Bussi, Javier
Marí, Gonzalo
Méndez, Fernanda
Instituto de investigaciones Teóricas y Aplicadas, Escuela de Estadística

EL DESAFÍO DEL BIG DATA EN ESTADÍSTICAS OFICIALES EN ARGENTINA[1]

Resumen:

La generación de datos de todo tipo se ha incrementado en los últimos años, ya sea en cuanto a cantidad, frecuencia y disponibilidad, lo que ha dado lugar a la incorporación del término Big Data. La abundancia de estas fuentes de datos representa un desafío y una oportunidad extremadamente interesante en el terreno de las Estadísticas Oficiales. Muchas agencias nacionales de estadística de distintos países han implementado divisiones dedicadas específicamente al tema y la Organización de las Naciones Unidas a través de su Comisión de Estadística creó el Grupo de Trabajo Mundial en el año 2014. Las recomendaciones sobre el uso de Big Data en esta área se centran en varios aspectos y se han vuelto más específicas con el avance de los años, pero pueden resumirse en los siguientes puntos: a) el acceso a datos que pertenecen a otros organismos b) establecimiento de una asociación exitosa y colaborativa con los proveedores de datos, c) desarrollo de actividades prácticas a través de proyectos piloto y d) construcción de metodología apropiada para la utilización de Big Data en el proceso de generación de Estadísticas Oficiales. El Instituto Nacional de Estadística y Censos (INDEC) ha dado importantes primeros pasos en el área siguiendo estas recomendaciones a través de la firma del acuerdo sobre cooperación en temáticas de innovación estadística con el Central Bureau of Statistics (CBS) de los Países Bajos en el año 2017. Aun así, el desafío del uso del Big Data en Estadísticas Oficiales en Argentina sigue siendo inmenso. Demanda intenso trabajo metodológico y técnico, y debe atender temas tales como la capacitación de personal en las metodologías necesarias y la creación de puestos específicos para la incorporación de fuentes de Big Data en la producción de Estadísticas Oficiales.

Palabras claves: Big Data, Estadísticas Oficiales, INDEC

Abstract:

The generation of all kinds of data has risen in recent years, in reference to quantity, frequency and availability, which has led to the incorporation of the term Big Data. This abundance of data represents a challenge but also an extremely interesting opportunity in the field of Official Statistics. Several national statistics agencies have implemented divisions specifically dedicated to

[1] Este trabajo se elaboró en el marco del Proyecto 1ECO199 Titulado “Métodos Estadísticos en el Ámbito Oficial”, dirigido por Gonzalo Marí.

this area and the United Nations through its Statistics Commission created the Global Working Group in the year 2014. The recommendations for the use of Big Data in Official Statistics are focused on several aspects which have become more specific through the years but can be summarized in the following points: a) the access to data that belong to other organizations, b) the implementation of a successful and collaborative partnership with data providers, c) the development of practical activities through pilot projects and d) the development of appropriate methodology for the use of Big Data in the generation process of Official Statistics. The Argentine National Institute of Statistics and Censuses (INDEC) has taken important first steps in the area through a collaborative agreement in the area of innovation in statistics with the Central Bureau of Statistics (CBS) from the Netherlands in 2017. Nonetheless, the challenge of using Big Data in Official Statistics remains enormous. It demands intense methodological and technical work and it should address topics such as building capacities in those methodologies and creating specific positions for the incorporation of Big Data sources in the generation of Official Statistics.

Keywords: Big Data, Official Statistics, INDEC

1. Introducción

La generación de datos de todo tipo se ha incrementado en los últimos años, ya sea en cuanto a cantidad, frecuencia y disponibilidad, lo que ha dado lugar a la incorporación del término Big Data. Este término se ha acuñado en español como Macrodatos en algunas traducciones (Naciones Unidas, 2015a). La abundancia de estas fuentes de datos representa un desafío y una oportunidad extremadamente interesante en el terreno de las Estadísticas Oficiales. Muchas agencias nacionales de estadística de distintos países han implementado divisiones dedicadas específicamente al tema y la Organización de las Naciones Unidas a través de su Comisión de Estadística creó el Grupo de Trabajo Mundial en el año 2014. El Instituto Nacional de Estadística y Censos (INDEC) en el año 2017, en el marco de la firma del acuerdo para la cooperación en temáticas de innovación estadística con el Central Bureau of Statistics (CBS) de los Países Bajos, celebró en Buenos Aires una conferencia sobre la implementación de Big Data en las estadísticas públicas, con exposiciones de representantes de ambos países y miembros del sector académico y privado.

En este artículo se intenta brindar una visión general del estado actual de la investigación sobre el uso de Big Data para estadísticas oficiales. En la siguiente sección se presentan algunas recomendaciones o lineamientos sobre el uso de Big Data en el área de Estadísticas Oficiales. En la sección 3 se describen los avances producidos en Argentina al respecto. En las secciones 4 y 5 se presentan la discusión y las conclusiones del artículo.

2. Recomendaciones acerca del uso de Big Data en el ámbito de las Estadísticas Oficiales

El Grupo de Trabajo Mundial de Naciones Unidas, a partir de su creación en el año 2014, propuso una estrategia para un programa mundial de utilización de Big Data en Estadísticas Oficiales (Naciones Unidas, 2015a). La misma propone la utilización práctica de fuentes de datos, estimulando paralelamente la capacitación en el área y el intercambio de experiencias. Incluye además la promoción de la confianza pública en la utilización de datos del sector privado. Una de las prioridades establecidas fue la utilización de estos datos en los indicadores de la Agenda 2030 para el Desarrollo Sostenible de las Naciones Unidas.

Se establecieron distintos equipos de trabajo asignados a tareas específicas que incluían: telefonía celular, imágenes satelitales y datos de las redes sociales con el fin de desarrollar actividades prácticas a través de proyectos piloto. Se realizaron tareas destinadas a estrechar lazos entre el sector privado y otras comunidades trabajando en el desarrollo de acuerdos provisorios generales para el acceso a los datos con proveedores de esta información. Al mismo tiempo se promocionaron los beneficios de la utilización de Big Data y se impulsó la participación de países en vías de desarrollo en los proyectos piloto. También se establecieron lineamientos referidos a la capacitación, ya que la misma era necesaria en las oficinas de Estadísticas Oficiales, en conjunto con lineamientos referidos a cuestiones metodológicas y de calidad que involucraban a varios sectores. Los miembros de este Grupo de Trabajo participaron en varios proyectos piloto y continuaron elaborando un inventario de tales proyectos.

Entre los progresos obtenidos por el Grupo, se generaron varios productos de promoción para la utilización de estos datos y se establecieron un conjunto de principios referidos al acceso a las fuentes de Big Data. Se establecieron clasificaciones preliminares de estas fuentes y también la definición de un marco de calidad para este tipo de datos. Se realizaron proyectos que

involucraron datos de telefonía celular y datos de redes sociales. Se apoyaron también iniciativas destinadas a la utilización de imágenes satelitales y datos espaciales para desarrollar métodos que puedan dar respuesta a los objetivos propuestos.

En 2015 se realizó una encuesta a nivel mundial sobre Big Data en las Estadísticas Oficiales con el objetivo de conocer los avances hasta el momento de las distintas oficinas de estadística con respecto a la utilización de estos datos. Las oficinas nacionales de estadística fueron interrogadas sobre su estrategia y experiencia práctica en esta área. El cuestionario contenía preguntas sobre la gestión, promoción y comunicación del uso de Big Data, acceso, privacidad y necesidades técnicas y de capacitación respecto a los datos como así también necesidades primordiales de las oficinas de estadística con respecto a su utilización. Se incluyeron además preguntas específicas sobre los distintos proyectos de Big Data a aquellas oficinas que los hubieran implementado.

Un total de 93 países respondieron la encuesta. Alrededor de la mitad de los países señalaron llevar a cabo proyectos de Big Data. Los principales beneficios de la incorporación de Big Data según las oficinas consultadas resultaron ser “Estadísticas más rápidas y oportunas”, “Reducción de la carga para el encuestado” y “Modernización del proceso de producción de estadísticas”, seguidos de “Nuevos productos y servicios” y “Reducción de los costos”. En cuanto a los 115 proyectos mencionados, 42 resultaron ser sobre datos de telefonía celular y 31 de datos extraídos de la web. La mayoría de las oficinas establecieron relaciones con institutos públicos y organismos académicos y de investigación, siendo menos las vinculaciones con empresas proveedoras de datos. En aquellos proyectos que pasaban a la parte de producción, el Big Data complementaba una fuente de datos existente. Estos proyectos fueron analizados a través de métodos estadísticos tradicionales lo que refleja que la exigencia para el análisis de Big Data no era tan alta o que las oficinas aún no contaban con la capacitación suficiente para el tratamiento de esta información. Las competencias técnicas por adquirir más mencionadas por las oficinas de Estadísticas Oficiales resultaron ser “experto en metodología sobre cuestiones relacionadas con los Macrodatos”, “especialista en ciencia de los datos” y “especialista en modelos matemáticos”, mientras que las capacidades relacionadas con tecnología de la información fueron menos mencionadas. Es claro que se necesitan nuevas técnicas avanzadas para el análisis pero aún no se contrata ni se capacita personal en el área. Las conclusiones respecto de esta encuesta establecen la necesidad de fomentar y brindar mayor capacitación en el área y promover la generación de proyectos piloto, donde participen países en desarrollo. En consecuencia, el Grupo de Trabajo Mundial estableció como prioridades principales la capacitación, la metodología y los marcos de calidad de Big Data.

La primera Conferencia Internacional sobre Big Data en las Estadísticas Oficiales realizada en 2014 en Beijing, China (Big Data UN Working Group, 2014), tuvo como objetivo promover el uso práctico de estas fuentes, tratando de encontrar soluciones a sus desafíos, de estimular la construcción de capacidades y compartir experiencias al respecto. La segunda Conferencia Internacional realizada en 2015 en Abu Dabi, Emiratos Árabes Unidos (Big Data UN Working Group, 2015), mostró los avances en los distintos proyectos. Estos incluían datos de telefonía móvil, de las redes sociales y satelitales. Se presentaron los progresos respecto a la divulgación de ciertas necesidades: capacitación, acceso a los datos, establecimiento de vínculos con otras entidades, calidad y metodología adecuada. También se lograron avances sobre la forma de comunicar la importancia del uso del Big Data con mayor eficacia.

En la tercera Conferencia Global, realizada en Dublín, Irlanda en 2016 (Big Data UN Working Group, 2016), se avanzó un paso más focalizando específicamente en tres aspectos: a) el ac-

ceso a datos que pertenecen a otros organismos y el establecimiento de una asociación exitosa con los proveedores de datos, b) estrategias para la construcción de capacidades para la utilización de Big Data en el proceso de generación de estadísticas y c) el uso de estos datos en la compilación de indicadores para el Desarrollo Sostenible de la Agenda 2030 (Naciones Unidas, 2015b). El 25 de septiembre de 2015, esta agenda fue adoptada unánimemente por todos los países miembros de las Naciones Unidas, constituye un plan de acción universal para las personas, el planeta y la prosperidad y contiene 17 Objetivos de Desarrollo Sostenible (ODS) y 169 metas relacionadas a estos para ser alcanzados en el año 2030. La misma está referida a las necesidades de todas las personas a través del planeta y demanda además su contribución para alcanzar estos objetivos. Para asegurar el cumplimiento de esta agenda, es necesaria la construcción de un conjunto sólido de indicadores de ODS, lo que demanda intenso trabajo metodológico y técnico que representa un desafío inclusive para el más avanzado Sistema Estadístico Nacional. El Grupo Interagencial de Expertos (IAEG) de las Naciones Unidas está integrado por especialistas en cada uno de los temas que incumben a los ODS y sus metas, y son los encargados del desarrollo de la metodología de cada uno de los 232 indicadores que servirán para evaluar y monitorear el cumplimientos de las metas hasta el 2030. Luego el IAEG es el encargado de la publicación de los metadatos de cada uno de los indicadores, de la compilación de las mediciones realizadas por los países, de la comparación entre los mismos, y de la agregación a nivel regional y global. De esta forma, se establece que son los países los que serán los encargados de la medición de los indicadores. Esto requiere que las oficinas de Estadísticas Oficiales deban modernizarse incrementando la capacidad de estas entidades para dar respuesta de una manera más flexible, efectiva y eficiente a los nuevos requerimientos y desafíos, incluyendo el monitoreo de estos indicadores de los Objetivos de Desarrollo Sostenible. Esto incluye la incorporación de nuevas fuentes de datos, en particular el uso del Big Data, que ha sido poco utilizado en la producción de Estadísticas Oficiales.

En la cuarta Conferencia Global, realizada en Bogotá, Colombia en noviembre de 2017 (Big Data UN Working Group, 2017), se estableció que nuevos avances eran necesarios para que el trabajo realizado hasta la fecha pueda ser utilizado por la comunidad estadística, y esto implica que las fuentes de datos, ya sea Big Data, registros administrativos o fuentes tradicionales de datos de Estadísticas Oficiales sean tratados de manera conjunta desde un enfoque multisectorial. El desafío en sí mismo no es utilizar las fuentes de Big Data, si no lograr una colaboración en el uso de los datos que sea confiable, y esto implica una estrecha colaboración con el sector privado, la comunidad académica y la sociedad civil. Los temas de las conferencias abordaron cuestiones tales como la forma en que pueden colaborar las oficinas de estadística, las empresas de tecnología y las compañías que poseen datos de manera de que todas se beneficien y se logren resultados en un mundo cambiante donde los datos son la fuente más importante para le creación de bienestar y desarrollo. También se enfocaron en la experiencia obtenida en los proyectos de colaboración, en relación a la cobertura, la inclusión (y exclusión) de participantes, actividades, gerenciamiento y financiamiento. Otros temas de interés se refirieron a la forma en que se puede compartir información sensible en una nube confederada de información dado los marcos regulatorios de privacidad y confidencialidad y también a cómo utilizar nuevas herramientas y servicios y al mismo tiempo adaptar los perfiles de los nuevos puestos de trabajo necesarios para estas tareas en las oficinas de Estadísticas Oficiales.

3. Avances en la utilización de Big Data en el ámbito de las Estadísticas Oficiales en Argentina

En el año 2017, el Instituto Nacional de Estadística y Censos (INDEC) firmó un acuerdo con la oficina gubernamental de Estadísticas Oficiales de los Países Bajos, el Central Bureau of Statistics (CBS) para la cooperación entre ambas agencias en temáticas de innovación estadística. Este convenio representa un importante primer paso hacia el cumplimiento de las recomendaciones acerca del uso de Big Data en las Estadísticas Oficiales. El CBS ha participado en proyectos piloto que involucran la utilización de Big Data y la colaboración con otros organismos, principalmente organismos generadores de la información. La oficina de estadística holandesa ha participado en dos de estos proyectos, cada uno de los cuales es un estudio de caso (Daas et al, 2015).

El primer estudio correspondía a la utilización de información que medía la intensidad del tránsito; cada medición registraba la cantidad de vehículos por minuto que circulaba por un lugar determinado, la velocidad y el largo de los mismos. Este trabajo fue el resultado de una participación colaborativa entre el CBS y distintas oficinas gubernamentales en donde los datos fueron almacenados en el Centro Nacional de Datos para Información de Tránsito (NDW). Los resultados incluyeron la determinación de patrones de comportamiento del tránsito para distintos tamaños de vehículos y representó un desafío para el futuro en cuanto a la obtención de estimaciones para todas las carreteras (no solo las que cuentan con puntos de medición) y la obtención de estimaciones de la variabilidad que reflejen la precisión de los estimadores.

El otro estudio de caso consistió en involucrar datos provenientes de las redes sociales. Esta información se consideró útil en el terreno de las Estadísticas Oficiales ya que la misma refleja diferentes aspectos de la vida diaria de las personas. Se analizaron dos cualidades en los mensajes: contenido y sentimiento, a partir de distintas redes sociales analizadas tales como Twitter, Facebook, Google+, Hyves y Linkedin como así también varios blogs y foros de discusión en idioma holandés. Se estudió la relación entre sentimiento y el índice de confianza del consumidor detectándose una correlación positiva entre ambos.

La experiencia en estos dos proyectos puede ser compartida de manera colaborativa gracias al acuerdo firmado con el INDEC, lo que puede generar proyectos pilotos locales que utilicen Big Data, cumpliendo con recomendaciones que sugieren involucrar en los mismos a países en desarrollo.

En el marco de la firma de este acuerdo con el CBS, se celebró en Buenos Aires una conferencia sobre la implementación de Big Data en las Estadísticas Oficiales, con exposiciones de representantes de ambos países y miembros del sector público y privado (INDEC, 2017). En la misma expusieron autoridades de ambas agencias de estadísticas oficiales y hubo una presentación sobre la experiencia holandesa. A continuación, lo hicieron representantes del sector privado (Telefónica Argentina y SAS Argentina), con experiencia en el uso de Big Data en el país. Posteriormente, representantes del ámbito académico (UdeSA/CONICET; Fundación Sadosky y UNGS), evaluaron las ventajas y desventajas de la aplicación de estas tecnologías en la construcción de estadísticas públicas. Esta conferencia constituye un primer paso para lograr una utilización de datos que sea confiable y que estimule una estrecha colaboración con el sector privado, la comunidad académica y la sociedad civil, otra de las recomendaciones establecidas para la utilización de Big Data en Estadísticas Oficiales.

4. Discusión

Uno de los principales conceptos referidos al uso de Big Data en Estadísticas Oficiales es complementar y mejorar las estimaciones de los métodos actualmente utilizados, suministrando estadísticas más rápidas y oportunas, y en muchos casos, reduciendo la carga sobre el encuestado. Pero existen dificultades que deben ser enfrentadas y para las cuales debe encontrarse una solución en los años venideros. Las fuentes de Big Data están disponibles para ser utilizadas, pero no están diseñadas por las oficinas de Estadísticas Oficiales y por lo tanto su estructura debe ser revisada y comprendida previamente a su uso para el análisis estadístico. Existen otros potenciales desafíos, como la presencia de datos perdidos, que no es ajena a las fuentes de Big Data, por ejemplo, aquellos producidos por la falla de un sensor, cortes de energía o la caída de servidores. El tiempo de procesamiento puede extenderse demasiado ante la necesidad de proveer estadísticas de publicación frecuente. En muchas ocasiones pueden surgir problemas de validación de las estimaciones, como por ejemplo que la población bajo estudio no represente a la población objetivo. En muchos casos este problema puede ser evaluado a través de comparaciones de características entre la población cubierta y la población objetivo, lo que resulta difícil ya que raramente esas características estén disponibles en los Macrodatos. Nuevas capacidades son requeridas y la descripción de los puestos necesarios deben redefinirse, ciertamente es necesario un alto nivel de especialización en el tratamiento de las fuentes de datos para que sean utilizables para el análisis estadístico. El término "científico de datos" se utiliza para describir a una persona con las habilidades mencionadas combinadas con conocimientos de metodología estadística (Schutt y O'Neil, 2013). Las diferencias estrictas entre metodología, ingeniería de software y especialización en hardware de tecnología de información se vuelven menos claras (Daas et al, 2015). Otro tema para tener en cuenta es el referido a la confidencialidad de los datos, que en muchos casos es difícil de definir y establecer límites. En muchas ocasiones existe un vacío legal y es recomendable en las mismas guiarse por la percepción pública. Muchas fuentes de datos pertenecen a organismos públicos, aun así, pueden surgir temas como la seguridad que impidan su utilización. En otros casos, pueden surgir cuestiones referidas a la propiedad de los datos y propósito de las publicaciones. Una cuestión adicional es que habitualmente las fuentes de Big Data tienden a modificarse rápidamente y en el caso de Estadísticas Oficiales es necesario contar con información que permita evaluar y comparar situaciones por un periodo prolongado de años, hecho que puede verse afectado por estas constantes modificaciones.

5. Conclusiones

La incorporación de datos provenientes de Big Data representa una gran oportunidad en el terreno de las Estadísticas Oficiales, pero también un gran desafío. Las recomendaciones sobre el uso de Big Data en esta área se centran en varios aspectos y se han vuelto más específicas con el avance de los años, pero pueden resumirse en los siguientes puntos: a) el acceso a datos que pertenecen a otros organismos b) establecimiento de una asociación exitosa y colaborativa con los proveedores de datos, c) desarrollo de actividades prácticas a través de proyectos piloto y d) construcción de metodología apropiada para la utilización de Big Data en el proceso de generación de Estadísticas Oficiales. El INDEC ha dado importantes primeros pasos en el área siguiendo estas recomendaciones a través de la firma del acuerdo sobre cooperación en temáticas de innovación estadística con el CBS en el año 2017, lo que puede fomentar la generación de proyectos piloto en el área. La conferencia realizada en el marco de la firma

del convenio sobre la implementación de Big Data en las Estadísticas Oficiales, con exposiciones de especialistas del sector público y privado, donde participaron organismos generadores de Big Data contribuye ciertamente al establecimiento de una asociación colaborativa con los proveedores de datos y la posibilidad de acceder a los mismos a través de acuerdos entre los distintos organismos. Aun así, el desafío del uso del Big Data en Estadísticas Oficiales en Argentina sigue siendo inmenso. Demanda intenso trabajo metodológico y técnico, y debe atender temas tales como la capacitación de personal en las metodologías necesarias, la creación de puestos específicos para la incorporación de fuentes de Big Data en la producción de Estadísticas Oficiales desde un enfoque multisectorial donde complemente el uso de registros administrativos y fuentes tradicionales de datos.

REFERENCIAS BIBLIOGRÁFICAS

Big Data UN Working Group, 2014. International Conference on Big Data for Official Statistics. Recuperado de: https://unstats.un.org/unsd/trade/events/2014/Beijing/default.asp

Big Data UN Working Group, 2015. International Conference on Big Data for Official Statistics. Recuperado de: https://unstats.un.org/unsd/trade/events/2015/abudhabi/default.asp

Big Data UN Working Group, 2016. International Conference on Big Data for Official Statistics. Recuperado de: https://unstats.un.org/unsd/bigdata/conferences/2016/default.asp

Big Data UN Working Group, 2017. International Conference on Big Data for Official Statistics. Recuperado de: https://unstats.un.org/unsd/bigdata/conferences/2017/default.asp

Daas, P. J.H.; Puts, M.J.; Buelens, B. and van den Hurk, P.A.M., 2015. *Big Data as a Source for Official Statistics*, Journal of Official Statistics, Vol. 31, No. 2, pp. 249–262

INDEC, 2017. *Estadísticas oficiales, buenas prácticas y big data*. Recuperado de: https://www.indec.gob.ar/gacetillasdeprensa_detalle.asp?id=143

Naciones Unidas, 2015a. *Informe del Grupo de Trabajo Mundial sobre los Macrodatos en las Estadísticas Oficiales*. E/CN.3/2016/1.

Naciones Unidas, 2015b. *Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible*. A/RES/70/1.

Schutt, R. and O'Neil, C., 2013. *Doing Data Science: Straight Talk from the Frontline*. Sebastopol, CA: O'Reilly Media.