Buenas Prácticas⁸, que no es sino un conjunto de 15 principios que rigen la producción en el SEE. De este código se deriva el marco de garantía de la calidad del SEE.

La cuestión clave ahora es cerciorarse de si el marco de calidad sigue siendo válido con las nuevas fuentes de información *Big Data*, en especial, debido al empleo de la nueva metodología, no solo en cuanto a la acuracidad, sino también en cuanto a otras dimensiones de la calidad como la puntualidad, la oportunidad, la comparabilidad, la carga de respuesta,...

OTROS RETOS

Con carácter más transversal, existen otros retos que merece la pena mencionar. Las oficinas de estadística están involucradas en un proceso de industrialización y modernización de su producción sobre la base de la estandarización internacional de métodos y procesos⁹. Nos parece evidente la urgencia de acelerar este proceso ante la creciente capacidad de generación de información a partir de estas nuevas fuentes *Big Data*. Si los productores de estadísticas oficiales no emplean estas fuentes ni tampoco industrializan su producción, existe un riesgo real de volverse irrelevantes ante la proliferación de estadísticas generadas por otras organizaciones que pueden utilizar estas fuentes a mayor velocidad.

En la misma línea, es imperante que el perfil profesional del estadístico oficial integre de manera unificada formación estadística e informática para satisfacer las demandas que los nuevos procesos de producción requerirán. Como ejemplo, los cursos del *European Training Statistical Program*¹⁰ ponen de relieve esta tendencia.

En definitiva, la Estadística Oficial debe superar un número de retos para incorporar las nuevas fuentes de información *Big Data* a su producción. Estos esfuerzos se enmarcan en el proceso de modernización e industrialización de las oficinas públicas de estadística. En este contexto, Eurostat, en común con sus socios del Sistema Estadístico Europeo (SEE), elaboró una respuesta estratégica para hacer frente a estos retos de modernización: la Visión 2020¹¹. Esta visión estratégica, adoptada por el Comité del SEE en mayo de 2014, identifica cinco áreas clave de actuación (usuarios, calidad, nuevas fuentes de datos, eficiencia del proceso de producción y difusión).

Para su implementación, se ha puesto en marcha un conjunto de proyectos europeos, entre los cuales se encuentra uno dedicado a los *Big Data* para la Estadística Oficial, centrado en la realización de proyectos piloto con fuentes *Big Data* concretas¹². El INE, además de en otros proyectos de la Visión 2020, participa en este como coordinador del paquete de trabajo sobre datos de telefonía móvil.

Referencias

1. D. Laney (2001). 3D data management: controlling data volume, velocity and variety. Gartner. Accesible en: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3-D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
2. Naciones Unidas (2014). Principios Fundamentales de las Estadísticas Oficiales. Resolución A/RES/68/261 del 29 de enero de 2014. Accesible en: <http://unstats.un.org/unsd/dnss/gp/FP-New-S.pdf>
3. E. Rosa-Pérez (2016). Improving the statistical process in the Hotel Occupancy Survey. European Conference on Quality in Official Statistics (Q2016). Madrid, 31 May- 3 June 2016. Accesible en: <http://www.ine.es/q2016/docs/q2016Final00112.pdf>
4. UN Global Working Group on Big Data for Official Statistics (2016). Draft Recommendations for Access to Data from Private Organizations for Official Statistics: The Way Forward. 3rd UN International Conference on Big Data for Official Statistics. 30 Aug- 1 Sept 2016. Accesible en: [http://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%20%20\(i\)%20b%20-%20Note%20on%20Draft%20Recommendations%20for%20Access%20for%20GWG%20meeting%20in%20Dublin.pdf](http://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%20%20(i)%20b%20-%20Note%20on%20Draft%20Recommendations%20for%20Access%20for%20GWG%20meeting%20in%20Dublin.pdf)
5. C.-E. Särndal and B. Swensson and J.H. Wretman (1992). Model assisted survey sampling Springer, New York.
6. T.M.F. Smith (1976). The foundations of survey sampling: a review. J. R. Stat. Soc. A 139, 183-204.
7. European Statistical System. Quality Assurance Framework of the European Statistical System, v1.2. Accesible en: <http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646>
8. European Statistical System (2011). Código de Buenas Prácticas de las Estadísticas Europeas. Accesible en: <http://ec.europa.eu/eurostat/documents/3859598/5922097/10425-ES-ES.PDF>
9. High-Level Group for the Modernization of Official Statistics. UNECE. Accesible en: <http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics>
10. Eurostat (2016). European Statistical Training Programme. Accesible en: <http://ec.europa.eu/eurostat/web/european-statistical-system/training-programme-estp>
11. Eurostat (2016). ESS Vision 2020: building the future of European Statistics. Accesible en: <http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020>
12. ESSnet on Big Data (2016). Accesible en: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data