OLIMPIADA ESTADÍSTICA 2017

El Instituto Nacional de Estadística (INE), la Facultad de Estudios Estadísticos (FEE) de la Universidad Complutense de Madrid y la Sociedad de Estadística e Investigación Operativa (SEIO) convocan la Quinta Olimpiada Estadística para estudiantes de Enseñanza Secundaria Obligatoria, Formación Profesional Básica, Bachillerato y Ciclos Formativos de grado medio.

La Olimpiada Estadística tiene como objetivos:

- Promover la curiosidad y el interés en la Estadística entre los estudiantes.
- Incentivar en los docentes el uso de nuevos materiales para la enseñanza de la Estadística fomentando el uso de datos reales y buscando aplicaciones de los conocimientos estadísticos adquiridos.
- Mostrar y acercar el protagonismo de la Estadística en distintos aspectos de la sociedad a estudiantes y docentes, dándola a conocer como estudio universitario.
- Promover el trabajo en equipo y la colaboración para conseguir objetivos comunes.

Plazo de inscripción: del 3 de octubre de 2016 al 27 de enero de 2017.

Realización de las pruebas de la primera fase: del 30 de enero al 13 de febrero de 2017.

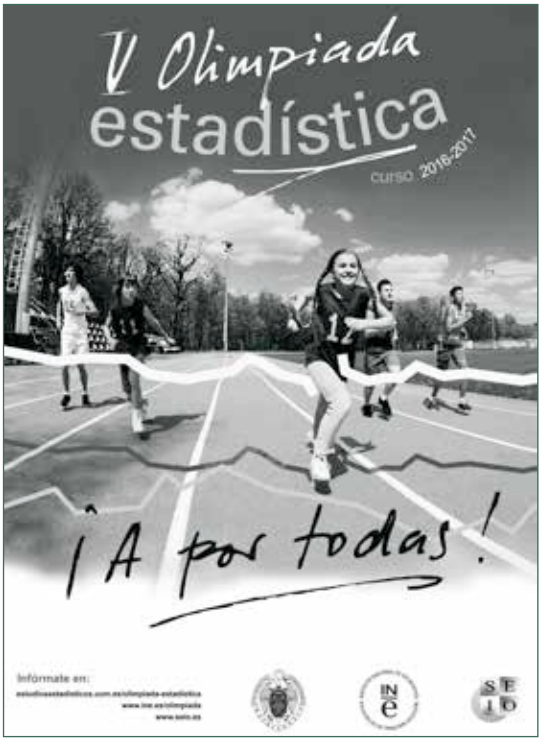
Publicación y comunicación de los resultados de la primera fase: 15 de febrero de 2017.

Realización del trabajo de la segunda fase: del 16 de febrero hasta el 9 de marzo de 2017.

Publicación de los equipos ganadores: 30 de marzo de 2017.

Para participar en la Olimpiada Estadística será necesario inscribirse utilizando el formulario habilitado al efecto en la página web del INE **www.ine.es/olimpiada**. Los participantes podrán plantear sus dudas a través del correo electrónico **oliestad@ucm.es**.

Bases de la convocatoria:
http://www.ine.es/explica/olimpiada2017_bases.pdf



EL INE TIENE PREVISTO PUBLICAR LAS CUENTAS DE GASTO EN PROTECCIÓN MEDIOAMBIENTAL Y ENERGÍA EN 2017

Desde hace años, el INE publica las cuentas de emisiones, impuestos y flujos de materiales y para el próximo ejercicio 2017 tiene previsto elaborar las cuentas de gasto en protección medioambiental y energía. Los tres nuevos módulos, regulados en el Reglamento sobre cuentas medioambientales europeas, son coherentes con el Sistema de Cuentas Medioambientales aprobado en la Comisión de Estadística de Naciones Unidas.

El conjunto de cuentas medioambientales permite analizar la contribución del medio ambiente a la economía, así como el impacto de la actividad económica sobre el medio; así, constituye una herramienta de base para la planificación estratégica y de análisis político sobre el desarrollo sostenible.

A efectos de elaborar la cuenta de flujos físicos de la energía es fundamental conocer y utilizar de forma adecuada las fuentes básicas de información. Por este motivo, el INE mantiene reuniones periódicas con el Ministerio de Industria, Energía y Turismo (MINETUR), responsable de la elaboración de las estadísticas de la energía. Además, los contactos con otros organismos como la Corporación de Reservas Estratégicas de Productos Petrolíferos (CORES) y el Instituto para la Diversificación y Ahorro de la Energía (IDEA) son periódicos. En este marco se analizan y comparten los avances realizados por el departamento de Industria y el Instituto Nacional de Estadística en la elaboración de las estadísticas básicas de la energía y el proyecto de la cuenta satélite en este ámbito. También se siguen las directrices marcadas por los trabajos que desarrolla la *Task Force* de Eurostat para mejorar el conocimiento de los consumos energéticos industriales y en la que participa MINETUR. Al igual que las marcadas por el INE, que lleva a cabo las actuaciones previas para elaborar la cuenta satélite de la energía en base a los cuestionarios de las estadísticas energéticas, de las cuentas nacionales y otras fuentes de información auxiliares.

BIG DATA: ¿A QUIÉN SIRVE?

La *era de la información* es fecunda en la tarea de alumbrar terminología altamente seductora. *Big Data* no solo alude a una realidad en el tratamiento actual de la información, sino que posee además la capacidad de sugerir al lego la existencia de un nuevo saber al que debería prestar su atención y más frecuentemente su apoyo financiero. Pero, antes de entrar en detalles, quizá será mejor fijar algunos puntos acerca de qué puede entenderse por *Big Data*. Resulta curioso que el rasgo más relevante y en el que parecen coincidir todos los expertos sea presentar este nuevo dominio del saber por oposición a otros ya tradicionales. Así entendemos por *Big Data* la recogida, estudio y explotación de datos que por su magnitud, variedad o complejidad no pueden ser procesados por medios tradicionales. Es inevitable que esta definición recuerde en algo a la de la ya vieja *patafísica* dedicada, en este caso, al estudio de las excepciones que la ciencia tradicional no contempla. Quizá por eso, sea mejor considerar al *Big Data* como la recopilación y análisis de grandes cantidades de información del tipo más variado producida en la interacción de los usuarios, tanto civiles como institucionales a través de la Red.

El *Big Data* no se explica sin la Red y esto es algo que debemos tener en cuenta para entender su funcionamiento y aplicaciones, veámoslo con algún ejemplo. Supongamos que deseo poner en el mercado un producto orientado a los aficionados a los eventos deportivos, en particular el fútbol. Como es harto frecuente en la economía especulativa que viene imperando en los últimos años, mi objetivo no se centra en el producto, me es indiferente, solo deseo saber qué tipo de bien ha de ser, qué precio puedo pedir por él, en qué lugar debe ser distribuido, etc. Hasta no hace mucho, este tipo de trabajo recibía el nombre de *estudio de mercado* y se basaba tanto en el juicio experto de los analistas como en estudios estadísticos basados en pequeñas muestras representativas. Veamos cómo hace el *Big Data*. En este caso el estudio se orientaría no a obtener una muestra pequeña pero significativa, sino que intentaría ser lo más exhaustiva posible analizando toda la información que producen en la Red los usuarios de una cantidad considerable de eventos deportivos. A través del estudio de las correlaciones entre todo género de información producida en los momentos críticos, se obtendrían las conclusiones que servirían para adoptar una decisión de mercado u otra.

En el límite, imaginemos toda la información producida a nivel global en la Red, esto es, aquella que producimos en nuestras redes sociales, mediante dispositivos móviles, cuando adquirimos un producto, cuando efectuamos un trámite administrativo, cuando buscamos un dato o consultamos una información y simplemente sometámosla a una búsqueda intensiva de patrones y correlaciones. La sugerencia, porque todavía lo es, de los defensores de este tipo de estudios es que posiblemente hallaríamos conexiones funcionales entre comportamientos que solo pueden ser apreciados a esa escala y que no hubieran sido descubiertos simplemente postulando una hipótesis inicial, que muy difícilmente se nos habría llegado a ocurrir.

Esto lleva a otro punto que merece atención. Son muchos –me podría incluir a un nivel muy modesto– los analistas que son capaces de rastrear cantidades notables de información en ciertas redes sociales, Twitter sería el caso paradigmático, con el fin de obtener información valiosa acerca del comportamiento de sus usuarios. Las herramientas, como en el *Big Data*, son aplicaciones que extraen y procesan de forma automática cantidades de información que un agente humano no podría analizar, pero que tampoco requieren un tipo de maquinaria exclusiva. Estos análisis, que podríamos considerar como Not-so-*Big Data*, necesitan una hipótesis previa, un objetivo inicial y no se orientan a obtener la exhaustividad en el estudio de los casos relevantes, trabajan con datos previamente *estructurados*. Pero no son por eso menos interesantes.

Resulta sospechoso que la diferencia entre el *Big Data* y otros estudios de la información circulante en la Red sea a veces el tamaño y los recursos. El mensaje que indirectamente se manda no puede ser más claro: solo aquellos que disponen de los recursos exclusivos de que nosotros disponemos pueden obtener conclusiones de calidad; este no es un negocio al alcance de cualquiera, así que no confíes en aficionados. La cuestión es ahora averiguar qué hay de cierto en esto y para ello quizá convenga hacer notar que para desarrollar un proyecto de *Big Data* no solo hace falta una

gran maquinaria en términos de hardware y software, sino algo tan elemental como los propios datos, detalle en el cual rara vez se insiste. ¿Qué empresas, entidades o corporaciones poseen los datos que, no lo olvidemos, *producimos* como usuarios? La respuesta no es muy compleja por fortuna: aquellas que ofrecen búsquedas, como Google o Yahoo, las redes sociales, Facebook, Twitter o Whatsapp, las que dominan el sector del comercio electrónico como Ebay y Amazon y las centradas en el ámbito del entretenimiento, Youtube, Spotify entre otras. A estas se pueden añadir muchas otras de menor perfil y repartidas por casi todos los ámbitos de la actividad económica en el mundo desarrollado. Obsérvese que todas ellas son empresas privadas movidas, como es obvio, por un afán de lucro y cuyo principal capital es la información que producimos a diario y que queda a buen recaudo en sus bases de datos. Es cierto que para procesar tamaña información no basta con un equipo medio, sino que se requiere la potencia de cálculo de grandes ordenadores que muy pocas entidades poseen. Por tanto, será del comercio, entre las compañías que atesoran nuestros datos y aquellas otras, frecuentemente estatales, que poseen los recursos, de donde surja el *Big Data* tal y como se nos presenta en la actualidad.

¿Debemos temer esta suerte de conjunción entre los propietarios de los datos y los de la maquinaria de análisis? Esta cuestión, que en el fondo es la clave, no tiene, como todo lo importante, una respuesta fácil. Para los *neoluditas* y en general para un amplio colectivo que encaja con dificultad los cambios tecnológicos, la respuesta es obvia. Estaríamos no solo ante un *nuevo Leviatán*, sino ante el último y definitivo *Gran Leviatán*. En la línea de las mejores distopías tecnológicas nos veríamos avocados a un Estado omnipresente capaz de anticipar nuestros movimientos convirtiéndonos en títeres de una realidad alienada cuya trama quedaría finalmente fuera de nuestra vista. Las grandes corporaciones, en plena connivencia con los poderes públicos, crearían dependencias haciendo de nosotros poco más que autómatas perfectamente alineados con sus cadenas de producción. Esta respuesta me parece fácil; es más, me parece repetir en muchos de sus aspectos la reacción que provocó la propia Revolución Industrial en el mítico Ned Ludd y sus seguidores en los albores del siglo xix. ¡Quememos las fábricas y los telares! ¡Acabemos con el maquinismo que arruina la dignidad del hombre! Bien, quémense, que otros aún más grandes y atroces vendrán a ocupar su lugar. Cuando la Historia se fija una pauta con la intensidad que otorgó en su día a la Revolución Industrial, o en la actualidad a la propia Sociedad de la Información, hay muy poco que pueda hacerse por resistir el oleaje. Pero que quede claro, la alternativa tampoco es dejarse llevar.

Cuando me toca abordar este punto, cosa que ocurre con cierta frecuencia, siempre recuerdo el análisis que en su día se hizo de la *fábrica* como representante máximo de la deshumanización y alienación de las masas. El capitalista pensó que había alcanzado la solución perfecta, una forma de incorporar la materia prima humana como un elemento más a su maquinaria. Tenerlos a todos juntos para evitar la autonomía del hombre libre, aumentar sus horarios hasta la extenuación gracias a la concentración de los cuerpos en espacios cerrados y vigilados era el sueño de todo capitalista. Pero nunca tuvo en cuenta que los que antes estaban separados y no se conocían llegaron a convertirse, como en la guerra, en hermanos de sangre primero y más tarde, no mucho, en *cama-*

radas afiliados a sindicatos y partidos de clase. Toda acción genera una reacción y nunca es posible prever todas las consecuencias de una acción, entre ellas las indeseadas para los principales actores de las mismas.

Si extrapolamos este análisis al presente, veremos que la misma fuente de nuestros temores puede ser la forma en que nos protejamos de sus consecuencias más abyectas. Los datos que las grandes corporaciones y gobiernos precisan para sus proyectos de *Big Data* se generan en redes sociales que deben mantener su atractivo para un público pretendidamente variado y representativo. Si mi Red solo interesa a sectores muy definidos simplemente no resultará atractiva. Y ese pequeño margen que queda para el ciudadano digital concienciado puede valer su peso en oro en momentos críticos. No es una conjetura, sino un hecho. Nuestro país, en el que solemos dar poco valor a todo lo que tiene que ver con la innovación, ha sido protagonista de algunos de los movimientos sociales basados en el uso de las redes y las nuevas tecnologías más importantes de los últimos años, sobre todo por su capacidad para alterar la realidad política de manera efectiva. Este espacio para lo dispar es constitutivo de las redes y no puede ser eliminado de ellas sin hacer que el capital en que basan su poder, los datos, se devalúe definitivamente.

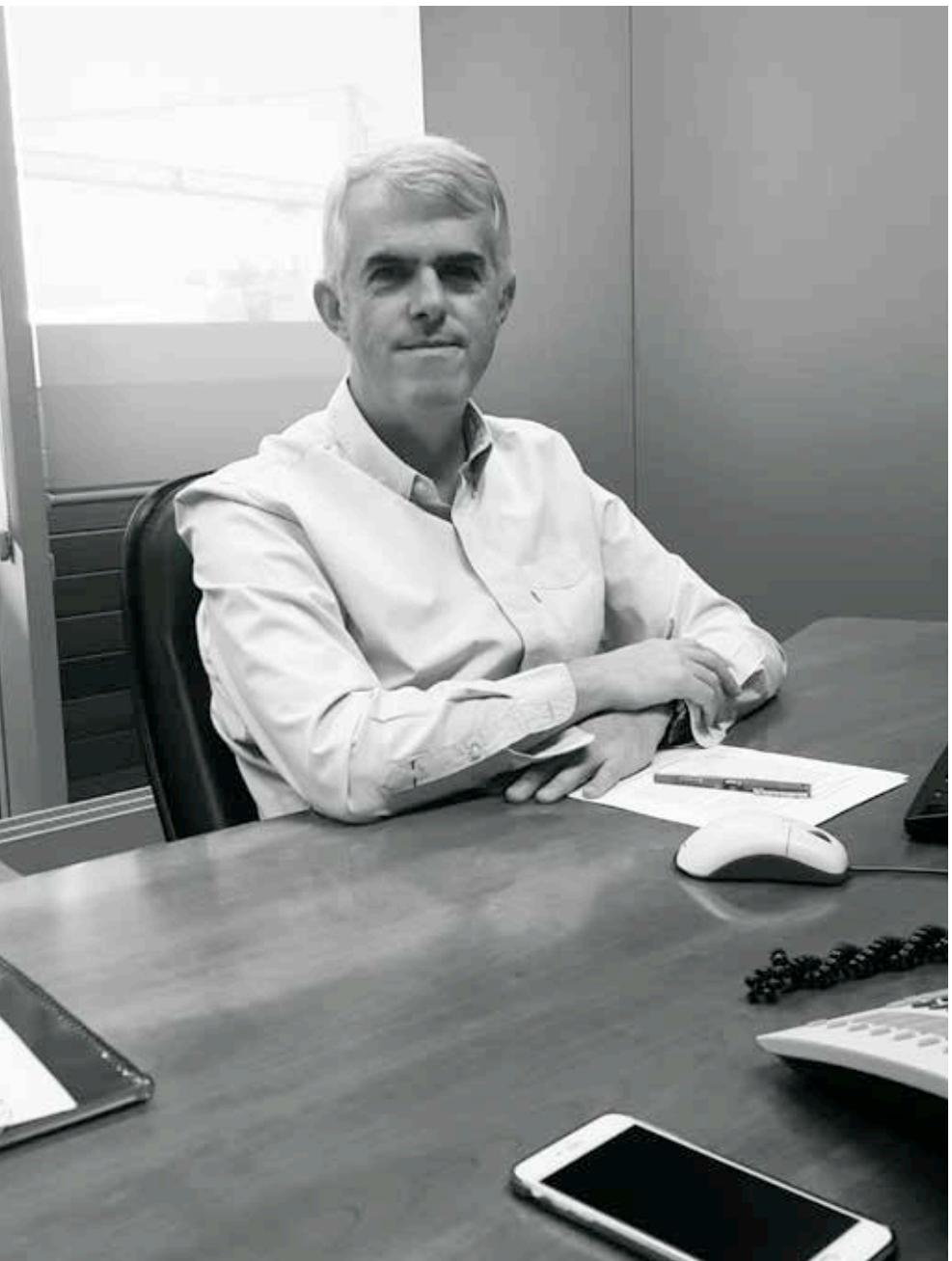
Por otra parte, también existen formas de defensa más activa, más cotidiana. Usando las similitudes, a veces fecundas, entre la realidad digital y material, podríamos decir que sin acumulación de capital no hay empresa, ni formación de plusvalía, auténtico motor de toda forma de capitalismo. El capital digital viene dado en este caso por los datos que los usuarios producimos a través de nuestro *trabajo*, digo bien, como usuarios de las más diversas redes. El trabajo realizado por estos usuarios requiere de su concurrencia activa en una plataforma, concurrencia sin la cual no se generará *plusvalía digital*, medida esta en términos del tiempo de uso de una Red. Para ello, es preciso que cada plataforma consiga dar con una fórmula de éxito capaz de atraer un número considerable de usuarios. La formación de capital digital es muy barata en términos de recursos –capital y maquinaria– materiales, pero requiere el despliegue de mucha habilidad para alcanzar el régimen de producción que permita la generación de plusvalía digital. Una Red que no cruce el rubicón del millón de usuarios registrados y activos apenas podría considerarse como una genuina *red social* de la era digital. Y esos valores no son sencillos de alcanzar aunque sí de perder. Un paso en falso, una propaganda negativa creíble, una mala práctica o un sesgo muy reconocible puede hacer que una Red pierda rápidamente usuarios a favor de una alternativa eficiente. Pero tampoco hay necesidad de que los *señores de la red* lo hagan mal, basta con comportarnos de forma promiscua buscando alternativas eficientes que ralenticen, compliquen o vuelvan inestables los procesos de formación de capital digital. No toda búsqueda tiene por qué depender de Google, ni tenemos por qué centrar toda nuestra vida social en una sola Red, ni tampoco asociar toda nuestra identidad a un único perfil de usuario. Porque para impedir la formación de capital digital o moldearla a favor de un interés más general, no solo hay que resistirse a concurrir donde de hecho también concurre la mayoría, también cabe actuar sobre la fiabilidad o el valor de los datos.

En fechas recientes ha sido muy comentado el fiasco de Google en torno a las predicciones de la incidencia de epidemias de gripe en Estados Unidos. Supuestamente era fácil hacerse una idea muy precisa prestando atención a las consultas que los usuarios practicaban en su buscador y en las redes sociales del gigante, pero no fue así. Quizá aquellos que se encuentran afectados no tengan ganas de encender el ordenador salvo para ver una película relajante o imaginar unas vacaciones de ensueño... quizá sea por otro motivo, pero lo cierto es que falló. También podemos imaginar que no lo hizo y que, al mejor estilo *conspiranoico*, se nos hizo creer que sus predicciones fueron fallidas para apartar sus éxitos de nuestra mirada, pero no lo creo. Quizá sea debido a mi formación en el campo de la lógica matemática, pero he visto muchas veces como los intentos de manejar una totalidad de entidades del tipo que sea manteniéndola bajo estrictos criterios de control suele llevar a paradojas y limitaciones que ponen nuestras ambiciones en su sitio. Aún no se ha demostrado que el *Big Data* encierre en sí mismo un proyecto imposible, quizá no se haga nunca, pero tampoco creo que llegue a dominar nuestras mentes y voluntades a través de la acción de un *Gran Hermano* tecnológico. Como en otras ocasiones en la Historia de la humanidad eso es algo que solo ocurrirá si *We the People* así lo permitimos, y quiero creer que seremos lo suficientemente maduros y conscientes como para que así lo hagamos. Sirva lo que viene a continuación para que cada cual saque sus conclusiones.

Enrique Alonso

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

El camino desde la Inteligencia Artificial al *Big Data*

Antonio Berlanga
Doctor en Ingeniería Informática.
Profesor Titular de Ciencias de la Computación e Inteligencia Artificial.
Dpto. de Informática de la Universidad Carlos III de Madrid.

Un recorrido por el desarrollo de la Inteligencia Artificial y sus paradigmas, en especial el Aprendizaje Automático, que va a proveer de muchos de los algoritmos que se aplican en el tratamiento automático de grandes volúmenes de datos no estructurados, conocido con el término de "*Big Data*".

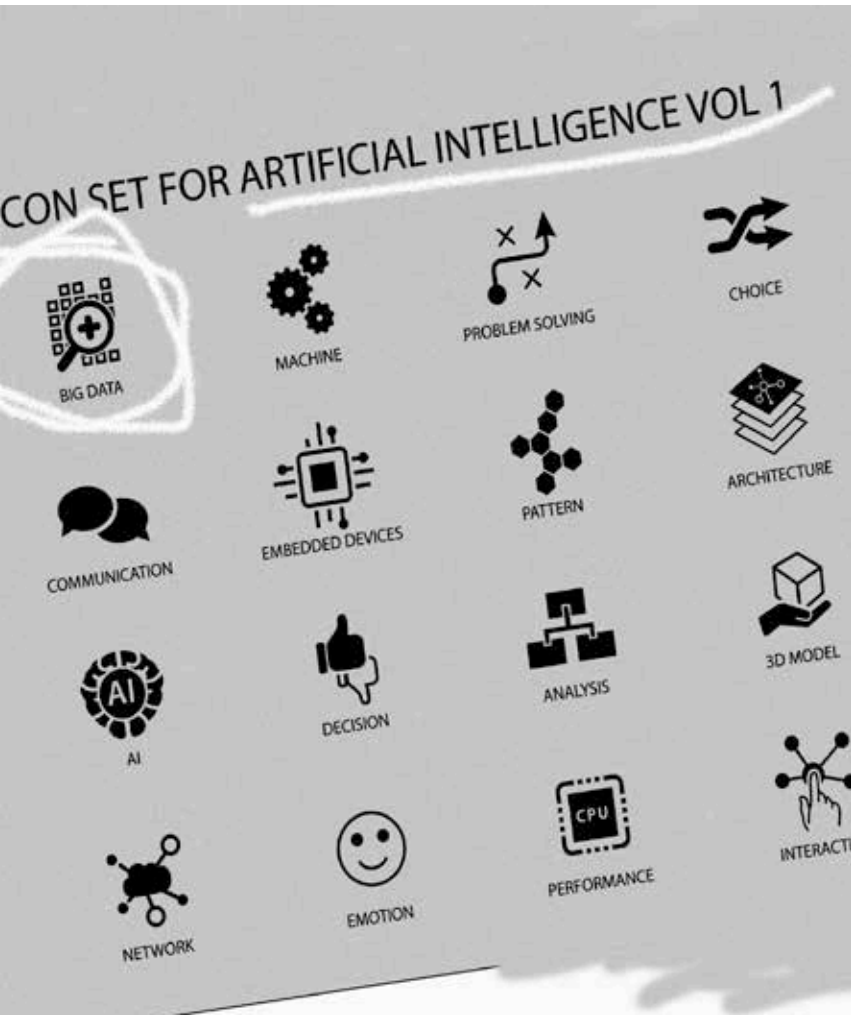
Se cumplen 60 años desde que John McCarthy acuñó el término "Inteligencia Artificial"[1]. Fue durante la reunión de 10 jóvenes investigadores, organizada por el propio McCarthy en Dartmouth, interesados en estudiar los aspectos de la inteligencia que podían simularse algorítmicamente[2]. Se anunció a su término, nada más y nada menos, el nacimiento de una nueva ciencia. El perfil académico y la carrera profesional de sus creadores marcaron su desarrollo posterior que alcanza el presente. Física, biología, economía, psicología cognitiva y en especial matemática e ingeniería, fueron las ramas de conocimiento embrionarias que han desarrollado una, todavía aún más interdisciplinaria ciencia. En las décadas siguientes, esta nueva disciplina experimentó un crecimiento extraordinario, fruto del interés que suscitó tanto académico como empresarial. Se intuían grandes posibilidades de aplicación y muy pronto también, se apreciaron sus limitaciones.

Uno de los criterios por los que se clasifican las técnicas de Inteligencia Artificial (IA) obedece a la forma cómo se construyen. Así, se distinguen entre las técnicas que utilizan una aproximación "*top-down*" de las que utilizan un esquema "*bottom-up*". Las técnicas "*top-down*" se basan en la llamada "*The physical symbol system hypothesis*"[3] que postula que un sistema físico puede realizar acciones inteligentes si está dotado de la apropiada representación simbólica de conocimiento. Mediante procesos, esta representación puede combinarse para obtenerse estructuras complejas de símbolos, estando estos procesos representados asimismo mediante símbolos. Este enfoque dio lugar a las aproximaciones basadas en la lógica y sistemas de reglas que tuvo en los sistemas expertos su máximo exponente[4]. Un sistema experto requiere de la incorporación del conocimiento que tienen especialistas en el dominio del problema que se va a resolver y de un sistema que automáticamente pueda modificar y ampliar este conocimiento. Los primeros inconvenientes para realizar una aplicación práctica de los sistemas expertos surgen de esos requerimientos. Por un lado, la forma de representar el conocimiento introducirá un sesgo sobre el alcance del razonamiento; por otro, los especialistas deben aportar conocimiento completo y libre de errores, ya que el sistema será incapaz de corregirlos salvo en casos de ambigüedades triviales. Otro problema surge de la dificultad para evaluar el desempeño de los sistemas expertos[5]. Pero el problema crítico se encuentra en la construcción del algoritmo que permite extraer nuevo conocimiento. El sistema experto razona con conocimiento de alto nivel, pero generalmente opera con datos de muy bajo nivel, a menudo con poca o ninguna estructura. Hacer la transición de los datos al cono-

> *Cuando las técnicas que se aplican sobre los datos proceden de la estadística cuantitativa y cualitativa clásica, se usa el término* "Data Analytics", *reservando* "Big Data" *para cuando se utilizan técnicas inductivas, las propias del Aprendizaje Automático.*

cimiento es una tarea sin cerrar y objeto de interés por parte de la comunidad de la inteligencia artificial simbólica y otras áreas de conocimiento afines. Este enfoque dominó la inteligencia artificial desde los años 50 hasta finales de los 80 del siglo pasado. Se desarrollaron muchos sistemas expertos, aplicándose a entornos donde la información está bien estructurada y el conocimiento puede organizarse, de forma natural, con sistemas de reglas, como por ejemplo en el área del diagnóstico médico, en ingeniería para el control y test de sistemas automáticos, en economía para la detección de fraude y cálculo de riesgo crediticio, etc.[6-10].

En el declive de la Inteligencia Artificial simbólica participó la emergencia del enfoque "bottom-up". La idea de modelar el razonamiento por deducción, similar al humano, es reemplazado por la imitación de pequeños comportamientos inteligentes, obtenidos por inducción, de los que pueden emerger comportamientos cada vez más complejos. Se rescata un campo de investigación formulado a finales de los años 60, el Aprendizaje Automático[11], que estudia la construcción de algoritmos que pueden extraer conocimiento a partir de conjuntos de datos. El resultado de un algoritmo de Aprendizaje Automático va a ser un modelo que explica las regularidades existentes en los datos. Los árboles y reglas de decisión, redes bayesianas, sistemas ocultos de Markov, clasificadores lineales, algoritmos evolutivos, máquinas de soporte vectorial, redes de neuronas artificiales son algunos de los algoritmos investigados y aplicados desde el Aprendizaje Automático. Todos comparten la característica de contar con métricas que cuantifican la bondad del modelo que construyen, ya sea para clasificar, agrupar o predecir. En función de esa medida reajustan los parámetros del modelo con el fin de optimizar su desempeño. Un elemento clave es el requisito de obtener modelos generales. Un modelo general inducido a partir de un conjunto de datos permite ser aplicado sobre otros datos distintos. Si tomamos como ejemplo la construcción de un sistema de reconocimiento de caracteres manuscritos, el modelo deberá reconocer otros caracteres en textos que no han sido utilizados para su construcción. Por tanto, se presenta la cuestión acerca de cómo debe ser el conjunto de datos, tanto en tamaño y características. Siguiendo con el ejemplo del reconocedor de caracteres, utilizar textos en castellano sesgará el modelo, reconociendo con diferente acierto los distintos caracteres, ya que hay letras con una frecuencia de aparición muy alta frente a otras de muy baja. Respecto del tamaño de la muestra, se observó que una técnica pobre con un conjunto de datos grande tendría mejor comportamiento que una buena técnica con pocos datos[12].



A finales de los años 80 no se disponía de grandes recursos para computación y el almacenamiento de volúmenes masivos de datos a bajo coste, lo que en gran parte (también aparecieron problemas formales en algunas técnicas que posteriormente se han superado) hizo que las técnicas de Aprendizaje Automático no se popularizasen más allá del entorno académico. Simultáneamente, se desarrollaron metodologías para poder extraer conocimiento de las bases de datos. El estudio de esas metodologías adoptó el nombre de Descubrimiento de Conocimiento en Bases de Datos, o más conocido por sus siglas en inglés KDD *(Knowledge Discovery in Databases)*. Un paso en las metodologías implicaba la aplicación de técnicas que revelasen estructuras y relaciones entre los datos, fue llamado Minería de Datos[13]. La Minería de Datos comparte una gran cantidad de técnicas con el Aprendizaje Automático, a tal forma que resulta, hoy día, difícil realizar una distinción clara entre ambas. El consenso actual es considerar al Aprendizaje Automático como el es-

tudio de técnicas, con las características que han sido mencionadas anteriormente, que pueden incorporarse a un proceso de Minería de Datos[14-15]. Durante las décadas siguientes, la Minería de Datos fue incorporándose, primero desplazando, en las áreas en las que se habían aplicado, a los sistemas expertos y, posteriormente con la explosión en los 90 de internet y los sistemas de información, a una gran variedad de actividades[16].

En 1997, se publica el primer artículo en una conferencia internacional en el que se realiza una definición del término "*Big Data*"[17]. Hace referencia al problema de tener que procesar un conjunto de datos con un tamaño superior a la memoria del ordenador y al del almacenamiento en disco, incluso si este es remoto. Empiezan a emerger situaciones en las que la acumulación de datos es tan grande que es necesario definir nuevos procedimientos para poder aplicar las técnicas de extracción de conocimiento. Pocos años después, en 2001 , se realiza una definición para "*Big Data*" que ha sido ampliamente aceptada[18], aunque ampliada y revisada con posterioridad. Conocida como las "3Vs", hace referencia a las características de volumen, velocidad y variedad de los datos. Cuando las técnicas que se aplican sobre los datos proceden de la estadística cuantitativa y cualitativa clásica, se usa el término "*Data Analytics*", reservando "*Big Data*" para cuando se utilizan técnicas inductivas, las propias del Aprendizaje Automático.

Es a partir de 2010 cuando el interés acerca del "*Big Data*" crece de forma exponencial[19]. El coste de almacenamiento ha caído en 10 años casi a su milésima parte y el de la computación a la centésima. Han surgido nuevos conceptos de aplicación que hacen un uso masivo de datos no estructurados; "*Smart Cities*", "Internet de la cosas", "*Smart Health*", "Industria 4.0" son solo algunos ejemplos[20-22]. Las grandes corporaciones internacionales anuncian aplicaciones y servicios basados en "*Big Data*"; cualquier usuario de redes sociales, sin saberlo, está haciendo uso de las facilidades y recursos que proporcionan estas técnicas.

El "*Big Data*" está llamado a revolucionar el mundo como lo hizo Internet, tendrá que evolucionar, como lo hizo la red de redes. Hay muchos desafíos a futuro[23], uno muy importante será el de crear especialistas en este campo con una formación académica híbrida entre la estadística, las ciencias de la computación, los sistemas de información, la computación de altas prestaciones; en definitiva multidisciplinar, tal como lo fue en su origen la Inteligencia Artificial.

Referencias

1. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach. Third edition*. 2014.
2. R. R. Kline, "Cybernetics, automata studies, and the dartmouth conference on artificial intelligence," *IEEE Ann. Hist. Comput.*, vol. 33, no. 4, pp. 5–16, 2011.
3. A. Newell, "Physical Symbol Systems*," *Cogn. Sci.*, vol. 4, no. 2, pp. 135–183, Apr. 1980.
4. P. V. S. Ponnapalli, "Introduction to expert systems. 2nd edn, P. Jackson, Addison-Wesley, Wokingham, 1990, ISBN 0-201-17578-9, xi + 526pp. £21.95," *Int. J. Adapt. Control Signal Process.*, vol. 6, no. 1, pp. 65–67, Jan. 1992.
5. J. Biegel and U. G. Gupta, "Expert systems in manufacturing: Promises and perils," *Comput. Ind. Eng.*, vol. 19, no. 1–4, pp. 127–130, Jan. 1990.
6. P. SZOLOVITS, R. S. PATIL, and W. B. SCHWARTZ, "Artificial Intelligence in Medical Diagnosis," *Ann. Intern. Med.*, vol. 108, no. 1, p. 80, Jan. 1988.
7. Z. Z. Zhang, G. S. Hope, and O. P. Malik, "Expert systems in electric power systems-a bibliographical survey," *IEEE Trans. Power Syst.*, vol. 4, no. 4, pp. 1355–1362, 1989.
8. B. K. Wong and J. A. Monaco, "Expert system applications in business: A review and analysis of the literature (1977–1993)," *Inf. Manag.*, vol. 29, no. 3, pp. 141–152, Sep. 1995.
9. Shu-Hsien Liao, "Expert system methodologies and applications—a decade review from 1995 to 2004," *Expert Syst. Appl.*, vol. 28, no. 1, pp. 93–103, 2005.
10. T. Gupta and B. K. Ghosh, "A survey of expert systems in manufacturing and process planning," *Comput. Ind.*, vol. 11, no. 2, pp. 195–204, Jan. 1989.
11. T. M. Mitchell, *Machine Learning*. McGraw-Hill, Inc., 1997.
12. P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012.
13. J. Han, M. Kamber, and J. (Computer scientist) Pei, *Data mining: concepts and techniques*. Elsevier/Morgan Kaufmann, 2012.
14. I. Bose and R. K. Mahapatra, "Business data mining — a machine learning perspective," *Inf. Manag.*, vol. 39, no. 3, pp. 211–225, 2001.
15. I. H. (Ian H. . Witten and E. Frank, *Data mining : practical machine learning tools and techniques*. Morgan Kaufman, 2005.
16. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.
17. M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, pp. 235–244, 1997.
18. D. Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety," 2001.
19. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, pp. 137–144, 2015.
20. M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *Eur. Phys. J. Spec. Top.*, vol. 214, no. 1, pp. 481–518, Nov. 2012.
21. M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.
22. S. Yin and O. Kaynak, "Big Data for Modern Industry: Challenges and Trends [Point of View]," *Proc. IEEE*, vol. 103, no. 2, pp. 143–146, Feb. 2015.
23. W. Fan and A. Bifet, "Mining big data," *ACM SIGKDD Explor. Newsl.*, vol. 14, no. 2, p. 1, Apr. 2013.

La revolución del *Big Data*

Esperanza Ibáñez Lozano
Manager. Public Policy and Gov't Relations. Google

A diferencia de épocas pasadas, donde el valor de las innovaciones estaba en la sustitución de la fuerza del hombre por la de la máquina, el siglo XXI se caracteriza por tener como base de desarrollo el conocimiento. El núcleo de ese conocimiento son los datos masivos o *Big Data*, entendido como el análisis de grandes volúmenes de datos, generalmente dispersos y poco estructurados, que provienen de muy diversas fuentes (sensores, comentarios, imágenes, textos...), y que sirven a la sociedad para, de forma novedosa, resolver problemas que se plantean en diferentes ámbitos y/u "obtener percepciones útiles o bienes y servicios de valor significativo"[1].

Un ejemplo de la aplicación práctica de los datos fue *Google Flu Trends*. Basándose en las búsquedas que la gente realizaba en Internet relacionadas con los síntomas de la enfermedad, un equipo de ingenieros de Google trabajó en el desarrollo de un modelo para predecir cómo se propagaría la gripe en Estados Unidos. Para ello, buscaron correlaciones entre las búsquedas de información y los avances de la gripe por el territorio americano, en función del tiempo. Después de utilizar millones de modelos matemáticos, encontraron la fórmula precisa: las predicciones que obtenían utilizando este método y las cifras que las instituciones públicas ofrecían sobre la enfermedad eran prácticamente idénticas. Con una diferencia: si bien este modelo permitía predecir qué podía ocurrir con anterioridad a que sucediera, las autoridades públicas no podían obtener los datos hasta que la enfermedad se había propagado durante al menos una o dos semanas.

No se trata de la cantidad de información de la que se disponga. Lo importante es cómo se materializan esos datos en productos y servicios, públicos o privados, y si aportan un verdadero valor a los ciudadanos

Aunque en este caso el ejemplo corresponde a Google, si algo caracteriza a este tipo de innovaciones es que no es exclusivo de grandes empresas. La innovación en base a los datos es posible y accesible para todo aquél que tiene interés. En primer lugar, no son un bien privativo y exclusivo de aquél que los utiliza por primera vez. Si en el siglo XX el petróleo que un fabricante utilizaba para su industria impedía a otros hacer uso de ese mismo recurso, en el siglo XXI los datos que uno emplea no limitan a otros para hacer uso de esos mismos datos.

En segundo lugar, los datos pueden procesarse una y otra vez, para unos u otros fines, sin que por ello se desgasten, o pierdan valor.

Y en tercer lugar, como ponen de manifiesto las profesoras Lambrecht de la London Business School y Tucker del MIT, no se trata de la cantidad de información de la que se disponga. Lo importante es cómo se materializan esos datos en productos y servicios, públicos o privados, y si aportan un verdadero valor a los ciudadanos[2]. Ellas citan como ejemplo de innovación creativa a WhatsApp, y hablan de cómo una empresa que a priori no disponía de la cantidad de información que sí tenían competidores que llevaban operando mucho tiempo en el sector de la mensajería instantánea, es capaz de abrirse hueco al ofrecer nuevos beneficios a los consumidores.

La capacidad de almacenamiento vs. coste también ha evolucionado de manera considerable durante los últimos años. Si antes este tipo de servicios estaban reservados para quienes podían costearlo, hoy almacenar grandes cantidades de información tiene un coste muy bajo, lo que se traduce en que cualquier institución, sin perjuicio de su naturaleza o tamaño, tiene la posibilidad de convertir datos en conocimiento para elaborar y desarrollar sus estrategias.

Los beneficios económicos que de todo ello se derivan son considerables, hasta el punto de que algunos estiman que este tipo de innovaciones podrían mejorar el PIB europeo en un 1,9%, que es el equivalente a un año de crecimiento económico en la UE[3]. Asimismo, la creciente adopción del *Big Data* tiene impacto en el empleo. Según datos de la Comisión Europea, de aquí al año 2020 habrá más de 825.000 puestos de trabajo sin cubrir como consecuencia de la falta de profesionales que reúnan las competen-

cias que la digitalización de la economía demanda[4], siendo la analítica de datos uno de ellos.

En definitiva, todos los sectores (telecomunicaciones, bancos, moda, viajes, alimentación...) necesitan personas capaces de trabajar la maraña de datos que existen, analizarlos y convertirlos en información relevante, en información útil. También se necesitan profesionales que velen porque estos procesos se realicen siempre de acuerdo con la normativa de privacidad, respetando los derechos fundamentales de los ciudadanos. La Asociación Internacional de Profesionales de la Privacidad calcula que serán unos 28.000 nuevos puestos de trabajo los que se creen ligados al desarrollo del Reglamento de Protección de Datos Europeo. Y esto no ha hecho más que empezar.

Los beneficios sociales que de la aplicación de los datos se derivan también son relevantes. Como relatan los profesores Mayer-Schönberger y Kenneth Cukier en su obra "*Big Data*. La revolución de los datos masivos", hace unos años, IBM realizaba un experimento con los coches eléctricos. Entre otras muchas necesidades logísticas, este tipo de vehículos necesitan ser recargados. Pero el suministro no es infinito, por lo que se precisa una importante labor para que las cargas no desequilibren la red de suministro. Utilizando un complejo modelo predictivo, basado en numerosos factores como batería del coche, localización, hora del día... IBM recopiló y analizó grandes cantidades de datos, históricos y en tiempo real, de numerosas fuentes para determinar los momentos y lugares óptimos para que los conductores cargasen las baterías de sus vehículos, de forma sostenible y sin riesgos ni para las ciudades, ni para los ciudadanos.

Los datos son también la base de la denominada "Inteligencia Artificial". Máquinas con capacidad de computación, programadas por el hombre, que se nutren de información y aprenden a base de repeticiones, de prueba-error, y que son capaces de realizar un trabajo de forma autónoma. El profesor Grilo, en el estudio titulado "Game Changers. Surfing the wave of technology disruption", cifra en un impacto económico anual de aquí a 2025, de entre 5,2 y 6,7 billones de euros los próximos desarrollos en la materia, ligados principalmente a incrementos de productividad de los trabajadores.

Y es que, a pesar de que el debate no es pacífico, como ha ocurrido en otros momentos de la historia, cada vez que una innovación disrumpe en la escena, la tecnología trae consigo nuevas formas de relación entre el trabajador del conocimiento y la máquina, que requiere nuevas capacitaciones y formación, como antes se ha referido.

El desarrollo de la Inteligencia Artificial se ha materializado en aplicaciones como instrumentos y equipos que garantizan la perfección en cirugías que no admiten error ni desviación alguna. Ha permitido la creación de exoesqueletos, facilitando que personas que no podían caminar, vuelvan a andar. El análisis de los datos ha permitido entender cómo reaccionan distintos enfermos frente a diferentes medicamentos, identificando el más efectivo según el caso.

> *Hoy almacenar grandes cantidades de información tiene un coste muy bajo, lo que se traduce en que cualquier institución, sin perjuicio de su naturaleza o tamaño, tiene la posibilidad de convertir datos en conocimiento para elaborar y desarrollar sus estrategias*

A lo largo de este artículo se ha hecho referencia a algunos de los importantes avances que se han producido en los últimos años ligados a la explotación del *Big Data* y a la Inteligencia Artificial, así como a las características de esta nueva revolución –abundancia, no exclusividad, accesibilidad de las herramientas para su explotación– que no hacen más que reforzar lo que Yossi Vardi vino en denominar "la democratización del derecho a innovar". Nuevas formas de desarrollo que se encuentran en un estadio muy inicial y que seguirán traduciéndose en beneficios económicos y sociales, para lo que será imprescindible repensar la formación de los trabajadores actuales y de las futuras generaciones, además de resultar en nuevos retos que exigirán soluciones creativas que permitan conjugar evolución y marco legal aplicable de una manera eficiente y en beneficio de todos.

Referencias

1. Mayer-Shönberger, V. & Cukier, K. *Big Data. La revolución de los datos masivos.*
2. Lambrecht, Anna y Tucker, Catharine (2015), *Can Big Data Protect a Firm from Competition?*
3. European Big Data Value Strategic Research & Innovation Agenda, V. 99 (2014).
4. Grand Coalition for Digital Jobs, EC (2016): https://ec.europa.eu/digital-single-market/en/grand-coalition-digital-jobs .

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 estructura[1].

El primero de los Principios Fundamentales de las Estadísticas Oficiales de Naciones Unidas[2] 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 alto[3].

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

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.

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.

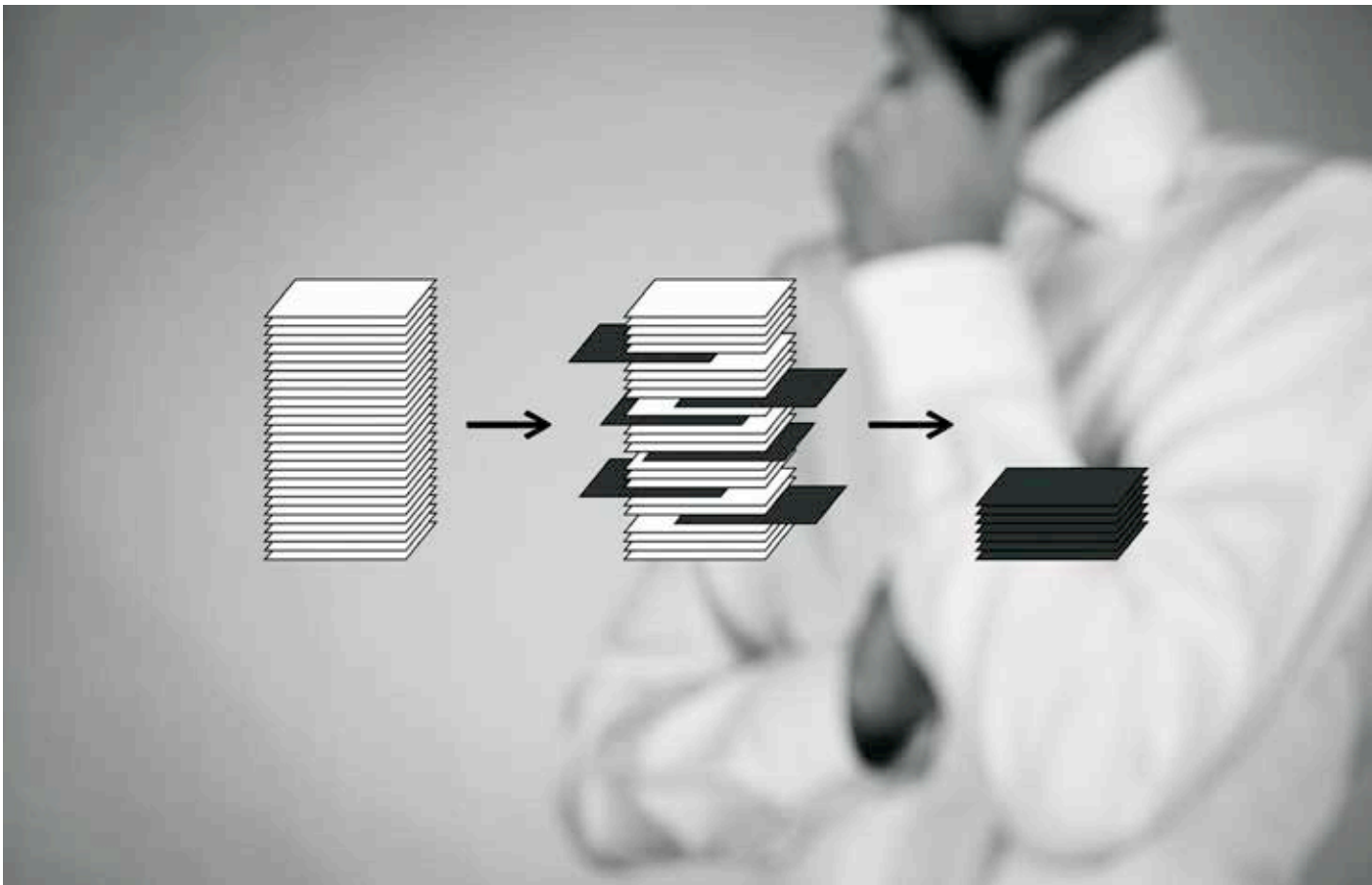
La mayoría de los *Big Data* son generados, almacenados y procesados por corporaciones privadas como resultado de su actividad de negocio, siendo datos (i) con información sobre terceras personas (usualmente sus clientes), (ii) no preparados para la explotación estadística sin un procesamiento previo, 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 auxiliar[5].

La ventaja del uso de los diseños muestrales reside 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 interés[6]. 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 Prácticas[8], 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 procesos[9]. 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 2020[11]. 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* concretas[12]. 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/ad949-3D-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%202%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

Aspectos legales del *Big Data*

Carlos Pérez Sanz
Socio. Ecija Abogados

INTRODUCCIÓN

Concepto de *Big Data*

Son muchas las definiciones que podemos encontrar sobre el concepto de *Big Data*, pero todas ellas tienen en común describir a este fenómeno tecnológico como el conjunto de recursos que permiten la gestión y análisis de cantidades ingentes y masivas de datos, con un alcance y dimensiones en constante crecimiento y sin parangón en la historia de la tecnología.

En la definición del *Big Data*, también se acepta comúnmente su caracterización a través de las denominadas "las 3 uves": volumen, variedad y velocidad.

De entre los recursos que integran el fenómeno *Big Data*, los siguientes son imprescindibles:

- **Recursos humanos**, cuya aportación creativa es muy variada. Esta requiere, en primer lugar, personas con amplios conocimientos técnicos, que pueden suponer, desde las tareas más elevadas, las del denominado "*data scientist*" o científico de datos, cuya labor consiste en la concepción y desarrollo de algoritmos que aportan valor añadido diferencial al proyecto de *Big Data* en cuestión, hasta la mera tarea de depuración de datos, pasando por la definición de objetivos y finalidades del proyecto o por la elección de las mejores técnicas de combinación o de anonimización de datos. En segundo lugar, el proceso de tratamiento de datos y la implementación de un proyecto *Big Data* requiere personas con un gran conocimiento del negocio y del sector en el que el proyecto vaya a operar.
- **Infraestructura tecnológica**, integrada por arquitectura hardware y software con dimensión y potencia suficientes y que, actualmente, suelen incluir soluciones de computación en la nube, programas de ordenador de gestión, selección y presentación de los datos, así como los algoritmos que aportan la lógica diferencial del proyecto.
- **Fuentes de datos**, cuya variedad, accesibilidad y capacidad de combinación e interrelación es, en el estadio tecnológico actual, de dimensiones enormemente superiores a las disponibles en estadios tecnológicos anteriores. Dichas fuentes incluyen los sistemas de captación de información, las bases de datos y repertorios históricos de datos propios del impulsor del proyecto *Big Data*, así como fuentes de datos ajenas a dicho impulsor (Internet, redes sociales, información puesta a disposición general por las administraciones públicas, etc...).

La privacidad es uno de los aspectos fundamentales que deben considerarse a la hora de analizar los proyectos Big Data *desde un punto de vista jurídico*

Elemento diferencial

El elemento diferenciador del fenómeno de *Big Data* es su **capacidad predictiva**. Hasta la aparición y consolidación del fenómeno *Big Data*, y por las limitaciones propias del estado de la técnica del momento, los proyectos relativos a inteligencia de datos se limitaban, en gran medida, a analizar la información histórica propia disponible, dándole orden y sentido, con el fin de poder diseñar de modo más eficaz los propios procesos productivos o de toma de decisiones.

Por contra, los proyectos *Big Data*, por el enorme desarrollo, potencia y capacidad combinativa de los diversos recursos involucrados, permiten a su impulsor anticiparse a acontecimientos futuros y, en algunos casos, predecirlos con poco margen de error. Siguiendo con el ejemplo anterior, un proyecto *Big Data* ayudaría a la organización a analizar el carro de la compra de sus

clientes y predecir si uno de ellos comenzará a aumentar su consumo de cerveza y pizza en función de su estilo de vida, o si por el contrario sus compras anteriores indican que vive en pareja y podrá renovar pequeños electrodomésticos como la batidora. Este conocimiento tan detallado permite a nuestra compañía, por ejemplo, enviar promociones personalizadas antes que los competidores, adelantándose así al deseo de compra de los clientes y aumentando las ventas anuales.

Objetivos principales

Las finalidades perseguidas por un proyecto *Big Data* pueden ser muy variadas. Ahora bien, la práctica totalidad de dichos proyectos pueden incluirse en alguna de las siguientes categorías, en función del objetivo principal perseguido:

- **Inteligencia comercial**, que engloba a todos aquellos proyectos *Big Data* que persiguen un conocimiento profundo y predictivo de los clientes que interactúan con el impulsor del proyecto. Estos proyectos buscan mejorar la experiencia del cliente en su relación con el impulsor, prevenir y evitar que el cliente corte la relación con el impulsor, generar patrones de comportamiento y segmentaciones avanzadas de clientes o la oferta personalizada de productos, servicios y precios.
- **Fraude y riesgo**: consistentes en la previsión y predicción de posibles actividades fraudulentas que redunden en perjuicio del impulsor del proyecto, desde las de sus propios clientes y proveedores, hasta de otros elementos externos (ciberseguridad o prevención de fallos de sistema).
- **Eficiencia operativa**: encaminados a optimizar los procesos productivos o de toma de decisiones, incluyendo proyectos asociados al Internet de las Cosas, eficiencia en la gestión logística y de flotas, o de optimización en la gestión de reclamaciones y devoluciones.
- **Monetización**: por último, están los proyectos *Big Data* que persiguen un beneficio económico directo, a través de la generación de nuevos negocios en entorno digital, apps, empresas *fin-tech*, intermediación en *e-commerce*, etc...

Sobre las anteriores premisas conceptuales, se analizarán los aspectos más importantes del fenómeno *Big Data* desde el punto de vista jurídico.

> *Existen muchas y muy variadas metodologías de anonimización, y es muy posible que sea necesario combinar varias de ellas para asegurar la completa disociación en un proyecto* Big Data

ASPECTOS JURÍDICOS PRINCIPALES

Privacidad y protección de datos personales

Es indiscutible que la privacidad es uno de los aspectos fundamentales que deben considerarse a la hora de analizar los proyectos *Big Data* desde un punto de vista jurídico, se analizan a continuación los aspectos más importantes de este ámbito normativo y su impacto en proyectos de esta naturaleza.

Finalidades legítimas

Uno de los principios orientadores de la normativa de privacidad es el de adecuación a las finalidades de tratamiento, según el cual los datos personales no pueden tratarse para finalidades incompatibles para aquellas que fueron informadas a los interesados en el momento de recabar sus datos.

Buena parte de los proyectos *Big Data* corre el riesgo de incumplir dicha salvaguarda para la privacidad personal, bien porque la combinación de los datos ya disponibles con nuevas fuentes de datos adicionales permite generar perfiles personales sin que exista **consentimiento previo** para ello, bien porque los objetivos del proyecto conducen a usos de datos que no eran razonablemente previsibles para los interesados en el momento inicial de obtención de sus datos. Estas situaciones de transgresión a la privacidad conllevan el correspondiente riesgo sancionador, incrementado de forma muy sustancial con el nuevo Reglamento Europeo de Protección de Datos que, si bien no será de aplicación hasta el 25 de mayo de 2018, prevé sanciones administrativas muy superiores a las actuales, que pueden alcanzar los 20 millones de euros o el 4% de la facturación total anual de la compañía.

Para superar estas situaciones de riesgo, las autoridades europeas en materia de protección de datos recomiendan someter el proyecto *Big Data* a un test de incompatibilidad, en el cual se sometan a prueba todos los elementos en juego:

finalidades informadas, contexto de obtención de datos, finalidades ulteriores del propio proyecto *Big Data*, tipología de datos involucrada y su impacto sobre los interesados conforme a las finalidades ulteriores del proyecto *Big Data*, etc. Dicho test se superará con éxito si se cumple alguna de las siguientes condiciones:

- Que las finalidades del tratamiento de datos del proyecto *Big Data* se ajusten a lo informado a los interesados en el momento inicial de recabar sus datos; o bien
- Que las finalidades del tratamiento de datos del proyecto *Big Data* sean razonablemente previsibles para los interesados, aun no habiendo sido explícitamente informados en el momento de obtener sus datos; o bien
- El tratamiento de datos resultante del proyecto *Big Data* está justificado por otras causas de legitimación previstas en la normativa de privacidad (como son, por ejemplo, el interés legítimo del responsable del tratamiento, el cumplimiento de obligaciones legales, contractuales, o en atención al interés vital de los interesados).

En caso de superarse con éxito este test de incompatibilidad, el proyecto de *Big Data* podrá considerarse conforme a la normativa de protección de datos, sin perjuicio del cumplimiento de otras obligaciones previstas por dicha normativa (como son la aplicación de medidas de seguridad, o la realización de análisis previos de impacto y su aprobación por las autoridades nacionales de protección de datos si el proyecto implica la generación de perfiles personales).

Ahora bien, si el test de incompatibilidad es negativo, el proyecto deberá sujetarse a información y consentimiento previos de los interesados involucrados, o bien someterse a procesos de anonimización y disociación de los datos.

Anonimización de datos

No todos los proyectos *Big Data* involucran tratamientos de datos de carácter personal, vinculados a personas físicas identificadas o identificables. Algunos no tienen como objetivo en absoluto a personas ni individuos, en otros la identidad de los individuos involucrados es completamente irrelevante, y en otros se hace necesario prescindir de la identidad de la persona por no haberse superado con éxito el test de incompatibilidad antes descrito.

Para las dos últimas situaciones, la aplicación de **técnicas de anonimización y disociación de datos son la mejor solución** para asegurar que el proyecto *Big Data* queda fuera del ámbito de aplicación de la normativa de protección de datos y de cualquier riesgo sancionador por incumplimiento de dicha normativa.

Ahora bien, para asegurar lo anterior, es necesario que la técnica de disociación aplicada garantice que se producen los siguientes efectos:

- **Aislamiento**, de forma que los registros de información que identifican a las personas quedan totalmente aislados y separados del resto de información objeto del proyecto *Big Data*;
- **No vinculación**, de forma que, tras el aislamiento, sea imposible volver vincular la información global del proyecto con los registros que aportan la identificabilidad de la persona; y
- **No inferencia**, de forma que no sea posible llegar a deducir la identidad de la persona a la cual corresponden los datos objeto de tratamiento, tras haberse asegurado el aislamiento y la no vinculación.

Existen muchas y muy variadas metodologías de anonimización, y es muy posible que sea necesario combinar varias de ellas para asegurar la completa disociación en un proyecto *Big Data*. En cualquier caso, debe tenerse en cuenta que la anonimización debe asegurarse de principio a fin, es decir, en todos y cada uno de los procesos y tratamientos de información específicos asociados al proyecto en cuestión, por lo que la selección de la mejor metodología de anonimización dependerá de las circunstancias particulares de cada proyecto *Big Data*.

Propiedad intelectual de las bases de datos

Existe otro aspecto importante relacionado con estos proyectos, y es el relativo a la propiedad, titularidad y protección jurídica del proyecto *Big Data* en su conjunto, como de los diferentes elementos intangibles que lo integran.

La normativa sobre propiedad intelectual incluye dos niveles de derechos de propiedad intelectual para las bases de datos:

- **La protección idéntica al resto de derechos de autor**, y que se confiere a las bases de datos cuya especial selección y disposición de los elementos que las integran puedan ser consideradas en sí mismas como creaciones intelectuales; y

- El derecho "*sui generis*", que protege la inversión sustancial que realiza el fabricante de la base de datos, ya sea en medios financieros, empleo de tiempo, esfuerzo, energía u otros similares, para la obtención, verificación o presentación de su contenido, con independencia de que la selección o disposición de sus elementos pueda ser considerada o no como una creación intelectual.

En un proyecto *Big Data*, ambos tipos de titularidad y propiedad intelectual pueden llegar a coexistir, y pueden darse situaciones de conflicto de titularidad. Por un lado, el derecho *sui generis* sobre las bases de datos resultantes de un proyecto *Big Data* corresponderá, casi siempre, al impulsor del proyecto, por ser quien realiza el esfuerzo de financiación de todos los recursos asociados al mismo. Pero ello puede no ser así respecto al esfuerzo intelectual creativo relativo a la especial selección y disposición de los datos asociados al proyecto *Big Data*, esfuerzo que, por aplicación del principio de autoría consagrado por la normativa de propiedad intelectual, puede corresponder a la empresa consultora externa que realice dicha aportación creativa al proyecto.

Por dicho motivo, si el impulsor de un proyecto *Big Data* quiere asegurarse de que le corresponden en exclusiva todos los derechos de propiedad intelectual sobre todos y cada uno de los elementos y resultados del mismo, será recomendable que asegure dicha titularidad por vía contractual, en particular, en relación con empresas consultoras externas que participen en el diseño y ejecución del proyecto.

Protección jurídica de los algoritmos

Resulta paradójico que los algoritmos, que constituyen uno de los elementos que puede aportar mayor creatividad a un proyecto *Big Data*, **no dispongan de una protección jurídica específica**.

Los algoritmos no pueden considerarse programas de ordenador, puesto que no están expresados en secuencias de instrucciones en lenguaje de programación. Los algoritmos constituyen la secuencia ordenada de operaciones que se pretende realizar para conseguir un resultado concreto o resolver un problema específico, y, por lo tanto, constituyen las ideas o principios en los que se basan los programas de ordenador asociados a los proyectos *Big Data*. Dichas ideas y principios están expresamente excluidos de ser considerados como propiedad intelectual.

Por otra parte, y conforme a diversas decisiones adoptadas por la Oficina Europea de Patentes y Marcas, los algoritmos tampoco reúnen las condiciones para ser protegidos como patente.

Por lo tanto, la única forma de protección posible para los algoritmos es la disponible para los **secretos industriales y comerciales**, sometiendo su conocimiento a obligación de confidencialidad (aplicable tanto al personal propio del impulsor del proyecto *Big Data*, como a los consultores y proveedores externos que participen en el mismo), y recurriendo a las acciones civiles y penales disponibles en nuestro ordenamiento en caso de acceso, apropiación o difusión no autorizados. En este ámbito específico de protección es de esperar que la reciente Directiva 2016/943 de Protección de los Secretos Comerciales, y su futura incorporación a nuestro derecho, venga a reforzar los recursos judiciales disponibles para la adecuada protección jurídica de los algoritmos que formen parte de los elementos integrantes de proyectos *Big Data*.

El problema de la dimensionalidad

José A. Guerrero
CEO Datrik Intelligence, S.A.

Las tres características que mejor definen al *Big Data* son el volumen, la heterogeneidad y la velocidad de generación de los datos. Existe la creencia de que, en lo que respecta al volumen, siempre 'más datos es mejor', y suele ser cierto, aunque debemos matizar dicha aseveración: Un mayor número de observaciones redunda en un mejor modelo, pero un mayor número de variables no necesariamente lo hace.

Para entender el problema debemos reflexionar sobre el proceso de ajuste de un modelo predictivo. La palabra que mejor define este proceso es 'equilibrio'. Equilibrio entre la complejidad y la capacidad de generalización. Un modelo excesivamente complejo será capaz de captar toda la información pero inevitablemente hará lo mismo con el ruido existente. Como resultado, las predicciones que se realicen sobre un nuevo conjunto de datos serán mediocres. A esto se le denomina *overfitting* o **sobreajuste**.

Para evitar el sobreajuste fijaremos una serie de restricciones a la complejidad de los modelos, de forma directa, limitando la profundidad de los árboles de clasificación o regresión, o indirecta, añadiendo un factor de penalización que se sume al término de error de ajuste, como se hace en Regresión de Ridge o en máquinas de vectores de soporte. Entendemos por **regularización** a la inclusión de cualquier tipo de restricción a la complejidad de un modelo, y la selección del nivel de estas penalizaciones se determina mediante validación cruzada.

El **problema de la dimensionalidad** lo definiremos como los potenciales efectos negativos derivados del aumento del número de variables respecto a las observaciones. Una gran dimensionalidad se traduce frecuentemente en sobreajuste, ya que aumentan los grados de libertad del sistema. También, en el caso de que haya colinealidad entre los predictores, puede impactar en la convergencia de los algoritmos y la estabilidad de las soluciones. La regularización suele ser necesaria para acotar el problema, aunque a veces no es suficiente, necesitándose actuaciones directas sobre la dimensión de los predictores.

El problema de la dimensionalidad lo definiremos como los potenciales efectos negativos derivados del aumento del número de variables respecto a las observaciones

Realizaremos a continuación una revisión de los métodos de abordaje del problema. Estos se dividen en métodos de selección y de extracción. Los primeros seleccionan un subconjunto de variables con cierto criterio, mientras los segundos generan nuevas variables mediante una transformación de las originales.

MÉTODOS DE FILTRADO

Se encuentran entre los métodos de selección y se caracterizan por utilizar un criterio para seleccionar las variables que es independiente del algoritmo con el que se ajuste el modelo. Los ejemplos más básicos son los que utilizan criterios univariables, como la correlación o asociación de cada predictor con la variable respuesta, criterios de información mutua o test de contraste de hipótesis, uni o multivariables, para calcular el nivel de significación de la relación entre las variables.

Estos métodos tienen la ventaja de que son rápidos de ejecutar. Entre sus inconvenientes están que, en el caso de los univariables, rechacemos una variable cuyo efecto principal no sea significativo pero que pudiera tener una interacción con otra variable. Otra situación frecuente es que se incluyan en la selección varios predictores que estén muy correlacionados entre sí. Para evitar esto último se han desarrollado dos métodos que tienen en cuenta no solo la relación de las variables con la respuesta sino con el resto de variables.

Una de estas técnicas es **mRMR (mínima redundancia, máxima relevancia).** El objetivo es medir no solo la información mutua entre la variable y la respuesta (relevancia) sino también la información mutua entre los predictores (redundancia).

Otro método que comparte concepto aunque utiliza diferente técnica para su resolución es **QPFS**

(*Quadratic Programming Feature Selection*). La idea es transformar la selección de variables en un problema de optimización cuadrática, donde los términos de segundo orden corresponden a la interrelación de los predictores (matriz de covarianzas) y los términos lineales a las correlaciones con la variable respuesta.

MÉTODOS ENVOLVENTES

Son otros métodos de selección que, a diferencia de los métodos de filtrado, intentan seleccionar un subconjunto de variables que obtenga el mejor ajuste posible con un algoritmo determinado. Para intentar que estos subconjuntos sean lo más generalizable posible se utilizan algoritmos base que sean robustos y cuyos parámetros sean fáciles de sintonizar, en concreto *random forest* o modelos lineales.

Los clásicos **procedimientos de *stepwise*** se encuadrarían en esta categoría. Partiendo de todas las variables se determina en cada paso el candidato que menos contribución aporta al modelo, y se elimina. Alternativamente se puede construir de forma creciente, comenzando sin ninguna variable e incluyendo en cada iteración la que más aporte. El criterio para decidir qué variable entra o sale puede ser la importancia relativa por permutación en el caso de *random forest* o un F-test (Snedecor), basado en la descomposición de la suma de cuadrados (varianza explicada) en el caso de regresión múltiple.

Boruta es un método que utiliza *random forest* como algoritmo subyacente. La idea es generar en cada iteración una serie de variables sombra a partir de los predictores, copiando cada uno de ellos y permutando entre sí los elementos de cada nueva columna. Se ajusta un modelo por *random forest* y se calculan las importancias relativas de cada variable. Si una variable sistemáticamente queda por debajo de las sintéticas (ruido), será indicativo de que su aportación al modelo será dudosa y por tanto se elimina. El proceso continua hasta que todas las variables son aceptadas, rechazadas o se alcanza un número de iteraciones límite.

Otra alternativa que podemos encuadrar en esta categoría es la ***Regresión Lasso***. Cuando hacemos una regresión múltiple con un gran número de variables los coeficientes tienen tendencia a aumentar de tamaño, lo cual es signo de sobreajuste. La regresión con regularización de Ridge utiliza un término de penalización sobre la norma L2 de los coeficientes de manera que se controla dicho sobreajuste. La regresión de Lasso es similar a la de Ridge, pero utiliza la norma L1 para la penalización. Uno de los efectos que tiene este cambio es que conforme aumentamos el factor de penalización, algunos de los coeficientes de las variables se optimizan en cero, y a partir de ese punto, si seguimos aumentando el factor, no vuelven a tomar valores no nulos. Esta interesante propiedad hace que la regresión de Lasso se utilice más como método de selección de variables más que como un modelo propiamente dicho.

MÉTODOS DE EXTRACCIÓN

El enfoque de los métodos de extracción es radicalmente distinto: transformar el conjunto de variables iniciales en un conjunto de menor dimensión que sea capaz de retener la mayor parte de información.

Los métodos de extracción tienen una gran ventaja y un pequeño inconveniente. La ventaja es que, como veremos, no utilizan información de la variable respuesta, y por tanto se pueden utilizar datos no etiquetados (análisis semisupervisado). Haciendo esto conseguimos una mejor representación de la información, ya que los modelos de extracción aprenden relaciones entre predictores que luego serán de utilidad para la fase del modelado predictivo.

A cambio, el inconveniente es que tras la transformación de los datos estos dejan de tener cualquier interpretación sencilla, ya que de hecho ni siquiera van a tener expresiones funcionales de las variables originales, salvo las combinaciones lineales que obtendremos en PCA.

- **PCA (Análisis de componentes principales)** es sin duda el método clásico más conocido. De forma intuitiva, dado un conjunto de puntos en un espacio multidimensional, PCA realiza un cambio del sistema de coordenadas de manera que las primeras dimensiones en dicho sistema recojan la mayor variabilidad posible de los datos. PCA asume que los datos siguen distribuciones normales y que tienen una representación lineal en cierta base. En la práctica la relajación de dichas hipótesis suele funcionar relativamente bien.
- **tSNE (*t-distributed stochastic neighbor embedding*)** es otro más reciente método para reducción de dimensionalidad. A diferencia de PCA, es una técnica no lineal, cuyo objetivo es que puntos similares o próximos en el espacio multidimensional queden repre-

sentados como próximos en el espacio reducido. tSNE modela dos distribuciones de probabilidad en ambos espacios y minimiza la divergencia de Kullback–Leibler de estas distribuciones. El resultado es una potente herramienta que suele obtener mejor separabilidad que PCA en proyecciones gráficas en dos o tres dimensiones y que mejora también su rendimiento cuando se utiliza como extractor de variables para modelos predictivos.

- **K-means** es de sobra conocido como un método de *clustering* por su potencia y escalabilidad, pero su aplicación como técnica de reducción de dimensionalidad no lo es tanto. Los métodos de *clustering* actúan agrupando casos en base a una medida de similitud o una métrica. Una vez que los clústeres están calculados podemos representar cada observación por su distancia, o una función de esta, a los centroides de los mismos. En la práctica obtenemos una reducción no lineal de la dimensionalidad. El uso de la distancia es suficiente para la aplicación de técnicas basadas en árboles, dado que estas son invariantes respecto de transformaciones monótonas de los predictores. Sin embargo, para la aplicación de otros métodos de ajuste, será necesaria la transformación de las distancias. Podemos, por ejemplo, utilizar el inverso de la distancia al cuadrado y posteriormente normalizar respecto a la suma de los valores para cada observación. El resultado final es una representación con tanta dimensiones como clústeres y que puede ser interpretada como una ponderación en base a los centroides de estos.

Las redes neuronales son actualmente, junto a los métodos basados en árboles, uno de los grupos de técnicas más utilizadas en aprendizaje automático. Los ***autoencoders***, una de sus posibles configuraciones, se pueden utilizar para la reducción de dimensionalidad. Básicamente un *autoencoder* consiste en una red neuronal cuya entrada es igual a la salida, e internamente tiene varias capas ocultas, la central de las cuales tiene un número de neuronas sensiblemente inferior a la de la entrada. Esta estructura fuerza la reconstrucción de los datos de entrada después de pasar por la capa central, de manera que las neuronas centrales deben aprender a capturar el máximo de información posible. En una variante de esta estructura, los ***denoising autoencoders***, se añade ruido aleatorio a los datos de entrada manteniéndose el objetivo de recuperar la señal previa a esta perturbación. Con esta técnica se consigue que la codificación de la información sea más robusta. La capa central del *autoencoder* puede ser utilizada como una nueva representación (comprimida) de los datos.

Un modelo bastante extendido para la clasificación de textos es **LDA (*Latent Dirichlet Allocation*).** La hipótesis de LDA es que cada texto pertenece con una distribución de probabilidad a un conjunto de categorías, y que la probabilidad de aparición de las palabras en un texto depende de dichas categorías. La distribución a priori de las categorías sigue una distribución de Dirichlet y de ahí deriva el nombre. En cierta forma LDA es una reducción de dimensionalidad, ya que la representación de un texto como *BoW (bag of words)* se traduce a una ponderación del conjunto de categorías que suele ser varios órdenes de magnitud menor. Aunque el método surge y se aplica en *text mining*, nada impide usarlo con datos que puedan ser representados matricialmente de una manera similar (matrices binarias o matrices de frecuencia, por ejemplo).

Incluso ***random forest*** se puede utilizar también como técnica para la reducción de dimensionalidad. La idea es utilizar el nodo final en el que recae cada observación como una variable categórica más: una nueva variable por cada árbol con tantos niveles como nodos finales. Limitando el número de nodos a un número razonable y fijando el número de árboles como la dimensión final deseada obtendremos una nueva representación del conjunto de datos.

REFLEXIÓN

Hemos visto hasta aquí que no solo podemos abordar el problema con técnicas específicas, sino que, con ciertas dosis de imaginación, es posible usar métodos que originariamente se pensaron para otros problemas.

Ideas como regularizar los coeficientes de una regresión multivariable de una manera no homogénea, sino teniendo en cuenta aspectos como la dependencia entre sí de las variables (regularización por bloques), *subsampling* de variables y ensamblado de los modelos resultantes, *clustering* de variables como alternativa al *subsampling* aleatorio..., pueden leerse en foros especializados; algunas con más fundamento, otras más alocadas, pero en definitiva plantean retos a la comunidad científica que sin duda harán que en el futuro próximo dispongamos de más herramientas para abordar el problema de la dimensionalidad y, por extensión, del *Big Data*.

Uso de información complementaria en las estadísticas de turismo

Fernado Cortina García, María Izquierdo Valverde, Jesús Prado Mascuñano y María Velasco Gimeno
Subdirección de Estadísticas de Turismo, Ciencia y Tecnología. INE

INTRODUCCIÓN

Según la Cuenta Satélite del Turismo de España, en 2014 el turismo representó un 10,9% del PIB y generó un 12,7% del total de puestos de trabajo. Hablamos, por tanto, de un sector cuyo seguimiento recibe una amplia atención por parte de los agentes económicos y sociales a nivel nacional e internacional, pero también en el ámbito regional y local.

España cuenta con un completo sistema de estadísticas de turismo conformado por las estadísticas de oferta (principalmente las Encuestas de Ocupación en Alojamientos Turísticos y sus índices asociados) y las estadísticas de demanda, que estiman los flujos de visitantes internacionales que recibimos cada mes (FRONTUR) y su gasto asociado (EGATUR), así como la actividad turística de los residentes tanto dentro de España como en el extranjero (ETR/Familitur). La Cuenta Satélite del Turismo sintetiza además ambas perspectivas.

El componente interregional del turismo hace de la comparabilidad un rasgo esencial de la calidad de los datos. Nuestras estadísticas oficiales están plenamente adaptadas al marco de trabajo desarrollado por los manuales y recomendaciones de la Organización Mundial del Turismo[1,2] y de Eurostat[3], por lo que cuentan con una fuerte base metodológica y un nivel de calidad ampliamente reconocido.

Sin embargo, en un contexto europeo de gran movilidad de la población, de ausencia de fronteras y de cambios en el comportamiento de los consumidores, las necesidades de información son muchas y el sistema debe prepararse para incorporar las nuevas fuentes de información disponibles.

Los datos generados por fuentes no puramente estadísticas ofrecen la posibilidad de mejorar la relevancia, oportunidad y puntualidad de los productos ofrecidos bajo el paraguas de la calidad de la estadística oficial. Ejemplos de estas nuevas fuentes de información son los dispositivos de control de tráfico, los datos de posicionamiento de dispositivos móviles o la actividad de las tarjetas de crédito durante un viaje.

Todas ellas han sido exploradas en mayor o menor medida en el marco del sistema español de estadísticas de turismo y en este artículo se resumen las distintas experiencias.

EL USO DE LA INFORMACIÓN DE DISPOSITIVOS DE CONTROL DEL TRÁFICO

Una de las fuentes de *Big Data* cuyo potencial está siendo explorado en diferentes campos de la estadística oficial son los dispositivos de control de tráfico que posibilitan el cálculo automático del número de vehículos que pasan por un determinado punto.

En España, desde hace años se incorpora la información proporcionada por la Dirección General de Tráfico (DGT) en el cálculo del número de turistas internacionales que acceden a nuestro país por carretera. Esta fuente de información es imprescindible desde que, por el Acuerdo de Schengen, en vigor desde 1995, se eliminaron los controles en las fronteras internas de los países firmantes del mismo. Es bien conocido que Portugal, Francia y España permiten la libre circulación de personas entre sus fronteras.

Las espiras son unos dispositivos colocados bajo el pavimento que registran el número de vehículos que pasan sobre ellas y permiten clasificarlos según su longitud. Es la propia DGT la que procesa los registros generados en las espiras colocadas en la frontera y cada mes proporciona al INE los microagregados necesarios para el proceso de estimación.

Esta información se completa con la procedente de cámaras de control del tráfico equipadas con un lector de matrículas, también propiedad de la DGT. Las situadas en los puntos fronterizos proporcionan un registro completo, debidamente anonimizado, de los vehículos que visitan nuestro país clasificados por nacionalidad de la matrícula.

Combinando estas dos fuentes de información con otras operaciones de campo complementarias que permiten conocer la ocupación media de los vehículos y el tipo de visitante, se obtiene el número de turistas y de excursionistas internacionales que entran por carretera cada mes en España.

Este es un ejemplo de éxito en el uso estadístico de información generada automáticamente por dispositivos diseñados y utilizados para otros fines. Una comunicación ágil y una voluntad decidida de cooperación a nivel institucional son esenciales en el funcionamiento de este proyecto, que depende del correcto funcionamiento de los dispositivos para asegurar la disponibilidad de la información.

Actualmente se están explorando las posibilidades que ofrecen las bases de datos de cámaras en la identificación del excursionismo, tanto en el caso de los visitantes internacionales como en las salidas al extranjero de los residentes en España. El seguimiento de una matrícula debidamente anonimizada a través de su huella en la base de datos podría servir para determinar el tipo de visitante. Disponer de información longitudinal y el seguimiento de la matrícula más allá de la frontera ayudaría a determinar el país de residencia habitual así como el entorno habitual del usuario del vehículo.

ANÁLISIS PRELIMINAR DE LOS DATOS DE TELEFONÍA MÓVIL

Una de las fuentes de datos más populares cuando se habla de *Big Data* son las bases de datos de los operadores de telefonía móvil, donde se almacenan las interacciones de los dispositivos móviles con las antenas que componen las redes de telefonía. Cada dispositivo tiene asociado un número de identificación cuyo seguimiento, debidamente encriptado, puede aportar valiosa información sobre la movilidad del portador del aparato. Estos movimientos se podrían asociar además al perfil del usuario, tomando en todo momento las adecuadas medidas de protección de la confidencialidad.

No cabe duda de que para el análisis del turismo esta información ofrece un potencial enorme, ya que permitiría obtener datos con un grado de detalle geográfico que las encuestas no pueden alcanzar. En el escenario más favorable, se podrían reducir las muestras utilizadas en las operaciones actuales, que se limitarían a recoger información sobre el alojamiento utilizado, las actividades realizadas o el gasto efectuado.

Ante esta situación, Eurostat encargó un estudio, cuyos resultados se publicaron en 2014, sobre la viabilidad del uso de los datos de telefonía móvil en las estadísticas de turismo[4]. Lamentablemente, una de las principales conclusiones del estudio

fue la constatación de las dificultades para acceder a esta información, principalmente por restricciones legales. En un gran número de países de la Unión Europea los Institutos Nacionales de Estadística están trabajando conjuntamente con los operadores de telefonía móvil para tratar de solventar estas dificultades y encontrar una forma de colaboración que resulte ventajosa para todos los actores implicados. También se están haciendo esfuerzos a nivel europeo en este sentido.

En España, el INE ha realizado un primer análisis de los datos de telefonía móvil a partir de un conjunto de agregados solicitados a un operador nacional. El objetivo era estudiar si las cifras derivadas de este primer análisis resultaban comparables con las estimaciones oficiales. En concreto, los indicadores objeto de estudio fueron la distribución del total de turistas por origen y destino y la duración media de la estancia.

El INE obtuvo datos referidos a un mes determinado, clasificados por comunidad autónoma o país de origen y por destino.

La complejidad de los datos capturados por una antena requiere transformarlos en un conjunto de información válido para el tratamiento estadístico. Este proceso lo llevó a cabo el operador de telefonía móvil siguiendo sus propios algoritmos. A partir de ahí, el INE trató de adaptar las definiciones de turismo internacionalmente aceptadas a las posibilidades de la base de datos. Este es un trabajo que precisa una interacción continua con el operador, el análisis detallado de los resultados, la detección de diferencias sistemáticas y la elaboración de conclusiones que permitan el refinamiento de las condiciones inicialmente establecidas para empezar de nuevo el ciclo hasta alcanzar resultados satisfactorios. Este primer paso es fundamental para un futuro aprovechamiento de los datos.

Comparando los datos de los residentes obtenidos con los datos de posicionamiento móvil con los de la encuesta FAMILITUR, se observó una estructura similar en los principales destinos, pero en algún caso se encontraron diferencias importantes. Un análisis más detallado mostró que la proporción de viajes dentro de la comunidad autónoma es sistemáticamente superior en los datos de telefonía móvil que en la encuesta, pero en mayor medida en la comunidad donde se daban las mayores diferencias.

También se detectaron diferencias sistemáticas en la variable estancia media. En los viajes internos realizados por los residentes, los datos de telefonía móvil estiman una estancia media más de un 50% más larga que la cifra estimada por FAMILITUR.

En el caso de los no residentes que nos visitan, la distribución de turista por país de origen también resultó muy similar en ambas fuentes, con alguna discrepancia un poco más acentuada en algún mercado concreto. Sin embargo, la estancia media estimada fue sistemáticamente inferior en los datos de telefonía móvil que en las cifras oficiales.

Estos hechos parecen indicativos de la necesidad de mejorar la determinación de la residencia habitual, de identificar posibles ruidos en las zonas fronterizas y en las áreas metropolitanas, etc. Conviene ahondar también en las diferencias conceptuales que puede haber tras estas discrepancias y analizar críticamente cada una de las fases del proceso de extracción de información de los datos de posicionamiento móvil.

APROXIMACIÓN A LOS DATOS DE LOS SISTEMAS DE PAGO CON TARJETA DE CRÉDITO

Otra de las nuevas fuentes de información de gran interés son los registros de las transacciones efec-

Las nuevas fuentes de información ofrecen un gran potencial en el ámbito del turismo. De ellas se podrán obtener datos con un nivel de detalle necesario para los usuarios pero que una operación por muestreo no puede cubrir

Los primeros análisis realizados ponen de manifiesto la necesidad de establecer un marco metodológico y conceptual que dote de validez a los nuevos indicadores y que establezca la convivencia de los mismos con los datos actuales

tuadas con tarjetas de crédito, como son los pagos en los terminales de venta (TPV) de distintos establecimientos o extracciones de efectivo en cajeros automáticos.

El INE solicitó a un operador nacional una serie de agregados obtenidos de las transacciones efectuadas por sus clientes con las tarjetas de la entidad para un periodo de dos años. El objetivo del análisis era la comparación del gasto medio por viaje de los residentes.

Como en el caso anterior, el trabajo directo con la base de datos fue realizado por el operador, que proporcionó al INE resultados agregados. De nuevo, la traducción de las definiciones de turismo en un algoritmo fue, y sigue siendo, el principal reto a desarrollar.

Teniendo en cuenta que la cobertura del gasto recogido no es la misma en las dos fuentes, cabría esperar que de los datos de tarjetas se derivase un gasto medio inferior al de la encuesta. En el análisis se observaron diferencias sistemáticas pero en dirección opuesta a la esperada.

WEB SCRAPING

Otra técnica de obtención masiva de datos es el webscraping, que consiste en la recopilación de información de un sitio web de forma automática. El INE está desarrollando un proyecto piloto para la obtención de precios de los paquetes turísticos, por un lado, y del alojamiento y el transporte incluido en cada paquete si se adquiriesen de forma independiente. El objetivo es la descomposición del precio de los paquetes turísticos en sus principales componentes: alojamiento y transporte.

Este desglose es exigido a nivel europeo en las estadísticas de turismo de los residentes como consecuencia de las necesidades de la Cuenta Satélite del Turismo[5] y de la Balanza de Pagos. La incorporación de información auxiliar en los procedimientos actuales posibilitaría obtener información detallada por destinos e incluso distinguir la nacionalidad del proveedor del servicio, permitiendo estimar el porcentaje del importe del paquete turístico que repercute en la economía de origen o de destino.

CONCLUSIONES

Las nuevas fuentes de información ofrecen un gran potencial en el ámbito del turismo. De ellas se podrán obtener datos con un nivel de detalle necesario para los usuarios pero que una operación por muestreo no puede cubrir.

Los primeros análisis realizados ponen de manifiesto la necesidad de establecer un marco metodológico y conceptual que dote de validez a los nuevos indicadores y que establezca la convivencia de los mismos con los datos actuales. Las estadísticas oficiales han alcanzados unos estándares de calidad muy altos que deben preservarse y, cómo no, actualizarse. Los esfuerzos dedicados a esta actualización y a la incorporación de estas nuevas fuentes de datos deben realizarse de forma coordinada en el Sistema Estadístico Europeo.

Este trabajo de conceptualización requiere, además, una gran colaboración entre las oficinas de estadística y los proveedores de datos. La información proporcionada por los mismos debe estar acompañada de una exhaustiva documentación auxiliar y la comunicación fluida entre todos los actores resulta esencial para salvaguardar la calidad de los resultados, especialmente en estas fases iniciales.

Referencias

1. OMT, Recomendaciones internacionales para las estadísticas de turismo, (RIET, 2008).
2. OMT, International Recommendations for Tourism Statistics 2008 Compilation guide, (2014).
3. Regulation (EU) No 692/2011 of the European Parliament and of the Council of 6 July2011 concerning European statistics on tourism.
4. EUROSTAT, Ahas R., Armoogum J., Esko S., Ilves M., Karus E., Madre JL., Nurmi O., Potier F., Schmücker D., Sonntag U., Tiru M. (2014), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report, Publications Office of the European Union.
5. United Nations, World Tourism Organization, Eurostat, OECD (2010), Tourism Satellite Account: Recommended Methodological Framework 2008, United Nations publication.

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).

ENCUESTA CONTINUA DE HOGARES (ECH) 2015

Información detallada en INEbase: www.ine.es

La Encuesta Continua de Hogares (ECH) es una encuesta continua que ofrece información anualmente sobre las características demográficas básicas de la población y de los hogares que componen (tipología y tamaño), así como de las viviendas que habitan.

La muestra anual efectiva es de unos 57.000 hogares. Con la muestra de un año (t) se obtienen resultados a nivel nacional y de comunidad autónoma y los datos corresponden al valor medio del periodo. La muestra acumulada de dos años (t y t-1) permite también desagregaciones a nivel provincial y los datos corresponden al valor medio del periodo, por lo que se refieren a 1 de enero del año t.

DEFUNCIONES SEGÚN LA CAUSA DE MUERTE 2014

Información detallada en INEbase: www.ine.es

La Estadística de Defunciones según la Causa de Muerte constituye una de las fuentes de información más importantes en el campo de la Sanidad. Se realiza siguiendo los criterios establecidos por la OMS en la Clasificación Internacional de Enfermedades (CIE), que recoje más de 12.000 enfermedades.

Los datos sobre causas de muerte se recogen a través de tres cuestionarios: Certificado Médico de Defunción/Boletín Estadístico de Defunción, Boletín Estadístico de Defunción Judicial y Boletín Estadístico de Parto.

Esta estadística proporciona información sobre la mortalidad atendiendo a la causa básica de la defunción, su distribución por sexo, edad, residencia y mes de defunción. También ofrece indicadores que permiten realizar comparaciones geográficas y medir la mortalidad prematura: tasas estandarizadas de mortalidad y años potenciales de vida perdidos.

DIRECCIONES Y TELÉFONOS DE INTERÉS

INE-Pº de la Castellana, 181 y 183 - 28046 Madrid.
www.ine.es

Atención a usuarios
Tfno.: 91.583.91.00
Fax: 91.583.91.58
Consultas: www.ine.es/infoine
Lunes a jueves de 9 a 14 y de 16 a 18 horas
Viernes de 9 a 14:30 horas

Índice-Librería del INE
Tfno.: 91.583.94.38
Fax: 91.583.45.65
E-mail: indice@ine.es
Lunes a viernes de 9 a 14:30 horas

Biblioteca
E-mail: biblioteca@ine.es

PUBLICACIONES EDITADAS POR EL INE DE ABRIL A JUNIO DE 2016

INEbase. Mayo 2016
Descarga gratuita a través de la web del INE
Contenido:
Boletín Mensual de Estadística (BME). Mayo 2016
Contabilidad Nacional Trimestral de España. 1º trimestre 2016
Encuesta de Condiciones de Vida. Módulo 2015. Participación social
Encuesta de Condiciones de Vida. 2015
EPA. Variables Submuestra. Serie 2006-2015
Estadística de Juzgados de Paz. Serie 2006-2015
Estadística de Profesionales Sanitarios Colegiados. 2015

INEbase. Abril 2016
Descarga gratuita a través de la web del INE
Contenido:
Boletín Mensual de Estadística (BME). Abril 2016
Encuesta Continua de Hogares. 2015
EPA. Flujos de la Población Activa. Serie 2005 - 1º trim. 2016
EPA. Resultados trimestrales. 1º Trimestre 2016
Estadística del Padrón Continuo. A 1 de enero de 2016. Datos provisionales
Indicadores de Confianza Empresarial. 2º Trimestre 2016

INEbase. Marzo 2016
Descarga gratuita a través de la web del INE
Contenido:
Boletín Mensual de Estadística (BME). Marzo 2016
Contabilidad Regional de España. Serie 2010-2015. 2015 estimación
Contabilidad Regional de España. Serie homogénea. Serie 2000-2015. 2015 estimación
Defunciones según la Causa de Muerte. 2014
Elecciones a Cortes Generales. 20 Diciembre 2015
Encuesta Coyuntural sobre Stock y Existencias. 4º trimestre 2015
Encuesta de Comercio Internacional de Servicios. 4º trimestre 2015
Encuesta de Turismo de Residentes. 4º trimestre 2015
Encuesta Trimestral de Coste Laboral. Serie 1º trim. 2008 - 4º trim. 2015
Estadística de Ejecuciones Hipotecarias. 4º Trimestre 2015
Indicadores de Confianza Empresarial. Módulo 2013-2015. Entorno empresarial
Índice de Coste Laboral Armonizado. ICLA. Serie 1º trim. 2000 - 4º trim. 2015
Índice de Precios de Vivienda (IPV). 4º trimestre 2015
Índice de Precios del Sector Servicios. 4º trimestre 2015
Padrón de Españoles Residentes en el Extranjero (PERE). A 1 de enero de 2016
Tablas de Mortalidad de la Población de España. Serie 1975-2014

Anuario Estadístico 2016
Papel. 591 páginas. Contiene CD-Rom. 18,00 € (IVA incluido)
Descarga gratuita a través de la web del INE

Estadística Española nº 189
Volumen 58, enero-abril 2016
Papel. 117 páginas. 16,61 € (IVA incluido)
Descarga gratuita a través de la web del INE

España en cifras 2016
54 páginas. 2,58 € (IVA incluido)
Descarga gratuita a través de la web del INE