Big data y estadística oficial

MIGUEL ÁNGEL MARTÍNEZ VIDAL
Instituto Nacional de Estadística

La forma más popular de definir lo que se entiende por big data sigue más o menos la definición dada por Gartner, que lo identifica con grandes volúmenes de datos a los que se asocian características como variedad y velocidad y se pueden añadir otras "v" como veracidad, volatilidad, valor o visualización.

Pero quizá la parte más relevante para identificar que un conjunto de datos es big data esté en que su análisis requiera de formas innovadoras para tratar la información que sean mucho más eficientes que las clásicas. Por ejemplo, el uso de herramientas para distribuir los procesos entre múltiples servidores o aplicar técnicas de análisis de datos para identificar patrones aparentemente ocultos en los datos, por citar un caso de la parte tecnológica y otro de la parte estadística o analítica.

El rastro digital que ciudadanos y empresas vamos dejando a lo largo de nuestra vida habitual supone un caudal de información digitalizada ingente. Se estima que el orden de magnitud de la cantidad de información que se genera anualmente es de zettabytes (10 elevado a 21 bytes, aunque ya hay centros con una capacidad de procesamiento para un orden de magnitud superior, yottabytes).

Y la tendencia es de un crecimiento exponencial, basta con pensar lo que aportará la generalización del internet de las cosas (IoT), por citar solo un ejemplo. Según un estudio reciente de la Comisión Europea el número de dispositivos conectados en el año 2020 será de entre 26 mil y 50 mil millones[1].

Así que la cantidad de información susceptible de ser analizada no tiene comparación con nada similar del pasado y no parece que podamos aprovecharla usando las mismas herramientas tecnológicas y estadísticas que hace 20 o 30 años cuando los datos digitales generados en todo el mundo a lo largo de un año cabrían en un pen-drive actual.

No solo tenemos que evolucionar las herramientas, también, y quizá sea lo más difícil y lo más trascendente, la forma de enfrentarse a los problemas. Hasta ahora el paradigma se basaba en acumular datos para responder a preguntas formuladas previamente. Sin duda eso seguirá siendo así, pero los conjuntos de big data abren también la puerta a una nueva manera de entender la producción, ya que pueden contener respuestas a cuestiones que no estaban formuladas cuando se inició su acumulación.

Las diversas fuentes de big data pueden tener un impacto muy relevante en prácticamente todas las áreas de la producción de la estadística oficial. Estas fuentes pueden usarse para estimar variables en dominios muy diversos, desde el ámbito de las encuestas de turismo, al de las estadísticas de consumo, mercado laboral o globalmente a encuestas dirigidas a empresas o a población. Prácticamente todos los sectores podrían enriquecerse con estas nuevas fuentes de información.

El potencial es enorme, sin embargo estamos comenzando a analizar sus posibilidades y por tanto hay que ser prudentes. Hay pilares de la estadística oficial que deben seguir identificando nuestra producción: la preservación de la confidencialidad, la independencia y la calidad de nuestros resultados. Y para garantizar todo ello hay una serie de retos importantes a superar: el acceso a los datos, la infraestructura tecnológica, la metodología para el análisis, la construcción de nuevos indicadores de calidad para estos productos, etc.

El Instituto Nacional de Estadística (INE) ya tiene acumulada experiencia en la producción estadística basada en múltiples fuentes. De hecho la combinación de datos procedentes de registros administrativos y su integración con información procedente de encuestas forma parte de nuestro sistema habitual de producción. Esta ha sido una evolución muy relevante en la producción estadística que venía demandada por razones de eficiencia, reducción de la carga estadística a ciudadanos y empresas y de costes para la propia institución.

De la misma forma, el INE va a afrontar el reto de aprovechar la información de big data. En la actualidad ya estamos trabajando para ello, en proyectos propios y en coordinación con el trabajo que en el Sistema Esta-

1 http://www.internet-of-things-research.eu/about_iot.htm

dístico Europeo se está realizando. No hay que olvidar que todos los retos citados anteriormente son comunes a todas las oficinas de estadística de la Unión Europea, y es mucho más eficiente avanzar de forma conjunta que hacerlo cada oficina de estadística por su cuenta. La estrategia elegida es impulsar algunos proyectos piloto asociados a distintas fuentes big data (telefonía móvil, contadores eléctricos, *web scraping*, etc.) con los que mostrar al mismo tiempo la utilidad de estas fuentes y resolver poco a poco los retos para integrar este tipo de información en la producción.

No cabe pensar que el big data suponga una revolución en la estadística oficial. Pero sí que habrá una evolución. Y debería ser rápida, porque los big data no están distribuidos en múltiples puntos, como lo están los datos que habitualmente recopila el INE en sus encuestas, sino que, en ocasiones, se concentran en pocos propietarios. Así que las posibilidades de explotación están en manos de varios agentes: la estadística oficial y esos propietarios de los datos. Las preguntas para las que la sociedad necesita respuestas han de ser respondidas, con eficiencia y calidad y, efectivamente, así serán resueltas ya sea por unos o por otros. Si nosotros no somos capaces de afrontar este reto con diligencia, quienes tienen esos acúmulos de datos sí proveerán esos resultados. Y no da igual quien responda. Es importante recordar que nosotros, la estadística oficial, aportamos a la sociedad valores como la independencia y calidad de la información. Así que deberemos adaptarnos lo antes posible.

Hay muchas cuestiones aún por resolver para la producción de estadísticas oficiales utilizando big data: acceso a estas fuentes, metodología para su tratamiento, infraestructura tecnológica, etc. Una de ellas es la de los indicadores de calidad. Actualmente nuestros usuarios disponen de información sobre la calidad de nuestras operaciones basados en errores de muestreo, de cobertura e incluso estudios sobre errores de medida. Para estas nuevas fuentes habrá que desarrollar nuevas medidas que sean comprendidas por los usuarios.

En cuanto a los métodos estadísticos que se deben utilizar en el futuro, en general, no va a ser suficiente hacer inferencia de la información de big data apoyándonos en pilares clásicos de la estadística como población objetivo, marco poblacional, muestra, estimación o errores de muestreo. Y en esa renovación metodológica está el reto al que tenemos que dar respuesta en los próximos años. Las técnicas de *machine learning*, la estimación basada en modelos y las estimaciones bayesianas tendrán un papel relevante.

Sin duda las capacidades para el tratamiento de big data se convertirán en un requisito en la formación de los profesionales futuros. Cada vez se hace más evidente la necesidad de superar la dicotomía estadístico o informático. De hecho esto ya ha sucedido en las grandes empresas privadas que están explotando big data para sus procesos internos o como productos de negocio, es una cuestión que ya no está en discusión.

Nosotros también necesitaremos las dos competencias en una sola persona. Es lo que se conoce ya en toda la industria y la comunidad educativa como científico de datos. A medida que vayamos trabajando con fuentes de datos de este tipo, se requerirá un perfil que aúne informática y estadística. Eso tendrá que tener reflejo necesariamente en las capacidades exigidas para el ingreso en los cuerpos estadísticos del estado. Ya hemos comenzado a dar pequeños pasos en la adaptación de los programas de las oposiciones. Y también en la formación del personal que ya está trabajando en el INE.

La máxima utilidad del uso de big data revertirá en aquellos ámbitos que sean capaces de desarrollar estrategias razonables, en el sentido de usar la razón, para responder a preguntas previamente formuladas o, mejor aún, a conclusiones derivadas de los datos sobre cuestiones aún no planteadas. Esperemos que la estadística oficial sea uno de los agentes que estén en este grupo.

Hay muchas cuestiones aún por resolver para la producción de estadísticas oficiales utilizando big data: acceso a estas fuentes, metodología para su tratamiento, infraestructura tecnológica, etc.

Organizaciones flexibles en sus esquemas de producción y análisis de la información, sensibles a la innovación y a las inversiones en alto valor añadido como el conocimiento, sociedades dirigidas por la cultura y la educación que tengan la capacidad de desechar informaciones interesadas y buscar conocimiento en los datos serán las más beneficiadas.

Por lo que respecta a la estadística oficial los ciudadanos ya conocen que nuestro compromiso con la preservación del secreto estadístico es esencial para nosotros. Ninguna información que divulgue el INE permite que se identifique ni directa ni indirectamente a una persona, un hogar o una empresa. Ninguna información individual administrativa o estadística que reciba el INE es trasmitida fuera del ámbito estadístico. Así que en este aspecto la sociedad puede estar segura de que con el big data actuaremos de la misma manera.