José Manuel Robles, J. Tinguaro Rodríguez, Rafael Caballero y Daniel Gómez (2020): Big data *para científicos sociales. Una introducción*. Centro de Investigaciones Sociológicas (Cuadernos Metodológicos, 60), 299 pp. ISBN: 978-84-7476-843-5.

Alba Taboada Villamarín
*Universidad de Huelva*

La digitalización que caracteriza a nuestras sociedades es responsable del aumento sin precedentes de grandes volúmenes de datos que se recogen y almacenan como recursos de información. Los datos masivos representan hoy uno de los mayores bienes de explotación para empresas e instituciones, que ven en estos la posibilidad de tomar decisiones eficientes o adquirir conocimientos en profundidad, permitiendo en muchas ocasiones la anticipación a problemas y su resolución.

El *big data* es uno de los fenómenos paradigmáticos que acompañan al siglo xxi. Esto ha sido posible gracias al desarrollo y abaratamiento de costos en computación, tanto a nivel de *software* como de *hardware* que, en la actualidad, nos permite almacenar y explotar grandes conjuntos de datos. Sin embargo, las características del *big data* no solo apelan a su tamaño, sino que atienden a la ya conocida regla de las tres V: *volumen, variedad* y *velocidad* y adicionalmente, *veracidad* y *valor*.

Lo que representan estas características, es la transformación que sufre el «dato» en su concepción clásica. En lo referido a la *variedad, big data* comprende registros que van desde señales de sensores a notas de voz, pasando por imágenes, *emails*, etc. En segundo lugar, esta variedad de formatos también transforma el modo en el que se almacenan y consultan los mismos, pasando de bases de datos estructuradas en las clásicas columnas y filas —bases de datos relacionales— a otras que complejizan su estructura y anidan la información —bases de datos no relacionales—. La *velocidad*, por otra parte, se ha convertido en condición esencial para el tratamiento y uso de los mismos, esto es posible gracias a técnicas de escalabilidad horizontal, que aúnan la potencia de múltiples ordenadores compartiendo el esfuerzo a la hora de computar registros.

Por último, la *veracidad* y *valor* refieren al contenido de estos datos y a la extracción de conocimiento que se puede hacer de los mismos. En ambos casos,

estas características resaltan por su complejidad, la necesidad de largos procesos de limpieza y validación, y equipos de experto capaces de dar significado y contexto a metadatos que no se crean con un objetivo preestablecido o fin único.

En consecuencia, los científicos en general; y las ciencias sociales en particular, encuentran en los datos masivos una fuente de recursos inagotables que en muchas ocasiones tienen una naturaleza social y política o generan preguntas que interpelan directamente a nuestras áreas de conocimiento. Como añadido, el *big data* entendido como concepto, es también objeto creciente de discusiones sociológicas, ya que este viene cargado de debates complejos que afectan a la ética de la procedencia de los datos, a la propiedad de los mismos o a su validez en el estudio de fenómenos y comportamientos sociales, entre otros.

Ante esta tesitura, Big data *para científicos sociales. Una introducción* viene a dar cuenta de los desafíos tanto metodológicos como epistémicos que tendrán lugar en el presente y futuro de las investigaciones sociales. Por ello, la razón de ser de este cuaderno con vocación de manual, es señalar una necesidad, a la vez que constata la primera zancada de un largo trayecto.

Entre sus múltiples bondades, es destacable que la autoría corresponde a un equipo multidisciplinar compuesto por el sociólogo J. M. Robles, el matemático J. Tinguaro Rodríguez, el informático R. Caballero y el estadístico D. Gómez. Todos ellos con una sonada experiencia en este campo de investigación y en la defensa de la transdisciplinariedad que requiere el trabajo con datos masivos.

Así queda declarado en el inicio del libro, que viene a plantear el rol del investigador social en este tipo de equipos de investigación. Este planteamiento, sin embargo, lo resuelven de forma acertada en tres tipos de posiciones. En primer lugar, el científico social como (1) personal especializado en una tarea multidisciplinar. Cuando se lleva a cabo una investigación de estas características, si la naturaleza del estudio o los datos adquiridos son de carácter sociocultural o político, es importante que haya investigadores sociales presentes tanto en la confección de las preguntas, en la supervisión de la recogida de datos, como finalmente en la interpretación de los resultados.

Se han dado a conocer múltiples investigaciones que pecan de una comprensión laxa de sentido y correlación, por constreñirse a conocimientos únicamente computacionales o matemáticos que resuelven problemáticas complejas en estrecheces realmente alarmantes para el científico social —véanse los múltiples ejemplos expuestos en C. O'Neil (2017): *Armas de destrucción matemática*, Capitán Swing)—. Lo que hace necesario la apertura al diálogo entre diferentes ramas de conocimiento y el esfuerzo por reunir equipos especializados en los temas a investigar. Para esto, es indispensable que las ciencias sociales comprendan estas herramientas y concilien con el idioma del científico de datos, sin exigir la comprensión total de estos conocimientos.

Ante tal necesidad, los autores confeccionan todo un manual metodológico para principiantes, que anima a otra de las posiciones que puede tomar el o la investigadora social, (2) ser protagonista de una investigación con datos masivos para fenómenos sociales. En este caso, conociendo las posibilidades que las técnicas *big data* ofrecen, pudiendo encontrar en ellas, un camino para el descubrimiento de las cuestiones que plantean en sus investigaciones o el estudio de temáticas en las que son necesarias técnicas *big data* para la recolección y el análisis de datos. De esta forma, se anima a incluir nuevas metodologías en las investigaciones sociales que en muchos casos pueden resultar altamente eficientes y reveladoras.

Es aquí donde las reflexiones dejan paso al método, deteniéndose en las diferentes etapas y elementos necesarios para aquel o aquella que desee introducirse en estas técnicas. Los autores apuestan por el entorno de Jupyter en lenguaje python y ofrecen el código necesario para llevar a cabo todos y cada uno de los ejercicios. Aunque con un lenguaje técnico, cada uno de los ejemplos se exponen de forma asequible para noveles, pudiendo llevar a la práctica la teoría que de forma amena se imparte.

El trabajo se inicia con la distinción de tres novedosas fuentes de datos donde los investigadores pueden acudir: 1. Redes sociales, 2. Datos incluidos en páginas web y 3. Ficheros disponibles para descarga de distintos formatos. Este libro nos permite conocer la forma de acceder a datos de fuentes secundarias no usadas de forma clásica en las ciencias sociales, dando un valor añadido a las investigaciones que pueden recoger testimonios y perspectivas que de otro modo sería muy costoso o difícil de obtener.

En segundo lugar, se ofrecen al lector alternativas para el almacenamiento de estos datos, resolviendo problemas de espacio y velocidad, probables cuando se trata con datos masivos. Aunque formatos como Excel siguen siendo válidos, hospedar las bases de datos en la nube o acudir a bases de datos no relacionales puede incrementar notablemente la eficiencia a la hora de explotar la información recabada.

El grueso del libro, sin embargo, lo ocupa el capítulo sobre tratamiento y análisis computacional de los datos. Aquí se lleva a cabo un gran esfuerzo por sintetizar las técnicas de explotación para datos masivos, que como resaltan, es aplicable a datos estándar, pero con el añadido de poder escalar a grandes registros o datos de diferentes características. Es destacable la forma en la que se exponen conceptos preliminares y el posterior desarrollo de las diferentes técnicas encuadradas en el aprendizaje automático o *machine learning*. En un recorrido apto y comprensible para científicos sociales, que difícilmente se podrá encontrar en otros manuales que atienden al análisis estadístico de datos masivos. En este sentido, es el manual por excelencia para aquellos investigadores sin una base matemática o estadística que quieran implementar estas metodologías en sus investigaciones.

A pesar de que este libro concentra en pocas páginas toda la información y los pasos necesarios para una investigación exitosa, lo cierto es, que tal y como advierten los autores se trata mínimamente de una toma de contacto. Aunque las explicaciones se extienden con gran exactitud, elementos como la limpieza de datos —el 70 % del trabajo a llevar a cabo en una investigación de este calado— se atiende de forma muy débil; por ello, permite al investigador social entender el potencial del *big data* y su idioma para una tarea multidisciplinar, pero se trata sólo de alimentar la curiosidad para aquellos que quieran realizar un análisis de manera autónoma.

De igual forma, se echa en falta que las bases de datos y objetivos de los ejercicios prácticos, se orienten a temáticas más próximas a las ciencias sociales, de forma que el lector consiga alguna pista de qué tipo de problemáticas puede resolver con cada técnica. Esto es, sin embargo, una tarea compleja para un ejercicio de introducción por lo que con una alta probabilidad, se resuelva en trabajos posteriores.

Por último, aunque queda señalado en las primeras páginas, es altamente destacable y de gran valor, que los autores ofrezcan un espacio a la reflexión del encaje que detenta el *big data* como fenómeno en la investigación sociológica. El paradigma epistémico que mejor puede acoger este tipo de metodologías, lo consagran a la sociología analítica, caracterizada por alejarse de elementos reflexivos de explicación teórica, acudiendo al análisis de tendencia para generar marcos explicativos de los fenómenos.

Si bien, ahondar en estos temas no es objeto del libro y por tanto de esta reseña, se agradece la invitación a que los y las científicas sociales tomen la tercera posición posible: (3) estudiar desde la teoría sociológica el *big data* como fenómeno, tanto en las formas particulares que intervienen y problematizan los estadios sociales, como el debate teórico epistémico que debe ser alimentado por la reflexión profunda sobre el futuro de estas metodológicas en las ciencias sociales.

Este es, por tanto, un libro que viene a poner la primera semilla en un terreno relativamente virgen. Caracterizado por su increíble novedad y atrevimiento, que no sorprenderá ver en las aulas de metodología en los próximos años, y al que, cuando el paso del tiempo haya permitido la digestión de estos debates, con mucha probabilidad, gran parte de las ciencias sociales le ofrezca el justo reconocimiento.