El método *scanner data* para la utilización de bases de datos de empresas en el IPC

Ignacio González Veiga
Subdirector de Precios y Presupuestos Familiares. INE

El Sistema Estadístico Europeo se ha marcado una serie de objetivos y retos para desarrollar en los próximos años, dirigidos a incrementar la eficiencia y la relevancia para la sociedad de las estadísticas oficiales en su conjunto. Uno de estos grandes retos es el aprovechamiento de los grandes volúmenes de información que surgen de la digitalización de la vida personal y económica que caracteriza a la sociedad actual: es lo que se conoce como *Big Data*.

La utilización de fuentes de *Big Data* en la estadística oficial permitiría reducir los costes de las encuestas, tanto desde el punto de vista de los que las producen como desde el punto de vista de los informantes, que los plazos de publicación se acorten y que se puedan aportar nuevos datos sobre aspectos sociales y económicos aún no estudiados.

Dentro del ámbito de *Big Data*, el *scanner data* juega un papel destacado y relevante. Precisamente por ello, Eurostat decidió en su momento dedicar un proyecto específico para esta fuente de información de manera que el impulso a los trabajos fuese más directo. Este documento describe este proyecto en el que están participando todos los países de la Unión Europea.

¿QUÉ ES *SCANNER DATA*?

La mayor parte de la información utilizada para el cálculo del Índice de Precios de Consumo (IPC) se obtiene mediante la visita del personal del Instituto Nacional De Estadística (INE) a los establecimientos representativos de cada sector, seleccionados previamente en cada provincia.

Este sistema de recolección de la información, junto con una muestra de establecimientos y productos significativa, garantiza la calidad de los resultados. Sin embargo, como sucede en la mayor parte de las estadísticas incluidas en el Plan Estadístico Nacional (PEN), el INE trabaja permanentemente para reducir la carga que conlleva responder a sus requerimientos por parte de los informantes (en este caso, los establecimientos) y, de paso, mejorar la precisión de sus estimaciones.

En esta línea de trabajo, el INE ha implantado en los últimos años la utilización de nuevos métodos y técnicas para la obtención de la información, basados en la explotación de registros administrativos y en el uso de dispositivos electrónicos de recogida.

En el caso del IPC, la recogida de los precios en los establecimientos mediante dispositivos electrónicos será una realidad el próximo año, lo que sin duda supondrá una ganancia en la precisión de esta estadística y una mayor eficiencia en los procesos de producción.

Por su parte, la utilización de bases de datos de las empresas informantes es algo que se ha comenzado a explorar recientemente y en la actualidad está en proceso de desarrollo. Es lo que se denomina en el ámbito internacional *scanner data*.

Básicamente, este método consiste en **utilizar la información registrada por las compañías de comercio minorista en la línea de caja de cada uno de sus establecimientos**. Habitualmente, esta información consiste en el número de unidades vendidas y los ingresos para cada uno de los productos comercializados, clasificados según criterios propios por cada compañía.

El *scanner data* ya está siendo utilizado en algunos países de nuestro entorno, ya que se trata de una alternativa más eficiente, precisa y completa de medir la inflación. Por ello, la oficina de estadística europea, EUROSTAT, promueve su utilización en el ámbito de la armonización de los índices de precios de los estados miembros de la UE. Como

no podía ser de otra manera, España se ha sumado a la propuesta y en 2014, el INE inició un proyecto piloto con el objetivo de evaluar todos los aspectos sobre la posible implantación en el cálculo del IPCA y, consecuentemente, del IPC.

A lo largo de los últimos dos años el INE, basándose en la experiencia de otros países, ha desarrollado el modelo metodológico más adecuado para el tratamiento de la información proveniente de las cadenas de supermercados e hipermercados y su posible integración en el cálculo del IPC.

En la actualidad, se ha concluido la fase de diseño metodológico y se están realizando las primeras pruebas con datos reales, para lo cual es imprescindible la colaboración de las compañías comercializadoras de productos. A continuación se detallan las características principales del método, así como la información requerida para desarrollarlo.

PROCEDIMIENTO PARA LA UTILIZACIÓN DE *SCANNER DATA*

La implantación de *scanner data* en la metodología de cálculo del IPC supone un cambio trascendental en la concepción de este indicador. Como se ha dicho, hasta ahora la producción se basa en la recogida de precios en los establecimientos y el cálculo de índices a partir de las medias de los mismos. Sin embargo, la utilización de las bases de datos de las empresas conlleva la gestión de un volumen de información incomparablemente mayor que hasta ahora, y un cambio radical en el procedimiento de cálculo del IPC.

Por tanto, se puede considerar que el proyecto tiene que salvar dos escollos importantes: uno, relativo a la disposición de las empresas a proporcionar la información requerida; el otro, relacionado con la propia utilización de la información, sus dificultades y sus consecuencias.

A) Obtención de la información

La información que se precisa para el desarrollo del proyecto no debe suponer una carga adicional para la empresa. Al contrario, la filosofía de partida del método *scanner data* es, precisamente, el aprovechamiento de las bases de datos disponibles en cada compañía. No se precisa, pues, una elaboración específica ni para modificar su contenido ni para cambiar su estructura.

Habitualmente, la información contenida en las bases de datos de las empresas comercializadoras de alimentación, perfumería y productos de limpieza es, para cada producto comercializado, la siguiente: ingresos, cantidades, denominación del producto, descripción (si existe algún campo donde se distinga), código (que puede ser uno propio de la empresa para su uso interno u otro adecuado a la clasificación internacional EAN). Esta información es suficiente como para plantearse su utilización en la producción del IPC.

No obstante, para que este método sea válido a efectos de su integración en el IPC, la información facilitada debe referirse a todos y cada uno de los productos con código asignado (ya sea el propio de la empresa o el código de barras del producto, o ambos) y en cada uno de los establecimientos. Asimismo, es imprescindible la regularidad y la continuidad de los envíos de esta información.

El formato de las bases de datos, el sistema de transmisión de la misma y los demás aspectos relacionados con la disposición de la información los debe decidir la empresa, para que el coste y el esfuerzo que conlleve su elaboración sea la menor posible.

B) Aspectos conceptuales fruto de la utilización de *scanner data*

Las primeras cuestiones que surgieron al comienzo del proyecto se refirieron, sobre todo, a los aspectos conceptuales. Especialmente, porque la incorporación de información sobre ventas de las compañías exige cambios en los métodos y las definiciones utilizados tradicionalmente en la metodología de cálculo de este indicador; asimismo, la integración de ambos tipos de información (la propia del IPC y la proveniente de las bases de datos) requiere procesos específicos que las homogeneice. Los principales retos conceptuales son los siguientes:

Diferencias entre los conceptos precio y valor unitario

El IPC mide, por definición, la evolución de los precios de los bienes y servicios adquiridos por los hogares. Se recoge, por tanto, el precio de venta al público en cada establecimiento. La utilización de *scanner data*, sin embargo, cambia esta filosofía ya que exige que para cada código de producto, se utilice su valor unitario (total de ingresos dividido por el total de unidades vendidas), pero no el precio propiamente dicho.

En realidad, el valor unitario no se corresponde con una única transacción real sino que representa a todas las realizadas a lo largo de un periodo de tiempo fijado. Esto supone un cambio importante en la definición del IPC y en los distintos tratamientos aplicados, como los de descuentos y ofertas.

Diferencias entre producto y la gama completa de variedades

La utilización de las bases de datos permite disponer de la información de todas las variedades vendidas de un producto. Esto difiere del procedimiento habitual del IPC que, por su concepción, realiza el seguimiento de precios de una única variedad en cada establecimiento.

Por tanto, el problema metodológico que suscita la incorporación de estas bases de datos al cálculo del IPC es doble: qué criterios utilizar para calcular los valores unitarios y cómo integrar los resultados con los datos sobre precios que viene utilizando el IPC tradicionalmente.

C) Aspectos metodológicos relevantes a tener en cuenta

Volumen de información

Otro aspecto a tener en cuenta cuando se aborda la utilización de *scanner data*, es el volumen de datos significativamente superior al que se obtiene con la recogida tradicional de precios. Por ello, además de los requisitos técnicos para el tratamiento de dicha información, también hay que introducir técnicas para determinar qué variedades deben formar parte del cálculo y cómo proceder ante los cambios de su contenido a lo largo del tiempo (productos que se venden un mes, pueden dejar de hacerlo en el futuro).

Proceso de cálculo

El objetivo primordial del proyecto es hacer compatible la información contenida en las bases de datos con los datos de precios utilizados en el cálculo habitual del IPC. El aspecto conceptual comentado anteriormente acerca del uso de valores unitarios frente a precios, no es el único obstáculo a salvar, sino que en el proceso que se debe seguir hasta llegar a obtener índices para cada conjunto de productos, se deben ir adoptando decisiones orientadas a poder integrar las dos fuentes de información.

Algunos de los temas más relevantes son, por ejemplo, los siguientes:

- Clasificaciones. Cada empresa tiene su propia clasificación, lo que obliga a establecer una relación entre estas y la utilizada por el IPC.
- Seguimiento de los productos. Un producto, o conjunto de productos, puede figurar en la base de datos porque haya sido vendido durante un periodo de tiempo, pero desaparecer en un momento determinado. Asimismo, la empresa puede cambiar el código de alguno de los productos, lo que dificulta su seguimiento. Esto supone un problema para la medición mensual de las tasas de precios que exige el IPC.
- Detección de valores atípicos. A diferencia de la recogida de precios presencial en los establecimientos, las bases de datos pueden contener valores atípicos cuyo origen no siempre es posible conocer. Puede suceder porque se hayan producido cambios en el contenido o porque haya habido alguna promoción para aumentar las ventas de los mismos. En cualquier caso, es preciso establecer normas para el control de estas situaciones antes de incorporarlo al cálculo del IPC.
- Integración de datos. La información de *scanner data* debe pasar finalmente a integrarse con los precios de IPC. Para ello, es necesario establecer el método de agregación, así como los pesos con los que los productos deben entrar a formar parte del IPC.

EL FUTURO DEL PROYECTO

Una vez encauzados los principales problemas metodológicos, los trabajos ahora se centran en conseguir la colaboración continuada de las principales empresas comercializadoras.

En principio, el proyecto se ha enfocado hacia las grandes empresas de supermercados e hipermercados. La colaboración inicial es la que mayor carga puede suponer a estos informantes, ya que se trata de establecer una línea directa de trabajo en la que se decida la estructura de la base de datos, el proceso de envío y, sobre todo, adquirir una rutina de colaboración que permita su incorporación en el IPC. Solo entonces se puede plantear la posibilidad de proceder al cálculo de este indicador con este método de obtención de la información.

De cara a futuro, además, está previsto extender la utilización de la información de *scanner data* para otros proyectos. Así, por ejemplo, la información contenida en las bases de datos tiene un enorme potencial para seleccionar los artículos representativos de la cesta de la compra del propio IPC, y para calcular ponderaciones a niveles de máxima desagregación, además de su posible utilización para otras estadísticas incluidas en el PEN y que elabora el INE, como las Paridades del Poder Adquisitivo (PPA).