Ciencias administrativas
REVISTA DIGITAL FCE - UNLP

Análisis de calidad de los datos en las estadísticas públicas y privadas, ante la implementación del Big Data.
Fernando Ariel Manzano| Daniela Avalos
Ciencias Administrativas | Año 11 | N° 22 Julio – Diciembre 2023
https://doi.org/10.24215/23143738e119
ISSN 2314 – 3738 | http://revistas.unlp.edu.ar/CADM

ARTÍCULO CIENTÍFICO

ANÁLISIS DE CALIDAD DE LOS DATOS EN LAS ESTADÍSTICAS PÚBLICAS Y PRIVADAS, ANTE LA IMPLEMENTACIÓN DEL BIG DATA

QUALITY ANALYSIS OF DATA IN PUBLIC AND PRIVATE STATISTICS, IN LIGHT OF BIG DATA IMPLEMENTATION

FERNANDO ARIEL MANZANO | https://orcid.org/0000-0002-1513-4891 | fernando14979@hotmail.com | Instituto de Geografía, Historia y Ciencias Sociales. Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina.

DANIELA AVALOS | https://ordic.org/0000-0002-0762-8437 | danielasoledad.av@gmail.com | Facultad de Ciencias Sociales, Universidad de Buenos Aires, Argentina.

Resumen / Abstract

El incremento del almacenaje y explotación de los datos a partir del avance de las tecnologías de mediación digital generó cambios en la gestión de los datos, así como también presentó nuevos retos en relación a su calidad. A partir de una amplia búsqueda de información especializada, el presente estudio, de carácter exploratorio, tiene como objetivo describir las diferentes valoraciones respecto a la calidad de los datos en el entorno de la producción de las estadísticas públicas y en el sector empresarial privado, examinando en particular los cambios producidos en el marco de la calidad estadística en ambos contextos, a partir de las potencialidades de la gestión del uso del Big Data.

Palabras clave: Big Data; calidad de los datos; estadísticas oficiales; estadísticas privadas; privacidad; privacidad de los datos.

The increase of data storage and exploitation since the progress of digital mediation technologies, generated changes in data management, as well as presented new challenges in relation to the quality of the data. Based on a broad search for specialized information., this current exploratory study aims to describe the different assessments regarding the quality of the data in the production of public statistics environment and in the private business sector through the examination in particular of the changes produced in the framework of statistical quality in both contexts, based on the potentialities of managing the use of Big Data.

Key Words: Big Data; data quality; official statistics; private statistics; privacy; data privacy.

| Clasificación JEL: Y10, C00 | Recibido: 29/12/2021 | Aprobado: 22/06/2022 | Publicado: 31/03/2023 |

Introducción

El desarrollo tecnológico ha tenido una gran incidencia en los procesos de producción. En la actualidad la revolución 4.0, la tecnología de la información, la inteligencia artificial[1], entre otras tecnologías digitales (Cortés et al., 2017; Serna, 2021) abrieron el camino para el tratamiento de grandes volúmenes de datos, mediante el concepto denominado Big Data, utilizando para esto algoritmos matemáticos (Lecuona, 2019). El desarrollo de Big Data y la analítica de datos –como sus disciplinas asociadas– (Jauregi-Maza, 2021) incluyen al conjunto de herramientas, técnicas y sistemas destinados a extraer todo el valor de los datos para enriquecer y complementar sistemas con capacidades predictivas (Mejías Alonso y Soler Jiménez, 2017). Para el sector privado los datos constituyen una nueva clase de activo económico y, por lo tanto, son demandados por actores del mercado (Alonso-Arévalo y Vázquez Vázquez, 2016). Además, los adelantos digitales en el sector empresarial alteraron las formas de tomar decisiones –muchas empresas toman decisiones basadas en el Big Data (Zúñiga, 2019)–, así como dieron lugar a nuevos modelos de negocios en el mercado digital (Labbé Figueroa, 2020).

Los avances en la recolección y análisis de datos poseen limitaciones conceptuales y operativas. Asumir altos niveles de errores estadísticos y que la confiabilidad de los datos pueda ser violada en cualquier punto en su ciclo de vida son características propias del Big Data (Galimany Suriol, 2015). Por ejemplo, frente a las tradicionales muestras probabilísticas –donde todos los elementos de la población tienen la misma probabilidad de ser seleccionados–, en Big data difícilmente se pueda contar con muestras de este tipo, dado que no hay forma de saber cuán incompletos se encuentran los datos que son almacenados por empresas (Burgess y Bruns, 2012; Gualda, 2022). Respecto a la calidad de los datos, otra limitación es que no existe información contrafáctica que los respalde; por definición los datos de Big Data son observacionales e incomparables con experimentos científicos (Carretero y Velthuis, 2018; Sosa Escudero, 2021), siendo el principal reto para las organizaciones –públicas y privadas– gestionar la calidad dado el crecimiento en la escala de dichos datos (Ontiveros y López Sabater, 2017).

No todos los datos son accesibles, esto depende de las políticas de encriptación de las compañías dueñas de los servidores (Rojo y Sánchez, 2019). La disponibilidad de información por parte de terceros depende de si las empresas desean compartir la información y del precio que fijen para acceder a los datos (Labbé Figueroa, 2020; Manovich, 2012; Stroe, 2020). Además, no toda la información que se encuentra en la red es fiable. Durante el procesamiento de grandes volúmenes de datos, se producen anomalías, debido a las siguientes causas: datos inconsistentes o incompletos, registros duplicados, valores perdidos, representaciones no convencionales de datos, entre otras (Aldana et al., 2018; Koudas et al., 2006). Dado que la calidad de los datos está directamente relacionada con la calidad del conocimiento que pueda extraerse –mediante la utilización de algoritmos de extracción de conocimiento– (García et al., 2016), los resultados erróneos impactan en la eficiencia de las empresas, elevando los costos y disminuyendo los beneficios del análisis e interpretación (López, 2011; López Porrero et al., 2010). Por tanto, las empresas llevan adelante mejoras de la calidad de los datos, limpiando los datos de baja calidad (Garzón Arredondo, 2015).

Por otra parte, la estadística nacional representa un bien público estratégico, basado en el trabajo de los institutos de estadística, que tienen la responsabilidad de generar, difundir y resguardar información estadística de calidad para la planificación y ejecución de las políticas públicas (Gauna et al., 2020). La calidad de la información se rige por normas que regulan su función –como el secreto estadístico que regula la protección de datos de carácter personal– y poseen compromisos con la aplicación del Código de Buenas Prácticas estadísticas (Tourís López, 2019). Para esto, se realizan tareas de comprobación de coherencia, estandarización de conceptos y métodos, desarrollo de marcos metodológicos, procesamiento de datos, entre otros procesos, a los fines de prevenir diferentes fuentes de invalidez y errores que afectan la calidad de los datos –guiando el modo de corregirlos– (Brackstone, 2003; Giusti y Massé, 1997; Naciones Unidas, 2011).

Para la gestión en beneficio público de grandes datos como fuente secundaria que no fueron generados para fines estadísticos, como el Big Data, los sistemas de estadísticas oficiales deben asegurar que la información recopilada sea de calidad, imparcial y su acceso sea posible para la comunidad de forma

1. El concepto de Inteligencia artificial se asocia a sistemas de computación que recaban información de diferentes fuentes, con capacidad para automatizar actividades y, mediante el uso de algoritmos, realizar operaciones análogas al aprendizaje y toma de decisiones, y evolucionar con o sin la intervención humana (Rouhiainen, 2018, p. 17).

simultánea, satisfaciendo sus necesidades, también que sea útil y pertinente para la toma de decisiones (Fernández y Ferrer, 2016; Galimany Suriol, 2015; Molina y Mérida, 2021). Por el contrario, las empresas privadas están dispuestas a utilizar datos de baja calidad, sacrificando precisión estadística de los macrodatos (Garzón Arredondo, 2015; Zarzky, 2016). Así, buscan maximizar la generación de valor económico de la información generada (Garzón Arredondo, 2015).

A partir de una amplia revisión de información especializada, el presente estudio, de carácter exploratorio, tiene como objetivo describir las diferentes valoraciones respecto a la calidad de los datos en el entorno de la producción de las estadísticas públicas y en el sector empresarial privado, examinando en particular los cambios producidos en el marco de la calidad estadística en ambos contextos a partir de las potencialidades de la gestión del uso del Big Data.

La calidad de los datos en las estadísticas oficiales

Si bien no existe una definición de calidad de la información estadística acordada internacionalmente, en las Oficinas de Estadística tradicionalmente se ha tenido en cuenta la exactitud como el aspecto principal de la calidad (Arribas et al., 2003).

A partir del año 1994 la Organización de las Naciones Unidas (ONU) emitió los Principios Fundamentales de las Estadísticas Oficiales, incluyendo criterios para la pertinencia, el profesionalismo y la apertura de los datos (Arribas et al., 2003; Brackstone, 2003;). Se ha ido ampliando el concepto tradicional de calidad, enfocándose en los criterios de relevancia, exactitud, oportunidad y puntualidad, accesibilidad y claridad, comparabilidad, coherencia, completitud y la satisfacción de las expectativas de los diferentes usuarios (Arribas et al., 2003; Elvers, 2002; Naciones Unidas, 2004). El principio 5 de los Principios Fundamentales expresa que los datos para fines estadísticos pueden obtenerse de todo tipo de fuentes, pero no considera explícitamente la situación del *Big Data* (Salgado, 2017). Para abordar las nuevas fuentes de información, se creó un Grupo de Trabajo Mundial de Naciones Unidas (United Nations Global Working Groupon Big Data, 2016). En el año 2014 en Beijing este grupo realizó la primera Conferencia Internacional sobre Big Data en las Estadísticas Oficiales. Con el objetivo de promover el uso de estas fuentes, siendo el desafío contar con datos confiables, se estimuló la colaboración con el sector privado, la comunidad académica y la sociedad civil (Bussi et al., 2017; Lindenboim, 2011; Naciones Unidas, 2010;). Además, se emprendió la búsqueda de metodologías adecuadas desde un enfoque multidimensional (Naciones Unidas, 2015). Este contempla el grado de calidad, la oportunidad, el nivel de precisión, el costo y la carga que implica el uso de las fuentes de Big Data. Dentro de la comunidad estadística en los últimos años se vienen realizando actividades de capacitación y divulgación de experiencias, dirigiendo los esfuerzos hacia el marco de calidad de los macrodatos (Bussi et al., 2017; Lineros, 2017).

Por otra parte, en la generación de datos primarios en las estadísticas oficiales –nacionales o internacionales–, obtenidos a través de censos o encuestas, se deciden cuáles son los agregados o índices que desean conocerse antes del diseño y ejecución de la operación estadística apropiada. Previo a la realización de un operativo estadístico, se examinan los aspectos metodológicos para evitar la presencia de conceptos no claros o erróneos. Los institutos de estadísticas oficiales informan a las personas o empresas que han sido seleccionadas en la muestra. Los informantes tienen la obligación de responder y el Estado la obligación de guardar el secreto estadístico de los datos relevados.

Durante el proceso de evaluación –de las fuentes primarias– se analiza la información a los fines de detectar diferentes errores, como ser omisiones o duplicaciones de unidades de relevamiento –sean estas de población, vivienda, hogar–; errores de los encuestadores o de los declarantes; no respuestas de preguntas; inconsistencias entre variables de un mismo registro o entre diferentes registros de una misma unidad de relevamiento; equivocaciones en la codificación y la introducción de datos; errores en la revisión manual o informatizada de los datos; entre otros (Arribas et al., 2003).

Se identifican las características de los errores y sus magnitudes, para luego aplicar correcciones siguiendo criterios establecidos (Giusti y Massé, 1997; CEPAL, 2020). Es responsabilidad de los institutos de estadística informar a los usuarios, a través de documentos metodológicos y de metadatos, cuáles son los distintos errores existentes y las modificaciones realizadas (Giusti y Massé, 1997; CEPAL, 2020; Instituto

Nacional de Estadística de España, 2015). Cabe destacar que durante este proceso de revisión deben conservarse los originales de esta fuente primaria en la mayor medida posible (CEPAL, 2011).

En el caso del uso de fuentes secundarias seleccionadas por las estadísticas oficiales, su tratamiento es diferente a las fuentes primarias. Por ejemplo, el uso de registros administrativos requiere de un proceso de conversión estadística, centrado en el análisis de parámetros de calidad que, por la cantidad y diversidad de objetivos, variables, métodos de registro, tratamiento y archivo de los registros –ajenos a los fines estadísticos– resulta complejo (Gauna et al., 2020).

Una arquitectura de Big Data[2] como fuente de datos secundaria implica para los institutos de estadísticas lidiar con un nivel de calidad aun inferior a los registros administrativos, dado que estas fuentes de información se caracterizan por un bajo nivel de estructuración de los datos[3] –resulta difícil poder asociarlos a una población o fenómeno de referencia–, alta diversidad de contenidos y fuentes –aportan elementos negativos como ruido, valores perdidos, inconsistencias, datos superfluos, etc.– y datos fuertemente dependientes y referidos a poblaciones marcadamente heterogéneas –lo contrario de las condiciones de independencia y distribución idéntica que intenta garantizar la estadística tradicional– (García et al., 2016; Monleon-Getino, 2015; Sosa Escudero, 2021). Además, el acceso a estas fuentes contradice el principio universal de no pagar por los datos para fines estadísticos oficiales (United Nations Global Working Group on Big Data, 2016). En este sentido, cobra relevancia la generación de estudios de viabilidad convincentes para que la comunidad estadística se posicione de forma unificada en relación a la utilidad de macrodatos en las estadísticas públicas.

El Big Data y la cadena de valor del dato en el sector privado

En la actualidad, es improbable un modelo de negocios que no se relacione con alguna de las formas de usar, almacenar, analizar o distribuir datos (Labbé Figueroa, 2020). Por lo tanto, el Big Data se transformó en un nuevo y significativo activo que alimenta la economía de la información (Rubinfeld y Gal, 2017). Las empresas capturan el mayor volumen de datos, para luego clasificarlos y transformarlos, con el fin de comercializarlos –proceso denominado datificación de la realidad[4]–. Un aspecto relevante es conocer si la información externa e interna de la organización resulta de utilidad para los diferentes consumidores, en función de esta última se determinará el grado de valor del Big Data (Labbé Figueroa, 2020).

Dado que a medida que crece la diversidad de los volúmenes de datos recolectados, aumenta la probabilidad de que se encuentren problemas en los datos y el desconocimiento de la naturaleza de estos (Nuñez-Arcia et al., 2016), se produce un viraje, dando prioridad a la calidad del dato por sobre la cantidad. Pasando de una definición de Big Data basada en las confluencias de solo 3Vs –volumen, velocidad y variedad (Laney, 2001)– a otra más amplia en donde se adicionan tres nuevas Vs, que hacen referencia a las características de variabilidad, veracidad y valor. Este nuevo paradigma da prioridad a la precisión de los datos, la capacidad de adaptarse al entorno cambiante de los negocios y la disponibilidad para ser utilizados en la toma de decisiones (García et al., 2016; Pereira Villazón et al., 2019). En los últimos años, emergió el término Smart Data, vinculado al proceso de filtrar el ruido de los datos almacenados y transformarlos en datos de calidad. Pero los métodos de análisis de calidad de los datos –junto a los *softwares* utilizados en este proceso– se transformaron en la principal limitación económica para las organizaciones (Cárdenas, 2009; Stucke y Grunes, 2016;).

Para abordar el fenómeno de los datos abiertos, la principal disciplina es la Ciencia de Datos. Su objetivo es estudiar el ciclo de vida del dato en distintos campos de aplicación (Pérez-Rave et al., 2019). Existen múltiples versiones de la cadena de valor del dato; entre las más habituales cabe mencionar:

-Fase de provisión del dato: se releva la información que genera la empresa, y se combina con información externa –recopiladas de fuentes muy heterogéneas– (Galimany Suriol, 2015). Esta selección

2. Una arquitectura Big Data es el diseño de sistemas y modelos que relacionan e integran las nuevas tecnologías y herramientas para procesar grandes volúmenes de datos con la infraestructura de una empresa u organización (Joyanes Aguilar, 2013).

3. Los datos pueden diferenciarse en estructurados (bases de datos relacionales), semi-estructurados (archivos HTML, XML, etc.) y no estructurados (fotos, videos, audios, textos, etc.) (Joyanes Aguilar, 2013).

4. Mayer-Schönberger y Cukier (2013) utilizan el término *datificación* para referirse al proceso por el cual se recopila la información, ya sean localizaciones de individuos, vibraciones de los motores o páginas web visitadas para luego transformarla en datos que sean cuantificables para así poder aplicar el análisis predictivo y obtener algún valor de la información.

depende de los objetivos del negocio y de las dificultades en la inserción de nuevos datos (Nuñez-Arcia et al., 2016; Timarán-Pereira et al., 2016). Dada la diversidad de datos, estos tienen diferencias en el tipo –estructurados, semiestructurados o desestructurados–, la frecuencia en que presentan –tiempo real, *near real time* o *batch–,* los niveles de exactitud y calidad de los mismos (Molina y Mérida, 2021). Por lo tanto, los arquitectos de datos seleccionan la tecnología adecuada y los ingenieros de datos elaboran el *software* para integrar la información de los sistemas en la plataforma Big Data, para que pueda ser utilizada de forma correcta por los científicos de datos.

-Fase de transformación: Una vez almacenada la información en el sistema de Big Data, se aplican las técnicas para mejorar la calidad del dato, como tareas de limpieza de datos desconocidos, nulos, duplicados, inconsistentes, imputación de datos faltantes y la reducción del número de variables en función de los algoritmos de extracción (Hernández y Rodríguez, 2008; Timarán-Pereira et al., 2016). El tiempo de limpieza necesario para que los datos puedan ser procesados –denominado *data wrangling*– absorbe el 80% del tiempo de trabajo de los científicos de datos, reduciendo el tiempo dedicado a su actividad principal, que es la analítica de datos (García Del Río y López Contreras, 2018; Mons, 2020; Villao Balón, 2021).

-Fase de descubrimiento y modelado: proceso de análisis por parte de los analistas de negocio y los científicos de datos con el objetivo de extraer el máximo valor transformando los datos en información (García et al., 2016). Por lo tanto, la analítica de datos constituye, en sí misma, un activo de valor muy importante, al proveer la obtención de patrones sobre los datos, obteniendo así predicciones y prevenciones de eventos (Moreno y Calderón, 2016). Para llevar a cabo las tareas de extracción de conocimiento –o información con valor de negocio– de los grandes volúmenes de datos, se utiliza la minería de datos (Martínez, 2001; Moine et al., 2011). Este proceso consiste en un conjunto de técnicas, basadas en estadística e inteligencia artificial[5], que permite encontrar información oculta o implícita en los patrones de comportamiento de los datos, que no es posible obtener mediante métodos estadísticos convencionales (Timarán-Pereira et al., 2016; Zamorano Ruiz, 2018). Los algoritmos generados para los diferentes análisis –dependiendo del tipo de datos utilizados– de hechos ocurridos o que están sucediendo incluyen seleccionar los modelos más apropiados. Los modelos pueden ser predictivos o descriptivos. Los primeros buscan medir los valores futuros o desconocidos de variables de interés, con el objetivo de estimar eventos y comportamientos, dando lugar a los posibles cambios en las estrategias de negocio (Matteucci, 2020; Morello, 2018; Timarán-Pereira et al., 2016), mientras que los últimos permiten observar patrones que explican los datos, explorando sus propiedades (Timarán-Pereira et al., 2016).

-Fase de exposición: Dado que no siempre es posible extraer valor de la información generada, las tres fases anteriores se realizan de la manera más económica. Sin embargo, el principal inconveniente es la manipulación de los datos en los entornos analíticos. El sistema denominado Gobierno de Datos (DG) –*Data Governance*, en inglés– ha surgido para comunicar eficientemente, dentro de la organización, las definiciones, políticas y normas de datos –estandarización de formatos–, en línea con maximizar el beneficio de la actividad empresarial (Carretero y Velthuis, 2018; Wiseman, 2018), siendo el principal responsable de establecer las bases de la gobernanza de datos el *chief data officer* (CDO) (Fernández, 2020). El CDO debe establecer mejoras en la calidad de los datos, optimizando la utilización de los activos de datos existentes, para incrementar la eficiencia operativa de la empresa (Teerlink et al., 2014).

Finalmente, se consolida el conocimiento descubierto para reportarlo a las partes interesadas, aplicando herramientas gráficas de visualización, traduciendo la información, para presentar resultados significativos y didácticos para los distintos usuarios (Cortés Rodríguez, 2020; Sancho et al., 2014; Timarán-Pereira et al., 2016).

Discusión

En los últimos años, el uso intensivo de las tecnologías de la información y de las comunicaciones (TIC) aumentó la demanda de los usuarios de la apertura de datos para ser reutilizados, conduciendo al surgimiento de los

5. El diseño de algoritmos de aprendizaje automático –*machine learning*, en inglés– es una rama de la Inteligencia artificial, que permite a las máquinas aprender y tomar decisiones con futuros datos proporcionados de forma automática (Zamorano Ruiz, 2018).

datos públicos abiertos (DA) u *open data* –por su denominación en inglés– (Christodoulou et al., 2018). Es fácil argumentar el potencial valor de los macrodatos para la optimización de las operaciones de las empresas del sector privado. No obstante, el Big data puede facilitar ciertas conductas anticompetitivas en la industria digital (Labbé Figueroa, 2020).

Un gran número de investigaciones remarcan que las fuentes de información de una estructura Big Data en el sector público y para la elaboración de estadísticas oficiales contienen diversas problemáticas, muchas de ellas vinculadas entre sí, como la falta de evidencia del fenómeno relevado (Luo et al., 2019), la metodología estadística para realizar inferencias respecto de las poblaciones de interés (Salgado, 2017), la preeminencia del volumen de información por sobre la precisión analítica (Del-Fresno-García, 2014), como también la falta de reglas implementadas, el incorrecto manejo de los datos cambiantes, los problemas de duplicación, errores tipográficos e información falsa o basada en percepciones subjetivas. Así mismo, la integración de fuentes diversas conlleva, en sí misma, problemas de calidad (Monleon-Getino, 2015; Nuñez-Arcia et al., 2016; Rodríguez et al., 2017), como la ausencia de los metadatos necesarios para usar la información con propósitos estadísticos (Salgado, 2017).

Estas dificultades originan una baja calidad de una arquitectura de Big Data –elevada proporción de datos no estructurados y erróneos (Paliotta, 2018)–, expresada en la poca usabilidad de esta (Pérez-Rave et al., 2019). También existe la dificultad de acceso institucional a los datos –cambios legales en los países–. Si bien los institutos nacionales de estadísticas poseen potestad de solicitar información a toda persona legal, física o jurídica –pública o privada–, la información de las distintas fuentes de Big Data corresponde a terceras personas. Ella desempeña un elemento central en la actividad económica del proveedor de datos –servicios de salud, telecomunicaciones, entre otros, que suelen tener legislaciones específicas– (Salgado, 2017).

Por último, muchas nuevas fuentes de datos digitales no se adecuan a las responsabilidades de las Estadísticas Oficiales. Estas últimas, en el marco de la gestión de DA, deben velar por la calidad de los datos –cumpliendo con los estándares de calidad y confidencialidad– (Salvador y Ramió, 2020; Vásquez Valdivia, 2021), cobrando especial importancia atender a la evidencia, la seguridad y privacidad del contenido de los nuevos datos (Christodoulou et al., 2018). Además, dado los principios éticos elementales del sector público, se debe garantizar la transparencia e identificar la existencia de errores de estos grandes volúmenes de datos (Paliotta, 2018).

Conclusiones

En el sector empresarial privado los adelantos digitales alteraron las formas de tomar decisiones, así como también nuevos modelos de negocios en el mercado digital vinculada a la comercialización de datos. En este sentido se busca diseñar distintos algoritmos predictivos que permitan generar información con mayor valor económico. Esta potencialidad depende directamente de la calidad de los datos.

La infraestructura Big Data no es una fuente diseñada para fines estadísticos. Se caracteriza por el alto contenido de errores que afecta la calidad de los datos. Ante el aumento constante del flujo de los repositorios de datos, son necesarios cada vez métodos más sofisticados para generar datos de utilidad para los consumidores, destacándose un aumento de los costos vinculado a la limpieza de los datos, las políticas de encriptación, y un valor económico oscilante en función de los diferentes tipos de uso de los consumidores. En un mercado digital altamente cambiante, dada la dificultad de generar ganancias de la información generada, se suele sacrificar el nivel de precisión estadística del contenido de los nuevos datos.

Los institutos de estadística se rigen por normas nacionales que regulan su funcionamiento y principios internacionales de buenas prácticas, dirigidas a resguardar la calidad estadística, siendo responsabilidad de los institutos de estadística informar a los usuarios a través de los metadatos el grado de precisión y confiabilidad de los datos generados. El componente de la calidad estadística oficial se fue ampliando en los últimos años, en relación a la vialidad de utilización de los macrodatos, la recomendación señala la necesidad de generar un marco de calidad especifico. Esta fuente secundaria requiere una conversión estadística, al igual que los registros administrativos. Pero a diferencia de estos últimos, las limitaciones conceptuales y operativas para su utilización son muy superiores.

Referencias

Aldana, H. S. M., Rivas, J. D. C. e Hidalgo, J. M. V. (2018). Big Data, el futuro de las predicciones certeras. *Revista Avenir*, *2*(2), 10-16.

Alonso-Arévalo, J. y Vázquez Vázquez, M. (2016). Big Data: La próxima "gran cosa" en la gestión de la información. *BID: textos universitaris de biblioteconomia i documentació*, *36*, 1-3.

Arribas, C., Casado, J. y Martínez, A. (2003). *Gestión orientada a asegurar la calidad de los datos en los Institutos Nacionales de Estadística*. Repositorio digital de la Comisión Económica para América Latina y el Caribe. https://repositorio.cepal.org/bitstream/handle/11362/16455/S034227_es.pdf

Brackstone, G. (2003). *Gestión de la calidad de los datos en un organismo estadístico*. Repositorio digital de la Comisión Económica para América Latina y el Caribe. http://hdl.handle.net/11362/16464

Burgess, J. y Bruns, A. (2012). Twitter Archives and the Challenges of "Big So-cial Data" for Media and Communication Research. *M/C Journal*, *15*(5). https://doi.org/10.5204/mcj.561

Bussi, J., Marí, G. P. y Méndez, F. (2017). *El desafío del big data en estadísticas oficiales en Argentina*. Facultad de Ciencias Económicas y Estadística de la Universidad de Rosario.

Cárdenas, G. C. (2009). Las herramientas de software para el análisis cuantitativo de información: Estudio preliminar de aplicaciones desarrolladas en Cuba. *Bibliotecas. Anales de investigación*, (5), 53-62.

Carretero, A. I. G. y Velthuis, M. P. (2018). Importancia de la calidad de los datos en la transformación digital. *RUIDERAe: Revista de Unidades de Información*, (13), ISSN-e 2254-7177.

CEPAL (2011). *Guía para asegurar la calidad de los datos censales* (Serie N°74). Repositorio digital de la Comisión Económica para América Latina y el Caribe. http://hdl.handle.net/11362/5515

CEPAL (2020). *Ley Genérica sobre Estadísticas Oficiales para América Latina* (LC/CEA.10/8). Repositorio digital de la Comisión Económica para América Latina y el Caribe. http://hdl.handle.net/11362/45253

Christodoulou, P., Decker, S., Douka, A. V., Komopoulou, C., Peristeras, V., Sgagia, S., Tsarapatsanis, V. y Vardouniotis, D. (2018). Data Makes the Public Sector Go Round. En P. Parycek, O. Glassey, M. Janssen, H. J. Scholl, E. Tambouris, E. Kalampokis, S. Virkar (eds.), *Electronic Government. EGOV 2018. Lecture Notes in Computer Science* (vol. 11.020). Springer. https://doi.org/10.1007/978-3-319-98690-6_19

Cortés, C. B. Y., Landeta, J. M. I. y Chacón, J. G. B. (2017). El entorno de la industria 4.0: implicaciones y perspectivas futuras. *Conciencia tecnológica*, (54), 33-45.

Cortés Rodríguez, K. I. (2020). *Calidad de datos contextual en Big Data: calidad de datos de Twitter. Escuela de Ingeniería y Gestión* [Tesis de grado, Instituto Tecnológico de Buenos Aires (ITBA)]. http://ri.itba.edu.ar/handle/123456789/3184

Del-Fresno-García, M. (2014). Haciendo visible lo invisible: Visualización de la estructura de las relaciones en red en Twitter por medio del análisis de redes sociales. *El Profesional de la Información*, *23*(3), 246–252. https://doi.org/10.3145/epi.2014.may.04

Elvers, E. (Agosto de 2002). *Comparison of Survey and Register Statistics*. The International Conference on Improving Surveys, Denmark, University of Copenhagen.

Fernández, A. (2020). El papel del Big Data en el reporting y la toma de decisiones. *Revista de Contabilidad y Dirección*, *31*, 21-36.

Fernández, Y. A. y Ferrer, D. C. (2016). Big Data: una herramienta para la administración pública. *Ciencias de la Información*, *47*(3), 3-8.

Galimany Suriol, A. (2015). *La creación de valor en las empresas a través del Big Data* [Tesis de Grado, Universidad de Barcelona]. Dipòsit Digital de la Universitat de Barcelona. http://hdl.handle.net/2445/67546

García, S., Ramírez-Gallego, S., Luengo, J. y Herrera, F. (2016). Big Data: Preprocesamiento y calidad de datos. *Novática*, *237*,17, 17-23.

García Del Río, A. y López Contreras, I. R. (2018). *Implementación de herramientas de extracción, transformación y carga de datos estructurados en Big Data* [Tesis de Grado, Universidad Autónoma de ciudad Juárez]. http://hdl.handle.net/20.500.11961/4665

Garzón Arredondo, A. (2015). *Evolución e impacto de Big Data en empresas grandes de diferentes industrias del sector corporativo en Antioquia* [Tesis de Doctorado, Universidad EAFIT]. Repositorio Institucional Universidad EAFIT

Gauna, N., Roggi, C. y Zuloaga, N. (2020). Los registros administrativos en la construcción y consolidación del Sistema Estadístico de la Ciudad. *Población de Buenos Aires*, *17*(29), 43-49.

Giusti, A. y Massé, G. (1997). Aspectos conceptuales relativos a la evaluación de calidad. En INDEC, *Evaluación de la calidad de datos y avances metodológicos* (Serie J n° 2). Secretaría de Programación Económica, Ministerio de Economía y Obras y Servicios Públicos.

Gualda, E. (2022). Social big data, sociología y ciencias sociales computacionales. *Empiria: Revista de metodología de ciencias sociales*, (53), 147-177.

Hernández, C. y Rodríguez, J. E. R. (2008). Preprocesamiento de datos estructurados. *Revista vínculos*, *4*(2), 27-48.

Instituto Nacional de Estadística de España (INE) (2015). *Política de revisión del Instituto Nacional de Estadística*. https://www.ine.es/ine/codigobp/politica_revision.pdf

Jauregi-Maza, L. (2021). Big Data: la revolución de los datos masivos en la Administración Pública. *Inguruak. Revista Vasca de Sociología y Ciencia Política*, (71), 79-100. http://dx.doi.org/10.18543/inguruak-71-2021-art05.

Joyanes Aguilar, L. (2013). *Big Data: Análisis de grandes volúmenes de datos en organizaciones*. Alfaomega.

Koudas, N., Sarawagi, S. y Srivastava, D. (2006). Record linkage: similarity measures and algorithms. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM*, 802-803.

Labbé Figueroa, M. F. (2020). Big Data: Nuevos desafíos en materia de libre competencia. *Revista chilena de derecho y tecnología*, *9*(1), 33-62.

Laney, D. (2001). *3D Data management: controlling data volume, velocity y variety*. META Group. https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...#

Lecuona, I. D. (2019). Evaluación de los aspectos metodológicos, éticos, legales y sociales de proyectos de investigación en salud con datos masivos (big data). *Gaceta Sanitaria*, *32*, 576-578.

Lindenboim, J. (2011). Las estadísticas oficiales en Argentina ¿Herramientas u obstáculos para las ciencias sociales? *Trabajo y Sociedad*, *15*(16), 19-38.

Lineros, E. M. (2017). El trinomio dato-información-conocimiento. En P. Díaz (Ed.) M*anual sobre utilidades del big data para bienes públicos* (pp. 35-48). Edimema.

López, B. (2011). Limpieza de Datos: Reemplazo de valores ausentes y Estandarización [Tesis de Doctorado, Universidad Central "Marta Abreu" de Las Villas)]. http://dspace.uclv.edu.cu:8089/handle/123456789/7213

López Porrero, B., Pérez Vázquez, R. y Batule Domínguez, M. (2010). Las reglas de asociación ordinales en la detección de errores en los datos. *Revista Cubana de Ciencias Informáticas, 4*(1-2), 47-52.

Luo, J.-D., Liu, J., Yang, K. y Fu, X. (2019). Big data research guided by sociological theory: A triadicdialogue among big data analysis, theory, and predictive models. *The Journal of Chinese Sociology*, *6*(11). https://doi.org/10.1186/s40711-019-0102-4

Manovich, L. (2012). Trending: ThePromises and theChallenges of Big Social Data. En M. Gold (Ed.), *Debates in the Digital Humanities*. University of Minnesota Press.

Martínez, G. (2001). Minería de datos: Cómo hallar una aguja en un pajar. *Ingenierías*, *14*(53), 53-66.

Matteucci, M. A. (2020). ¿Es posible el uso de Big data en materia tributaria? Instituto de Extrapolítica Y Transhumanismo (IET) de Lima, Perú.

Mayer-Schönberger, V. y Cukier, K. (2013). *Big data: la revolución de los datos masivos*. Turner.

Mejías Alonso, E. y Soler Jiménez, J. (2017). *La vigilancia y el control de la población a través de la gestión, la conservación y la explotación de datos masivos*. Universidad Autónoma de Barcelona.

Moine, J. M., Haedo, A. S. y Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. En *XIII Workshop de Investigadores en Ciencias de la Computación*. Universidad Tecnológica Nacional, Rosario.

Molina, V. H. Á. y Mérida, A. F. (2021). Datificación crítica: práctica y producción de conocimiento a contracorriente de la gubernamentalidad algorítmica. Dos ejemplos en el caso mexicano. *Administración Pública y Sociedad (APyS)*, (11), 211-231.

Monleon-Getino, A. (2015). El impacto del big data en la sociedad de la información. Significado y utilidad. *Historia y Comunicación Social*, *20*(2), 427-445. https://doi.org/10.5209/rev_HICS.2015.v20.n2.51392

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, *578*, 491. https://doi.org/10.1038/d41586-020-00505-7

Morello, F. (25 de enero de 2018). Analítica avanzada para la transformación digital. https://www.df.cl/opinion/columnistas/analitica-avanzada-para-la-transformacion-digital

Moreno, L. P. y Calderón, C. C. A. (2016). Empleo de Big Data en la gestión de las Telecomunicaciones. *Tono, Revista Técnica de la Empresa de Telecomunicaciones de Cuba SA*, *13*(2), 48-57.

Naciones Unidas. (2004). *Manual de Organización Estadística*. https://www.cepal.org/es/publicaciones/3976-manual-organizacion-estadistica-funcionamiento-organizacion-oficina-estadistica

Naciones Unidas. (2010). *Principios y recomendaciones para los censos de población y habitación. Revisión 2* (Informes Estadísticos Serie M No. 67/Rev.2). Departamento de Asuntos Económicos y Sociales. División de Estadística. https://unstats.un.org/unsd/publication/seriesm/seriesm_67rev2s.pdf

Naciones Unidas. (2011). *Manual de revisión de datos de los censos de población y vivienda. Revisión 1*. Departamento de Asuntos Económicos y Sociales. División de Estadística.

Naciones Unidas. (2015). *Informe del Grupo de Trabajo Mundial sobre los Macrodatos en las Estadísticas Oficiales* (E/CN.3/2016/1). https://www.cepal.org/es/publicaciones/3976-manual-organizacion-estadistica-funcionamiento-organizacion-oficina-estadistica

Nuñez-Arcia, Y., Díaz-de-la-Paz, L. y García-Mendoza, J. L. (2016). Algoritmo para corregir anomalías a nivel de instancia en grandes volúmenes de datos utilizando MapReduce. *Revista Cubana de Ciencias Informáticas*, *10*(3), 105-118.

Ontiveros, E. y López Sabater, V. (2017). *Economía de los Datos. Riqueza 4.0*. Ariel y Fundación Telefónica. Barcelona

Paliotta, A. P. (2018). Nuevas profesiones y técnicas de web data mining en Argentina: el caso del Data Scientist. *Revista del Centro de Estudios de Sociología del Trabajo (CESOT)*, (10), 63-94.

Pereira Villazón, T., Portilla Manjón, I. y Rodríguez Salcedo, N. (2019). Big data y Relaciones Públicas: Una revisión bibliográfica del estado de la cuestión. *Revista de comunicación*, *18*(1), 151-165.

Pérez-Rave, J., Correa Morales, J. C., y González Echavarría, F. (2019). Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín. *Ingeniare. Revista chilena de ingeniería*, *27*(3), 495-509.

Rodríguez, P., Palomino, N. y Mondaca, J. (2017). *El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe*. Banco Interamericano de Desarrollo.

Rojo, I. D. J. P. y Sánchez, A. A. C. (2019). Reinsurgencia de la etnografía en la era del Big Data: apuntes desde el sur global. *Virtualis*, *10*(19), 42-56.

Rouhiainen, L. (2018). *Inteligencia artificial*. Alienta Editorial.

Rubinfeld, D., Gal, M. (2017). Access barriers to big data. *Arizona Law Review. 59*(2), 339-382. http://doi.org/10.2139/ssrn.2830586

Salgado, D. (2017). Big Data en la Estadística Pública: retos ante los primeros pasos. *Economía industrial*, (405), 121-129.

Salvador, M. y Ramió, C. (2020). Capacidades analíticas y gobernanza de datos en la Administración pública como paso previo a la introducción de la Inteligencia Artificial. *Reforma Democr. Rev. CLAD*, *77*, 5-36.

Sancho, J. V., Ochoa, B. M., y Domínguez, J. C. (2014). Aproximación a una taxonomía de la visualización de datos. *Revista Latina de Comunicación Social*, (69), 486-507.

Serna, M. S. (2021). Inteligencia artificial y gobernanza de datos en las administraciones públicas: reflexiones y evidencias para su desarrollo. *Gestión y Análisis de Políticas Públicas*, (26), 20-32.

Sosa Escudero, W. (2021). *Big data y ciencia de datos: conceptos, oportunidades y desafíos* [Nota informativa]. Organización Internacional del Trabajo.

Stroe, M. J. (2020). Internet de *las Cosas: propiedad de los datos y su tratamiento como secretos empresariales*. Universidad de Zaragoza, Departamento de Derecho de la Empresa, España.

Stucke, M. y Grunes, A. (2016). *Big data and competition policy*. Oxford: Oxford University Press.

Teerlink, M., Sigmon, P. W., Gow, B. y Banerjee, K. (2014). *El nuevo héroe del Big Data y la analítica de datos. El director de datos* [Informe ejecutivo]. IBM Institute for Business Value. IBM Corporation, Estados Unidos de América.

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado Pérez, J. C. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*. Ediciones Universidad Cooperativa de Colombia

Tourís López, R. M. (2019). Secreto estadístico y protección de datos de carácter personal como límites al acceso a la información pública. La estadística de criminalidad. *Derecom*, (26), 75-97.

United Nations Global Working Group on Big Data (2016). *Recommendations for access to data from private organizations for Official Statistics.* United Nations.

Vásquez Valdivia, A. (2021). *Apertura y uso de datos para hacer frente al Covid-19 en América Latina*. Gestión Pública, n. 88. Santiago, Comisión Económica para América Latina y el Caribe.

Villao Balón, A. J. (2021). *Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela Lic. Angélica Villón*. Universidad Estatal Península de Santa Elena.

Wiseman, J. M. (2018). *Data-Driven Government: The Role of Chief Data Officers.* IBM Center for the Business of Government, Washington, D.C.

Zarzky T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology & Human Values, 41*(1), 118-132.

Zamorano Ruiz, J. (2018). *Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea*. Universidad Complutense de Madrid.

Zúñiga, G. (2019). Big data y los desafíos que plantea al abuso de posición de dominio. *Revista de Actualidad Mercantil*, (6), 208-226.