El futuro del estudio de la brecha digital: el *Big Data*

David Salgado* y José Manuel Robles**

RESUMEN

Los *Big Data* se han transformado en una fuente de datos de especial relevancia para las distintas disciplinas de las ciencias sociales. El volumen de datos que se ponen a disposición de los investigadores, la velocidad con la que estos se generan y pueden ser analizados, así como la variedad de formatos hacen que muchos especialistas consideren que estamos ante una transformación trascendental para la investigación científica. Sin cuestionar estas expectativas, este trabajo se pregunta qué efectos tiene el uso de estas nuevas fuentes de datos sobre un campo de investigación que, como la brecha digital, requiere bases de datos representativas de la población objeto de estudio para, de esta forma, fundamentar las políticas que persiguen el correcto desarrollo de la sociedad de la información, así como para realizar análisis que ofrezcan una visión ajustada de la penetración de este fenómeno en una población concreta. En este trabajo planteamos "el problema de la inferencia" como una cuestión abierta para la aplicación sistemática de los *Big Data* al estudio de cuestiones sociales, y mantenemos que este problema afecta, de manera especial, al estudio de la brecha digital.

1. Introducción

Con "brecha digital" nos referimos a las diferencias en los porcentajes de penetración del uso de Internet entre unos ciudadanos y otros, así como entre unos grupos sociales y otros (DiMaggio y Hargittai, 2001). Como extensión de este interés por los efectos negativos del desarrollo de las Tecnologías de la Información y la Comunicación (TIC), los especialistas han comenzado a usar el término "desigualdades digitales" para referirse a las diferencias existentes entre aquellos ciudadanos que realizan usos de Internet que generan ventajas competitivas y aquellos otros que no están en disposición de hacerlo (Van Dijk, 2006).

Tras casi dos décadas estudiando este fenómeno, la comunidad académica cuenta con un volumen importante de evidencia empírica sobre qué variables y factores subyacen a estos dos fenómenos. Aunque, ciertamente, existe una importante reducción de la brecha digital en muchos países desarrollados, el nivel de estudio de la población, así como la situación laboral o los recursos económicos continúan incidiendo, en países como España, significativamente sobre ser o no un usuario de Internet (Torres-Albero *et al*., 2013). De la misma forma, las habilidades digitales, así como la percepción de la utilidad de la tecnología, son factores clave para predecir la desigualdad digital (Torres-Albero *et al*., 2017).

Los expertos en la medición de la brecha digital y la desigualdad digital han usado, fundamentalmente, encuestas dirigidas a pobla-

* Instituto Nacional de Estadística (INE) (david.salgado.fernandez@ine.es).

** Universidad Complutense de Madrid (jmrobles@ccee.ucm.es).

ción general que ofrecen información sobre, por ejemplo, el porcentaje de ciudadanos que usan Internet en una determinada comunidad, así como sobre qué usos realizan de esta herramienta y qué actitudes expresan hacia esta tecnología. En el caso de España contamos, como las más destacadas, con las encuestas realizadas por el Instituto Nacional de Estadística (INE)[1] y el Observatorio de las Telecomunicaciones y de la Sociedad de la Información (ONTSI)[2]. Esta tendencia se debe, fundamentalmente, a que el estudio de la brecha digital persigue conocer la penetración de Internet en comunidades determinadas, sean estas coincidentes con países o regiones, y para ello es necesario contar con el factor de la representatividad estadística. Igualmente, el uso de encuestas representativas de la población general ha permitido informar las políticas públicas para el correcto desarrollo de la sociedad de la información (SI) en España. No obstante, los expertos en la brecha digital han usado otro tipo de técnicas que, como los métodos de observación directa, especialmente los experimentales, han generado información relevante sobre determinados comportamientos relacionados con este fenómeno. Entre estos, destacan los estudios dirigidos a analizar las habilidades digitales (Van Deursen y Van Dijk, 2011).

En los últimos años, no son pocos los expertos que han advertido sobre las potencialidades del uso del *Big Data* para el estudio de distintos fenómenos sociales (Lin, 2015). Sin poner en duda las potencialidades de esta nueva fuente de datos, es importante considerar tanto sus beneficios como sus limitaciones para el análisis social (Hargittai, 2015). Especialmente, y en el estudio de la brecha digital, debemos ser conscientes de qué supone usar estas fuentes para los objetivos que persiguen las instituciones públicas comprometidas con el desarrollo de la SI y los propios académicos interesados en este tema. En este trabajo defenderemos que, en campos de estudio tan cercanos a las políticas públicas como el caso de la brecha digital, el uso del *Big Data* plantea problemas relevantes muy relacionados con la dificultad de hacer generalizaciones sobre poblaciones. Esto no implica –debe quedar claro– que este recurso no sea valioso para las ciencias sociales, sino que conviene ser prudente y observar en qué contextos y de qué forma debe usarse. De esta forma, defenderemos que, mientras las encuestas que hasta ahora han sido utilizadas para medir la brecha digital permitían hacer apreciaciones de carácter general como, por ejemplo, el porcentaje de usuarios de Internet en España, los análisis basados en *Big Data* permiten, sin minimizar su valor, realizar afirmaciones de carácter fundamentalmente exploratorio.

En este artículo procederemos de la siguiente manera. En primer lugar, realizamos una breve descripción del estudio de la brecha digital en un contexto de emergencia del uso académico de los *Big Data*. En segundo lugar, planteamos el problema de la inferencia como uno de los desafíos a los que debe enfrentarse cualquier experto que desee usar los *Big Data* como fuente de datos para el análisis social. Por último, en las conclusiones, discutimos el efecto de este problema sobre el uso potencial de los *Big Data* en la investigación sobre la brecha digital.

2. Marco teórico: el estudio empírico de la brecha digital y el desafío del *Big Data*

El estudio de la brecha digital está, en términos metodológicos, fragmentado en aproximaciones directas e indirectas. Mientras las primeras son de carácter experimental y tratan de medir en laboratorio o en un contexto controlado cuestiones como los usos de Internet o las habilidades digitales de diferentes segmentos de la población (Van Deursen y Van Dijk, 2011; Hargittai, 2010), los métodos indirectos se centran en la recogida de datos generales a través de fuentes como las encuestas (Robles *et al*., 2010).

Los métodos directos, como todo método experimental o cuasiexperimental, ofrecen importantes y múltiples ventajas para el estudio social. Entre estas, destacaríamos la posibilidad de tener un mayor control sobre la incidencia de las variables que se desea analizar, así como una observación real de las capacidades y el tipo de uso que los usuarios dan a Internet (Hargittai, 2010); es decir, la posibilidad de observar los comportamientos de los sujetos de estudio y

[1] *Encuesta de equipamiento y uso de TIC en los hogares* – Año 2016 (http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176741&menu=ultiDatos&idp=1254735576692).

[2] Véase: http://www.ontsi.red.es/ontsi/

no, como ocurre con otros métodos, lo que el sujeto "dice que hace". No obstante, este tipo de métodos encuentra limitaciones importantes en lo que se refiere a, por una parte, el número de comportamientos que pueden observarse simultáneamente, y, por otra, la posibilidad de generalizar los resultados. Aunque existen estudios de tipo experimental que permiten una cierta generalización de los resultados, como es el caso del estudio de Van Deursen y Van Dijk (2011) para medir las habilidades digitales de los holandeses, generalmente estos métodos han generado información específica sobre un tipo concreto de comportamiento digital. No obstante, y gracias a los estudios sobre la brecha digital, sabemos que comportamientos tan complejos como el uso de Internet tienen un carácter multidimensional y requieren un conjunto importante de variables para controlar el efecto del azar sobre nuestros pronósticos.

En este terreno encuentran su mayor fortaleza los métodos indirectos, como las encuestas. La principal ventaja del uso de la encuesta como herramienta de medición de la brecha digital es que nos permite recoger información sobre muchos y variados factores para, posteriormente, realizar análisis que, como el modelo de ecuaciones estructurales, facilitan la compresión de la complejidad del fenómeno. Igualmente, las encuestas generan información con un alto grado de fiabilidad sobre la población objetivo de estudio. Esto es especialmente importante en el caso de campos de investigación que, como la brecha digital, tienen una estrecha vinculación con el desarrollo de las políticas públicas. Este tipo de estrategias requieren información generalizable a un territorio concreto y, para esto, la encuesta resulta especialmente útil. No obstante, y como principal contrapartida, sería necesario destacar que este método no ofrece información sobre las acciones de los agentes, sino su valoración sobre lo que sabe/puede hacer o su opinión sobre lo que ha hecho. Esta circunstancia es un obstáculo importante cuando, como es el caso del estudio de la brecha digital, se trata de avanzar en la compresión de, por ejemplo, las capacidades reales de uso de Internet que poseen los ciudadanos de un determinado país (De Marco *et al*., 2014).

En los últimos años, y al igual que en el resto de campos de investigación en las ciencias sociales (Mützel, 2015), ha surgido un nuevo y prometedor recurso para la medición de la brecha digital. Dicho recurso se denomina *Big Data*. Se trata de un método indirecto, ya que toma datos, la huella digital, que los usuarios dejan registrados al hacer uso de cualquier dispositivo electrónico. Sin embargo, y a diferencia del resto de métodos indirectos como la encuesta, no recoge información de lo que los ciudadanos dicen hacer, sino de lo que los ciudadanos hacen cuando se conectan a estos dispositivos. Es decir, los datos conocidos como *Big Data* son registros de acciones y de opiniones no inducidas por el investigador. Esto constituye, a todas luces, una importante ventaja a la hora de realizar investigación social. Otros factores importantes son, como es bien sabido, el volumen, la variedad y la velocidad con la que podemos recoger los datos; en otras palabras, lo que se ha venido llamando las tres "V" del *Big Data* (IBM, 2013).

Aún no contamos con un volumen suficiente de estudios empíricos sobre la brecha digital que hagan un uso sistemático de datos procedente de grandes bases de datos (*Big Data*). Sí existe, sin embargo, un creciente número de investigaciones usando *Big Data* en otras áreas[3], así como un importante volumen de trabajos académicos que nos informan sobre las posibilidades potenciales de esta fuente de datos para la investigación empírica en ciencias sociales (Jenkins *et al.*, 2016). Así, por ejemplo, los expertos han señalado, al menos, tres argumentos en favor del uso del *Big Data* en esta disciplina. En primer lugar, se ha asegurado que esta nueva fuente de datos hará posible sondear procesos y aspectos del comportamiento social que, con las fuentes de información tradicionales, no era posible rastrear (Mayer-Schönberger y Cukier, 2014). Por otra parte, se mantiene que el *Big Data* ofrece un acercamiento más dinámico y a tiempo real de los fenómenos que interesan a las ciencias sociales (Sharon y Zandbergen, 2016). Por último, se supone que, gracias al *Big Data*, no será necesario utilizar muestras de la población, ya que, una vez conectados todos los ciudadanos a la red, tendremos datos directos de toda la población (Lin, 2015).

No han sido pocos, sin embargo, los autores que han alertado sobre un conjunto de

[3] Véanse, por ejemplo, los números monográficos publicados en revistas como *Journal of Communication* (2014, Vol. 64, 6), *International Journal of Sociology* (2016, Vol. 46, 1), *Annals of the American Academy of Political and Social Science* (2015, Vol. 659, 1) o *Political Science and Politics* (2015, Vol. 48, 1).

incertidumbres que se ciernen sobre el uso del *Big Data* en las ciencias sociales. También, en este caso, se puede resumir el debate en tres grandes líneas argumentativas. En primer lugar, encontraríamos el llamado efecto *messiness* (Harggitai, 2015); es decir, la incidencia que la variedad, el desorden y la cantidad de datos irrelevantes que contiene los *Big Data* puede tener sobre la calidad de los datos usados para el análisis social. En segundo lugar, algunos autores alertan sobre el uso "frívolo" de la correlación estadística. Se ha observado cómo, gracias al gran volumen de datos disponibles, existe una creciente tendencia a explorar relaciones entre variables sin que, previamente, los investigadores cuenten con un modelo o con hipótesis explicativas cuya justificación tenga un fundamento teórico (Nagler y Tucker, 2015). Por último, se ha planteado en qué medida el uso académico del *Big Data* tiene implicaciones normativas de gran complejidad como, por ejemplo, aquellas relacionadas con la mercantilización de la información (Crain, 2016).

No obstante, en este trabajo nos centramos en otra importante línea de debate sobre el uso de *Big Data* en ciencias sociales que afecta directamente al estudio de la brecha digital: la emergencia y la creciente importancia que adquieren, gracias al uso del *Big Data*, los métodos exploratorios en detrimento de modelos explicativos. Tal y como trataremos de mostrar, los datos recogidos a través de la huella digital no permiten realizar inferencias estadísticas, por lo que cualquier generalización queda descartada como base de la explicación. Esto conduce irremediablemente, según nuestro argumento, a análisis de carácter exploratorio. En el siguiente apartado daremos razones que justifican esta limitación de los datos recogidos a través de la huella digital y, en las conclusiones, debatiremos sobre sus efectos potenciales en el estudio de la brecha digital.

3. El *Big Data*, la inferencia estadística y la brecha digital

Con la expresión "huella digital" no nos referimos únicamente a la huella que, directa o indirectamente, dejan los usuarios en su interacción con alguna aplicación informática conectada a Internet. Para los propósitos de este artículo, queremos abarcar, además, aquellas situaciones en las que alguna actividad humana de cualquier naturaleza es registrada directa o indirectamente en un sistema de información digital.

Uno de los ejemplos paradigmáticos son las redes celulares de telefonía móvil (Sauter, 2014). Para dar servicio a un usuario de telefonía móvil, el sistema (la red celular) registra en qué célula del territorio geográfico se encuentra el teléfono móvil (por tanto, la persona). Y esto ocurre tanto si el usuario establece una conexión (llamada, SMS, conexión a Internet, etcétera) como si no lo hace (la posición es monitorizada por cuestiones operativas de las redes). Esta información no está nunca en Internet, pero sí es registrada en un sistema de información digital exclusivamente privado con un alto grado de protección física e informática. Otro ejemplo, que no entra en la categoría de *Big Data*, se encontraría en las estadísticas actuales de alojamientos turísticos (European Statistical System, 2012). Hoy día una buena parte de estos establecimientos poseen un sistema informático de registro de sus clientes. Este registro, no obstante, no es público ni accesible vía Internet. La proliferación de este tipo de sistemas de información digitales para ejecutar o bien asistir en la ejecución de un número cada vez mayor de actividades ofrece un enorme potencial para la producción de información estadística, así como, potencialmente, para la investigación social.

Estos sistemas informáticos permiten la utilización del método llamado "recogida de datos automática" que optimiza este proceso. Consiste en disponer de una pequeña aplicación informática controlada y supervisada exclusivamente por la unidad estadística, sin intervención de personal de la oficina estadística, y que toma los datos necesarios de este registro informático para configurar de modo automático un fichero donde se dispone de toda la información requerida para la operación estadística. Este fichero es enviado telemáticamente a la oficina de estadística.

Los beneficios son evidentes (Rosa-Pérez, 2016). Aparte de la automatización del proceso, que evita la cumplimentación manual del cuestionario, disminuyendo así la llamada "carga al informante"[4], este procedimiento permite

[4] Véase el principio 9 del *Código de Buenas Prácticas Estadísticas del Sistema Estadístico Europeo* (European Statistics Code of Practice, 2011).

alcanzar un mayor grado de desagregación de la información (por ejemplo, detallando todas las nacionalidades de los huéspedes del establecimiento), supone la entrada automática de datos en el sistema de información de la oficina de estadística, con el aumento de la eficiencia por coste del proceso de producción[5], y posibilita el control *in situ* sobre posibles errores de cumplimentación que, de otro modo, obligarían a aumentar la carga al informante haciendo necesario un recontacto para comprobar los datos cumplimentados potencialmente erróneos. Este ejemplo reproduce, a pequeña escala, el potencial que la información digitalizada de la actividad humana supone para la producción estadística y para el análisis social.

Como hemos señalado más arriba, el análisis social de la brecha digital, en particular, y la producción de datos estadísticos, en general, pueden hacer un uso provechoso y positivo del *Big Data*. No obstante, este uso debe estar sujeto a un conjunto de consideraciones que limiten, en la medida de lo posible, efectos no deseados. Uno de los retos más importantes del uso de los *Big Data* para la producción estadística oficial es el conjunto de métodos estadísticos necesarios para su procesamiento y, en especial, para realizar las inferencias respecto de las poblaciones de interés (humanas, de empresas, de establecimientos, etcétera).

Consideremos un ejemplo relacionado esta vez con nuestro tema de reflexión en este artículo: las habilidades digitales. A través de distintas fuentes digitales, como teléfonos móviles, herramientas y servicios de Internet, etc., podemos obtener datos sobre las habilidades digitales de los españoles. Estaríamos hablando de grandes cantidades de datos que seguramente serían muy útiles para estudiar este fenómeno concreto, pero ¿cómo representan los datos recogidos a la población de análisis y, en particular, cómo se relacionan dichos datos con la muestra (que, aunque sea muy grande, sigue siendo una muestra)? Este es, desde nuestro punto de vista, un problema capital para el uso de este tipo de datos en ámbitos que, como la brecha digital, suelen referirse a poblaciones concretas. Llamamos a esta circunstancia "el problema de la inferencia".

[5] Véase el principio 10 del *Código de Buenas Prácticas Estadísticas* (European Statistics Code of Practice, 2011).

3.1. Inferencia basada en diseños muestrales

La metodología de la estadística oficial, en la que se basan las encuestas que suelen usarse para el análisis de la brecha digital en España (Eurostat, INE y ONTSI), solucionó formalmente el problema de la inferencia en los años treinta y cuarenta del pasado siglo, pero *a priori* esta solución no puede aplicarse al uso del *Big Data*. Ilustraremos este reto con un ejemplo concreto.

El problema esencial resuelto formalmente por la metodología estadística oficial es el de la estimación en poblaciones finitas (Särndal, 1992). Básicamente, este problema consiste en, dada una magnitud Y en una población de unidades estadísticas (por ejemplo, el número total de personas paradas, un índice de precios de productos industriales, el número total de fumadores, etc.), se desea proporcionar una estimación lo más precisa posible de Y empleando los datos correspondientes debidamente recogidos de una muestra de la población. Obsérvese que en la formulación del problema no existe ningún elemento aleatorio o que haga referencia al azar.

La solución formal se basa en la selección de una muestra probabilística, esto es, seleccionada mediante un diseño muestral que otorga una probabilidad a cada muestra posible, y seleccionando aleatoriamente una de ellas. Posteriormente, conociendo esta probabilidad y las probabilidades de selección derivadas asociadas a cada unidad estadística de la muestra, se construye un estimador $\hat{Y}$ a partir de estas probabilidades de selección y los datos recogidos. En la práctica, indudablemente el procedimiento es más complejo, pero la esencia matemática es esta.

Para su implementación práctica, esta solución requiere un listado de todas las unidades estadísticas del que pueda extraerse la muestra probabilística seleccionada. Este es el papel de los registros de población (o poblaciones marco) que constituyen una pieza central de las oficinas productoras de estadísticas oficiales. Son registros de poblaciones humanas, empresariales, de cuentas de cotización, de establecimientos turísticos, etcétera.

Una propiedad esencial de esta solución (Smith, 1976) es que los estimadores $\hat{Y}$ se

construyen, sin hacer hipótesis *a priori*, sobre la distribución de los valores de las variables en la población. Además, existen procedimientos matemáticos contrastados que permiten la construcción de estos estimadores con las siguientes dos propiedades esenciales. En primer lugar, son estimadores (asintóticamente) insesgados; esto es, en promedio sobre todas las muestras posibles, las estimaciones $\hat{Y}_s$ provenientes de cada muestra *s* posible coinciden (asintóticamente) con la magnitud *Y* que desea estimarse. En segundo lugar, explotando información auxiliar disponible, existen técnicas para que las variaciones en las estimaciones entre todas las muestras posibles sean muy pequeñas. La calidad de las estadísticas oficiales se fundamenta en estas propiedades. No obstante, no debe concluirse que esta solución adolece de ciertas deficiencias que no se detallan aquí (Valliant *et al*., 2000).

3.2. Inferencia basada en modelos estadísticos

Esta no es la única solución al problema de estimación en poblaciones finitas. Alternativamente puede construirse un modelo estadístico para los valores de las variables de interés en toda la población, así como también el correspondiente estimador usando los datos recogidos en la muestra y los valores que el modelo predice para los datos no recogidos en ella (Valliant *et al*., 2000; Chambers, 2012). En principio, el mecanismo de selección de la muestra ahora es irrelevante.

Para nuestros propósitos aquí, debe señalarse que esta solución necesariamente requiere hipótesis *a priori* sobre la distribución de los valores de las variables en la población (la elección de los modelos estadísticos). La ventaja de esta solución radica en que en general puede obtenerse una mayor precisión (menor variación de las estimaciones) si los modelos estadísticos escogidos son correctos (Hansen, 1987).

La teoría estadística contiene técnicas para analizar si los modelos son correctos, aunque también hay ejemplos concretos que ilustran los graves problemas en las estimaciones si las especificaciones de los modelos no son correctas (Hansen, 1987). De hecho, existen técnicas para hacer robustas las estimaciones mediante el control del mecanismo de selección de la muestra: son las llamadas muestras equilibradas (Valliant *et al*., 2000; Chambers, 2012).

Desde la década de los setenta hasta mediados de los ochenta tuvo lugar un intenso debate sobre la solución que debía emplearse para la producción de estadísticas oficiales (Smith, 1994). Finalmente, el argumento que prevaleció puede ilustrarse en la siguiente posición de Hansen *et al*. (1983): "parece deseable, siempre que sea factible, evitar estimaciones o inferencias que precisen ser defendidas como juicios de los analistas que llevan a cabo la encuesta". Debe señalarse que la independencia profesional de la estadística oficial es, de hecho, el primer principio del *Código de Buenas Prácticas Estadísticas* (European Statistics Code of Practice, 2011). Disponer de una metodología matemática que prescinde de hipótesis *a priori* refuerza este principio, y así ha sido entendido por la estadística oficial en todo el mundo.

3.3. La inferencia en los *Big Data: machine learning*

Uno de los retos metodológicos más importantes para el uso de los *Big Data* en el análisis social consiste en responder a la siguiente pregunta garantizando, en particular, el cumplimiento del marco de calidad de las estadísticas: ¿Cuál de las soluciones encontradas por la estadística oficial es la más adecuada para los *Big Data*?

Las dificultades para la aplicación de las técnicas tradicionales descritas más arriba aparecen a diversos niveles. En primer lugar, uno de los usos actuales de los *Big Data* es la búsqueda de patrones en los datos a través de técnicas de *data mining*. Este uso va más allá del problema de estimación en poblaciones finitas. Por tanto, es necesario identificar y formular en términos precisos qué problema estadístico quiere resolverse en cada caso.

En segundo lugar, incluso restringiéndonos al problema clásico en poblaciones finitas, las características de los datos tienen conse-

cuencias notables. Los datos, en general, no están identificados y, por tanto, los registros de población no son útiles. Esto quiere decir que no puede emplearse *stricto sensu* la solución basada en diseños muestrales.

De hecho, las técnicas empleadas son técnicas de *machine learning* (Murphy, 2012), que hacen un uso extensivo de modelos estadísticos y, en muchos casos, incluso de estadística bayesiana. Esto es algo que en la estadística tradicional no se hace porque requiere modelizar a varios niveles. Por tanto, para aplicarlo, sería necesario revisar el paradigma de la inferencia.

Por último, el uso de estas técnicas no se circunscribe exclusivamente a la cuestión de la inferencia, sino que aparece también al procesar e interpretar los datos. En contraposición a los datos oficiales tradicionales, que están generados con un sistema de metadatos normalizados que establecen el significado y las propiedades de cada variable, los *Big Data* son generados para fines no estadísticos y carecen de este sistema de metadatos estadísticos. Así pues, la conexión con los conceptos estadísticos de interés no está clara. Esta conexión puede establecerse a través de diversas técnicas de *machine learning*. Por ejemplo, al analizar datos de telefonía móvil para estimar la movilidad humana entre el hogar y el centro de trabajo, los datos son atributos espacio-temporales que no establecen si la persona se encuentra en el hogar, en el centro de trabajo o en algún otro lugar. Esto debe deducirse estudiando los patrones de los mismos datos.

Como ejemplo ilustrativo de las dificultades que implican todos estos factores, mencionamos el impactante ejercicio de estimación de la prevalencia de la gripe en Estados Unidos realizada por Google a partir exclusivamente de las búsquedas de términos relacionados en su buscador de Internet (Butler, 2013). Son las llamadas *Google Flu Trends*. Las estimaciones se realizaron mediante estas técnicas de modelización estadística y, posteriormente, se compararon con las cifras oficiales (Olson *et al*., 2013).

Durante las cinco primeras temporadas (2003/2004-2007/2008), la estimación básicamente coincidió con las cifras oficiales. Sin embargo, en la siguiente temporada (2008/2009), los modelos subestimaron los datos reales. La causa es difícil de identificar a ciencia cierta, pero matemáticamente la razón estriba en que las hipótesis *a priori* sobre la búsqueda de términos ya no fueron correctas. Al corregir los modelos (es decir, al cambiar las hipótesis), se recuperó la coincidencia hasta 2011/2012. En 2012/2013, sin embargo, volvió a suceder algo similar, ahora en sentido contrario: el modelo sobreestimó las cifras reales. Nuevamente, las hipótesis *a priori* sobre la búsqueda de términos fallaron (Butler, 2013).

Desde el punto de vista metodológico, este ejemplo ilustra claramente el reto que para la estadística oficial y tradicional tiene el cambio de paradigma en la inferencia: ¿deben descansar las cifras oficiales sobre hipótesis *a priori*? ¿Puede hacerse esta dependencia de las hipótesis más robusta ante fallos en la especificación de los modelos? En nuestra opinión, este ejemplo muestra no solo la necesidad de profundizar en la investigación de las nuevas técnicas en la producción estadística, sino que también sugiere indirectamente cómo la colaboración público-privada puede traer beneficios para ambas partes y, sobre todo, para la sociedad mediante la combinación de diversas fuentes de datos.

4. Discusión

Tal y como hemos mostrado en el apartado anterior, las ciencias sociales y, entre ellas, la estadística, se enfrentan a un importante desafío. El uso de los *Big Data* para el análisis social, en general, y para el estudio de la brecha digital, en particular, está lleno de posibilidades. Entre ellas, cabe destacar el volumen de datos que se pone a disposición de los investigadores, así como la posibilidad de analizar el comportamiento de los ciudadanos directamente y no a través de lo que estos reconocen hacer al ser preguntados en encuestas o a través de otros métodos indirectos.

No obstante, como también se ha señalado, el estudio de la brecha digital es un ejemplo de estudio social con una doble dimensión: académica y pública. Los investigadores que han centrado sus estudios en las distintas características de la brecha digital tienen presente que están identificando, describiendo y analizando un fenómeno que debe ser controlado para un

correcto desarrollo de la SI. Por este motivo el estudio de la brecha digital nos ha conducido, casi inexorablemente, a tener en mente poblaciones amplias en las que este tipo de desigualdad es, siguiendo a Goldthorpe (2017), una regularidad estadística de la población; es decir, el objeto de la sociología entendida como una ciencia de la población.

Esta característica concreta del estudio de la brecha digital, genera, desde nuestro punto de vista, el principal problema relacionado con el uso de los *Big Data* en este campo de estudio. Hemos llamado a este problema "el problema de la inferencia". Así, consideramos que, en primer lugar, las técnicas y procedimientos que la estadística oficial ha desarrollado desde la década de los treinta para resolver este problema no son aplicables al estado actual del uso de los *Big Data*. Igualmente, la propia naturaleza de estos datos, no pensados para el análisis estadístico, genera problemas a la hora de procesar e interpretarlos, ya que, a diferencia de los producidos por encuestas basadas en muestras representativas, los *Big Data* no están generados con un sistema de metadatos normalizados que establecen el significado y las propiedades de cada variable.

El escenario que crean estas características de los *Big Data* es doble. Por una parte, se requiere profundizar en las técnicas y procedimientos que permitan a estadísticos y expertos en ciencias sociales usar estos datos con mayores garantías. Esto pasa, naturalmente, por una mayor colaboración entre los nuevos agentes generadores de datos, las empresas privadas y las instituciones públicas que, tradicionalmente, han producido información estadística (institutos nacionales de estadística, Eurostat, etc.). En otras palabras, se requiere más investigación.

El estado actual del "problema de la inferencia", aplicado a los *Big Data*, no genera una imposibilidad metodológica respecto a esta fuente de datos. Todo lo contrario; el problema únicamente debe ser ponderado en relación a las múltiples posibilidades y ventajas que ofrecen los *Big Data*. Algunas de estas ventajas han sido apuntadas aquí y se refieren a cuestiones como la posibilidad de observar, sin la mediación del investigador, las acciones de los ciudadanos, así como las opiniones o preferencias que estos ofrecen en contextos de interacción social reales.

Lo que, a nuestro juicio, sí genera "el problema de la inferencia", dado el estado actual de los *Big Data*, es una importante dificultad para la realización de estudios analíticos que ofrezcan generalizaciones sobre una población dada. Gracias a los avances de la estadística oficial desde los años treinta para resolver el mencionado problema, los expertos en ciencias sociales pueden describir el comportamiento de los agentes individuales y colectivos de una determinada población y, gracias a las técnicas estadísticas avanzadas, generar modelos explicativos de dichos comportamientos. Los estudios basados en los *Big Data* propician, fundamentalmente, análisis exploratorios pensados para buscar tendencias que deben ser posteriormente justificadas a través de otras técnicas y para los que los criterios de inferencia estadística no resultan imprescindibles.

Los estudios exploratorios de este tipo pueden ser de gran relevancia para el estudio de la brecha digital al permitir a los especialistas analizar nuevas tendencias de uso de Internet, así como para registrar patrones de comportamiento social en espacios sociales como las redes sociales digitales. No obstante, para que las tendencias observadas a través del análisis de *Big Data* puedan formar parte destacada de los diagnósticos sobre la brecha digital, así como de las políticas públicas dirigidas al desarrollo de la SI, estas deben ser refrendadas con herramientas que, como las encuestas, permitan comprender el peso real de dichas tendencias en la población, así como su distribución entre los distintos grupos sociales que forman parte de la comunidad estudiada.

Tal y como señala Goldthorpe (2017), el objetivo de la sociología es, en primer lugar, hacer de las tendencias sociales algo transparente. La estadística tradicional ha permitido a los sociólogos realizar grandes avances en esta dirección. El segundo objetivo de la sociología sería comprender los mecanismos subyacentes a este tipo de tendencias y, por lo tanto, entenderlas en su complejidad. Los estudios exploratorios que pueden llevarse a cabo en el momento presente con los *Big Data* constituyen un paso previo, pero tremendamente poderoso, para mejorar nuestra capacidad de hacer trasparentes los procesos más significativos de la sociedad de la información.

Bibliografía

Butler, D. (2013), "When Google got flu wrong", *Nature*, 494: 155–156.

Chambers, R. L., y R. G. Clark (2012), An introduction to model-based survey sampling with applications, Oxford, Oxford University Press.

Crain, M. (2016), "The limits of transparency: Data brokers and commodification", *New Media & Society*, 7: 1-17.

De Marco, S.; Robles J. M., y M. Antino (2014), "Digital skills as a conditioning factor for digital political participation", *Communications: The European Journal of Communication Research*, 39(1): 146–167.

DiMaggio P., y E. Hargittai (2001), "From the Digital Divide to Digital Inequality. Studying Internet use as penetration increase", *Centre for Arts and Cultural Policy Studies*, 15: 1-23.

European Statistical System (2012), *ESSnet on Automated Data Collection and Reporting in Accommodation Statistics* (https://ec.europa.eu/eurostat/cros/content/tourism).

European Statistics Code of Practice (2011), *Eurostat and ESS* (http://ec.europa.eu/eurostat/web/products-manuals-andguidelines/-/KS-32-11-955).

Goldthorpe, J. H. (2017), *La sociología como ciencia de la población*, Madrid, Alianza editorial.

Hansen, M. H. (1987), "Some history and reminiscences on survey sampling", *Statistical Science*, 2: 180–190.

Hansen, M. H.; Madow, W. G., y B. J. Tepping (1983), "An evaluation of model-dependent and probability sampling inferences in sample surveys", *Journal of the American Statistical Association*, 78: 776–793.

Hargittai, E. (2010), "Digital na(t)ives? Variation in Internet skills and uses among members of the 'Net Generation'", *Sociological Inquiry,* 80(1): 92-113.

— (2015), "Is bigger always better? Potential biases of Big Data derived from social network sites", *Annals of the American Academy of Political and Social Science*, 659(1): 63–76.

IBM (2013), "What is Big Data? – Bringing Big Data to the enterprise", (https://www-01.ibm.com/software/in/data/bigdata/).

Jenkins, J. C.; Slomczynski, K. M., y J. K. Dubrow (2016), "Political behavior and Big Data". *International Journal of Sociology*, 46(1): 1–7.

Lin, J. (2015), "On building better mousetraps and understanding the human condition: Reflections on Big Data in the social sciences", *Annals of the American Academy of Political and Social Science*, 659(1): 33–47.

Mayer-Schönberger, V., y K. Cukier (2014), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Nueva York, Mariner Books.

Murphy, K. P. (2012), *Machine learning: A probabilistic perspective*, Massachusetts, The MIT Press.

Mützel, S. (2015), "Facing Big Data: Making sociology relevant", *Big Data & Society*, 2(2): 115-128.

Nagler, J., y J. A. Tucker (2015), "Drawing inferences and testing theories with Big Data", *Political Science and Politic*s, 48(1): 84–88.

Olson, D. R.; Konty, K. J., Paladini, M., Viboud, C., y L. Simonsen (2013), "Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales", *PLOS Computational Biology,* 9: 156-170.

Robles, J. M.; Torres, C., y O. Molina (2010), "Brecha digital. Un análisis de las desigualdades tecnológicas en España", *Sistema*, 218: 3-22.

Rosa-Pérez, E. (2016), "Improving the statistical process in the hotel occupancy survey", *European Conference on Quality in Official Statistics* (http://www.ine.es/q2016/docs/q2016Final00112.pdf).

Sharon, T., y D. Zandbergen (2016), "From data fetishism to quantifying selves: Self-

tracking practices and the other values of data", *New Media & Society*, 60: 1-17 .

Särndal, C. E.; Swensson, B., y J. Wretman (1992), *Model assisted survey sampling*, Berlín, Springer.

Sauter, M. (2014), *From GSM to LTE – advanced: An introduction to mobile networks and mobile broadband*, Nueva Jersey, Wiley.

Smith T. M. F. (1976), "The foundations of survey sampling: A review", *Journal of the Royal Statistical Society*, 139: 183–204.

— (1994), "Sample surveys 1975-90; an age of reconciliation?", *International Statistical Review*, 62: 5–34.

Torres-Albero, C.; Robles, J. M., y S. De Marco (2013), "Inequalities in the Information Society: From the Digital Divide to Digital Inequality", en A. López Peláez (Ed.), *The robotics divide. A new frontier in the 21st Century?*, Berlín, Springer: 173-194.

— (2017), "Revisión analítica del modelo de aceptación de la tecnología. El cambio tecnológico", *Papers, Revista de Sociología*, 102 (1): 5-27.

Valliant, R.; Dorfman, A. H., y R. M. Royall (2000), *Finite population sampling and inference: a prediction approach*, Nueva Jersey, Wiley.

Van Deursen, A. J. A. M., y J. A. G. M. Van Dijk (2011), "Rethinking Internet skills: The contribution of gender, age, education, Internet experience, and hours online to medium- and content-related Internet skills", *Poetics*, 39: 124-144.

Van Dijk, J. (2006), "Digital Divide research. Achievements and shortcomings", *Poetics*, 34: 221–235.