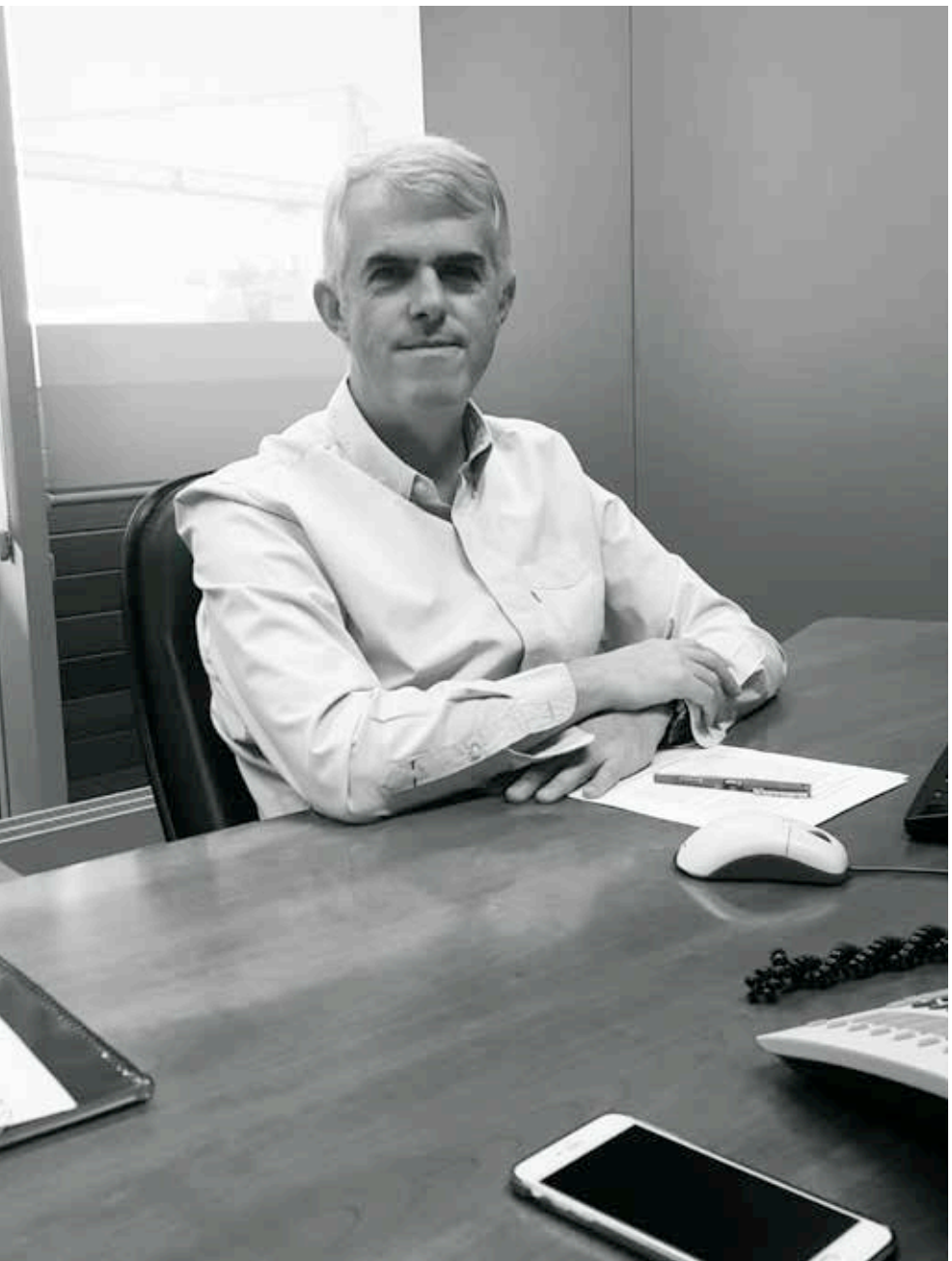
Miguel Ángel Martínez Vidal

"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"



El tratamiento del *Big Data* ha supuesto una revolución absoluta en el análisis de datos hasta el punto de que el concepto ha tenido un enorme calado social. En casi todas partes se habla del *Big Data* pero, ¿cabría concebir una definición exacta de este concepto?

La forma más popular 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 de la definición esté en que su análisis requiere 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 ordenadores 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 análisis de datos masivos se nos presenta, sin duda, como un instrumento de extraordinaria potencia pero,

¿qué resultados podrá generar el manejo del *Big Data* que habrían sido imposibles desde paradigmas anteriores? ¿Qué se hace verdaderamente posible a través del *Big Data*?

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 exponencial, basta con pensar lo que aportará la generalización del internet de las cosas, por citar solo un ejemplo.

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 un año cabrían en un *pendrive* 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 datos. 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 se encuentra en permanente actualización con respecto al desarrollo de nuevos procesos estadísticos. ¿En qué modo se integrará el tratamiento de datos masivos en el INE? ¿Cabe esperar a corto plazo una verdadera revolución en el mundo de la estadística?

El INE ya tiene acumulada experiencia en la producción estadística basada en diversas 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 Estadístico 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 solo tenemos que evolucionar las herramientas, también, y quizá sea lo más difícil y lo más trascendente, la forma de enfrentarse a los problemas*

No creo que haya ninguna revolución. Pero sí que habrá una evolución en la estadística oficial. 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 se concentran en pocos propietarios. Así que las posibilidades de explotación están en manos de varios agentes. Las necesidades de información de la sociedad serán satisfechas por unos o por otros y nosotros aportamos valores como la independencia y calidad de la información. Así que deberemos adaptarnos lo antes posible.

Desde el punto de vista estadístico, en general, no va a ser posible 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.

La complejidad del *Big Data* está permeando planes de estudio, títulos universitarios, postgrados privados… ¿Cree que las habilidades relativas al manejo del *Big Data* se convertirán en un requisito en la formación de los profesionales futuros?

Sin duda, creo que así será. 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.

Las diversas fuentes de Big Data *pueden tener un impacto muy relevante en prácticamente todas las áreas de la producción de la estadística oficial*

Nosotros también necesitaremos las dos competencias en una sola persona. Es lo que se viene llamando 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.

Las aplicaciones del *Big Data* parecen en principio ilimitadas hasta el punto de que su desarrollo podría cambiar para siempre nuestra manera de interpretar la realidad. En el ámbito empresarial, en la investigación científica, en los procesos de comunicación… ¿cuál cree que es el ámbito en el que revertirá la máxima utilidad e impacto del *Big Data*?

Los datos por sí solos no nos dicen nada, no responden a ninguna pregunta ni resuelven ningún problema. Se necesita, previamente, convertirlos en información. Hoy en día ya somos bombardeados con un montón de esa información. En muchas ocasiones los estadísticos podemos comprobar cómo una misma estadística es usada en argumentos totalmente opuestos para apoyar puntos de vista establecidos a priori.

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.

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.

La historia ha demostrado que toda experiencia de progreso acelerado entraña sin duda un riesgo, ¿sabemos cuáles son los peligros derivables de algo tan potente como la gestión de datos masivos? Aspectos como la privacidad o la autonomía en la toma de nuestras decisiones podrían verse comprometidas en un escenario de excesiva transparencia como el que parece posibilitar el *Big Data*. ¿Hay algo de lo que preocuparse o estas cautelas no son más que la expresión de un temor tecnofóbico?

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.

Acabamos nuestras entrevistas pidiendo a los encuestados un esfuerzo de imaginación. ¿Cómo ves la sociedad española dentro de 20 años? Danos un temor, una prioridad y un deseo para nuestro país.

El temor sería que no desarrollemos un sentido profundo de la autocrítica, que pienso que hoy en día no lo ejercemos demasiado ni en el plano individual ni como sociedad.

La prioridad, sin duda, la educación y la innovación. No podemos sobrevivir compitiendo en actividades de bajo valor añadido. El futuro está en el conocimiento.

El deseo es que se reconozca el esfuerzo. El que realizan personas y entidades que alcanzan objetivos admirables y el que debemos realizar para tener las condiciones de vida que todos deseamos.

Diego S. Garrocho