ENTREVISTA

Carlos Ballano Fernández

"LAS TÉCNICAS ESTADÍSTICAS VAN A SER ABSOLUTAMENTE NECESARIAS PARA LA EXPLOTACIÓN DE LAS NUEVAS FUENTES DE DATOS"

Actualmente, la actividad digital de individuos y organizaciones genera una gran cantidad de información que vemos cómo cada día es aprovechada por distintas empresas en la generación de nuevos productos y servicios. ¿En qué medida la estadística oficial puede aprovechar también estos datos para mejorar los productos que proporciona actualmente a la sociedad?

Hay que tener en cuenta que la estadística oficial tiene una importante experiencia en la utilización de información no procedente de encuestas para la generación de información. Desde hace años se utiliza la información procedente de ficheros administrativos con una finalidad estadística. Estos usos se extienden a distintos tipos de operaciones estadísticas:

- Para la formación de directorios, como es el caso del Padrón Municipal de Habitantes para el Marco de Población, o los ficheros que se utilizan para la formación del DIRCE.
- Otros para completar variables en encuestas, como es el caso de los datos de renta en la Encuesta de Condiciones de Vida, o de salario y discapacidad en el caso de la Encuesta de Población Activa.
- Sustituir la recogida directa. En el Censo Agrario 2020 la información de más de 650.000 explotaciones ha procedido de la explotación de la información de registros, fundamentalmente el FEGA del Ministerio de Agricultura, investigándose el resto mediante recogida directa; o en el caso de la Estadística Estructural de Empresas, en la que se utiliza la infor-

mación administrativa para disminuir el tamaño muestral.

- El Censo de Población y Viviendas 2021, que consta de una investigación exhaustiva a partir de ficheros administrativos y una operación en campo para estudiar las características esenciales de la población.
- Para el tratamiento estadístico de la no respuesta, como es la calibración a las cifras de población en las encuestas dirigidas a los hogares.

La utilización de esta información no es directa, sino que exige un trabajo de adecuación a los criterios estadísticos que, en el caso de operaciones importantes, puede significar un trabajo de varios años, como es el caso del uso en los Censos de Población y Viviendas o del Censo Agrario.

También exige en algunas ocasiones la utilización de técnicas estadísticas no estándares en las encuestas habituales por muestreo a través de marcos de lista; un ejemplo es el empleo de muestreo indirecto para la obtención de información a nivel de empresa estadística a partir de la información procedente de las unidades legales.

En la actualidad a la información anterior se añade la mayor disponibilidad de información procedente directamente del registro que queda de la actividad de los individuos, como es la procedente de telefonía móvil, tarjetas de crédito, reservas por internet de hoteles o apartamentos, compras por internet... El potencial estadístico de toda esta información se podría clasificar en dos categorías:

- Para proporcionar nuevos indicadores.
- Para, combinada con encuestas ya existentes, facilitar los desgloses temporales, espaciales o de otro tipo, proporcionando una mayor granularidad de la información generada por encuestas.

Como en el caso de los registros administrativos, es necesario un trabajo arduo para aproximar los conceptos estadísticos de interés en base a la información externa disponible. Es, por esto, que han aparecido en el ámbito de la estadística oficial las 'estadísticas experimentales', que permiten un acercamiento de 'aprender mientras se hace', de forma que mientras se va ajustando los procedimientos se va ofreciendo un producto a la sociedad hasta que es un producto maduro que puede ser parte de la estadística oficial.

Como muestra de las posibilidades que ofrece esta información me referiré solo a las estadísticas experimentales del INE: la información sobre movilidad en base a la información de la telefonía móvil, estimación de la ocupación en establecimientos turísticos en base a la información de las plataformas digitales, estimación de los gastos de visitantes extranjeros en base a la información procedente de las tarjetas bancarias, o el estudio de viviendas turísticas en base a técnicas de *web scraping*.

La utilización de esta información no es directa, sino que exige un trabajo de adecuación a los criterios estadísticos que, en el caso de operaciones importantes, puede significar un trabajo de varios años

Además de estas estadísticas experimentales el INE está trabajando en otros proyectos como es la obtención de precios directamente de las grandes superficies de España, para su utilización en el IPC.

Obviamente, lo anteriormente mencionado no agota las posibilidades de utilización y las oficinas de estadística de nuestro entorno trabajan en proyectos en los que las nuevas fuentes de información ofrecen posibilidades de mejora en la estadística oficial. Un ejemplo relevante es el proyecto Big Data for European Statistics (BDES), lanzado por Eurostat, en el que el INE ha colaborado en un subproyecto, y en el que se han investigado en profundidad la potencialidad estadística de distintas fuentes de Big Data, como la telefonía móvil, las ofertas de trabajo *on line*, la obtención de informa-

ción de las empresas a través de sus páginas web, de los hogares a partir de los contadores eléctricos, de las transacciones financieras de distintos instrumentos y plataformas de pago, de la información de la tierra procedente de satélites y fuentes para el turismo.

Las oficinas de estadística de nuestro entorno trabajan en proyectos en los que las nuevas fuentes de información ofrecen posibilidades de mejora en la estadística oficial

Sin duda el aprovechamiento de las nuevas fuentes de datos supone un reto para las instituciones estadísticas. ¿Puede señalar los aspectos que en su opinión son los más importantes y novedosos en relación al uso de estas nuevas bases de datos?

En mi opinión son importantes los siguientes aspectos:

- Legal: al ser un fenómeno emergente hay que ver cómo se considera desde la legislación estadística, tanto europea como nacional. Es un asunto que se está tratando en ambos niveles y que tiene que ver con la gestión de la información que se considera un bien público desde el punto de vista de la estadística oficial.
- De infraestructura: es necesario pasar a infraestructuras informáticas que permitan mucha mayor capacidad de almacenamiento y de tratamiento de la información. Existe un proyecto de la Secretaría General de la Administración Digital en este sentido.
- De metodología estadística: sobre los procedimientos para transformar y utilizar los datos disponibles a información estadística relevante. Aquí hay que tener en cuenta la experiencia de las oficinas de estadística en el uso de información auxiliar, tanto para estimaciones por calibración como en las técnicas estadísticas para estimación en pequeñas áreas.
- De medida de la calidad de la información: la información producida debe ser de utilidad para la toma de decisiones. En este aspecto hay que tener en cuenta que el control del proceso en las encuestas puede servir para corregir los sesgos de cobertura de las nuevas fuentes de datos. En algunos casos, las fuentes externas también pueden ayudar a detectar sesgos en las encuestas.
- De seguridad y preservación de la confidencialidad estadística y de protección de datos: los sistemas deben de satisfacer todas las condiciones que se establecen en el Esquema Nacional de Seguridad, la legislación nacional y europea sobre protección de datos y secreto estadístico.

Durante mucho tiempo las técnicas de muestreo han sido nucleares para el diseño de las operaciones estadísticas. Estas técnicas contribuyen a disponer de información relevante, midiendo su calidad de manera científica y permiten asignar los recursos disponibles de forma óptima. En la época de las fuentes Big Data, ¿cree que disminuye la importancia del muestreo en favor de otras técnicas o seguirá siendo en el futuro un referente principal del trabajo de instituciones como el INE?

Como he comentado más arriba, mi opinión es que las técnicas estadísticas van a ser absolutamente necesarias para la explotación de las nuevas fuentes de datos. Y no me refiero solamente a que en algunos casos la magnitud de la información disponible haga necesario tomar muestras para su explotación. Aunque habrá indicadores que se deriven directamente de dichas fuentes, en muchos otros casos será la combinación de las nuevas fuentes y encuestas las que se deban utilizar para la producción de la información.

Por otra parte, considero adecuada la cita de Deming de 1950 que utiliza Lohr: 'El muestreo no es el reemplazo de una cobertura parcial por otra total, sino la ciencia y el arte de controlar y

medir la fiabilidad de la información estadística a través de la teoría de la probabilidad'.

Desde este punto de vista más amplio, creo que las nuevas fuentes de datos van a provocar avances en la teoría y práctica del muestreo.

Se entiende que el mayor aprovechamiento de los datos existentes redundará en una menor carga directa de solicitudes de información por parte de las instituciones estadísticas a los hogares y sobre todo a las empresas. Pero mientras tanto, ¿qué medidas adopta el INE para minimizar esta carga?

El INE es consciente de la carga estadística que genera. Por este motivo lleva trabajando desde hace décadas en técnicas para mitigar y distribuir esa carga.

En lo que atañe al diseño muestral aplicamos una técnica llamada de Números Aleatorios Permanentes, junto a una función de carga estadística, para obtener las muestras estructurales dirigidas a las empresas de la manera más equitativa posible, en el sentido de distribuir la carga estadística a las empresas, medida en el número de encuestas que recibe y en el tiempo estimado de realizar los cuestionarios, lo más uniformemente posible. Si una empresa es seleccionada de manera aleatoria para realizar una encuesta en un determinado año, dicha empresa tiene mucha menos probabilidad de que vuelva a ser seleccionada en otra encuesta dirigida a empresas en ese año, si por el tamaño no se investiga de forma exhaustiva.

También se lleva a cabo un control del número máximo de estadísticas en las que colabora una empresa en función de su tamaño.

Asimismo se utiliza la información administrativa disponible tal como he comentado más arriba. Un caso relevante es el uso del Impuesto de Sociedades para precargar los datos en la cumplimentación a través de la web de la Estadística de Estructura Económica. Por motivos de calendario solo podemos utilizar la información en la muestra de esta encuesta que se lanza en septiembre, pero esperamos que en breve pueda extenderse el uso al total de la muestra de esta encuesta.

Acabamos nuestros encuentros pidiendo a los entrevistados un esfuerzo de imaginación. ¿Cómo ve la sociedad española dentro de 20 años? Denos un temor, una prioridad y un deseo para España.

Una forma de ver el futuro es mirar al pasado y proyectar la evolución percibida. Hace 20 años el INE contrató más de 40.000 personas para hacer el Censo de Población y Viviendas. En 2021 el presupuesto y el personal se han reducido sustancialmente debido a la disponibilidad de fuentes administrativas y de Big Data.

Si proyecto esta evolución 20 años en el futuro concibo que lo que habrá será unos sistemas de información integrados y georreferenciados que proporcionarán información a un nivel de granularidad temporal y espacial suficiente, una vez identificadas las potencialidades y metodologías de combinación de fuentes más adecuadas, con unos criterios de calidad también adecuados y preservando el secreto estadístico.

Un temor, mejor dos: la inmovilidad por un lado, y la adopción de novedades no suficientemente maduras para la producción, por otro.

Una prioridad: identificar las áreas de mejora que nos sitúen en una mejor posición para afrontar el futuro.

Un deseo: que sepamos transmitir a los nuevos estadísticos que se incorporarán en las próximas oposiciones la necesidad de mejorar el INE, de la misma forma que en su momento nos lo transmitieron a nosotros los que nos precedieron. ●

CARLOS BALLANO FERNÁNDEZ

Subdirector del Departamento de Control de la Producción Estadística y Muestreo del INE.

Funcionario del INE desde 1985. Ha trabajado fundamentalmente en asuntos relacionados con el muestreo y la recogida de datos.
Licenciado en Ciencias Matemáticas y en Ciencias Físicas.
Diploma de Estudios Avanzados en 'Sistemas Estocásticos y su Control Óptimo'.