El reto Big Data para la estadística pública

ESTADO DEL ARTE

Jesús Alberto González Yanes

TFM Ingeniería de Sistemas de Decisión

Universidad Rey Juan Carlos

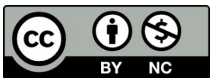
Madrid, 2017-2018

Trabajo Fin de Máster de Ingeniería de Sistemas de Decisión: El reto Big Data para la estadística pública. Estado del arte.

Alumno: Jesús Alberto González Yanes.

Director: Este trabajo ha sido dirigido por José Felipe Ortega Soto, Profesor Ayudante Doctor en el Departamento de Teoría de la Señal y Comunicaciones y Sistemas Telemáticos y Computación de la URJC

Agradecimientos: A Mara por no permitir mi rendición. A los compañeros del Instituto Canario de Estadística (ISTAC) y de otras Oficinas Estadísticas de las Comunidades Autónomas, así como a los de la Unidad Mixta de Investigación en Estadística Pública ISTAC-ULL, por soportarme críticamente y por aprender colectivamente.



INDICE

Capítulo 1. Introducción

1.1. La sociedad datificada

La sociedad de finales de Siglo XX y principios del Siglo XXI está cambiando rápidamente en muchos aspectos, entre ellos los vinculados al mundo de la información. Vivimos en la época ***SMAC*** (Social, Mobile, Analytics, Cloud) donde las personas, muchas de ellas denominadas nativas digitales, no conciben su vida sin un dispositivo móvil a través del que se relacionan con el mundo. Este estilo de vida, al que ya se llama digital, genera un tsunami de cambios y una verdadera montaña de datos en flujo constante.

A esto se suma lo que Kevin Ashton denominó ***Internet de las Cosas*** (IoT, por sus siglas en inglés), concepto que se refiere a la interconexión digital de objetos cotidianos con internet. La idea subyacente es que los objetos se equipan con sensores, que generan datos que se comunican por Internet. En esta línea nos encontramos con las ***Ciudades Inteligentes (Smart Cities)***, en las que los sistemas de iluminación, la señalización viaria y otros servicios públicos sensorizados son importantes generadores de datos públicos.

El nacimiento de estos nuevos fenómenos es producto del advenimiento de las computadoras, que trajo consigo equipos de medida y almacenaje que hicieron sumamente más eficiente el proceso de datificación. La incorporación de ordenadores a las empresas y a las administraciones públicas extendió el almacenamiento y tratamiento de datos durante los años ochenta y noventa del siglo pasado, dando lugar a la inteligencia de negocios aplicada tanto a la empresa como al sector público ***(Business Intelligence)***, y dando lugar también a implicaciones en la estadística pública con el surgimiento de la estadística basada en registros administrativos[1]. Este avance también se facilitó gracias al tratamiento y análisis matemático de datos, permitiendo descubrir su valor oculto y dando lugar a términos comerciales como ***Minería de Datos (Data Mining)*** que describe el uso de la estadística y de métodos matemáticos en el análisis de los datos empresariales.

Por lo tanto las empresas vienen desarrollando desde hace años sistemas de extracción, tratamiento y análisis de datos de sus sistemas de gestión. Además con el tiempo se ha extendido el acceso y la disponibilidad de datos, convirtiéndose en la base de nuevos modelos de negocio más allá del negocio

[1] Wallgren, Anders, and Britt Wallgren. *Register-Based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology. Chichester, England ; Hoboken, NJ: John Wiley & Sons Ltd, 2007.

tradicional -como por ejemplo el proyecto Smart Step de Telefónica[2]-, de negocios nuevos basados en datos a cambio de servicios -Google sería el ejemplo paradigmático-, o de negocios fundamentados en datos abiertos de origen público o privado -PriceStat[3] sería un buen ejemplo-.

Por otra parte, la puesta en valor de los datos para fines económicos o democráticos ha dado lugar a que la sociedad exija a sus Gobiernos a poner a disposición de ciudadanos y empresas los datos de los que dispone. Los gobiernos han reaccionado a las exigencias mediante la formulación de políticas sobre ***datos abiertos*** y de disponibilidad de información del sector público. En la Unión Europea se ha traducido en la promulgación de la *Directiva relativa a la reutilización de la información del sector público*[4], actualmente en revisión, en la que se especifica que los documentos elaborados por los organismos del sector público de los Estados miembros constituyen un conjunto amplio, diverso y valioso de recursos que pueden beneficiar a la economía del conocimiento:

> *"Las políticas de apertura de la información, que propician la disponibilidad y la reutilización generalizada de la información del sector público con fines privados o comerciales, con restricciones mínimas o nulas de carácter jurídico, técnico o económico, y que favorecen la circulación de la información no solo para los agentes económicos, sino también para el público, pueden desempeñar una función importante a la hora de impulsar el desarrollo de nuevos servicios basados en formas novedosas de combinar y utilizar esa información, estimular el crecimiento económico y promover el compromiso social." (Directiva 2013/37/EU del Parlamento Europeo y del Consejo)*

La nueva era de la datificación masiva, junto con el aumento de la potencia de procesamiento y capacidad de almacenamiento de datos, ha sido descrita por muchos autores, entre ellos por Mayer-Schönberger and Cukier (2013)[5]:

> *"Nos hallamos inmersos en un gran proyecto de infraestructura que, de alguna manera, rivaliza con los del pasado: de los acueductos romanos a la Encyclopédie de la Ilustración. No llegamos a advertirlo con claridad porque el proyecto de hoy es demasiado nuevo, porque estamos en mitad de él, y porque, a diferencia del agua que fluye por los acueductos, el producto de nuestras labores es intangible. El proyecto es la datificación. Como aquellos otros avances infraestructurales, traerá consigo cambios fundamentales en la sociedad." (Mayer-Schönberger and Cukier, 2013)*

[2] http://dynamicinsights.telefonica.com/blog/488/smart-steps-2
[3] http://www.pricestats.com/
[4] Unión Europea. *Directiva 2013/37/EU del Parlamento Europeo y del Consejo de 26 de Junio, por la que se modifica la Directiva 2003/98/EC relativa a la reutilización de la información del sector público*, n.d.
[5] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt, 2013.

En este contexto tecnológico surge el término ***Big Data***, entendido como el análisis y la gestión de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas de software habitualmente utilizadas para la captura, gestión y procesamiento de datos. Y a los datos procedentes de estos nuevos procesos de datificación se les conoce como datos de ***fuentes Big Data*** a los que se les asocia nuevas características: volumen, velocidad y variedad.

De acuerdo con la Wikipedia, Big Data es un concepto que hace referencia a la acumulación masiva de datos y a los procedimientos usados para identificar patrones recurrentes dentro de esos datos. También según la Wikipedia, la disciplina dedicada a los datos masivos se enmarca en el sector de las tecnologías de la información y la comunicación. Esta disciplina se ocupa de todas las actividades relacionadas con los sistemas que manipulan grandes conjuntos de datos, tales como la captura, almacenamiento, búsqueda, compartición, análisis, y visualización.

Este tipo de definiciones, con una perspectiva tecnológica, describen al Big Data como la gestión y el análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas de software habitualmente utilizadas para la captura, gestión y procesamiento de datos.

En definitiva, como hemos visto, el advenimiento de las computadoras trajo consigo equipos de medida y almacenaje que hicieron sumamente más eficiente el proceso de datificación. En ese sentido, una confluencia de tendencias tecnológicas, sociales y económicas están dando como resultado la generación de enormes flujos de datos.

Justificación de trabajo: Este cambio de contexto en el mundo de la información tiene implicaciones directas en la Oficinas de Estadística. Entre ellas encontramos muchas cuestiones prácticas vinculadas con aspectos tecnológicos, metodológicos, jurídicos o sociales, pero también existe una importante y estratégica respecto a la posición que las Oficinas de Estadística quieren ocupar en el futuro sociedad de la información. En los últimos años se han abordado estos problemas tanto por la comunidad internacional de la Estadística Pública como por parte de la Academia.

Este trabajo tiene como objetivo revisar y estructurar los riesgos, problemas y soluciones identificados para finalmente realizar una propuesta de acción para que las Oficinas Estadísticas, como principales

productoras de los datos que sostienen la toma de muchas de las decisiones públicas y privadas, integren las fuentes Big Data como parte de su parrilla de fuentes de datos.

1.2. El viejo-nuevo problema Big Data en la estadística pública

El problema de gestión de grandes volúmenes de datos es un problema al que ha tenido que enfrentarse la estadística pública desde hace muchos años. Por ejemplo, a medida que la población de Estados Unidos crecía, la US Census Bureau fue buscando estrategias para mejorar la velocidad y la precisión del proceso de levantamiento de censos. En 1790 para el primer censo, cuando los Estados Unidos contaban con 3,9 millones de residentes, el gran volumen de tabular a mano los resultados era uno de los mayores retos. A medida que el país creció, también lo hizo el desafío. Cuando contar los resultados del censo se hizo tan largo que casi duraba una década, la búsqueda de soluciones condujo necesariamente a la creación de la moderna tecnología de procesamiento de datos. El primer dispositivo para acelerar el conteo del censo fue creado en 1872 por el Oficial Mayor del Censo Charles W. Seaton. La máquina utilizaba rodillos para sumar las pulsaciones de teclado introducidas manualmente. Sin embargo, incluso con la máquina Seaton, el procesamiento del censo se alargó durante casi toda la década. Unos años después, en 1880 comenzó a realizarse nuevamente el censo en EEUU y debido a la cantidad de personas censadas tardó 8 años en finalizarse, incluso hubo variables que no llegaron a tabularse. Por este motivo, la US Census Bureau llevó a cabo un concurso en 1888 para encontrar un método más eficiente para procesar y tabular el gran volumen de datos que recogía. El resultado fue la tabuladora de Herman Hollerith, génesis de la fichas perforadas y de la empresa IBM[6].

Tal como señala Caballero en el libro *Las bases de Big Data*[7] el almacenamiento y procesamiento de datos ha sido una de las tareas asociadas a los ordenadores desde su aparición. El primer ordenador comercial, UNIVAC I, construido en 1951, fue adquirido por la Oficina del Censo de Estados Unidos para tratar la ingente cantidad de información obtenida en los censos, a la que había que sumar los datos que comenzaban a recopilarse a través de muchas otras fuentes: hospitales, escuelas, etc. Pronto, UNIVAC reveló su potencia a la hora de realizar cálculos y predicciones estadísticas imposibles hasta el momento. Uno de sus mayores éxitos fue la predicción del resultado de las elecciones presidenciales en 1952. A partir del recuento de tan solo un 1% del total de votos, UNIVAC predijo que el siguiente presidente sería Eisenhower, mientras que la mayoría de los comentaristas políticos daban por ganador a su rival, el hoy olvidado Stevenson.

[6] http://www.datosconinteligencia.blogspot.com.es/2015/09/el-viejo-problema-del-big-data-en-la.html
[7] Caballero Roldán, Rafael, and Enrique Martín Martín. Las bases de Big Data. Madrid: Los Libros de la Catarata: Universidad Complutense de Madrid, 2015.

Entonces, como diría el Bugs Bunny traducido *"¿Qué hay de nuevo, viejo?"*. En 2001 Douglas Lane[8] propuso tres características que distinguían a lo que ahora denominamos Big Data: **volumen, velocidad y variedad**. Tradicionalmente, como hemos visto, las Oficinas de Estadísticas se han enfrentado a los problemas de volumen, pero en la actualidad aparecen dos elementos nuevos: la velocidad y la variedad. Siguiendo esta dirección, el primer documento (UNECE, 2013) que estudia el problema Big Data en la estadística pública *What Does 'Big Data' Mean for Official Statistics?*[9] lo define como una variante de la propuesta de Douglas Laney:

> *"Big Data son las **fuentes de datos** que generalmente pueden ser descritas como de alto volumen, velocidad y variedad, que requieren formas rentables e innovadoras de procesamiento con el fin de mejorar los análisis y de apoyar las tomas de decisiones"*

Por lo tanto, para la estadística pública el problema Big Data se aborda como un problema de nuevas fuentes de datos. En esa dirección el problema se enfrenta considerando que estas fuentes de datos podrían complementar o sustituir las fuentes tradicionales utilizadas en la estadística pública, las encuestas y los registros administrativos, pero con algunas características peculiares:

1. La propiedad sobre las fuentes de datos generalmente no es pública, con los problemas derivados para el acceso, uso y mantenimiento de las fuentes.

2. La fuentes de datos no están pensadas para fines estadísticos con los problemas derivados de conceptualización y sesgos.

En el documento anteriormente citado se enumeran algunos de los retos derivados de las características señaladas: (1) **Legislativo**, p.e. respecto al acceso y uso de los datos (2) **Privacidad**, p.e. gestión de la confianza pública para la aceptación del uso de esas fuentes y su enlace con otras fuentes de datos (3) **Financiero**, p.e. coste-beneficio potencial de acceso a las fuentes de datos (4) **Gestión**, p.e. políticas y directivas sobre la gestión y protección de los datos (5) **Metodológico**, p.e. calidad de los datos e idoneidad de los métodos estadísticos (6) **Tecnológico**, p.e. temas relacionados con la tecnología de la información.

Estos cambios tienen diversos impactos sociales. Respecto a la *privacidad*, la datificación de buena parte de nuestras vidas, y sus posibilidades comerciales, genera actitudes diversas en la opinión

[8] http://www.gartner.com/analyst/40872/Douglas-Laney
[9] Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?" UNECE, March 10, 2013.

pública sobre el derecho a la intimidad. A algunos ciudadanos les preocupa si sus datos se reutilizan sin su consentimiento, por razones comerciales o de otro tipo, para fines distintos para los que fueron recabados. A otros no les importa tanto, si esto significa que a cambio se proporcionan servicios de forma gratuita. Muchas personas comparten voluntariamente información en las redes sociales sin preocuparse por la privacidad. El nuevo Reglamento del Parlamento Europeo y del Consejo, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, viene a cubrir parte de las preocupaciones ciudadanas al respecto.

Además la estadística pública seguramente tendrá que enfrentarse, aunque por ahora no está en ninguna agenda, a dos ideas que encontramos en el libro *Big Data: A Revolution That Will Transform How We Live, Work, and Think*[10] relacionadas con cambios epistemológicos aportados por las fuentes Big Data:

1. Big Data es mejor que Small Data. Anunciando la muerte de las muestras y la probabilidad, obviando que las grandes fuentes de datos también son muestras, con el problema añadido de ser muestras no probabilísticas. Muestras posiblemente sesgadas y por tanto olvidando unos de los componentes del error cuadrático medio, poniendo su atención únicamente en la varianza que es el elemento que se reduce con el aumento de muestra.

2. Big Data acaba con el estudio de la causalidad y lo sustituye por la correlación. Obviando que la correlación estadística puede ser producto del azar.

Por lo tanto, son muchos los caminos a recorrer por parte de la estadística pública para incorporar las fuentes Big Data a su producción estadística. Este trabajo es un acercamiento a esos caminos y a los avances acumulados hasta el momento; y para ello vamos a partir desde una revisión de la misión, principios y valores de la estadística pública para con ello situar correctamente el posicionamiento ante los retos señalados.

[10] Mayer-Schönberger, Viktor, and Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston: Houghton Mifflin Harcourt, 2013.

1.3. La estadística pública en la sociedad datificada

1.3.1. La estadística pública en la encrucijada de la pérdida de hegemonía

El cambio de contexto en el mundo de la información, que hemos señalado en el apartado anterior, tiene implicaciones directas en la Oficinas de Estadística. Entre ellas encontramos muchas cuestiones prácticas, pero también existe una importante y estratégica: **¿Qué posición quieren ocupar las Oficinas de Estadística en el futuro sociedad de la información?**

Hasta alrededor de la década de los 80 los datos fueron esencialmente un bien escaso por el alto precio de su adquisición. Antes de la era de la datificación, mucha información no estaba disponible y debía ser recogida para un propósito particular. La información estadística oficial, basada fundamentalmente en datos de encuestas o censos, tenía un valor único pues simplemente no había otra alternativa. Por ejemplo, los datos de los censos de población, recogidos puerta a puerta, eran inmensamente valiosos para los responsables políticos, investigadores y otros usuarios.

A partir de la década de los 90 los datos recogidos por las Administraciones Públicas fueron cada vez más accesibles para fines estadísticos, como consecuencia de la informatización de sus procedimientos. En este escenario la recopilación de datos estadísticos por medio de cuestionarios se complementó, e incluso se sustituyó, por fuentes de datos administrativas[11], con el fin de reducir costes y reducir la carga sobre los encuestados. Hoy en día algunos países no llevan a cabo amplios estudios poblacionales, y realizan su censo mediante la combinación y el análisis de datos de varias fuentes administrativas. Aún en este contexto, la información proporcionada por las Oficinas de Estadística seguía siendo única. Esta posición se reforzaba ante la posibilidad de combinar los datos de diferentes fuentes, ya que en muchos países no hay otra organización autorizada para realizar esas combinaciones.

Sin embargo la datificación está cambiando el entorno de las Oficinas de Estadística, dando lugar a que la escasez de datos se convierta en un problema menor. Para las Oficinas de Estadística hay beneficios potenciales en estas nuevas fuentes de datos, de las que surgen nuevas posibilidades tanto en la reducción de cargas a los encuestados y costes de producción, como en la producción de nueva información. Pero también da lugar a la **pérdida de la hegemonía de sus datos**, ya que otros

[11] Wallgren, Anders, and Britt Wallgren. *Register-Based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology. Chichester, England ; Hoboken, NJ: John Wiley & Sons Ltd, 2007.

jugadores en el mercado de la información pueden empezar, y de hecho han comenzado a hacerlo, a producir estadísticas que hasta el momento solo ejecutaban las Oficinas de Estadística.

Por ejemplo el *Billion Prices Project*[12] del Massachusetts Institute of Technology (MIT) dirigido por Alberto Cavallo y Roberto Rigobon, que en la actualidad se ha convertido en una propuesta comercial a través de la empresa *PriceStats*[13], nació como una iniciativa académica que utilizaba los precios recogidos diariamente en cientos de tiendas en línea de todo el mundo para llevar a cabo investigación económica. Este proyecto se fundamentó en la tesis doctoral de Cavallo, A (2009)[14] en la Universidad de Harvard, y también dió lugar en 2007 a la aparición del proyecto *InflacionVerdadera.com* creado para proveer índices de precios alternativos a los oficiales en Argentina, publicados por el Instituto Nacional de Estadística y Censos (INDEC).

Desde 2007 hasta 2012 se publicó un índice de alimentos y bebidas y otro de la Canasta Básica Alimentaria, utilizando los precios diarios en dos grandes supermercados de Buenos Aires y utilizando las mismas metodologías del INDEC. Los resultados del trabajo, cuyo objetivo era demostrar la manipulación de las estadísticas oficiales en Argentina, fueron publicados en el artículo académico *Online and official price indexes: Measuring Argentina's inflation*[15]. En Agosto del 2012 reemplazaron los índices originales de *InflacionVerdadera.com* por un Indice de Precios al Consumidor producido por *PriceStats*, comparable al IPC general del INDEC. El índice es publicado semanalmente en la revista The Economist[16] como alternativa a las estadísticas oficiales del INDEC.

Este es un claro ejemplo de cómo las fuentes Big Data pueden ser un instrumento al servicio del control externo del cumplimiento de los principios y valores de las Oficinas Estadísticas reconocidos internacionalmente. En esta nueva situación surgen diversas **cuestiones fundamentales** para una Oficina de Estadística y **el futuro de la estadística pública**:

1. ¿Cómo garantizar que las Oficinas de Estadística aporten valor añadido único en el futuro? y en ese sentido ¿las Oficinas de Estadística deben seguir haciendo estadísticas para las que existe una alternativa de mercado?

[12] http://bpp.mit.edu/
[13] http://www.pricestats.com/
[14] Cavallo, A. *Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization [recurso Electrónico]*. Harvard University, 2009.
[15] Cavallo, A. "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics* 60, no. 2 (2013): 152–65. doi:http://dx.doi.org/10.1016/j.jmoneco.2012.10.002.
[16] http://www.economist.com/node/21548242

2. ¿Pueden las Oficinas de Estadística asumir nuevas funciones o capacidades, en base a su posición institucional y a los conocimientos que han acumulado? Por ejemplo ¿se puede garantizar el acceso a fuentes de datos de propiedad privada?

3. ¿Sería mejor cambiar el papel de las Oficinas de Estadística pasando de la producción de información estadística hacia la validación de la información producida por los demás?

Desde un punto de vista práctico también surgen preguntas importantes respecto al uso potencial de las fuentes Big Data:

1. ¿En qué medida son útiles las fuentes Big Data para la producción y la mejora de las estadísticas públicas actuales? y ¿qué nueva información puede producir una Oficina Estadística mediante el uso de estas nuevas fuentes de datos?

2. ¿Cuál debe ser el marco jurídico de acceso a las fuentes Big Data para fines estadísticos? y si se tuviera acceso ¿cuáles son los riesgos de usar datos sobre los que no se controla su generación por parte de las Administraciones Públicas? además de ¿cómo asegurar la seguridad y confidencialidad de dichos datos?

3. ¿Cuáles son los requerimientos metodológicos y tecnológicos para el uso de fuentes Big Data?

4. El uso de estas fuentes ¿significa cambios de procedimientos? ¿es necesario aumentar la velocidad de producción/difusión de datos para aprovechar una de las principales características de estas fuentes?

1.3.2. Misión, principios y valores de la estadística pública en el contexto Big Data

Naciones Unidas reconoce a las estadísticas oficiales como un elemento indispensable en el sistema de información de una sociedad democrática pues proporcionan a los gobiernos, a la economía y a la ciudadanía datos de la situación económica, demográfica, social y ambiental de un país o de una región. En ese sentido considera que la información estadística es esencial para el desarrollo, pero también para el conocimiento mutuo y el comercio entre los Estados y los pueblos del mundo. Con este fin, NNUU indica que las Oficinas de Estadística han de compilar y facilitar, de forma imparcial,

estadísticas oficiales de comprobada utilidad práctica para que los ciudadanos puedan ejercer su derecho a mantenerse informados.

Pero para que los ciudadanos confíen en las estadísticas oficiales, los organismos estadísticos deben contar con un conjunto de valores y principios fundamentales. De acuerdo con Naciones Unidas, los principios generales son la (1) independencia, (2) la pertinencia o relevancia, (3) la credibilidad, así como (4) el respeto a los derechos de los informantes. Estos principios han sido desarrollados en los principios fundamentales de las estadísticas oficiales[17].

En coherencia con las líneas trazadas por Naciones Unidas, en el ámbito europeo, el Reglamento (CE) nº 223/2009, del Parlamento Europeo y del Consejo de 11 de marzo de 2009 relativo a la estadística europea, señala los siguientes principios en su artículo 2: (1) Independencia profesional, (2) imparcialidad, (3) fiabilidad, (4) secreto estadístico, (5) rentabilidad. Estos principios estadísticos se desarrollaron posteriormente en el Código de Buenas Prácticas de la Estadística Europea, que tiene por finalidad garantizar la confianza de la población en las estadísticas europeas mediante la determinación de la forma en que deben desarrollarse, elaborarse y difundirse las estadísticas con arreglo a los principios estadísticos europeos y a las mejores prácticas internacionales.

Las normas citadas desempeñan un papel vital en la obtención de la confianza en las estadísticas oficiales. A su vez estas normas se refuerzan con los códigos éticos de los estadísticos, destacando la *Declaración sobre Ética Profesional* del Instituto Internacional de Estadística (ISI), que además se complementan con diferentes códigos éticos elaborados por los distintos sistemas estadísticos nacionales.

En ese sentido, la pregunta que nos debemos hacer en un principio desde la estadística oficial es cómo el nuevo contexto Big Data encaja dentro de la misión, principios y valores que guía nuestra actividad pública. Para ellos vamos a realizar una revisión sintética a partir de la agrupación de los principios en tres grandes bloques:

1. Big Data y los principios asociados a las fuentes de datos para fines estadísticos
2. Big Data y los principios asociados al derecho de acceso y la protección de la intimidad
3. Big Data y los principios de objetividad política y científico-técnica

[17] Documentos Oficiales del Consejo Económico y Social, 1994, suplemento Nº 9 (E/1994/29), cap. V. Para más información véase el apéndice II o consúltese el sitio web <http://unstats.un.org/unsd/statcom/doc94/s1994.htm>

A) Big Data y los principios asociados a las fuentes de datos para fines estadísticos

En la Resolución sobre los Principios Fundamentales de las Estadísticas Oficiales aprobada por la Asamblea General de NNUU el 29 de enero de 2014, indica que los datos para fines estadísticos pueden obtenerse de todo tipo de fuentes, ya sea encuestas estadísticas o registros administrativos. Sorprende que no haya mención explícita a las fuentes Big Data, siendo una resolución del año 2014, pero de la esencia del principio podríamos extraer que la intención es establecer que la estadística pública pueda realizarse no sólo a partir de encuestas, sino con cualquier tipo de fuente de datos útil para sus fines.

Esta propuesta de pluralismo de fuentes se ordena en el principio mencionado, indicando que éstas se deben seleccionar considerando: su calidad, oportunidad, costo y carga que impondrá a los encuestados. Los criterios de oportunidad, costo y carga a los encuestados son también considerados en el Código de Buenas Prácticas de las Estadísticas Europeas y son fácilmente comprensibles; sin embargo el criterio de calidad necesita ser explicitado cuando se trabaja con datos no recopilados con fines estadísticos como pueden ser los datos administrativos o las fuentes Big Data. En ese sentido debemos referenciar una propuesta sobre marco de calidad para el uso de fuentes Big Data en la estadística pública, elaborada por UNECE Big Data Quality[18] e inspirada en el documento *Checklist for the Quality Evaluation of Administrative Data Sources*[19]. Este marco se estructura en tres hiperdimensiones, cada una con sus dimensiones de calidad, que a su vez se organizan en diversos factores a analizar.

Tabla 1. Dimensiones del marco de calidad para el uso de fuentes Big Data

Hiperdimensión	**Dimensiones de calidad**	**Factores a considerar**
Fuente	Entorno institucional	Sostenibilidad de la entidad proveedora de datos Confiabilidad general de los datos Transparencia e interpretabilidad de la entidad proveedora y de los datos
	Privacidad y seguridad	Legislación que afecta a los datos Restricciones de privacidad, seguridad y confidencialidad Percepción ciudadana sobre el uso de los datos
Metadatos	Completitud	Metadatos disponibles, interpretables y completos

[18] UNECE Big Data Quality Task Team. "A Suggested Big Data Quality Framework." UNECE, December 2014.
[19] Piet Daas, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Tóth. "Checklist for the Quality Evaluation of Administrative Data Sources." Discussion Paper. The Hague/Heerlen: Statistics Netherlands, 2009.

	Complejidad	Restricciones técnicas Datos estructurados, semiestructurados o no estructurados Legibilidad de los datos Presencia de jerarquías y anidamientos
	Accesibilidad y claridad	Accesibilidad de datos y metadatos Definiciones claras, explicaciones Conformidad con los estándares
	Relevancia	Grado en que los datos miden los conceptos que deben medirse para los usos previstos
	Usabilidad	Recursos adicionales necesarios para el tratamiento de los datos Análisis de los riesgos
	Tiempo	Oportunidad Periodicidad Cambios a través del tiempo
	Enlazamiento	Presencia y calidad de variables de enlace Niveles al que se puede realizar enlazamiento
	Coherencia y consistencia	Estandarización Disponibilidad de metadatos para variable clave
	Validez	Transparencia de métodos y procesos Solvencia de métodos y procesos
Datos	Exactitud y selectividad	Error total de la muestra Datos de referencia con los que comparar Selectividad. Problemas de cobertura
	Tiempo	Oportunidad Periodicidad
	Enlazamiento	Calidad de las variables de enlace
	Coherencia y consistencia	Coherencia entre los metadatos y los datos
	Validez	Coherencia de los procesos y métodos con los datos observados

En el capítulo 3 del presente trabajo realizaremos un mayor acercamiento al marco de calidad para el uso de fuentes Big Data en la producción de estadísticas oficiales.

B) Big Data y los principios asociados al derecho de acceso y la protección de la intimidad

El Código de Buenas Prácticas de las Estadísticas Europeas indica claramente, en su principio sobre recogida de datos, que las autoridades estadísticas deben tener un mandato jurídico claro para recoger la información destinada a la elaboración de estadísticas. Asimismo señala que a petición de las autoridades estadísticas, **se puede obligar por ley** a las administraciones, las empresas, los hogares y el público en general a permitir el acceso a los datos destinados a la elaboración de estadísticas europeas o a facilitar dichos datos.

En el apartado primero de este capítulo señalamos que una de las características peculiares de las fuentes Big Data es que generalmente la propiedad sobre las mismas no es pública. Asimismo, tal como veremos más adelante, muchas empresas han encontrado en estos datos un nuevo nicho de mercado que hasta el momento no habían explotado; donde los clientes potenciales identificados son tanto el sector privado como el sector público. Estos nichos de mercados se definen no tanto como acceso a datos sino como acceso a servicios a partir de datos, así tenemos por ejemplo el proyecto Smart Steps de Telefónica[20].

Evgeny Morozov, investigador sobre estudios políticos e implicaciones sociales de la tecnología, en una entrevista[21] de presentación de su último libro *La locura del solucionismo tecnológico*[22] en el diario El País señala que *"los datos son una de las más preciadas mercancías"*. A lo largo de la entrevista Morozov insiste en que en las últimas cinco décadas los datos se han convertido en una de las más preciadas mercancías:

> *"Tu seguro quiere saber qué posibilidades tienes de enfermar; tu banco quiere saber qué probabilidades tienes de no pagar tu hipoteca. Hay un mercado gigante de la venta de datos, no solo de tipo digital: si no miras lo que firmas cuando ofreces datos, es más que posible que acaben siendo agregados en una base administrada por un puñado de firmas norteamericanas." (Morozov, 2015)*

[20] http://dynamicinsights.telefonica.com/blog/488/smart-steps-2
[21] http://elpais.com/elpais/2015/12/17/eps/1450358550_362012.html
[22] Morozov, Evgeny. *La locura del solucionismo tecnológico*. Madrid; Móstoles, Madrid; Buenos Aires: Clave Intelectual ; Katz, 2015.

¿Y qué es lo que se debería hacer con ellos?, Evgeny Morozov plantea tres opciones:

1. Una es el statu quo: que un par de monopolios, Google y Facebook, continúen recopilando aún más información sobre nuestra vida para que pueda ser integrada en dispositivos inteligentes: mesas inteligentes, termostatos inteligentes; cualquier cosa que tenga un sensor generará un dato. Google Now es el paradigma de un sistema que intenta hacer acopio de todos esos datos para hacer predicciones y darte ideas. Si sabe que vas a volar te recuerda que hagas el check-in y te dice el tiempo atmosférico en el destino, actuando como un asistente virtual. Es el discurso de Google en términos de movilidad social: dar a los pobres los servicios que los ricos ya reciben.

2. La segunda es seguir a los disruptores. Hay compañías que chupan nuestros datos y los convierten en dinero. Una solución es que cada cual capture sus propios datos y los integre en un perfil, dando acceso a quien quiera y cobrando por ello. De ese modo, cada persona se convierte en un empresario.

3. Y la tercera opción aún no está muy articulada, pero debería ser perseguida según Morozov. Los datos, en un buen marco político, económico y legal, pueden llevarnos a servicios fantásticos. El único futuro del transporte público es una combinación de datos, algoritmos y sensores que determinan dónde está la gente y adónde quiere ir.

En ese sentido Evgeny Morozov indica que habría que oponerse a que el paradigma de la propiedad privada se extienda a los datos:

> *"Ha habido esfuerzos de comercializar hasta el aire, y hay que oponerse. Los datos, sin la capacidad de analizarlos, no son gran cosa. Hoy en día solo algunas grandes empresas son capaces de estudiarlos. Esa información debería estar bajo un control público, que no significa un control del Estado, sino de los ciudadanos. La reciente fascinación en Europa por esa idea del común, que no tiene nada que ver con la de los comunes, es un marco sano. La gente podría ceder esos datos voluntariamente, pero siendo propietaria de estos."*

Esta perspectiva contrasta con la aportaciones del European Big Data Value Partnership[23] en sus informes *European Big Data Value Strategic Research & Innovation Agenda*[24] en los que se ponen en

[23] http://www.bdva.eu/
[24] Big Data Value Europe. "European Big Data Value Strategic Research & Innovation Agenda". Big Data Value Association, January 2016.

valor la potencialidad económica de las fuentes Big Data y se define una agenda estratégica de investigación e innovación europea para su desarrollo.

Como vemos, hay un debate intenso sobre los datos, su propiedad y el derecho de acceso para fines públicos. Si bien la legislación estadística puede obligar a facilitar el acceso a las Oficinas Estadísticas, esta capacidad tendrá que convivir en la tensión de intereses público-privado contrapuestos; tensión que necesitará de espacios de cooperación con los proveedores. En esa línea se sitúan seminarios como el *Joint OECD - PARIS21 Workshop - Access to New Data Sources for Statistics: Business Models for Private-Public Partnerships*[25] para cuya preparación se elaboró el informe *Public-Private Partnerships for Statistics (*Klein, Jütting, and Robin, 2016) que es un buen análisis sobre el problema aquí planteado.

Por otra parte, el *Manual de organización estadística* nos recuerda que la potestad que confiere la legislación a las Oficinas de Estadística para recabar información no es de mayor utilidad a menos que todos los sectores de la sociedad estén dispuestos a cooperar. En ese sentido es importante señalar que la confidencialidad de la información individual es, probablemente, la mayor preocupación de los informantes; especialmente cuando se trata de gran acumulación de datos por parte del Estado, datos que en un principio han sido generados por los ciudadanos para otros fines distintos a los estadísticos.

Ante lo expuesto es importante señalar que existe el peligro de que entre la sociedad se genere una visión de las Oficinas de Estadísticas como instituciones orwellianas. Por ejemplo, tras la publicación del artículo denominado *Las operadoras seguirán el rastro de tu móvil para alimentar el censo de 2021*[26] en el que se hace público por parte del Instituto Nacional de Estadística (INE) de España el uso de datos de telefonía móvil para los estudios de movilidad del Censo de 2021, se desató un amplio debate en Menéame[27] contrario a su uso. Paralelamente se publicaron varios artículos en blogs especializados sobre la legalidad de la medida, como por ejemplo el artículo titulado *La ilegalidad de usar los datos del móvil para completar el censo*[28].

En el capítulo 3 del presente trabajo revisaremos las estrategias, riesgos y soluciones de fuentes Big Data en la producción de estadísticas oficiales; y entre ellos estudiaremos los riesgos relacionados con el accesos de datos.

[25] http://www.oecd.org/std/oecd-paris21-workshop-access-to-new-data-sources-for-statistics.htm
[26] http://www.eldiario.es/hojaderouter/tecnologia/moviles/censo-2021-INE-big_data-operadoras_0_493100796.html
[27] https://www.meneame.net/m/tecnolog%C3%ADa/operadoras-seguiran-rastro-tu-movil-alimentar-censo-2021
[28] http://derechoynormas.blogspot.com.es/2016/03/la-ilegalidad-de-usar-los-datos-del.html

C) Big Data y los principios de objetividad política y científico-técnica

El Manual de Organización Estadística elaborado por NNUU advierte que para tener credibilidad y desempeñar su función es preciso que las Oficinas de Estadística tengan una posición de independencia ampliamente reconocida. Sin la credibilidad derivada de un alto grado de independencia, los usuarios perderán la confianza en la exactitud y la objetividad de la información del organismo y quienes le proporcionan los datos estarán menos dispuestos a cooperar con él. Esta credibilidad se desarrolla en varios Principios Fundamentales de las Estadísticas Oficiales:

1. *Relevancia, imparcialidad y acceso equitativo:* Las estadísticas oficiales constituyen un elemento indispensable en el sistema de información de una sociedad democrática y proporcionan al gobierno, a la economía y al público datos acerca de la situación económica, demográfica, social y ambiental. Con este fin, los organismos oficiales de estadística han de compilar y facilitar en forma imparcial estadísticas oficiales de comprobada utilidad práctica para que los ciudadanos puedan ejercer su derecho a la información pública.

2. *Patrones profesionales, principios científicos y ética:* Para mantener la confianza en las estadísticas oficiales, las Oficinas de Estadística han de decidir con arreglo a consideraciones estrictamente profesionales, incluidos los principios científicos y la ética profesional, acerca de los métodos y procedimientos para la reunión, el procesamiento, el almacenamiento y la presentación de los datos estadísticos.

3. *Responsabilidad y transparencia:* Para facilitar una interpretación correcta de los datos, las Oficinas de Estadística han de presentar información conforme a normas científicas sobre las fuentes, métodos y procedimientos de la estadística.

4. *Uso de patrones internacionales:* La utilización por las Oficinas de Estadística de cada país de conceptos, clasificaciones y métodos internacionales fomenta la coherencia y eficiencia de los sistemas estadísticos a nivel oficial.

El Código de Buenas Prácticas de las Estadísticas Europeas es más exhaustivo respecto al conjunto de principios relacionados con la objetividad política y científico-técnica:

1. *Independencia profesional*. La independencia profesional de las autoridades estadísticas frente a otros departamentos y organismos políticos, reguladores o administrativos, y frente a los operadores del sector privado, garantiza la credibilidad de las estadísticas europeas.

2. *Imparcialidad y objetividad*. Las autoridades estadísticas desarrollan, elaboran y difunden estadísticas europeas respetando la independencia científica y de forma objetiva, profesional y transparente, de modo que todos los usuarios reciben el mismo trato.

3. *Metodología sólida*. Las estadísticas de calidad se apoyan en una metodología sólida, que requiere herramientas, procedimientos y conocimientos adecuados.

4. *Procedimientos estadísticos adecuados*. Las estadísticas de calidad se apoyan en procedimientos estadísticos adecuados, aplicados desde la recogida de los datos hasta la validación de los mismos.

5. *Precisión y fiabilidad.* Las estadísticas europeas reflejan la realidad de manera precisa y fiable.

6. *Coherencia y comparabilidad.* Las estadísticas europeas son consistentes internamente a lo largo del tiempo y comparables entre regiones y países; es posible combinar y utilizar conjuntamente datos relacionados procedentes de fuentes diferentes.

Revisando los principios y considerando que las fuentes Big Data, tal como hemos señalado anteriormente, en buena medida son de origen privado y que además no están diseñadas para fines estadísticos, se pueden dar algunos problemas que las Oficinas Estadísticas deben saber abordar. Por ejemplo:

1. Desconfianza de la ciudadanía en los resultados estadísticos, como producto de su desconfianza en las empresas cedentes de los datos y en la no manipulación de los mismos por parte de dichas empresas a favor de sus intereses económicos, o la ruptura de los acuerdos de cesión si los datos no les son favorables. En definitiva no es más que una nueva figura de desconfianza sobre la independencia profesional de la Oficinas Estadísticas frente a los operadores del sector privado.

2. Dificultad para armonizar distintas fuentes con diferentes objetivos, con la finalidad de poder proporcionar datos comparables entre regiones y países; y consistentes internamente a lo largo del tiempo.

3. Problemas metodológicos no triviales, al estar habitualmente ante grandes volúmenes de datos que no son datos censales, sino en todo caso muestras de una población o más genéricamente de eventos de una población. Por lo tanto nos encontramos ante la suma de las dificultades metodológicas producto de muestras no probabilísticas, a las que se deben sumar los problemas habituales de las estadísticas basadas en registros administrativos.

En el capítulo 2 del presente trabajo realizaremos una mejor aproximación a las capacidades informativas y los problemas metodológicos de las fuentes Big Data en contraste con otras fuentes de datos tradicionales en la estadística pública.

1.4. Los retos Big Data a los que se enfrenta la estadística pública

1.4.1. La respuesta de la estadística pública ante un nuevo escenario

En 2010, la Oficina de la Conferencia de Estadísticos Europeos[29] creó el Grupo de Alto Nivel *Modernisation of Statistical Production and Services (HLG)*[30] para supervisar y coordinar el trabajo internacional sobre modelos de negocio dentro de las oficinas de estadística. Dentro de este grupo se formó un equipo de trabajo de expertos, coordinados por la Secretaría de la UNECE[31], con el objetivo de producir un documento que explicara los problemas relacionados sobre el uso del Big Data por las Oficinas Estadísticas.

El Grupo de Trabajo publicó en marzo de 2013 el documento *What does "Big Data" for Official Statistics?*[32] que es el primer documento estratégico que analiza los principales desafíos en materia de legislación, privacidad, cuestiones financieras, gestión, metodologías y tecnología y que además ofrece algunas recomendaciones básicas para las Oficinas de Estadística. El documento señala desde un primer momento que la recolección de datos de fuentes Big Data y su incorporación al proceso de producción de estadísticas no es tarea fácil, y en ese sentido intenta abordar dos cuestiones

[29] http://www.unece.org/stats/cesbureau.html
[30] http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics
[31] http://www.unece.org/info/ece-homepage.html
[32] Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?" UNECE, March 10, 2013. http://www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184.

elementales: (1) En qué conjuntos de datos deben centrar su atención las Oficinas de Estadística (2) Cómo una Oficina de Estadística puede utilizar las fuentes Big Data y los retos asociados a su uso.

En esta dirección, reconociendo la necesidad de investigar más a fondo los beneficios y desafíos de las fuentes Big Data para las estadísticas oficiales, la Comisión de Estadística de Naciones Unidas acordó, en su 45ª sesión de marzo de 2014, crear el Grupo de Trabajo Global (GTG)[33] sobre Big Data y Estadísticas Oficiales. Este grupo de trabajo nació tras la celebración del seminario previo a la 44ª sesión de la Comisión de Estadística en 2013 sobre *Big Data for Policy, Development and Official Statistics*[34]. En este seminario oradores del sector privado y de Oficinas de Estadística llegaron a la conclusión de que las fuentes Big Data constituyen una fuente de información que no puede ser ignorada por la estadística pública y que los estadísticos oficiales debían organizarse y tomar medidas urgentes para explotar las posibilidades y abordar los retos asociados con eficacia.

Con la aprobación del grupo de trabajo, la comunidad estadística internacional reconoce el potencial de las fuentes Big Data para las estadísticas oficiales, entendiendo que éstas pueden ayudar a cumplir mejor el mandato de proporcionar estadísticas oportunas y coherentes sobre la economía, la demografía, la sociedad y el medio ambiente. Las labores del grupo de trabajo son fundamentalmente abordar los retos enumerados en el apartado anterior, con el fin de identificar estrategias en cuestiones relativas a la legislación, la privacidad, las finanzas, las gestión, la metodología y la tecnología. Estos retos se abordan actualmente por 8 equipos de trabajo: (1) Promoción y comunicación, (2) Big Data y los objetivos del desarrollo sostenible, (3) Acceso a datos y cooperación con propietarios, (4) Formación, (5) Cuestiones transversales, (6) Datos de telefonía móvil, (7) Imágenes de satélite y (8) Datos de redes sociales. Estos grupos de trabajo se han ido complementando con seminarios y conferencias internacionales:

2014 Octubre (Pekín) - International Conference on Big Data for Official Statistics[35]
2015 Marzo (Nueva York) - Big Data Seminar at the 46th UN Statistical Commission[36]
2015 Octubre (Abu Dhabi) - 2nd Global International Conference on Big Data for Official Statistics[37]
2016 Septiembre (Dublín) - 3rd Global International Conference on Big Data for Official Statistics[38]
2017 Noviembre (Bogotá) - 4th Global International Conference on Big Data for Official Statistics[39]

[33] http://unstats.un.org/unsd/bigdata/
[34] http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html
[35] http://unstats.un.org/unsd/trade/events/2014/beijing/default.asp
[36] http://unstats.un.org/unsd/statcom/statcom_2015/seminars/big_data/default.html
[37] http://unstats.un.org/unsd/trade/events/2015/abudhabi/default.asp
[38] http://unstats.un.org/unsd/bigdata/conferences/2016/default.asp
[39] https://unstats.un.org/unsd/bigdata/conferences/2017/default.asp

Una de las primeras tareas abordadas por el UN Global Working Group on Big Data for Official Statistics fue realizar la encuesta Big Data Project Survey[40] a Oficinas Estadísticas con el objetivo de identificar las actividades desarrolladas entorno a Big Data. Esta investigación se abordó desde los siguientes campos de interés: (1) Estrategia Big Data, (2) Gobierno, (3) Calidad, (4) Privacidad y (5) Habilidades.

El nacimiento del Grupo de Trabajo Global de Naciones Unidas vino precedido por el Scheveningen Memorandum[41] sobre "Big Data and Official Statistics" adoptado por el European Statistical System Committee (ESSC)[42], el 27 de septiembre de 2013. En el memorando, la Conferencia de Directores Generales de Institutos Nacionales de Estadística (DGINS)[43] reconoció que las fuentes Big Data representan nuevas oportunidades y retos para las estadísticas oficiales y animaba al Sistema Estadístico Europeo y sus socios a examinar el potencial de dichas fuentes. Los acuerdos incluidos en el memorandum son los siguientes:

1. *Reconocimiento.* Reconocer que el Big Data representa nuevas oportunidades y desafíos para las estadísticas oficiales, y por lo tanto animar al Sistema Estadístico Europeo y sus socios a examinar el potencial del Big Data en ese sentido.

2. *Necesidad de estrategia*. Reconocer que el Big Data es un fenómeno que está afectando a muchos ámbitos. Por tanto, es esencial desarrollar una "Estrategia de estadísticas oficiales basadas en Big Data" y examinar el lugar y las interdependencias de esta estrategia en el contexto más amplio de una estrategia global del gobierno a nivel nacional, así como a nivel de la UE.

3. *Legislar el acceso de datos*. Reconocer las implicaciones del Big Data en la legislación de protección de datos y derechos de las personas (por ejemplo, acceso a fuentes de datos en poder de terceros), implicaciones que deben ser abordadas apropiadamente como un asunto prioritario.

4. *Compartir experiencias*. Tener en cuenta que varios institutos nacionales de estadística están iniciando actualmente o considerando los diferentes usos del Big Data en un contexto nacional. Es necesario compartir las experiencias obtenidas en los proyectos Big Data concretos y colaborar dentro del Sistema Estadístico Europeo y a escala internacional.

[40] http://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf
[41] http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13
[42] http://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/essc
[43] http://ec.europa.eu/eurostat/web/ess/about-us/ess-gov-bodies/dgins

5. *Formación.* Reconocer que el desarrollo de las capacidades y habilidades necesarias para explorar con eficacia las fuentes Big Data es esencial para su incorporación en el Sistema Estadístico Europeo. Esto requiere esfuerzos sistemáticos, con cursos de formación adecuados y el establecimiento de comunidades de intercambio de experiencias y buenas prácticas.

6. *Cooperación.* Reconocer el carácter multidisciplinar del Big Data, lo que requiere sinergias y asociaciones entre los expertos y las partes interesadas de diversos dominios, incluyendo gobierno, universidades y titulares de las fuentes de datos privadas.

7. *Innovación metodológica y tecnológica.*. Reconocer que el uso de grandes volúmenes de datos en el contexto de las estadísticas oficiales requiere nuevos desarrollos metodológicos, de evaluación de la calidad y de abordaje de los problemas tecnológicos relacionados. El Sistema Estadístico Europeo debería hacer un esfuerzo especial para apoyar esos desarrollos.

8. *Plan de acción.* Los Directores coinciden en la importancia de dar seguimiento a la implementación del memorando, y por lo tanto consideran que es necesario adoptar un plan de acción y plan de trabajo del Sistema Estadístico Europeo para el uso de fuentes Big Data.

Por otra parte, la *European Statistical System - Vision 2020* [44] es una respuesta estratégica común, aprobada en mayo de 2014, del Sistema Europeo de Estadística (Eurostat, los Estados miembros de la UE y de la EFTA) a los desafíos a lo que se enfrentan las estadísticas oficiales. En la ESS-Vision 2020 se identifica como uno de los elementos clave para el sistema en el año 2020 la incorporación de nuevas fuentes de datos, en ese aspecto el sistemas se visiona para el 2020 de las siguiente manera:

> *"Basamos nuestros productos y servicios estadísticos en encuestas tradicionales y nuevas fuentes, incluyendo datos administrativos, geoespaciales y, cuando sea posible, fuentes Big Data. Las nuevas fuentes de datos complementan las ya existentes y nos ayudan a mejorar la calidad de nuestros productos. Vamos a trabajar juntos para conseguir el acceso a nuevas fuentes de datos, crear métodos y encontrar la tecnología adecuada con el fin de utilizar nuevas fuentes de datos para elaborar estadísticas europeas de una manera fiable."*

[44] http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255

La necesidad de elaborar un plan de acción fue uno de los elementos considerados en la convocatoria del *2014 ESS Big Data Event: Big Data in Official Statistics*[45]. Posteriormente en la 22ª Sesión del European Statistical System Committee (ESSC)[46], de 26 de septiembre de 2014, se aprobó el documento *Big Data Action Plan and Roadmap 1.0*[47] en el que se se plantea una visión para 2020 y post-2020; y además se enumeran actuaciones en varios tópicos: (1) Política, (2) Comunicación, (3) Identificación de fuentes Big Data, (4) Aplicaciones/pilotos, (5) Métodos, (6) Calidad, (7) Tecnología, (8) Habilidades, (9) Intercambio de experiencias, (10) Legislación, (11) Gobierno.

1.4.2. Los retos identificados por la estadística pública

Es evidente que el desafío del uso de datos de fuentes Big Data dentro de la estadística pública significa necesariamente la modernización de las Oficinas Estadísticas. Ese desafío requiere al abordaje de diferentes retos, que sintéticamente podemos resumir en los siguientes puntos:

Estrategia: Es necesario definir cómo integrar las nuevas fuentes Big Data en la actividad de las Oficinas Estadísticas. Esta estrategia puede estar dirigida tanto a la integración de las nuevas fuentes en la producción habitual de las Oficinas, como en la identificación de nueva información estadística basada en dichas fuentes.

Acceso: Existe un debate intenso sobre los datos, su propiedad y el derecho de acceso para fines públicos. Si bien la legislación estadística puede obligar a facilitar el acceso a las Oficinas Estadísticas, esta capacidad tendrá que convivir en la tensión de intereses público-privado contrapuestos; tensión que necesitará de espacios de cooperación con los proveedores.

Privacidad: La datificación de buena parte de nuestras vidas genera actitudes diversas en la opinión pública sobre el derecho a la intimidad. Sin embargo cuando se trata de gran acumulación de datos por parte del Estado la confidencialidad, proporcionalidad y fin de los mismos pasan a ser una importante preocupación ciudadana. En ese sentido existe el peligro de que entre la sociedad se genere una visión de las Oficinas de Estadísticas como instituciones orwellianas.

---

[45] https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf

[46] http://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/essc

[47] Eurostat Big Data Task Force. "Big Data Action Plan and Roadmap 1.0." Eurostat. Accessed February 28, 2016. https://ec.europa.eu/eurostat/cros/sites/crosportal/files/ESSC%20doc%2022_8_2014_EN_Final%20with%20ESSC%20opinion.pdf.

Por otra parte, la generación de gran cantidad de datos a gran velocidad pone sobre la mesa nuevos retos tecnológicos para cumplir el mandato del deber de secreto estadístico, que impide que a través de la información publicada por las Oficinas Estadísticas se pueda identificar directa o indirectamente a las unidades de análisis.

Calidad: La dimensiones de evaluación de la calidad de las fuentes Big Data para su integración en la actividad de las Oficinas Estadísticas deben ser identificadas, especialmente debido a que son datos recopilados para fines no estadísticos.

Metodología: Con las fuentes Big Data nos encontramos ante la dificultad de datos recopilados para fines no estadísticos, por lo tanto estamos ante problemas similares a los planteados con los registros administrativos, al menos en lo que respecta a los conceptos usados en la recolección de datos y su relación con las definiciones internacionalmente armonizadas. Además algunas de las fuentes Big Data son muestras, con el problema añadido de ser muestras no probabilísticas y posiblemente sesgadas por el método o por las cuotas de mercado del agente recolector.

Tecnología: La incorporación de fuentes Big Data a la actividad estadística requerirá de la incorporación de tecnología Big Data a las Oficinas Estadísticas. Definir arquitecturas, hardware y software requeridos es uno de los retos que debe ser abordado.

Formación: El desarrollo de las capacidades y habilidades necesarias para explorar con eficacia los Big Data es esencial para su incorporación a la actividad de la Oficinas Estadísticas. Esto requiere esfuerzos sistemáticos, como cursos de formación adecuados y el establecimiento de comunidades de intercambio de experiencias y buenas prácticas.

Bibliografía del Capítulo 1

Big Data Value Europe. “European Big Data Value Strategic Research & Innovation Agenda.” Big Data Value Association, January 2015. http://www.bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final.pdf#overlay-context=downloads%26page%3D1%3Fq%3Ddownloads%26page%3D1.

Big Data Value Europe. “European Big Data Value Strategic Research & Innovation Agenda.” Big Data Value Association, January 2016. http://www.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA__v2.pdf.

Borgman, Christine L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press, 2015.

Caballero Roldán, Rafael, and Enrique Martín Martín. *Las bases de Big Data*. Madrid: Los Libros de la Catarata : Universidad Complutense de Madrid, 2015.

Cavallo, A. *Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization [recurso Electrónico]*. Harvard University, 2009. https://books.google.es/books?id=3r_-ZwEACAAJ.

Cavallo, A. "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics* 60, no. 2 (2013): 152–65. doi:http://dx.doi.org/10.1016/j.jmoneco.2012.10.002.

Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?" UNECE, March 10, 2013. http://www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184.

Eurostat. "European Statistical System - Vision 2020." Eurostat. Accessed November 6, 2016. http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255.

Klein, Thilo, Johannes Jütting, and Nicholas Robin. "Public-Private Partnerships for Statistics: Lessons Learned, Future Steps." OECD Development Co-operation Working Papers, February 29, 2016. http://www.oecd-ilibrary.org/development/public-private-partnerships-for-statistics-lessons-learned-future-steps_5jm3nqp1g8wf-en.

Letouzé, E. "Big Data for Development: Challenges & Opportunities." UN Global Pulse, May 2012. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

Maeztu, David. "La ilegalidad de usar los datos del móvil para completar el censo." *Del derecho y las normas*. Accessed May 4, 2016. http://derechoynormas.blogspot.com.es/2016/03/la-ilegalidad-de-usar-los-datos-del.html?spref=tw.

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt, 2013.

Piet Daas, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Tóth. "Checklist for the Quality Evaluation of Administrative Data Sources." Discussion Paper. The Hague/Heerlen: Statistics Netherlands, 2009.

http://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60.

Reimsbach-Kounatze, Christian. "The Proliferation of 'Big Data' and Implications for Official Statistics and Statistical Agencies." OECD Digital Economy Papers, January 12, 2015. http://www.oecd-ilibrary.org/science-and-technology/the-proliferation-of-big-data-and-implications-for-official-statistics-and-statistical-agencies_5js7t9wqzvg8-en.

Struijs, Peter, Barteld Braaksma, and Piet JH Daas. "Official Statistics and Big Data." *Big Data & Society* 1, no. 1 (June 10, 2014). doi:10.1177/2053951714538417.

UNECE Big Data Quality Task Team. "A Suggested Big Data Quality Framework." UNECE, December 2014.

United Nations. *Manual de Organización Estadística. El Funcionamiento y la Organización de una Oficina de Estadística.* New York: United Nations Publications, 2005. http://www.cepal.org/publicaciones/xml/7/15497/lcw6e.pdf.

Unión Europea. *Directiva 2013/37/EU del Parlamento Europeo y del Consejo de 26 de Junio, por la que se modifica la Directiva 2003/98/EC relativa a la reutilización de la información del sector público*. Accessed February 20, 2016. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:ES:PDF.

Wallgren, Anders, and Britt Wallgren. *Register-Based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology. Chichester, England ; Hoboken, NJ: John Wiley & Sons Ltd, 2007.

Capítulo 2. Las fuentes Big Data en la estadística pública y sus problemas metodológicos

"El experto en estadística no puede eludir la obligación de tener claros los principios de inferencia científica, pero tampoco las demás personas pueden evitar tal obligación."

R.A. FISHER, *Design of Experiments*

2.1. Tipologías de fuentes de datos y sus características

2.1.1. Datos primarios y de primera-segunda generación: censos y encuestas

Las Oficinas de Estadísticas son los órganos de las Administraciones Públicas encargados de la producción y publicación de las estadísticas oficiales sobre un amplio abanico de temas de interés nacional o subnacional.

Hasta alrededor de la década de los 80 los datos fueron esencialmente un bien escaso por el alto precio de su adquisición, y este trabajo se le encomendaba a las Oficinas de Estadística como órganos especializados del Gobierno. Inicialmente se les encargó la realización de censos o encuestas completos, pues los métodos muestrales no existían o no estaban suficientemente avanzados, dando lugar a lo que podríamos denominar **fuentes de datos de primera generación**. Posteriormente, con el desarrollo de los métodos estadísticos y las teorías muestrales durante la primera mitad del Siglo XX y la aparición de la unidad estadística de Naciones Unidas, se impulsó el uso de encuestas dentro de las Oficinas Estadísticas a lo largo de las segunda mitad del Siglo XX; dando lugar a lo que podríamos denominar **fuentes de datos de segunda generación.**

Como vemos, antes de la era de la datificación mucha información no estaba disponible y debía ser recogida para un propósito particular. La información estadística oficial, basada fundamentalmente en datos de encuestas o censos, tenía un valor único pues simplemente no había otra alternativa. Por ejemplo, los datos de los censos de población eran inmensamente valiosos para los responsables políticos, investigadores y otros usuarios.

Por esta razón en la actualidad muchas de las estadísticas elaboradas por las Oficinas de Estadística están realizadas a partir de encuestas basadas en muestras o censos. Los datos coleccionados a través de estos procedimientos suelen definirse como **datos primarios**. Al comparar las encuestas por muestreo con los censos completos, en términos relativos, cada uno de ellos se revela más fuerte allí donde el otro es más débil. Por lo tanto podríamos considerarlos contrapuestos, pero insistiendo también sobre su naturaleza complementaria.

La realización de un censo completo de toda la población requiere la movilización de recursos financieros y humanos a gran escala, movilización que no puede mantenerse durante un periodo prolongado ni ser repetida con frecuencia. La necesidad de desplegar un operativo humano numeroso, y por lo tanto peor entrenado y supervisado, significa que el tipo de información que puede recogerse de forma adecuada en un censo tiene que ser de contenido relativamente sencillo. Por ello los censos son costosos y lentos, incluso con los actuales procedimientos modernos y eficientes se puede tardar varios años hasta que la mayor parte de los datos censales lleguen a mano de los usuarios. Éstas son las razones básicas por las que no se realizan censos con mayor frecuencia, o con una mayor profundidad y riqueza de datos.

Por consiguiente, el objetivo primario de un censo consiste, por regla general, en obtener un retrato detallado y completo del tamaño de la población y de sus características estructurales básicas, proporcionando un buen detalle para dominios pequeños y especialmente para áreas locales. Por el contrario, las investigaciones muestrales pueden diseñarse de modo que se obtenga una amplia variedad de datos para el estudio de interrelaciones y cambios. Además las encuestas por muestreo pueden diseñarse de modo flexible mediante métodos apropiados de acopio de datos, de modo que se acomoden a una gama amplia de necesidades. A su vez su elaboración es mucho más barata que la de un censo, por lo que pueden repetirse más a menudo con el fin de suministrar información sobre variables que cambian y fluctúan rápidamente.

Sin embargo, las limitaciones más importantes de las encuestas por muestreo es su incapacidad para suministrar detalles suficientes sobre dominios pequeños, y especialmente sobre áreas locales. Éste es el principal motivo por el que los censos generales siguen conservando su utilidad, aunque en algunos países o regiones se está planeando prescindir de su ejecución. Así por ejemplo la Unión Europea pretende que el censo 2021 ó el 2031 sea el último censo tal como lo conocemos. Además las muestras generalmente dependen de información de fuentes externas, tanto en lo que respecta a los marcos de selección de la muestra como también respecto a su uso en estimadores de razón y otros métodos parecidos.

2.1.2. Datos secundarios y de tercera generación: registros administrativos

A partir de la década de los 90 los datos recogidos por las Administraciones Públicas fueron cada vez más accesibles para fines estadísticos, como consecuencia de la informatización de sus procedimientos. En este escenario, la recopilación de datos estadísticos por medio de cuestionarios se complementó, e incluso se sustituyó, por fuentes de datos administrativas con el fin de reducir costes y reducir la carga sobre los encuestados; dando lugar a lo podríamos denominar f**uentes de datos de tercera generación**. Hoy en día algunos países no llevan a cabo amplios estudios poblacionales, realizando sus censos mediante la combinación y el análisis de datos de varias fuentes administrativas. En ese sentido algunas Oficinas Estadísticas han procedido a la combinación e integración de diferentes registros administrativos, constituyendo un sistema de datos integrados como base importante en su actividad estadística[48].

Una primera comparativa entre tipos de fuentes la podemos encontrar ya en 1979, en el artículo *Samples and Censuses* de Leslie Kish[49], donde se identifican ocho criterios para compararlos:

Tabla 2.- Ocho criterios para comparar muestras, censos y registros administrativos

Criterios	Muestras	Censos	Registros administrativos
Rico, complejo, variado, flexible	***		
Preciso, relevante, pertinente	*		?
Económico	*		***
Apropiado, oportuno, estacional	**		*
Preciso (grande y completo)		*	*
Detallado para dominios pequeños		**	*
Extenso (cobertura), creíble		*	?
Contenido poblacional	**	*	

Kish utiliza los asteriscos (*) para indicar las ventajas relativas de cada fuente, de acuerdo a cada criterio. El signo (?), usado para los registros administrativos, se refiere a las diferencias extremas de las distintas variables en diferentes situaciones. Como señala el autor, los registros de nacimientos y

[48] Wallgren, Anders, and Britt Wallgren. *Register-Based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology. Chichester, England ; Hoboken, NJ: John Wiley & Sons Ltd, 2007.

[49] Kish, Leslie. "Samples and Censuses." *International Statistical Review / Revue Internationale de Statistique* 47, no. 2 (August 1979): 99. doi:10.2307/1402563.

defunciones, servicios públicos (teléfono, electricidad), impuestos, etc., pueden ser ser precisos o deficientes. Esta advertencia acerca de la precisión se aplica tanto a su pertinencia[50] como a su cobertura. Hacer uso de ellos puede resultar muy económico (***) siempre que estén disponibles, ya que fueron otros departamentos ajenos a las Oficinas Estadísticas quienes soportaron sus costes. Pero rara vez poseen la riqueza y la diversidad de información requerida por la estadística pública; y su contenido poblacional puede estar limitado. Respecto al cuadro, Kish llama la atención en la naturaleza complementaria de las muestras, dominante en cinco criterios, y de los censos, que dominan en los otros tres. Su escepticismo respecto al uso de los registros administrativos posiblemente es consecuencia de la fecha en la que se elabora el artículo, cuando todavía la informatización de las Administraciones Públicas no estaba tan avanzada como en la actualidad.

Una comparativa más actual de las características entre tipos de fuentes la encontramos en el ya clásico libro *Register-Based Statistics: Administrative Data for Statistical Purposes* (Wallgren, Anders and Wallgren, Bitt 2007):

Esquema 1.- Similitudes y diferencias entre fuentes primarias y secundarias

<table>
<tr><th>Encuestas</th><th>Censos</th><th>Registros administrativos</th></tr>
<tr><td>No incluidos en sistema de datos integrados</td><td colspan="2">Incluidos en sistema de datos integrados. Pueden ser usados para diversas estadísticas</td></tr>
<tr><td colspan="3">Usan el sistema de datos integrados para definir la población de análisis y recuperar datos</td></tr>
<tr><td>Muestreo, estimadores, medidas del error</td><td colspan="2">La integración con otros registros administrativos es importante</td></tr>
<tr><td colspan="2">Recolección propia - cuestionarios propios</td><td>Uso de los registros administrativos de terceros</td></tr>
<tr><td colspan="2">Edición - posibilidad de recontacto con las unidades informantes</td><td>Posibilidad de contacto con la autoridad responsable del registro</td></tr>
<tr><td colspan="2">No respuesta</td><td>Valores perdidos. Infracobertura</td></tr>
<tr><td>Defectos de calidad: errores de muestreo, errores de medición</td><td>Defectos de calidad: errores de medición</td><td>Defectos de calidad: errores de relevancia, falta de comparabilidad</td></tr>
<tr><td>Tablas pequeñas. No hay estimaciones para pequeños grupos</td><td colspan="2">Tablas grandes con muchas celdas. Posibilidad de tener datos para pequeños grupos</td></tr>
</table>

[50] Pertinencia: Las estadísticas satisfacen las necesidades de los usuarios.

Como observamos las diferencias entre los distintos tipos de fuentes son fundamentalmente producto de la naturaleza para la cual se ha definido la datificación. Evidentemente las fuentes administrativas son diseñadas para la gestión de las políticas y procedimientos de la Administración Pública que los impulsa o gestiona; por lo tanto los datos son un reflejo de un procedimiento o política en una Administración.

Las diferencias en la materialización de políticas en los diferentes Estados y Regiones dan lugar a procedimientos distintos que se concretan en fuentes de datos distintas. Esas diferencias están referidas tanto a las unidades sobre las que se recopilan datos, como a la información recogida sobre las mismas; dando lugar a que el alineamiento con definiciones y conceptos internacionales sea una labor prácticamente imposible. Un ejemplo paradigmático en España son las diferencias entre las mediciones de paro en la Encuesta de Población Activa (EPA), alineada con las definiciones y conceptos recomendados por la Organización Internacional de Trabajo y con la metodología definida por Eurostat, y las mediciones de Paro Registrado elaboradas según los criterios definidos en la Orden de 11 de marzo de 1985 por la que se establecen criterios estadísticos para la medición del paro registrado[51]. Un situación similar nos la encontramos con los datos de población ocupada aportados por la EPA y los ofrecidos por la Estadística de Afiliación a la Seguridad Social.

Por lo tanto en el uso de fuentes administrativas para fines estadísticos nos encontramos dos problemas importantes vinculados con la cobertura: Por una parte el problema de la cobertura poblacional del registro, como consecuencia de la exclusión de unidades de observación[52], y por otra parte el problema de la cobertura conceptual de la información capturada en el registro. Ambos problemas impiden o dificultan el uso de fuentes administrativas en la elaboración de estadísticas internacionalmente comparables, dificultades que se han ido resolviendo con el tiempo a través de diferentes estrategias metodológicas.

Sin embargo, la informatización de la Administraciones Públicas trajo consigo la posibilidad de acceder más fácilmente y a menor coste a datos de las unidades de observación, genéricamente personas físicas y jurídicas, pero también trajo consigo la recolección de **datos de eventos**, generados por o sobre una población dada. Estos datos tienen la peculiaridad de registrar los eventos a los que se ha visto expuesto una población dada, dando lugar a un mayor volumen de información que la recopilada por encuestas o censos, pero sobre todo dando lugar a la posibilidad de la mejora de los

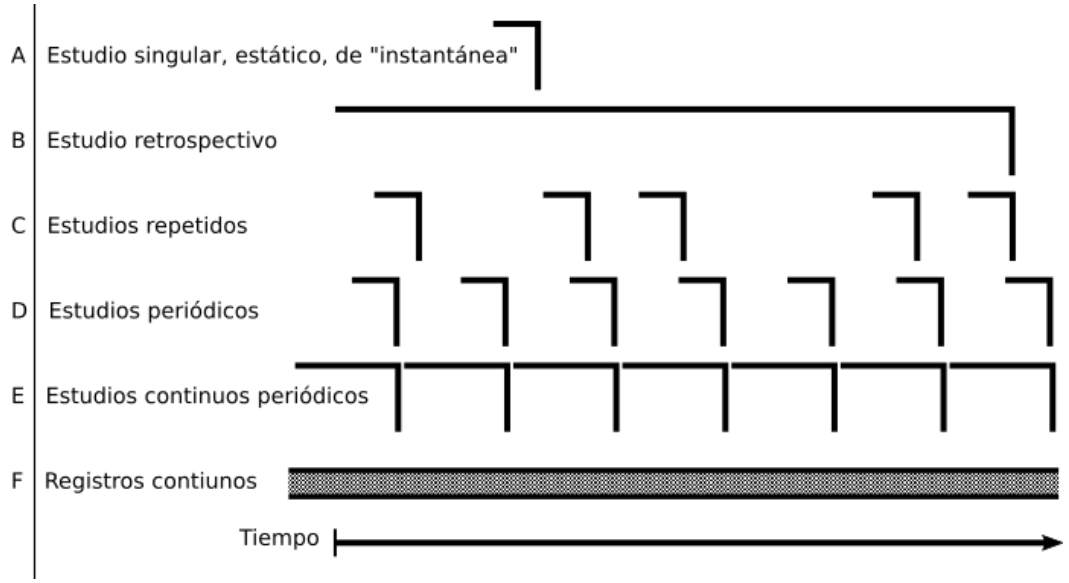
[51] https://www.boe.es/diario_boe/txt.php?id=BOE-A-1985-4112

[52] El problema de cobertura poblacional no sólo es un problema de exclusión de unidades, también puede ser de inclusión inadecuada. Genéricamente podemos definirlo con un problema de falta de coincidencia entre la población objeto de estudio y el marco poblacional utilizado para la selección de las unidades de observación.

estudios longitudinales. Para la representación de esta nueva capacidad respecto al tratamiento del tiempo en las investigaciones estadísticas, resulta pertinente recuperar el esquema de Leslie Kish en *"Diseño estadístico para la investigación"*[53] sobre cómo elaborar diseños para abarcar poblaciones a lo largo de periodos de tiempo:

Esquema 2.- Diseños para abarcar poblaciones a lo largo del tiempo

Las líneas verticales simbolizan las fechas de recogida de datos y las barras horizontales los periodos de referencia que se están cubriendo.



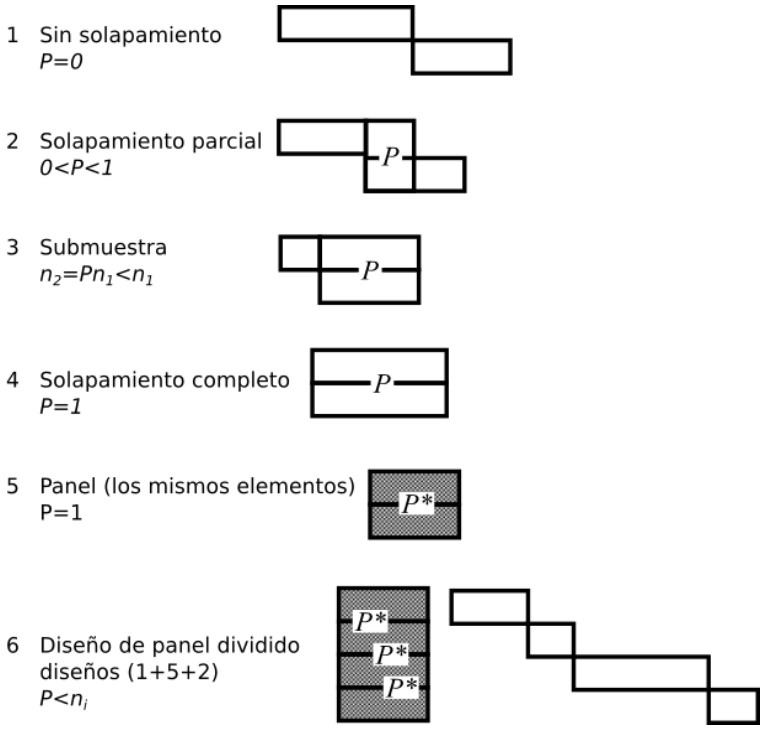
Los estudios retrospectivos abarcan periodos largos, quizás a lo largo de toda la vida de los sujetos, a partir de una fecha inicial de recogida. Los estudios repetidos recogen información a intervalos irregulares, y tiene diferentes periodos de referencia. Los estudios periódicos aparecen con periodos de acopio espaciados regularmente y tienen, por lo general, periodos de referencia similares. Los estudios periódicos continuos cubren todo el periodo de tiempo con períodos de referencia continuos, un ejemplo puede ser la Encuesta de Población Activa (EPA). Finalmente, podemos observar que los registros continuos cubren todo el periodo de tiempo, sin embargo habitualmente el acceso a registros administrativos no es a registros continuos sino a instancias o instantáneas de los registros en diversos momentos t fijados.

Habitualmente a todos ellos, excepto en el caso de los estudios singulares, se les ha denominado de un manera ambigua como **estudios longitudinales**. No obstante se ha utilizado el término estrictamente longitudinal para designar los diseños de tipo panel que se realizan sobre los mismo elementos,

[53] Kish, Leslie, and Centro de Investigaciones Sociológicas (España). *Diseño estadístico para la investigación*. Madrid: CIS : Siglo XXI, 1995.

diseños que Kish representa en el punto 5 del esquema 3 abajo representado. En ese sentido las instancias de los registros administrativos abrieron el camino para reducir los costes de los estudios estrictamente longitudinales tipo panel y hacerlos más viables técnicamente al reducir los casos de muerte o las negativas a responder de las encuestas panel.

Esquema 3.- Diseños de solapamientos



Por tanto, los registros administrativos ofrecen la posibilidad de crear cohortes de individuos para estudiar el cambio a lo largo del tiempo y para reunir información sobre las personas que han experimentado en su vida un evento determinado. Por ejemplo, quienes hayan vivido en una recesión o quienes hayan pasado por un sistema educativo.

A las necesidades generales de los investigadores han venido a sumarse las necesidades que tienen los planificadores sociales, los administradores públicos y los formuladores de políticas, de disponer de datos válidos y actuales sobre dominios pequeños y áreas locales. En esa dirección, otra de las cualidades de los registros administrativos es su capacidad de aportar datos para esos dominios, pero con la debilidad de no poder alinearlos con definiciones y conceptos internacionalmente aceptados.

Los datos administrativos pueden ser particularmente valiosos para la evaluación de las políticas y su análisis puede contribuir al desarrollo de las mismas. Su alta granularidad dominio-tiempo, que se traduce en su capacidad para facilitar análisis longitudinales y su capacidad para aportar información

en pequeños dominios, ha dado lugar a numerosos estudios para la evaluación de políticas sociales. Especialmente reseñable es la propiedad de los registros administrativos para el estudio de dominios pequeños, no sólo para el caso de datos locales, sino también para el análisis de poblaciones poco frecuentes o poblaciones raras. R. Connelly enumera en el artículo *The Role of Administrative Data in the Big Data Revolution in Social Science Research*[54] algunos ejemplos clarividentes sobre su uso en la investigación de políticas sociales.

Como hemos visto, por regla general, los datos censales están anticuados, los datos de registros administrativos pueden resultar inadecuados y los datos muestrales carecen de detalle para áreas pequeñas. **En la medida en que los puntos fuertes y débiles de estas fuentes de datos se complementan parece razonable tratar de combinar las ventajas de cada tipología para obtener estimaciones de dominios pequeños, pero también para reducir los costes y la carga de respuesta a las unidades informantes**.

En los países nórdicos existe una larga trayectoria en el uso de los registros administrativos para fines estadísticos, pero también en la combinación de encuestas y registros administrativos[55]. Las posibilidades de utilizar los registros administrativos en la producción de estadísticas oficiales fueron reconocidas por primera vez en Statistics Norway en 1969. Desde entonces, una parte cada vez mayor de la producción estadística total se ha basado en registros o en una combinación de registros y encuestas. A través de los años, la estrategia para un mayor uso de los registros ha tenido y aún tiene tres componentes principales:

1. A fin de aprovechar al máximo los registros administrativos en la producción de estadísticas oficiales, debe establecerse en el país un sistema de registros administrativos bien coordinados. Al mismo tiempo, debe establecerse un marco legislativo que gobierne el intercambio de información entre las agencias para asegurar la aceptación pública del sistema de registros y su uso. El objetivo de este sistema no es servir primordialmente a las necesidades estadísticas, sino hacer más eficaz la administración pública. Por lo tanto, el establecimiento de dicho sistema va mucho más allá de la responsabilidad de un organismo nacional de estadística, pero Statistics Norway ha desempeñado y sigue desempeñando un papel clave en esta evolución en Noruega. La Ley de Estadística asigna a Statistics Norway un papel de coordinación en relación con los registros públicos y administrativos.

[54] Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." *Social Science Research* 59 (September 2016): 1–12. doi:10.1016/j.ssresearch.2016.04.015.
[55] Thomsen and Holmøy, "Combining Data from Surveys and Administrative Record Systems. The Norwegian Experience."

2. Sea cual sea el éxito de Statistics Norway en influir en que los registros administrativos se conviertan en una herramienta poderosa en la producción estadística, siempre habrá problemas de calidad. Para muchos propósitos, incluyendo los estadísticos, es necesario tener conocimiento de la calidad de la información en un registro específico. Por supuesto, es responsabilidad del propietario del registro reunir esa información. Sin embargo, en muchos casos se identifican problemas de calidad al combinar un registro con otros registros o información estadística. Dado que esta combinación se realiza principalmente con fines estadísticos, Statistics Norway se encuentra en una posición especial para reunir y almacenar información relativa a la calidad de todo el sistema de registros administrativos.

3. La producción de estadísticas nunca puede ser la principal preocupación al establecer un sistema de registros administrativos. Por lo tanto, existe casi siempre la necesidad de transformar y/o combinar información de diversas fuentes, incluso estadísticas, antes de que pueda ser publicada como estadística oficial. En algunos casos esta transformación es simple, pero en otros casos puede llegar a ser bastante compleja.

2.1.3. Datos secundarios y de cuarta generación: Big Data

Tal como señalamos en el Capítulo 1, el advenimiento de la sociedad de la datificación en el Siglo XXI ha puesto a disposición de los estadísticos públicos nuevas fuentes de datos, conocidas como **fuentes Big Data**. Estas fuentes de datos genéricamente son **fuentes de eventos** y por lo tanto vienen a sumarse al valor que hasta ahora han aportado los registros administrativos en ese aspecto, incorporando como beneficio **la datificación de eventos que no tienen que estar vinculados a procedimientos administrativos de las Administraciones Públicas.** Bajo esa consideración, podríamos resumir en dos puntos la capacidad informativa de estas fuentes:

1. Capacidad de datificar en tiempo real actividades humanas de todo tipo.
2. Capacidad de ofrecer información con gran granularidad dominio-tiempo.

Estas características de las capacidades informativas de las fuentes Big Data nos pueden conducir a la fantasía de la **omnisciencia**, es decir, conocer todo lo que ha pasado, lo que ahora ocurre y lo que ocurrirá. Sin embargo estas fuentes no están ausentes de problemas metodológicos, sumando tanto los problemas de las encuestas como los de los registros administrativos, asociados a dos problemas de cobertura:

1. Cobertura conceptual: Las fuentes Big Data aportan datos que no han sido recopilados para fines estadísticos, con los consiguientes problemas derivados de la dificultad de obtener las variables requeridas en la estadística pública para el estudio de un fenómeno.

2. Cobertura poblacional: Las fuentes Big Data son muestras, con el problema añadido de ser muestras no probabilísticas.

La experiencia obtenida en las Oficinas Estadísticas en el uso de encuestas y de datos administrativos para fines estadísticos, y las similitudes entre las características específicas de estas fuentes y las fuentes Big Data, puede ayudar a mostrar el camino de cómo extraer información y cómo integrar las fuentes Big Data en las estadísticas oficiales. Un primer ejercicio en el que se comparan las características principales de las tres fuentes de datos lo encontramos en "*Will 'big data' transform Official Statistics"* (Florescu, 2014). Sobre dicho ejercicio encontramos una especificación más detallada en el artículo *"The Opportunities, Challenges and Risks of Big Data for Official Statistics"* (Kitchin, 2015) que detallamos a continuación:

Tabla 3.- Comparación de características según tipologías de fuentes de datos

	Encuestas y censos	**Registros administrativos**	**Big Data**
Especificación	Productos estadísticos definidos ex-ante	Productos estadísticos definidos ex-post	Productos estadísticos definidos ex-post
Propósito	Diseñados para fines estadísticos	Diseñados para otros propósitos	No diseñados, o diseñados para otros propósitos
Subproductos	Bajo potencial para generar subproductos	Gran potencial para generar subproductos	Gran potencial para generar subproductos
Métodos	Uso de métodos estadísticos clásicos	Uso de métodos estadísticos clásicos, dependiendo de los datos	Uso de métodos estadísticos clásicos generalmente no es posible
Estructura	Datos estructurados	Datos estructurados o semiestructurados	Datos semiestructurados o no estructurados
Comparabilidad	Comparabilidad buena entre países	Comparabilidad mala entre países	Comparabilidad buena entre países en fuentes internacionalizadas
Sesgos	Datos no sesgados	Datos posiblemente sesgados	Sesgo desconocido, posiblemente sesgados
Error	Típicos errores de muestreo y ajenos al muestreo	Errores no muestrales por ejemplo, datos ausentes, datos mal grabados o incoherentes, valores atípicos.	Incluyen tanto los típicos errores asociados a muestras como a los de registros administrativos. Nuevas fuentes de error
Persintencia	Persistentes	Posiblemente menos	Menor persistencia

		persistentes	
Volumen	Bajo volumen	Volumen medio	Alto volumen
Oportunidad	Lentas	Potencialmente rápidos	Potencialmente muy rápidos
Coste	Caras	Baratos	Potencialmente baratos
Carga a las unidades informantes	Alta carga de respuesta para las unidades informantes	No aumenta la carga de respuesta	No aumenta la carga de respuesta
Cobertura geográfica	Nacional	Nacional	Internacional
Cobertura demográfica	Población objetivo	Población sujeta a la gestón	Consumidores
Propiedad	Pública	Pública	Privada

Las características de las fuentes Big Data vienen determinadas por las características clásicas con las que éstas son conocidas (3Vs): Volumen, velocidad y variedad. Estas características se pueden estudiar desde una perspectiva tecnológica pero también desde un acercamiento estadístico.

1. **Volumen:** En términos estadísticos el volumen es una característica que se refiere no sólo a la acumulación de información de muchas unidades de observación o de eventos, sino también a la alta dimensionalidad de algunas fuentes Big Data. En términos tecnológicos, como vimos en el capítulo anterior, las Oficinas Estadísticas tienen experiencia en el tratamiento de grandes volúmenes de datos, pero para el tratamiento de las fuentes de cuarta generación posiblemente sea necesario evolucionar las capacidades de almacenamiento y computación de las oficinas.

2. **Velocidad:** La velocidad es una característica disruptiva respecto a la historia de la datificación oficial fundamentada en censo y encuestas. En su momento el acceso a datos administrativos mejoró la oportunidad de la información ofrecida por las Oficinas Estadísticas, sin embargo con las fuentes Big Data la estadística pública se enfrenta a la información en tiempo real.

   La posición de la estadística pública en un sistema de decisión se ha situado históricamente en los niveles analíticos o estratégicos y nunca en el operacional, que es donde se requiere información en tiempo real. Una incursión de la estadística oficial en el nivel operacional requeriría cambios tecnológicos y procedimentales.

Desde la perspectiva de los cambios tecnológicos, se necesitarían evoluciones en las arquitecturas de datos de las Oficinas de Estadística. En la actualidad estas arquitecturas están dirigidas al procesamiento supervisado de datos por lotes, en cambio las arquitecturas Big Data (como la Kappa o la Lambda) están dirigidas al procesamiento automático de datos en modo batch o tiempo real. En ese sentido, el paso al procesamiento automático conlleva cambios procedimentales importantes en muchos de los procesos definidos en el Generic Statistical Business Process Model (GSBPM).

3. **Variedad:** Otra de los retos que introducen las fuentes Big Data son la variedad de formatos y tipos de información disponible. Tradicionalmente la estadística pública ha extraído sus datos de formularios estructurados, sin embargo en las fuentes Big Data nos encontramos con recursos informativos de todo tipo, asociados a formatos de imagen, voz y texto. Estos nuevos formatos aportan mucha riqueza informativa, pero su tratamiento requiere de nuevos procedimientos y tecnologías en las Oficinas Estadísticas.

Como hemo visto las fuentes Big Data tienen importantes similitudes con las fuentes administrativas respecto a su capacidad informativa, incluso mejoran dicha capacidad. Estas características las convierten en fuentes atractivas para su inclusión en la parrilla de inputs de la estadística oficial. Sin embargo las fuentes Big Data, tal como se señala en la tabla 3, aumentan los problemas metodológicos respecto a las fuentes administrativas:

1. Fuentes posiblemente sesgadas y con dificultades para conocer el sesgo
2. Incluyen tanto errores ajenos al muestreo como nuevas fuentes de error
3. La falta de persistencia dificulta la comparabilidad temporal
4. El uso de métodos estadísticos tradicionales no siempre es posible

Estos problemas plantean cuestiones importantes relativas a la idoneidad de las fuentes Big Data para su uso en las estadísticas oficiales, y en ese sentido algunos autores proponen que inicialmente su uso se circunscriba exclusivamente a la mejora de la calidad de las estimaciones dentro de los marcos metodológicos actuales, mientras que estudian los niveles y causas de los errores de muestreo y no muestreo de estas fuentes. En el resto del capítulo revisaremos con mayor detalle los problemas metodológicos enumerados y realizaremos un acercamiento a estrategias metodológicas para el abordaje de su solución.

2.2. Las fuentes de error en las fuentes Big Data

Algunas de las preocupaciones de los estadísticos respecto al proceso de extracción de información a partir de los datos las señalan Kreuter y Peng, en su capítulo *Extracting Information from Big Data: Issues of Measurement, Inference and Linkage* del libro *Privacy, Big Data, and the Public Good: Frameworks for Engagement*[56]. Estas preocupaciones están ligadas por una parte a las cuestiones de la medición y por otra a las de inferencia, preocupaciones que señalamos a continuación:

1. Respecto a las cuestiones de medición son diversas las preguntas que se realiza el científico social: ¿Están recogidas en los datos todas las variables clave y las covariables necesarias para responder a las preguntas de una investigación? ¿Hay variables no observadas o no controladas que pueden confundir o perturbar un determinado análisis? ¿Hay errores sistemáticos de medición? ¿Hay falta de respuesta sistemática en alguna variable? ¿Cómo afecta el instrumento de medida en la medición? ¿Hay coherencia entre las mediciones?

2. Por cuestiones de inferencia nos referimos a: ¿Entendemos el proceso de muestreo? ¿Están todas las unidades de observación que necesitamos en el análisis? ¿Hay exclusiones sistemáticas de unidades de observación? ¿Por qué algunas unidades de observación aparecen varias veces? ¿Tenemos todas las medidas que necesitamos en todas las unidades de observación? ¿A quiénes y a cuántos representan las unidades?

Para las fuentes de datos que están diseñadas para responder a las preguntas específicas de una investigación científica, las respuestas a las cuestiones anteriores son relativamente fáciles de determinar, o mejor aún, se tienen en cuenta a la hora de la captura de los datos. Pero cuando los datos se utilizan para fines distintos a los que fueron recopilados, tal como sucede con las fuentes Big Data, las respuestas no son tan triviales. Las dos preocupaciones señaladas están ligadas directamente a los problemas de cobertura que enumeramos en el apartado anterior como características de las fuentes Big Data:

1. Asociado a la preocupación de la medición nos encontramos con el problema de **cobertura conceptual**: Las fuentes Big Data aportan datos que no han sido recopilados para fines

[56] Lane, Julia I., ed. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY: Cambridge University Press, 2014.

estadísticos, con los consiguientes problemas derivados de la dificultad de obtener las variables requeridas en la estadística pública para el estudio de un fenómeno.

2. Asociado a la preocupación por la inferencia nos encontramos con el problema de la **cobertura poblacional**: Las fuentes Big Data son muestras, con el problema añadido de ser muestras no probabilísticas, con coberturas poblacionales potencialmente sesgadas al propósito de la creación de la fuente.

Estos dos problemas de cobertura siempre han sido de interés para los estadísticos pues son problemas clave en las desviaciones entre el valor poblacional que se desea estimar y el valor que medimos a través de nuestro instrumento de medición. Este error es conocido en la estadística como **sesgo**, y lo podemos encontrar en todas las tipologías de fuentes de datos que revisamos en el apartado anterior. A su vez cuando las fuentes son muestras de la población nos encontramos con otro componente de error que es la **varianza**, este componente mide las desviaciones que obtenemos como consecuencia de usar una muestra entre todas las muestras posibles de una población dada. En el contexto del estudio de fuentes Big Data es importante señalar que cuando la muestra aumenta evidentemente la varianza disminuye, de tal manera que para una muestra que recopila datos de toda la población -o sea, un censo- la varianza es 0.

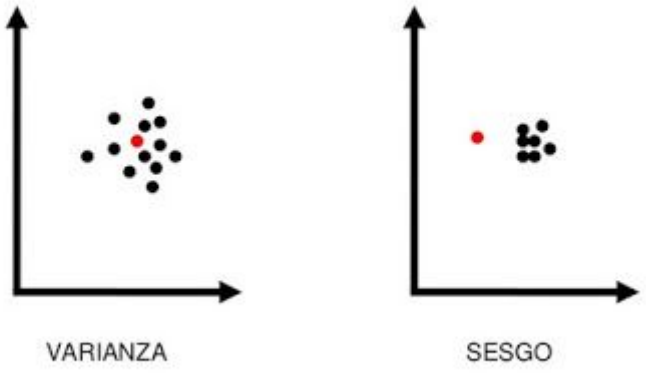
2.2.1. Sesgo y varianza en muestras no probabilísticas

Por lo tanto nos encontramos con dos componentes importantes a la hora de medir la bondad de la estimación de un dato. Por una parte tenemos el sesgo y por otra la varianza. Estas dos fuentes de error se sintetizan en el conocido error cuadrático medio que se representa como:

$$ECM(\widehat{\theta}) = B^2(\widehat{\theta}) + Var(\widehat{\theta})$$

Donde *b* representa el sesgo y *v* la varianza. En la representación gráfica del esquema 4 se visualizan mejor estas dos componentes del error de la estimación un valor. En color rojo representamos el valor real y no conocido, o sea el valor que queremos estimar a través de una muestra de datos, y en negro las estimaciones obtenidas por muestras diferentes entre todas las posibles muestras de una población. A la izquierda está representado un procedimiento de selección de muestras por el que obtenemos una estimaciones insesgadas pero con mucha variabilidad, en el gráfico de la derecha está representado un procedimiento que genera estimaciones muy concentradas, pero lejanas del valor objetivo, es lo que se conoce como estimaciones sesgadas.

Esquema 4.- Componentes del error cuadrático medio



Estas componentes del error cuadrático medio se alimentan por diversas de fuentes de error que en la literatura clásica se clasifican como **errores de muestreo** y **errores ajenos al muestreo**.

Los errores de muestreo son errores inherentes al diseño muestral, siendo consecuencia tanto del **método de selección de la muestra** como del **procedimiento de inferencia** utilizado para pasar de la muestra a la población. Ambos elementos, selección e inferencia, influyen en el sesgo y en la varianza. La situación ideal es que el investigador pueda controlar ambos procedimientos para minimizar las fuentes de error muestral, sin embargo debido a diferentes causas esta situación no siempre es posible en el caso de la selección de las unidades muestrales de observación.

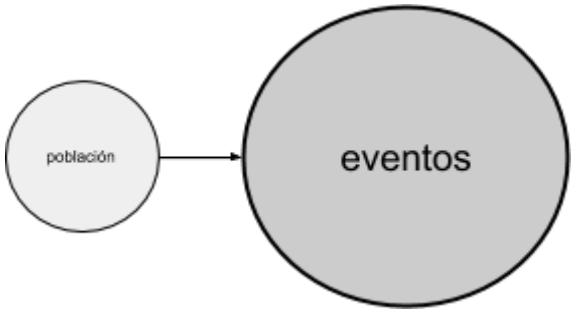
El método de selección de la unidades de observación determina la obtención de muestras probabilísticas o no probabilísticas, si a partir del mismo se puede calcular la probabilidad de selección de una unidad de observación. En el caso de muestreos polietápicos nos podemos encontrar con muestras semiprobabilísticas, cuando solo es posible determinar la probabilidades de selección de la unidades de algunas de las etapas.

Obtener muestras probabilísticas siempre ha sido un elemento clave en el diseño de muestras, pues la probabilidad de selección de una unidad de observación es la que marca su peso en la inferencia, y **a través de la misma se corrigen los sesgos debido a representaciones no equiprobables de ciertas unidades de observación**. Por lo tanto, las muestras no probabilísticas se enfrentan a problemas más complejos para asegurar la calidad de sus estimaciones; problemas y soluciones que se compendian

excelentemente en el documento de la American Association for Public Opinion Research sobre este particular[57].

En el caso que nos ocupa encontramos que una buena parte de las fuentes Big Data no son censos del fenómeno estudiado sino muestras del mismo. Asimismo, genéricamente los métodos de captura de los datos no permiten determinar la probabilidad de selección de las unidades muestrales, encontrándonos por tanto ante muestras no probabilísticas. Finalmente también resulta preciso señalar que siendo las fuentes Big Data habitualmente fuentes de eventos, nos enfrentamos a su vez a los problemas de las muestras polietápicas sobre determinación de varias probabilidades.

Esquema 5.- Las fuentes Big Data habitualmente son muestras polietápicas no probabilísticas



De acuerdo con el esquema anterior, la probabilidad de que un evento sea recogido en una fuente Big Data viene determinada por la probabilidad de que una unidad de análisis potencialmente generadora de eventos (unidades de primera etapa) forme parte de la fuente y en segundo lugar por la probabilidad de que esa unidad de primera etapa genere un evento (unidades de segunda etapa). Por ejemplo, la probabilidad de un tuit en Twitter depende en primer lugar de que una persona esté en la red social y en segundo lugar de que esa persona tuitee ese tuit; o sea, nos encontramos ante un probabilidad condicionada difícilmente calculable:

$$P(tuit) = P(tuitero)P(tuit|tuitero)$$

Por lo tanto en estos casos nos tropezamos con los problemas derivados de los **sesgos por autoselección**. En las estadísticas, el sesgo de autoselección surge en cualquier situación en la que las unidades de observación se seleccionan a sí mismas para formar parte de la muestra, dando lugar a una muestra sesgada con muestreo no probabilístico. En definitiva, con las **fuentes Big Data nos**

[57] Baker, Reg, Brick, J. Michael, Bates, Nancy A., Couper, Mick P., Dever, Jill A., and Tourangeau, Roger. "Report of the AAPOR Task Force on Non-Probability Sampling." AAPOR, June 2013.

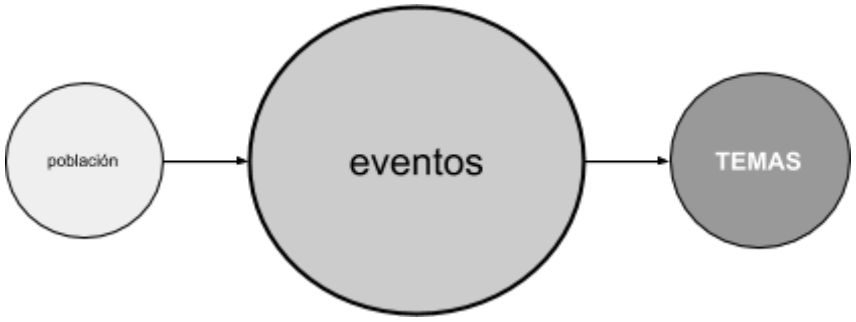
enfrentamos a muestras politetápicas no probabilísticas con las consiguientes dificultades para corregir los problemas de sesgos. En la sección sobre los problemas de inferencia trataremos este asunto con más detenimiento y presentaremos estrategias para abordar dichos problemas.

2.2.2. Cobertura de temas en fuentes Big Data de redes sociales

Como señalan muchos autores, los usuarios que publican contenido en redes sociales pueden ser atípicos en sus datos demográficos, pudiendo ser sus opiniones y comportamientos no representativos de toda la población (Boyd y Crawford, 2012; Baker et al. 2013; Couper 2013; Smith 2013). En otras palabras, las inferencias de los datos de redes sociales son intrínsecamente de calidad desconocida porque provienen de muestras no probabilísticas que no están diseñadas para cubrir la población.

Otros autores han observado que aunque los usuarios de redes sociales no cubren la población en estudio, sin embargo el análisis de su contenido predice razonablemente bien las estimaciones realizadas por encuestas (O'Connor et al., 2010; Tumasjan et al., 2010). En definitiva su planteamiento se fundamenta en que si bien en redes sociales no hay asegurada una buena cobertura poblacional, sí que existe una buena cobertura de temas de estudio asociados a los eventos de la población cubierta por la red social.

Esquema 6.- Cobertura de temas a partir de eventos en redes sociales



Una forma de expresar esto es decir que las propiedades de los datos de redes sociales, no sólo su magnitud, pueden anular su origen no probabilístico. Por lo tanto, los datos de redes sociales podrían terminar cubriendo adecuadamente los temas investigados, a pesar de que los individuos que contribuyeron al corpus de las redes no son muestreados de manera representativa.

Los escépticos sin embargo han considerado que las afirmaciones de los entusiastas son excesivamente optimistas (Couper 2013, Langer Research Associates, 2013, Smith, 2013). Además,

otros autores han observado que los datos de redes sociales pueden introducir nuevos tipos de sesgo y errores de medición (Biemer 2014, Tufekci 2014), lo que plantea la cuestión de si son suficientemente fiables para ser utilizados en las estadísticas oficiales o en las investigaciones sociales.

Empíricamente, las comparativas entre los análisis de contenido de las redes sociales con los resultados ofrecidos por encuestas indican que éstos no estiman con exactitud los fenómenos sociales. Algunos intentos de estimación, incluso en dominios que han mostrado previamente buena relación entre el análisis de contenido y las encuestas, no han tenido éxito (Kim et al., 2014), e incluso un estudio que suele citarse como un gran éxito (O'Connor et al., 2010) también se ha utilizado para discutir las limitaciones de los métodos (Smith, 2013). Así por ejemplo, las debilidades de los procedimientos dan lugar a que decisiones aparentemente menores, como incluir o no los tweets que mencionan a un pequeño partido político en un estudio de intención de voto, conduzcan a resultados muy diferentes (Jungherr, Jürgens y Schoen 2012).

Algunos autores (Schober et al., 2016) consideran que la distinción entre cobertura poblacional y cobertura temática es el elemento clave del potencial alineamiento entre las conclusiones de los datos de las encuestas y los análisis de las redes sociales. Para estos analistas, una adecuada cobertura temática puede lograrse sin una correcta cobertura poblacional, como consecuencia de los mecanismos de propagación de la información que caracterizan las dinámicas de las redes. Según estos autores, comprender cuándo se alcanza una buena cobertura temática, ya sea mediante una adecuada cobertura poblacional o no, es un problema científico central.

Estos acercamientos parten de la ausencia de intencionalidad en la generación de temas, tales como campañas a través de los llamados *influencers*[58] o las manipulaciones por parte de la red social, tal como presentó Facebook en un artículo académico en 2014 con los detalles sobre cómo había modificado con éxito los estados de ánimo de centenares de millares de usuarios a través de la manipulación del suministro de noticias, entradas y comentarios visibles para el individuo[59].

Por tanto las conclusiones generales al respecto (Schober et al., 2016) indican que en la actualidad no se disponen desarrollos metodológicos sólidos para conseguir resultados coherentes entre los análisis de contenido en redes sociales y las encuestas de opinión. De hecho algunos autores (Gayo-Avello, 2013) concluyen de manera pesimista indicando que lamentablemente, a medida que avanzamos en la lista de requisitos para equiparar las estimaciones basadas en redes sociales con las encuestas, las

[58] Un *influencer* es una persona que cuenta con cierta credibilidad sobre un tema concreto, y por su presencia e influencia en redes sociales puede llegar a convertirse en un prescriptor sobre una marca o un asunto determinado.
[59] Kramer, Guillory, and Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks."

tareas son cada vez más difíciles e incluso inalcanzables. Como señala Gayo-Arvello respecto al uso de Twitter en las predicciones electorales: *"Mejorar el análisis del sentimiento con respecto a los tweets políticos será difícil pero probablemente alcanzable. Combatir el sesgo demográfico es al menos concebible en principio. Sin embargo, explicar el sesgo de autoselección podría ser inviable, al menos, usando únicamente los datos de Twitter."*

En la estadística pública encontramos algunos estudios sobre el uso de los contenidos en redes sociales. El más conocido, quizás por ser de los primeros publicados, es el realizado por Statistics Netherlands sobre el uso de datos de redes sociales para la estimación del Índice de Confianza del Consumidor[60] [61]. En esta investigación se estudió la relación entre los cambios en la confianza de los consumidores holandeses y los mensajes de las redes sociales públicas holandesas, relevando una fuerte asociación entre la confianza del consumidor y el sentimiento en los mensajes públicos de Facebook. A su vez la inclusión del sentimiento de mensajes de Twitter aumentó dicha asociación. Los hallazgos descritos en el documento son consistentes con la noción de que los cambios en la confianza del consumidor y el sentimiento de las redes sociales se ven afectados por un fenómeno subyacente idéntico.

Sin embargo, algunos autores (Schober et al., 2016) sugieren que es prematuro aprobar o rechazar totalmente la idea de que los análisis de las redes sociales podrían reemplazar a las encuestas en la producción de estadísticas oficiales. Algunos proponen reemplazar encuestas con datos de redes sociales cuando haya un patrón externo de referencia (por ejemplo, reclamaciones de seguro de desempleo[62], como en los análisis de Antenucci et al. [2015]). En todo caso, si el patrón de referencia es una encuesta, entonces tendría que haber por lo menos encuestas ocasionales para calibrar las tendencias de las redes sociales. Se podría pensar, por ejemplo, en recopilar datos de encuestas bimestrales en lugar de mensuales y realizar estimaciones con los datos de redes sociales en los meses sin recogida de datos; pero sólo si con el tiempo las tendencias de las redes sociales demuestran una asociación suficiente con los resultados de la encuesta.

2.2.3. El tamaño como fuente de error

Una de las características importantes de las fuentes Big Data es su alto volumen de información, que se concreta en una alta dimensionalidad y en una gran cantidad de unidades de análisis. Evidentemente

[60] van den Brakel, Jan et al., "Social Media as a Data Source for Official Statistics; the Dutch Consumer Confidence Index."
[61] Daas, Puts, and European Central Bank, *Social Media Sentiment and Consumer Confidence.*
[62] Antenucci et al., "Using Social Media to Measure Labor Market Flows."

estos tamaños sí importan, en este apartado vamos a revisar cómo éste afecta a la calidad de la información.

Anteriormente señalamos que el tamaño en unidades de análisis es inversamente proporcional a la varianza, siendo igual a cero cuando la muestra coincide con la población. **Pero el aumento de tamaño no está ausente de problemas, pues nos arriesgamos a incrementar los sesgos producto de los errores ajenos al muestreo.** Los errores ajenos al muestreo son producto del proceso de recopilación de los datos, y por tanto al aumentar el número de unidades nos exponemos a aumentarlos.

Los errores ajenos al muestreo suelen clasificarse en errores de especificación, errores de no respuesta, errores de cobertura, errores de medida, errores de comunicación y errores de procesamiento. Veamos cada uno de ellos, y algunos ejemplos de cómo las fuentes Big Data están sometidas a los mismos:

1. **Error de especificación:** Un error de especificación surge cuando el concepto implícito por la pregunta de la encuesta difiere del concepto que debería haber sido medido en la encuesta. El error de especificación es a menudo causado por la mala comunicación entre el investigador (o experto en la materia) y el diseñador del cuestionario.

   Los errores de especificación son errores clásicos en las estadísticas basadas en fuentes secundarias como es el caso de las fuentes Big Data, y son hermanos de los problemas de cobertura conceptual de estas fuentes.

2. **Error de no respuesta.** Este error se produce cuando no se consigue obtener los datos de todas las unidades, siendo de dos clases: Unidad de no respuesta, cuando se pierde la información para todas las variables o ítem de no-respuesta, cuando se pierde al menos una, pero no todas, las variables.

   En una fuente Big Data por ejemplo puede ser como consecuencia de la ausencia de datos de algún sensor o antena, que haya dejado de recopilar información por su avería o falta de suministro eléctrico. Por lo tanto el aumento de sensores, o el aumento de la información recogida por los mismos, nos expone a un aumento de la no respuesta.

2. **Error de cobertura o de marco.** Este problema surge cuando algunas unidades de la población en estudio son excluidas o incluidas erróneamente en el conjunto de unidades de análisis.

   Por ejemplo, en un estudio sobre el turismo, la inclusión o exclusión errónea de unidades poblacionales a partir de fuentes como las tarjetas de crédito o la telefonía móvil es bastante probable.

3. **Error de medida.** Este tipo de error se refiere a las inexactitudes que aparecen debido a un mal instrumento de medida. Por ejemplo una mala calibración del cualquier instrumento de medida automática es un origen clásico de errores de medida en fuentes Big Data.

4. **Error de comunicación.** Estos se producen como consecuencia de problemas en la interacción entre la unidad sobre la que se recoge información y las unidades recopiladoras de la misma. Los problemas de comunicación surgen por motivos diversos y se traducen en transmisiones erróneas de datos de forma deliberada o no.

   Las fuentes Big Data no están ausentes de este tipo de problema. Por ejemplo la declaración de ciertos estados, como el estado civil, la edad o el nivel de estudios, así como las declaraciones religiosas y políticas suelen ser origen de error en fuentes Big Data asociadas a las redes sociales. Por otra parte, la pérdida de datos en la comunicación con sensores es una fuente de error más que conocida.

5. **Error de procesamiento.** El error de procesamiento de datos incluye errores en la edición, grabación, codificación, asignación de pesos de la encuesta y tabulación de los datos de la encuesta. Como veremos más adelante, en las fuentes Big Data una importante fuente de error de procesamiento será el método de inferencia utilizado para intentar corregir los problemas de autoselección, o para obtener información válida mediante métodos predictivos.

Por otra parte, Fan et al. (2014) identifican tres tipos de problemas asociados al tamaño masivo y la alta dimensionalidad de las fuentes Big Data: a) acumulación de ruido, b) correlaciones espurias y c) endogeneidad incidental. Los errores ajenos al muestreo señalados anteriormente sólo aumentarán estos problemas.

a) Acumulación de ruidos

Para ilustrar la acumulación de ruido, supongamos que un analista está interesado en clasificar a los individuos en dos categorías -C1 y C2- sobre la base de los valores de 1000 características en una fuente Big Data. Supongamos además que, desconocido para el investigador, el valor medio para las personas en C1 es 0 en todas las 1000 características mientras que las personas en C2 tienen una media de 3 en las primeras 10 características y un valor de 0 en las otras 990 características. Una regla de clasificación basada en las primeras $m \leq 10$ características funciona bastante bien con poco error de clasificación. Sin embargo, a medida que se incluyen más y más características en la regla, el error de clasificación aumenta porque las características no informativas (es decir, las características 990 que no tienen poder de discriminación) eventualmente abruman las señales informativas (es decir, las primeras 10 características). En Fan et al. (2014), cuando $m > 200$, el ruido acumulado excede la señal incorporada en las primeras 10 características y la regla de clasificación se convierte en una regla de clasificación equivalente a un lanzamiento de monedas.

b) Correlaciones espurias

La alta dimensionalidad también puede introducir correlaciones espurias. como consecuencia de que muchas características no relacionadas pueden estar altamente correlacionadas simplemente por casualidad, lo que produce falsos descubrimientos e inferencias erróneas. Por ejemplo, utilizando poblaciones simuladas y tamaños de muestra relativamente pequeños, Fan et al. (2014) muestran que con 800 características independientes, el analista tiene una probabilidad del 50% de observar una correlación absoluta que excede 0.4. Sus resultados sugieren que hay riesgos considerables de falsas inferencias cuando se usan datos de alta dimensionalidad.

c) Endogeneidad incidencial

Finalmente, un supuesto clave en el análisis de regresión es que las covariables del modelo no están correlacionadas con el error residual. La endogeneidad se refiere a una violación de esta suposición. Para los modelos de alta dimensionalidad, esto puede ocurrir puramente por casualidad, un fenómeno que Fan y Liao (2012) llaman "endogeneidad incidental". Los riesgos de endogeneidad incidental aumentan a medida que aumenta el número de variables en el proceso de selección del modelo. Por lo tanto, es una preocupación particularmente importante para los analistas de fuentes Big Data.

Diversos autores[63] sugieren métodos estadísticos robustos dirigidos a mitigar los riesgos señalados anteriormente. Sin embargo, como indicamos, estos problemas se agravan aún más cuando se introducen errores ajenos al muestreo. Biemer y Trewin (2012) muestran que los errores ajenos al

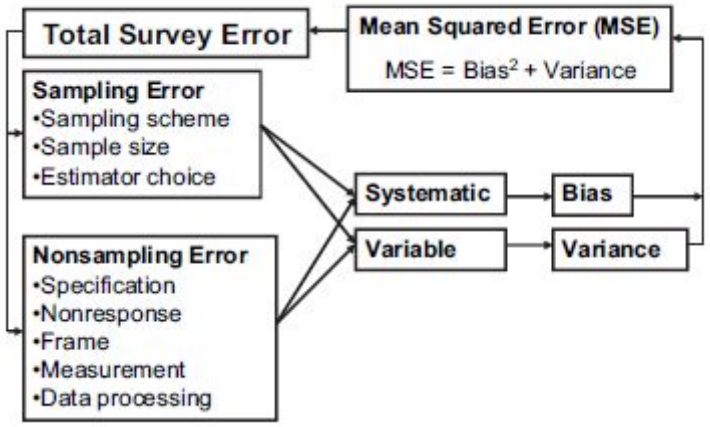
[63] Ver por ejemplo Fan et al. 2014, Stock y Watson 2002, Fan y al. 2009, Hall y Miller 2009, Fan y Liao 2012

muestreo sesgan los resultados e inflan la varianza de las estimaciones, de tal manera que son difíciles de evaluar o mitigar en el proceso de análisis de datos. Así, el alto volumen y la alta dimensionalidad de las fuentes Big Data combinada con los riesgos de errores ajenos al muestreo requieren enfoques nuevos y robustos para el análisis de dichos datos.

2.2.4. Total Error Framework para fuentes Big Data

La gestión de los riesgos que los errores pueden introducir en el análisis de Big Data se facilita a través de una mejor comprensión de las fuentes y la naturaleza de los mismos. Esta comprensión se obtiene a través de un conocimiento profundo del mecanismo generador de datos, su procesamiento y los métodos utilizados para crear un conjunto de datos específicos o las estimaciones derivadas de los mismos. Para los datos generados mediante encuestas, este conocimiento se recoge en un marco denominado en inglés "Total Survey Error (TSE)" que identifica todas las fuentes principales de error que contribuyen a la validez de los datos y la exactitud del estimador (ver, por ejemplo, Biemer 2010).

Esquema 7.- Total Survey Error, sus componentes y el Error Cuadrático Medio



El marco TSE sintentiza las fuentes de error que hemos revisado en los apartados anteriores. Aunque en la literatura estadística se han propuesto varias métricas aceptables para cuantificar la TSE, la métrica más común para el trabajo de encuestas es el Error Cuadrático Medio (ECM) -MSE en sus siglas en inglés-. Cada estimación calculada a partir de los datos de la encuesta tiene un ECM que resume los efectos de todas las fuentes de error en la estimación. Un ECM pequeño indica que el TSE es pequeño y bajo control; sin embargo un ECM grande indica que una o más fuentes de error están afectando adversamente la exactitud de la estimación.

Desafortunadamente, rara vez es posible calcular el ECM en situaciones prácticas porque esto generalmente requiere una estimación del parámetro estimado que esté libre de errores. Aún así, el concepto es bastante útil para entender cómo los efectos combinados de los errores de la encuesta reducen la precisión de la estimación. Además, los diseñadores de encuestas pueden beneficiarse del conocimiento de estos conceptos a través de una mejor comprensión de cómo sus decisiones de diseño afectan la calidad general de los datos de la encuesta.

En términos estadísticos, el ECM es la diferencia cuadrada esperada entre una estimación, $\widehat{\theta}$, y el parámetro que se pretende estimar θ, que puede escribirse como:

$$ECM(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2$$

O tal como vimos en una sección anterior de este capítulo, se puede descomponer en términos de sesgo y varianza

$$ECM(\widehat{\theta}) = B^2(\widehat{\theta}) + Var(\widehat{\theta})$$

Como se muestra en el esquema 7, el ECM refleja los efectos acumulativos de las diversas fuentes de error, tanto de muestreo como de no muestreo, en la estimación de la encuesta. Cada origen de error puede contribuir con un error aleatorio, un error sistemático o una combinación de ambos tipos. Los errores aleatorios repercuten en la varianza de la estimación, mientras que los errores sistemáticos afectan al sesgo.

Supongamos que queremos estimar una característica y de una unidad análisis en una encuesta. Los errores pueden hacer que el valor observado de y sea mayor o menor que su verdadero valor para un individuo. Matemáticamente, esto puede escribirse como:

$$y_i = \mu_i + \epsilon_i$$

Donde y_i y μ_i son los valores reales y observados para la unidad de análisis i, mientras que ϵ_i es el efecto de la acumulación de errores.

Este efecto puede ser positivo para algunas unidades de análisis y negativo para otras, de tal forma que si en el conjunto de la muestra el efecto es 0 entonces tenemos que el valor estimado $\widehat{\theta}$ será próximo al valor de θ, al margen de los correspondientes errores muestrales. Por ejemplo, supongamos que θ es

la media poblacional, que para una muestra aleatoria simple se estima por la media de la muestra y representada como $\overline{y}$. Si $E(\epsilon_i) = 0$ entonces $E(\overline{y}) = \mu$, o sea la esperanza de la media muestral coincide con la media poblacional, y por lo tanto se dice que $\overline{y}$ es insesgado para μ. En ese sentido los ϵ_i que cumplen esa condición, se denominan **errores aleatorios o variables** pues añaden variación a las observaciones pero no le suman sesgo. En otras situaciones los errores pueden ser **sistemáticos** pues la suma de los errores a través de la muestra tienen dominancia positiva o negativa.

Tal como se representa en el esquema 7, los errores variables y sistemáticos pueden ser alimentados tanto por parte de los errores muestrales (método de selección, tamaño de la muestra y selección de método de inferencia) como de los errores ajenos al muestreo (que ya relacionamos en apartados anteriores).

En el caso de las encuestas, el marco TSE brinda información útil sobre la manera en que los diversos procesos de generación y tratamiento de datos afectan a la estimación, siendo útiles para sugerir métodos para producir inferencias de mayor calidad. En esa dirección alguno autores como Biemer (2014) y Japec et al (2015) creen que se necesita un **marco de error total para Big Data**[64][65].

Sin embargo, el **marco Big Data Total Error (BDTE)** necesariamente deberá incluir tipos de errores adicionales que son exclusivos de las fuentes Big Data y que pueden crear sesgos e incertidumbres sustanciales en las estimaciones. Al igual que el marco de TSE, el marco BDTE ayudará a comprender las limitaciones de los datos, lo que conducirá a un mejor análisis y uso de las estimaciones. A su vez el marco también puede informar sobre líneas de investigación para reducir los efectos de los errores en el análisis de Big Data.

2.2.5. Google Flu Trends: un ejemplo paradigmático de nuevas fuentes de error

Las fuentes Big Data traen consigo nuevas oportunidades, pero también nuevos retos para los analistas de datos en materia de control y corrección de los errores de estas fuentes. Un ejemplo bien conocido de los problemas de las fuentes Big Data es el proporcionado por la serie de repercusión de la gripe elaborada por Google[66] a partir de las búsquedas sobre los síntomas, remedios y otras palabras clave

[64] Biemer, P. P., "Toward a Total Error Framework for Big Data. Presentation in American Association for Public Opinion Research (AAPOR) 69th Annual Conference."

[65] Japec et al., "Big Data in Survey Research."

[66] https://www.google.org/flutrends/about/

relacionadas con la enfermedad. Esta serie proporciona estimaciones, casi en tiempo real, de la repercusión de la gripe en los EEUU y en otros veinticuatro países del mundo.

En comparación con los datos ofrecidos por Centers for Disease Control and Prevention (CDC)[67] asociado al US Department of Health & Human Services, Google Flu Trends proporcionó indicadores muy precisos de la incidencia de la gripe en los Estados Unidos entre 2009 y 2011. Sin embargo, para la temporada de gripe 2012-2013, Google Flu Trends predijo más del doble de la proporción de visitas médicas por gripe contabilizadas por CDC.

Butler (2013)[68] y Lazer et al. (2014)[69] citan varias causas de este error: **(a) La arrogancia Big Data, (b) la modificación del comportamiento de usuarios en la generación de eventos y (c) la dinámica de algoritmos**.

El primero ocurre cuando el analista de fuentes de Big Data cree que el volumen de los datos compensa cualquiera de sus deficiencias, obviando así la necesidad de enfoques analíticos tradicionales y científicos. Como señalan Lazer et al. (2014: 2), la arrogancia Big Data no reconoce que *"(...) la cantidad de datos no significa que uno pueda ignorar los problemas fundamentales de la medición, construcción de validez, confiabilidad y dependencias entre los datos (...)"*.

Las estimaciones de Google Flu Trends a partir de julio de 2012 fueron demasiado altas para 100 de 108 semanas. Sin embargo esta no fue la primera vez en la que había ocurrido tal circunstancia. En 2009, Google Flu Trends tuvo que ajustar sus algoritmos después de que sus modelos subestimaran el número de pacientes en el inicio de la pandemia de la gripe H1N1 (gripe porcina) en los Estados Unidos, un error que fue atribuido a **cambios en el comportamiento** de búsqueda de las personas como resultado de la naturaleza excepcional de la pandemia[70].

Si bien es difícil determinar con precisión lo que causó el cambio en el comportamiento de las búsquedas relacionadas con la gripe, hay varias explicaciones posibles de por qué el modelo original de Google Flu Trends subestimó la actividad durante la pandemia pH1N1. En primer lugar, los usuarios estaban realizando menos consultas relacionadas con las complicaciones de la gripe, tales como bronquitis y neumonía. En segundo lugar, el virus pH1N1 surgió durante los meses de primavera y verano, en lugar de los meses de otoño e invierno típicos de la gripe estacional. Además,

[67] https://www.cdc.gov/
[68] Butler, "When Google Got Flu Wrong."
[69] Lazer et al., "The Parable of Google Flu."
[70] Cook et al., "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic."

las personas pueden buscar diferentes términos de consulta cuando se está enfermo con gripe en el invierno en comparación con el verano. Consultas tales como "gripe porcina" fueron populares durante la pandemia pH1N1 y probablemente representaron algunos de los cambios en el comportamiento de búsqueda. Sin embargo, este tipo de consultas específicas de la pandemia no se incluyeron en los modelos de Google Flu Trends pues no se correlacionaban bien con los datos del CDC en temporadas anteriores, ni necesariamente se esperaba que se correlacionaran con la actividad futura de la gripe estacional o no estacional.

Respecto a los problemas de estimación a partir de julio de 2012, varios investigadores sugieren que los problemas pudieron deberse a una amplia cobertura mediática durante la temporada de gripe de Estados Unidos en ese año, incluyendo la declaración de una emergencia de salud pública por parte del Estado de Nueva York. En ese sentido, la cobertura mediática también pudo haber afectado el comportamiento de búsqueda, disparado las relacionadas con la gripe por parte de personas que no estaban enfermas.

La tercera fuente de error, señalada por Lazer et al. (2014), es **la dinámica de algoritmos** que se produce cuando el motor de generación de datos se modifica de tal manera que los términos de búsqueda anteriormente altamente predictivos pierden su capacidad predictiva. Por ejemplo, cuando un usuario de Google buscaba "fiebre" o "tos", los otros programas de Google comenzaron a recomendar búsquedas de síntomas y tratamientos de la gripe. Por lo tanto, las búsquedas relacionadas con la gripe aumentaron artificialmente o de forma inducida. En la investigación mediante encuestas esta situación es similar al sesgo inducido por los entrevistadores, que sugieren a los encuestados que están tosiendo que podrían tener la gripe, y luego preguntarles si piensan que pueden tener gripe.

La dinámica de algoritmos no se limita a Google. Plataformas como Twitter y Facebook también están modificando con frecuencia sus algoritmos para mejorar la experiencia de los usuarios. Todas estas plataformas cambian sus algoritmos más o menos frecuentemente, con resultados ambiguos para cualquier tipo de estudio a largo plazo. Una lección clave proporcionada por Google Flu Trends es que los análisis exitosos de hoy pueden no producir buenos resultados mañana.

2.3. El problema de la inferencia

Tal como hemos visto en las secciones anteriores, las fuentes Big Data suelen ser muestras no probabilísticas extraídas de una población objetivo. La población objetivo se define como la colección

completa de unidades de observación, según las características especificadas en el diseño de la encuesta.

En una muestra probabilística una unidad de observación se vincula con la población objetivo por la probabilidad de que ésta sea seleccionada en la muestra. Con ello tenemos que, bajo el paradigma de la inferencia basada en diseños, estas probabilidades de selección permiten calcular estimaciones para la población objetivo. Sin embargo, en muestras no probabilísticas la ausencia de enlace entre muestra y población impide la inferencia basada en diseños. Esto ha dado lugar a que algunos investigadores mantengan la imposibilidad de estudiar las propiedades estadísticas de las estimaciones basadas en muestras no probabilísticas (Biemer y Lyberg 2003, Sección 9.2). Estas preocupaciones son válidas dentro del paradigma basado en el diseño. Sin embargo, no necesariamente implica que todas las inferencias estadísticas basadas en muestras no probabilísticas sean imposibles.

Como señalamos en la sección 2.2.1. siempre que hacemos inferencias de una muestra a una población objetivo, las propiedades estadísticas de las estimaciones asumen una gran importancia. Las propiedades particulares de interés son el sesgo y la varianza, que se sintetizan en el error cuadrático medio. El comportamiento del sesgo y la varianza a medida que aumenta el tamaño de la muestra es especialmente crítico porque queremos que las muestras grandes sean lo más precisas posible. Los investigadores han desarrollado una variedad de técnicas para mejorar las propiedades estadísticas de las estimaciones de población. En muestras probabilísticas, estas técnicas están bien documentadas en libros de texto y artículos de revistas de muestreo. Para las muestras no probabilísticas los resultados están más dispersos entre disciplinas, en parte porque hay varios métodos de muestreo no probabilístico.

El conjunto de procedimientos de estimación que se han utilizado para la inferencia con muestras no probabilísticas pueden clasificarse en: pseudo-estimación basada en el diseño y estimación basada en modelos. Además en las dos últimas décadas se ha avanzado en el uso de la inferencia basada en algoritmos.

La estimación basada en modelos y la inferencia algorítmica constituyen la base teórica de lo que se ha denominado aprendizaje estadístico. El aprendizaje estadístico[71] desempeña un papel clave en muchas áreas de la ciencia, las finanzas y la industria. Estos son algunos ejemplos de problemas de aprendizaje:

[71] Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*.

1. Predecir si un paciente, hospitalizado debido a un ataque al corazón, tendrá un segundo ataque al corazón. La predicción debe basarse en medidas demográficas, dietéticas y clínicas para ese paciente.
2. Predecir el precio de una acción en 6 meses a partir de ahora, sobre la base de medidas de rendimiento de la empresa y datos económicos.
3. Identificar los números en un código postal manuscrito, a partir de una imagen digitalizada.
4. Estimar la cantidad de glucosa en la sangre de una persona diabética, desde el espectro de absorción infrarroja de la sangre de esa persona.
5. Identificar los factores de riesgo para el cáncer de próstata, basados en variables clínicas y demográficas.

La ciencia del aprendizaje juega un papel clave en los campos de la estadística, la minería de datos y la inteligencia artificial, que se intersecan con áreas de ingeniería y otras disciplinas. En un escenario típico, tenemos una medición de resultados, por lo general cuantitativa (como un precio de las acciones) o categórica (como ataque cardíaco sí / ataque cardíaco no), que queremos predecir sobre la base de un conjunto de características. Tenemos un conjunto de datos de entrenamiento, en el que observamos las mediciones de resultados y características de un conjunto de objetos (como personas). Usando estos datos construimos un modelo de predicción, o aprendiz, que nos permitirá predecir el resultado de nuevos objetos invisibles.

Los ejemplos anteriores describen lo que se llama el problema de aprendizaje supervisado. Se llama "supervisado" debido a la presencia de la variable de resultado para guiar el proceso de aprendizaje. En el problema de aprendizaje sin supervisión, observamos solamente las características y no tenemos medidas del resultado.

A continuación realizamos un breve acercamiento a la estimación basada en diseño, la basada en modelos y la algorítmica.

2.3.1. Pseudo-estimación basada en el diseño

La inferencia basada en diseño es el enfoque clásico para obtener las estimaciones poblacionales a partir de una encuesta por muestreo. Una variable objetivo y es observada para un muestra de todas las unidades que componen una población. La selección de la muestra es aleatoria, pero de acuerdo con un método controlado. Por ejemplo, en la estadística pública suelen utilizarse habitualmente diseños estratificados, de conglomerados o multietápicos. En función del diseño utilizado, cada unidad

de la población i tienen una probabilidad π_i de ser seleccionada en la muestra. Con esta información, un estimador de la población total Y sería:

$$\widehat{Y} = \sum_{i \in S} \frac{1}{\pi_i} y_i$$

Siendo S la muestra seleccionada. Este estimador clásico es conocido como el estimador de Horvitz-Thompson (Horvitz and Thompson, 1952). Existen versiones mejoradas de este estimador, con el fin de corregir la no respuesta, incorporando para ello información auxiliar en el procedimiento de selección; Sarndal et al, (1992) es una referencia clásica en este contexto. Como vemos en las inferencias basada en el diseño no se utilizan variables auxiliares en la estimación, pero sí en el diseño muestral. Por ejemplo, en muestras estratificadas donde la probabilidad de inclusión varía entre estratos, la variable de estratificación es un ejemplo de variable auxiliar usada en el diseño y por tanto utilizada indirectamente en el estimador.

Una propiedad importante y altamente estimada por las Oficinas de Estadística, es que el estimador basado en diseños es insesgado, y por lo tanto la exactitud del mismo depende únicamente de la varianza. Con ello, un aumento de la muestra significa una reducción de la varianza y por lo tanto una mejora en la exactitud de la estimación.

En las muestras probabilísticas, la ponderación se utiliza para implementar una fórmula de estimación dada una serie de respuestas de una encuesta. Los pesos para las muestras probabilísticas comienzan con los pesos iniciales π_i (a veces denominados pesos de diseño o probabilidad inversa del peso de selección). A continuación, se aplican ajustes para mejorar la eficiencia o para abordar sesgos potenciales, donde los sesgos pueden deberse a errores de no respuesta y cobertura entre otros.

Como vemos, para los estimadores basados en el diseño necesitamos conocer la probabilidad de inclusión π_i. Por lo tanto, en muestras no probabilísticas donde no conocemos la probabilidad de inclusión, la aplicación de la inferencia basada en el diseño no es posible. En ese sentido, en el contexto de las muestras no probabilísticas, surge la **pseudo-estimación basa en el diseño**. El término pseudo se agrega porque, a diferencia de la estimación tradicional basada en el diseño, las probabilidades de selección son desconocidas y no definidas (no existe un enlace explícito entre la muestra y el marco). En lugar de la probabilidad conocida, en la pseudo-estimación se utiliza una probabilidad estimada para el cálculo de las estimaciones.

Existen diferentes métodos de estimación de la probabilidad de selección. Un método habitual es la **postestratificación**, que consiste en estimar la probabilidad mediante el cálculo de la razón entre el tamaño de la muestra y el total estimado de la población dentro de algunos estratos homogéneos que cubran toda la población objetivo.

Otra alternativa para la estimación de la probabilidad es **modelar la propensión** de las unidades a estar presentes en el conjunto de datos, tal como se hace en los procedimientos de modelización de la no respuesta. En esta caso el factor de ajuste es el inverso de las propensiones estimadas y con ello la estimación sería:

$$\widehat{Y}_{\rho} = \sum_{i \in S} \frac{1}{\widehat{\rho}_i} y_i$$

Donde $\widehat{\rho}_i$ es la propensión estimada de que la unidad i forme parte de la muestra. Los $\widehat{\rho}_i$ son habitualmente estimados por regresión logística, pero también se usan métodos probit o no paramétricos (Little 1986; Da Silva and Opsomer 2009; Phipps and Toth 2012).

En lugar de estimar las propensiones de respuesta individual, el enfoque que la mayoría de las encuestas usan es formar grupos y ajustar los pesos en cada grupo por el inverso de la propensión del grupo. Sarndal et al. (1992) los describen como grupos de homogeneidad de respuesta (RHGs por su siglas en inglés), también habitualmente conocidos como clases de ponderación. En este caso el estimador puede escribirse como:

$$\widehat{Y}_{\rho_g} = \sum_{g} \sum_{i \in S_g} \frac{1}{\widehat{\rho}_g} y_i$$

El procedimiento señalado se conoce en la literatura como de ajustes de puntuación de propensión **(Propensity Score Adjustments, PSA)** o como ponderación de regresión logística, y es uno de los diversos métodos de ponderación que intentan eliminar el sesgo tanto en muestras probabilísticas como en las no probabilísticas. Se introdujo por primera vez para estudios observacionales, pero se utilizan ampliamente en las encuestas probabilísticas para limitar el efecto de los sesgos de la no respuesta, bajo el supuesto de que la respuesta es un fenómeno aleatorio. En esta aplicación, los modelos de respuesta se desarrollan mediante regresión logística, que utilizan variables predictoras conocidas tanto para los encuestados como para los no respondientes. Las estimaciones de propensión de respuesta resultantes (es decir, la probabilidad condicional de respuesta) se utilizan para ajustar los

pesos iniciales de los encuestados por un factor que se supone que es su probabilidad de responder a la encuesta.

Los métodos de PSA también se han utilizado en muestras no probabilísticas para intentar ajustar los efectos combinados de los errores de cobertura, la falta de respuesta y el muestreo no probabilístico[72]. Para estimar la probabilidad condicional de respuesta bajo todas estas fuentes de error puede ser necesaria una encuesta de referencia generada a partir de una muestra probabilística. Utilizando los datos de ambas muestras se utiliza un modelo logístico para estimar la probabilidad de participar en el estudio no probabilístico.

Los **ajustes de calibración s**e han estudiado ampliamente para las muestras probabilísticas y se ha demostrado que reducen tanto el sesgo como la varianza de las estimaciones. Tres ejemplos de ajuste de calibración bien conocidos y ampliamente utilizados son la postestratificación, el ranking y el de ponderación de regresión generalizada (GREG).

En muchas muestras no probabilísticas la postestratificación es la única forma de ponderación posible, mientras que en las muestras probabilísticas la calibración es un ajuste adicional del peso generado a partir de los pesos iniciales y quizás de los pesos ajustados por no respuesta. Este tipo de metodología de ponderación puede ser la única herramienta útil para muestras no probabilísticas que no tienen información suficiente para construir un PSA (muestras sin una encuesta de referencia correspondiente).

Como vemos, ha existido voluntad investigadora entre la comunidad científica para intentar encontrar procedimientos que permitan trabajar con ciertos niveles de confianza con los pseudoestimadores basado en diseño. Sin embargo estos estudios no están cerrados, tal como se revisa en el *Report of the AAPOR Task Force on Non-Probability Sampling,* así nos encontramos que Dever, Rafferty y Valliant (2008)[73] encontraron algunos beneficios, aunque inconsistentes, del ajuste de GREG en la reducción del sesgo de no cobertura. Respecto a los PSA, Yeager, et al (2011)[74]encontraron que la postestratificación mejoró la precisión de las estimaciones de la muestra no probabilística, aunque nuevamente sus resultados fueron inestables. Las investigaciones realizadas por Lee (2006)[75] y Lee y

[72] Una extensa lista de citas se puede encontrar en Lee (2006)
[73] Dever, Rafferty, and Valliant, "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?"
[74] Yeager et al., "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples."
[75] Lee, "An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey."

Valliant (2009)[76] mostraron que los PSA o la calibración por sí solos no son suficientes para reducir los sesgos en las estimaciones de las encuestas no probabilísticas a niveles relativamente bajos.

2.3.2. Inferencia basada en modelos

Cuando disponemos de variables auxiliares x correlacionadas con la variable objetivo a estimar y, esta correlación se puede aprovechar a través de algún modelo (Van den Brakel and Bethlehem, 2008; Vaillant et al., 2000). El modelo más elemental es el modelo lineal:

$$y = \beta_0 + \beta x + \epsilon \qquad (1)$$

Donde y es una función lineal de x, siendo ϵ una medida del error por la falta de una relación perfecta. En los modelos se supone que este error sigue una distribución aleatoria normal y por ende la variable de interés y también sigue una distribución aleatoria normal.

En situaciones en las que las x son conocidas para toda la población, y sin embargo las y sólo están disponibles para una parte de ella, se suele aplicar la inferencia basada en modelos para obtener las estimaciones no disponibles. Para ello se construye el modelo con los datos conocidos y posteriormente se utiliza dicho modelo para estimar los valores de y desconocidos a partir de las x conocidas. La construcción del modelo consiste en la estimación de los parámetros del mismo, o sea la estimación de las β. Por lo tanto, con las estimaciones de $\widehat{\beta}_0$ y $\widehat{\beta}$, la estimación de Y basada en el modelo sería:

$$\widehat{Y} = \sum_{i \in S} y_i + \sum_{i \in R} (\widehat{\beta}_0 + \widehat{\beta} x_i)$$

Donde S es el conjunto de unidades para las cuales y es conocida, y R el complementario de S. Los valores y de las unidades que pertenecen a R son estimados mediante el modelo ajustado.

Como vemos, la estimación basada en modelos se sostiene en un modelo estadístico que describe la variable que se estima en la encuesta como una distribución normal. Por lo tanto, con este tipo de procedimiento de estimación se supone que la característica de interés (la variable y) es una variable aleatoria con una distribución propia, de modo que la aleatoriedad no proviene del proceso de

[76] Lee and Valliant, "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment."

generación de la muestra no probabilística. En estos procedimientos las observaciones se utilizan para ajustar el modelo y el análisis se lleva a cabo suponiendo que el muestreo puede ser ignorado. En otras palabras, **el resultado estimado a partir del modelo no está estadísticamente relacionado con el método de muestreo**.

Bajo estas consideraciones las estimaciones basadas en modelos parecen ser un buen instrumento metodológico para su uso en las fuentes Big Data. Sin embargo hay que señalar que varios investigadores han discutido este análisis basado en modelos y las condiciones bajo las cuales el método de muestreo puede ser ignorado para poder hacer inferencias válidas (por ejemplo, Little y Rubin 2002, Sugden y Smith 1984, Pfeffermann y Rao 2009).

En ese sentido, entre la comunidad estadísticas existe una preocupación lógica por el control del sesgo de autoselección. Como vimos en secciones anteriores el sesgo de autoselección ocurre cuando una parte de la población objetivo, con características particulares, es excluida del muestreo. El sesgo de autoselección aparece en diversas situaciones como por ejemplo: Cuando seleccionamos la muestra de forma no aleatoria, cuando la población objetivo no está bien definida, cuando no incluimos a toda la población objetivo en el universo muestral, cuando la tasa de no respuesta es relevante en la muestra o cuando la muestra está basada en participantes voluntarios.

Al presentarse el problema de la selección muestral los modelos de estimación deben recurrir, además de la ecuación objetivo que se pretende estimar, a una segunda ecuación que se le suele denominar **ecuación de selección**. La ecuación de selección corresponde a un modelo de variable dependiente discreta y mide la probabilidad de estar en la muestra. El ejemplo típico es el considerado por el Premio Nobel Heckman[77] en 1979 en su trabajo sobre el mercado laboral, pero también el método de Máxima Verosimilitud de Amemiya 1981.

Por otra parte, en la aplicación de la estimación basada en modelos es posible que no observemos datos de la variable dependiente o de las variables explicativas para toda la población. En este caso, tendremos muestras censuradas o truncadas según cómo sea el tipo de limitación en la información disponible. Una muestra está **truncada** si los datos sólo están disponibles para un subconjunto de la población total, o sea, los valores de las variables explicativas X sólo se observan cuando se observa Y [78]. En una muestra **censurada**, tenemos observaciones de las X de toda la población, pero el valor de

[77] Heckman, "Sample Selection Bias as a Specification Error."
[78] Por ejemplo, el gasto médico de una muestra de pacientes entrevistados después de someterse a un tratamiento dental. En este caso, sólo observamos a personas con gasto mayor que cero.

la Y se desconoce para un subconjunto de la población. Existen modelos en la literatura científica para abordar las situaciones anteriormente señaladas, como son los modelos Tobit[79].

Pero sin embargo los modelos que más interés suscitan entre los investigadores de las fuentes Big Data son los **modelos dinámicos** como instrumentos para realizar nowcasting o forecasting**.** En un modelo estático la variable tiempo no desempeña un papel relevante. En un modelo dinámico, por el contrario, alguno/s de los elementos que intervienen en la modelización no permanecen invariables, sino que se consideran como funciones del tiempo, describiendo trayectorias temporales.

Un artículo de referencia al respecto es *Predicting the Present with Google Trends*[80] que relata cómo se pueden construir modelos dinámicos para predecir a corto plazo la tendencia de muchas variables de interés para la estadística pública, tales como la venta de vehículos, la tasa de desempleo o la entrada de turistas a un territorio. En este mismo capítulo ya presentamos las estimaciones ofrecidas por Google Flu Trends. Este mismo planteamiento se encuentra en el ejercicio *Social Media Sentiment and Consumer Confidence*[81] de predicción del índice de confianza del consumidor a partir de datos de las redes sociales o en el artículo *Using Social Media to Measure Labor Market Flows*[82].

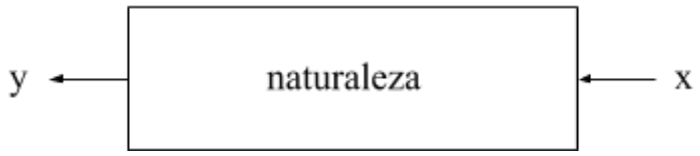
En estos trabajos se buscan modelos que relacionen datos externos ofrecidos por las Oficinas Estadísticas (variables dependientes) con los datos aportados por fuentes Big Data (variables independientes). Con estas correlaciones se pretende tanto poder realizar nowcasting (o sea, tener estimaciones anticipadas a las estimaciones oficiales), como estudiar la viabilidad de sustituir la encuesta oficial por fuentes Big Data. A su vez, con esta estrategia se elimina el debate sobre el problema del sesgo por autoselección; pues se buscan relaciones entre estimaciones de fuentes oficiales obtenidas por fuentes de primera, segunda o tercera generación con los de cuarta generación.

Uno de los problemas del uso de modelos dinámicos es que el mismo mantenga estimaciones ajustadas a lo largo del tiempo, pudiendo darse rupturas en la calidad de las estimaciones tal como explicamos en el apartado sobre Google Flu Trends.

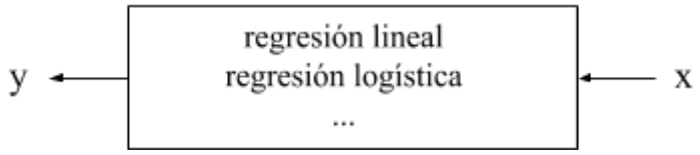
[79] https://es.wikipedia.org/wiki/Modelo_Tobit
[80] Choi and Varian, "Predicting the Present with Google Trends."
[81] Daas, Puts, and European Central Bank, *Social Media Sentiment and Consumer Confidence.*
[82] Antenucci et al., "Using Social Media to Measure Labor Market Flows."

2.3.3. Inferencia algorítmica

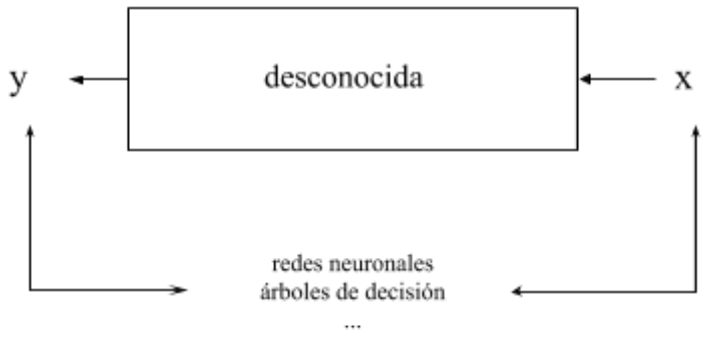
Una nueva cultura según Breiman[83], propone considerar los datos como datos generados por una caja negra en la que un vector de datos de las variables de entrada x (variables independientes) entran por un lado, y por el otro lado salen los datos de las variables de respuesta; de tal manera que dentro de la caja negra la naturaleza funciona asociando las variables predictoras con las variables de respuesta. Esquemáticamente podría representarse como:



En el análisis basado en modelos se asume que dentro de la caja negra funciona un modelo de datos estocásticos, por ejemplo un modelo de regresión lineal o de regresión logística:



Un acercamiento a la estimación basada en modelos desde una perspectiva diferente nos permite considerar un modelo como una función F que asigna valores a un valor desconocido y a partir de un valor conocido x, $F(x) = y$. Esta perspectiva de la estimación basada en modelos permite extender los métodos de inferencia a partir de la clases de funciones utilizadas. El análisis en esta cultura considera la caja negra como desconocida y compleja de conocer; siendo su objetivo encontrar una función F, que podría ser un algoritmo, capaz de asociar las entradas x con las salidas y.



[83] Breiman and others, "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)."

En el enfoque algorítmico, el conjunto de datos para los cuales se conoce tanto x como y se dividen en dos grupos. Una grupo se utiliza para ajustar el algoritmo, mediante aprendizaje o entrenamiento. El segundo conjunto de datos se utiliza para evaluar o probar las capacidades predictivas del algoritmo entrenado.

En general es imposible expresar los métodos algorítmicos analíticamente en términos de un modelo matemático, pero usando una similitud conceptual con el estimador basado en modelos, el estimador algorítmico se podría expresar como:

$$\widehat{Y} = \sum_{i \in S} y_i + \sum_{i \in R} F(x_i)$$

Donde S es el conjunto de unidades para las cuales se conoce y. Este es el conjunto que se divide en dos bloques en la fase de entrenamiento. Por otra parte, el conjunto R contiene las unidades de población con y desconocidas. La incertidumbre de este estimador surge de la capacidad imperfecta de estimación del algoritmo, y se evalúa usando alguna función de coste.

Como ejemplo, Breiman (2001) discute en su artículo dos algoritmos: los Árboles de Clasificación y Regresión (CART) y las Redes Neuronales Artificiales (ANN)[84]. El algoritmo CART[85] crea árboles de decisión basados en una serie de reglas aprendidas a partir de los datos. Son árboles binarios y el objetivo es determinar el valor de una variable que mejor diferencia un conjunto de datos para clasificarlos en dos grupos. Los datos se dividen de acuerdo con este valor, formando dos ramas. Cada rama se divide de nuevo siguiendo el mismo procedimiento, hasta que los grupos son suficientemente homogéneos. Una ANN es una red de funciones que permiten la aproximación de prácticamente cualquier función no lineal. Existen varios diseños de red y métodos de formación de las redes.

Una característica distintiva de la inferencia en las estadísticas oficiales, comparada con algunas otras aplicaciones en la minería de datos, es que las predicciones unitarias no son de interés. Lo que interesa es la cantidad de población obtenida a partir de los valores unitarios. Supongamos por ejemplo un algoritmo de predicción capaz de clasificar los correos electrónicos en "bueno" o "spam". En el contexto de la inferencia en la estadística oficial, el algoritmo se utilizaría para obtener la proporción de correos electrónicos no deseados recibidos en un período de 24 horas, sin necesidad de estudiar las predicciones sobre los correos electrónicos individuales.

[84] Detalles y referencias adicionales para estos métodos se pueden encontrar en Breiman (2001) y Hastie et al. (2009).
[85] Breiman et al., *Classification and Regression Trees. Reprint*.

2.3.4. Inferencia y tipologías de fuentes de datos

Con el riesgo de simplificar el problema, la tabla 4 proporciona un marco en el que se asocian tipologías de datos y métodos de inferencia. La inferencia basada en modelos y algorítmica es referida conjuntamente en la tabla como inferencia predictiva o aprendizaje estadístico (Hastie et al., 2009).

Tabla 4. Relación entre tipologías de datos y métodos de inferencia

	Encuestas	**Registros administrativos**	**Big Data**
Métodos basados en diseño	Común en las estadísticas oficiales: muestreo por encuestas y estimación basada en el diseño. Cuando las encuestas son no probabilísticas se puede usar pseudo-estimación basada en diseños.	Como no hay diseño, la estimación basada en el diseño no es aplicable. Como alternativa se puede usar pseudo-estimación basada en diseños.	
Métodos predictivos	La inferencia predictiva puede aplicarse en principio, pero generalmente no es necesaria.	Como los registros son a menudo (casi) completos, el modelado simple es suficiente en la mayoría de los casos.	Los datos de cuarta generación son a menudo altamente selectivos, la consideración de los métodos predictivos es beneficiosa.

Los métodos basados en diseño sólo son aplicables cuando existe un diseño conocido que sustenta los datos. Cuando falta información de diseño, el estimador de propensión basado en modelos se asemeja en cierta medida a un estimador basado en diseño.

Si bien los métodos predictivos pueden aplicarse a los datos de encuestas, no hay ninguna necesidad de hacerlo. En algunas situaciones el modelado puede ser útil, por ejemplo cuando el tamaño de la muestra es pequeño (Rao, 2003) o cuando los niveles de falta de respuesta son altos (Van de Brakel y Bethlehem, 2008).

En un contexto estadístico, los datos de tercera generación se refieren a los registros administrativos. Estos registros a menudo son de cobertura total, o por lo menos cubren una proporción tan grande de la población que se puede suponer razonablemente representativa de la población total. La estimación de las partes faltantes de los registros puede realizarse mediante métodos predictivos, siempre y cuando el tamaño del conjunto de unidades para las cuales la variable objetivo es desconocida es mucho menor que el conjunto de las que se conoce.

La situación es diferente para las fuentes Big Data, en este caso es esencial determinar el mejor método de inferencia, ya que puede afectar de manera crucial a la estimación de la población. Se podrían utilizar métodos de pseudo-estimación basada en diseños, pero fundamentalmente métodos predictivos. El alcance no debe limitarse al modelado lineal (generalizado), sino que debe incluir enfoques no paramétricos o algorítmicos.

Bibliografía del Capítulo 2

Amemiya, Takeshi. "Qualitative Response Models: A Survey." *Journal of Economic Literature* 19, no. 4 (1981): 1483–1536.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew Shapiro. "Using Social Media to Measure Labor Market Flows." Cambridge, MA: National Bureau of Economic Research, March 2014. http://www.nber.org/papers/w20010.pdf.

Bart Buelens, Harm Jan Boonstra, Jan van den Brakel, and Piet Daas. "Shifting Paradigms Official Statistics: From Design-Based to Model-Based to Algorithmic Inference." Discussion Paper. The Hague/Heerlen: Statistics Netherlands, 2012. https://www.cbs.nl/nl-nl/achtergrond/2012/38/shifting-paradigms-in-official-statistics-from-design-based-to-model-based-to-algorithmic-inference.

Baker, Reg, Brick, J. Michael, Bates, Nancy A., Couper, Mick P., Dever, Jill A., and Tourangeau, Roger. "Report of the AAPOR Task Force on Non-Probability Sampling." AAPOR, June 2013. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf.

Breiman, Leo, and others. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3 (2001): 199–231.

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees. Reprint*. Chapman & Hall/CRC, Boca Raton, FL, 1998.

Biemer, P. P. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74, no. 5 (January 1, 2010): 817–48. doi:10.1093/poq/nfq058.

Biemer, Paul P., and Dennis Trewin. "A Review of Measurement Error Effects on the Analysis of Survey Data." In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith De Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, 601–32. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012. http://doi.wiley.com/10.1002/9781118490013.ch27.

Biemer, P. P. “Toward a Total Error Framework for Big Data. Presentation in American Association for Public Opinion Research (AAPOR) 69th Annual Conference,” 2014.

Biemer, Paul P, and Lars E Lyberg. *Introduction to Survey Quality*. Vol. 335. John Wiley & Sons, 2003.

Boyd, danah, and Kate Crawford. “CRITICAL QUESTIONS FOR BIG DATA: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” *Information, Communication & Society* 15, no. 5 (June 2012): 662–79. doi:10.1080/1369118X.2012.678878.

Brick, J. Michael. “Unit Nonresponse and Weighting Adjustments: A Critical Review.” *Journal of Official Statistics* 29, no. 3 (January 1, 2013). doi:10.2478/jos-2013-0026.

Butler, Declan. “When Google Got Flu Wrong.” *Nature* 494, no. 7436 (February 13, 2013): 155–56. doi:10.1038/494155a.

Choi, Hyunyoung, and Hal Varian. “Predicting the Present with Google Trends.” *Economic Record* 88, no. s1 (2012): 2–9.

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. “The Role of Administrative Data in the Big Data Revolution in Social Science Research.” *Social Science Research* 59 (September 2016): 1–12. doi:10.1016/j.ssresearch.2016.04.015.

Cook, Samantha, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi. “Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.” Edited by Benjamin J. Cowling. *PLoS ONE* 6, no. 8 (August 19, 2011): e23610. doi:10.1371/journal.pone.0023610.

Couper, Mick P. “Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys,” 2013. doi:10.18148/srm/2013.v7i3.5751.

Daas, Piet J. H, Marco J. H Puts, and European Central Bank. *Social Media Sentiment and Consumer Confidence.* Frankfurt am Main: European Central Bank, 2014. http://bookshop.europa.eu/uri?target=EUB:NOTICE:QBBF14001:EN:HTML.

Dever, Jill A, Ann Rafferty, and Richard Valliant. “Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?” In *Survey Research Methods*, 2:47–60, 2008.

Fan, J., F. Han, and H. Liu. “Challenges of Big Data Analysis.” *National Science Review* 1, no. 2 (June 1, 2014): 293–314. doi:10.1093/nsr/nwt032.

Fan, Jianqing, Richard Samworth, and Yichao Wu. “Ultrahigh Dimensional Feature Selection: Beyond The Linear Model.” *J. Mach. Learn. Res.* 10 (December 2009): 2013–38.

Fan, Jianqing, and Yuan Liao. “Endogeneity in High Dimensions.” *The Annals of Statistics* 42, no. 3 (June 2014): 872–917. doi:10.1214/13-AOS1202.

Florescu, Denisa, Martin Karlberg, Fernando Reis, Pilar Rey Del Castillo, Michail Skaliotis, and Albrecht Wirthmann. “Will ‘big Data’ Transform Official Statistics.” In *Q2014–European Conference on Quality in Statistics*, 2014. http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.

Gayo-Avello, Daniel. “A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter Data.” *Social Science Computer Review*, 2013, 0894439313493979.

Hall, Peter, and Hugh Miller. “Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems.” *Journal of Computational and Graphical Statistics* 18, no. 3 (January 2009): 533–50. doi:10.1198/jcgs.2009.08041.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer, 2009.

Heckman, James J. “Sample Selection Bias as a Specification Error.” *Econometrica* 47, no. 1 (January 1979): 153. doi:10.2307/1912352.

Horvitz, Daniel G, and Donovan J Thompson. “A Generalization of Sampling without Replacement from a Finite Universe.” *Journal of the American Statistical Association* 47, no. 260 (1952): 663–85.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O’Neil, and Abe Usher. “Big Data in Survey Research: AAPOR Task Force Report.” *Public Opinion Quarterly* 79, no. 4 (2015): 839–80. doi:10.1093/poq/nfv039.

Jungherr, A., P. Jurgens, and H. Schoen. “Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. ‘Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment.’” *Social Science Computer Review* 30, no. 2 (May 1, 2012): 229–34. doi:10.1177/0894439311404119.

Kim, Annice, Joe Murphy, Ashley Richards, Heather Hansen, Rebecca Powell, and Carol Haney. “Can Tweets Replace Polls? A U.S. Health-Care Reform Case Study.” In *Social Media, Sociality, and Survey Research*, edited by Craig A. Hill, Elizabeth Dean, and Joe Murphy, 61–86. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013. http://doi.wiley.com/10.1002/9781118751534.ch3.

Kish, Leslie, and Centro de Investigaciones Sociológicas (España). *Diseño estadístico para la investigación*. Madrid: CIS : Siglo XXI, 1995.

Kish, Leslie. “Samples and Censuses.” *International Statistical Review / Revue Internationale de Statistique* 47, no. 2 (August 1979): 99. doi:10.2307/1402563.

Kitchin, Rob. “The Opportunities, Challenges and Risks of Big Data for Official Statistics.” *Statistical Journal of the IAOS* 31, no. 3 (2015): 471–81. doi:10.3233/SJI-150906.

Kramer, A. D. I., J. E. Guillory, and J. T. Hancock. “Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks.” *Proceedings of the National Academy of Sciences* 111, no. 24 (June 17, 2014): 8788–90. doi:10.1073/pnas.1320040111.

Lane, Julia I., ed. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY: Cambridge University Press, 2014.

Langer Research Associates. “Briefing Paper: Social Media and Public Opinion,” 2013. http://www.langerresearch.com/wp-content/uploads/Langer_Research_Briefing_Paper-Social_Media_and_Public_Opinion.pdf.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343, no. 6176 (March 14, 2014): 1203–5. doi:10.1126/science.1248506.

Lee, Sunghee. “An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey.” *Social Science Computer Review* 24, no. 4 (2006): 460–75.

Lee, Sunghee, and Richard Valliant. “Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment.” *Sociological Methods & Research* 37, no. 3 (2009): 319–43.

Little, Roderick J. A. “Survey Nonresponse Adjustments for Estimates of Means.” *International Statistical Review / Revue Internationale de Statistique* 54, no. 2 (August 1986): 139. doi:10.2307/1403140.

O’Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” *ICWSM* 11, no. 122–29 (2010): 1–2.

Pfeffermann, Danny, and Calyampudi Radhakrishna Rao. *Handbook of Statistics_29A: Sample Surveys: Design, Methods and Applications*. Vol. 29. Elsevier, 2009.

Phipps, Polly, and Daniell Toth. “Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data.” *The Annals of Applied Statistics* 6, no. 2 (June 2012): 772–94. doi:10.1214/11-AOAS521.

Rao, J. N. K. *Small Area Estimation*. Wiley Series in Survey Methodology. Hoboken, N.J: John Wiley, 2003.

Särndal, Carl-Erik and Jean-Claude, Deville. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87, no. 418 (1992): 376–82.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe, and Frederick G. Conrad. “Social Media Analyses for Social Measurement.” *Public Opinion Quarterly* 80, no. 1 (2016): 180–211. doi:10.1093/poq/nfv048.

Silva, Damião N da, and Jean D Opsomer. “Nonparametric Propensity Weighting for Survey Nonresponse through Local Polynomial Regression.” *Survey Methodology* 35, no. 2 (2009): 165–76.

Smith, T. W. “Survey-Research Paradigms Old and New.” *International Journal of Public Opinion Research* 25, no. 2 (June 1, 2013): 218–29. doi:10.1093/ijpor/eds040.

Stock, James H, and Mark W Watson. “Forecasting Using Principal Components From a Large Number of Predictors.” *Journal of the American Statistical Association* 97, no. 460 (December 2002): 1167–79. doi:10.1198/016214502388618960.

Sugden, RA, and TMF Smith. “Ignorable and Informative Designs in Survey Sampling Inference.” *Biometrika* 71, no. 3 (1984): 495–506.

Thomsen, Ib, and Ann Marit Kleive Holmøy. “Combining Data from Surveys and Administrative Record Systems. The Norwegian Experience.” *International Statistical Review* 66, no. 2 (1998): 201–221.

Tufekci, Zeynep. “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls.” *CoRR* abs/1403.7400 (2014). http://arxiv.org/abs/1403.7400.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.” *ICWSM* 10 (2010): 178–85.

Valliant, Richard, Alan H Dorfman, and Richard M Royall. “Finite Population Sampling and Inference: A Prediction Approach,” 2000.

Van den Brakel, J, J Bethlehem, and others. “Model-Based Estimation for Official Statistics.” *Statistics Netherlands Discussion Paper*, 2008.

Van den Brakel, Jan, Söhler, Emily, Piet JH Daas, and Buelens, Bart. “Social Media as a Data Source for Official Statistics; the Dutch Consumer Confidence Index.” CBS Statistics Netherland, January 2016. https://www.cbs.nl/en-gb/background/2016/07/social-media-as-a-data-source-for-official-statistics-the-dutch-consumer-confidence-index.

Yeager, David S, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang. “Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples.” *Public Opinion Quarterly*, 2011, nfr020.

Capítulo 3. Usos de fuentes Big Data y marco de calidad

3.1. Los potenciales roles de las fuentes Big Data en la estadística pública

Los profesionales de las estadísticas oficiales son bastante conservadores y muy cautelosos en el uso de nuevas tipologías de datos. En ese sentido las fuentes Big Data no son una excepción. Sin embargo se han identificado usos potenciales de las fuentes Big Data en la estadística pública, por similitud con las fuentes administrativas, que categorizamos en dos bloques:

1. Uso como fuentes de estimación estadística
2. Uso como información de apoyo en el diseño y elaboración de estadísticas

3.1.1. Uso como fuentes de estimación estadística

A) Sustitución de fuentes existentes por fuentes Big Data

La primera cuestión que se plantea al estudiar el uso de fuentes Big Data en la estadística pública consiste en considerar la posibilidad de sustituir fuentes primarias, como censos o encuestas, por dichas fuentes.

En el artículo *"Will big data transform official statistics?"* (Florescu et al. 2014)[86] se concluye, tras revisar las variables sobre las que se recopilan datos en las distintas encuestas del Sistema Estadístico Europeo, que las fuentes Big Data aún no ofrecen una alternativa para todas estas variables. La misma conclusión fue alcanzada por el proyecto *"Internet as a Data Source"* que evalúa su uso para sustituir las encuestas sobre las TIC (Karlberg y Skaliotis, 2013)[87]. A esta incapacidad de las fuentes Big Data para cubrir todas la variables que se estudian en la estadística pública, se unen las dificultades de las

[86] Florescu, Denisa, Martin Karlberg, Fernando Reis, Pilar Rey Del Castillo, Michail Skaliotis, and Albrecht Wirthmann. "Will 'big Data' Transform Official Statistics." In *Q2014–European Conference on Quality in Statistics*, 2014. http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.
[87] Karlberg and Skaliotis, "Big Data for Official Statistics—strategies and Some Initial European Applications."

estimaciones basadas en diseño para muestras no probabilísticas, tal como estudiamos en el Capítulo 2.

B) Sustitución parcial de fuentes existentes por fuentes Big Data

Como vimos en la sección sobre coberturas de temas en fuentes Big Data de redes sociales del Capítulo 2, algunos autores (Schober et al., 2016)[88] sugieren que es prematuro aprobar o rechazar totalmente la idea de que los análisis de las redes sociales podrían reemplazar a las encuestas en la producción de estadísticas oficiales. Sin embargo algunos investigadores proponen reemplazar encuestas con datos de redes sociales cuando haya un patrón externo de referencia (Antenucci et al. 2015)[89]. En todo caso, si el patrón de referencia es una encuesta, entonces tendría que haber por lo menos encuestas ocasionales para calibrar las tendencias de las redes sociales. Se podría pensar, por ejemplo, en recopilar datos de encuestas bimestrales en lugar de mensuales y realizar estimaciones con los datos de redes sociales en los meses sin recogida de datos; pero sólo si con el tiempo las tendencias de las redes sociales demuestran una asociación suficiente con los resultados de la encuesta.

En esa dirección la comunidad científica vinculada a la estadística pública, propone que las fuentes Big Data podrán sustituir a algunos resultados estadísticos, manteniendo sus definiciones sin cambios, siempre que se cumplan dos condiciones:

1. Satisfacen las necesidades evolutivas de información.

2. Existen otras fuentes no sesgadas de referencia para la comparación de los resultados ofrecidos por las fuentes Big Data y la posible corrección de sesgos.

En esta estrategia se buscarían modelos que relacionen datos insesgados de las encuestas con los datos potencialmente sesgados de las fuentes Big Data, y los modelos se aplicarían cuando no se realizan las encuestas. La repetición de encuestas permite validar y actualizar regularmente los modelos, evitando con ello la obsolescencia de los mismos.

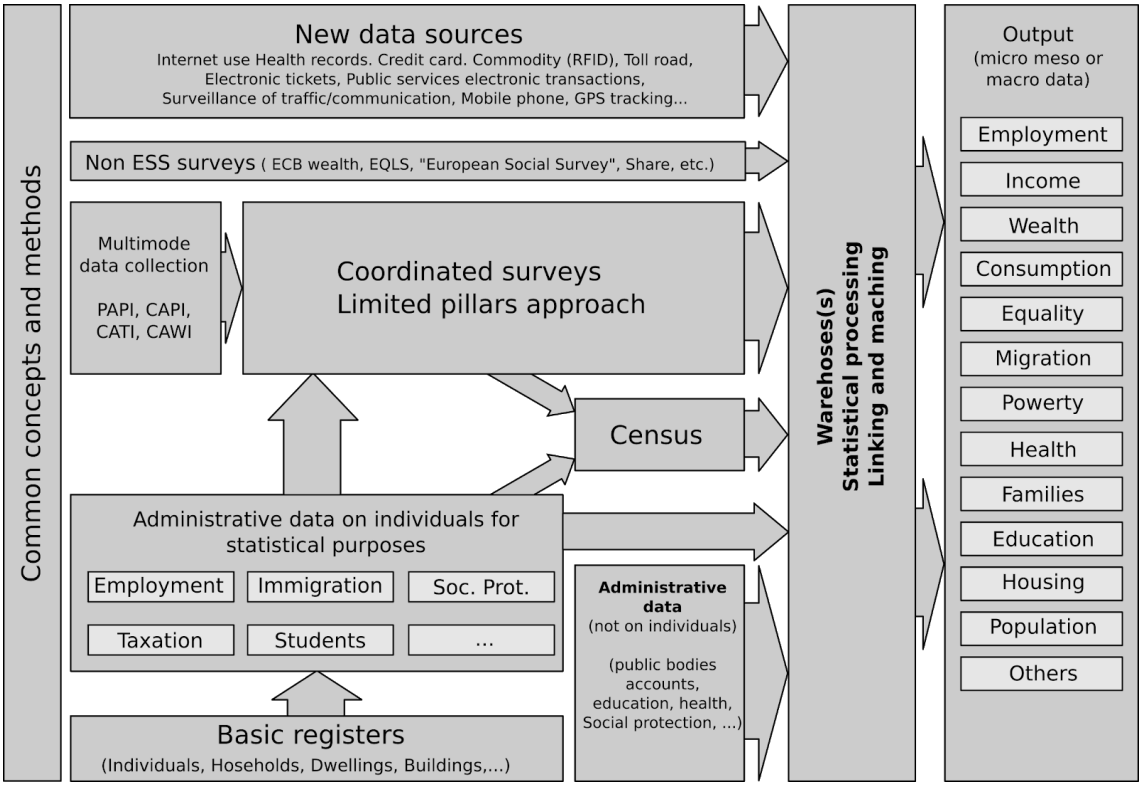
C) Incorporación de variables Big Data a otras fuentes: microintegración

Una de las características de las fuentes Big Data es su capacidad para obtener datos de eventos difícilmente recopilables por los métodos tradicionales. Por lo tanto estas fuentes aportan información complementaria a las fuentes usadas habituales en la estadística pública.

[88] Schober et al., "Social Media Analyses for Social Measurement."
[89] Antenucci et al., "Using Social Media to Measure Labor Market Flows."

Esta complementariedad da lugar a un ecosistemas de fuentes diversas al servicio de la mejora de las estadísticas públicas en una dimensión determinada. Por ejemplo, el documento "*The Modernisation of European Social Statistics"*[90] recoge la incorporación de las fuentes Big Data dentro de un ecosistema de fuentes diversas para la elaboración de las estadísticas sociales en Europa.

Esquema 8.- Ecosistema multifuentes para la modernización de las estadísticas sociales



La complementariedad requiere en algunos casos la integración de las fuentes Big Data con otras fuentes de datos tradicionalmente usadas en la estadística pública. **El emparejamiento estadístico (statistical matching) es un enfoque basado en modelos para proporcionar información conjunta sobre variables e indicadores recolectados a través de múltiples fuentes**. Los beneficios potenciales de este enfoque radican en la posibilidad de mejorar el uso complementario y el potencial analítico de las fuentes de datos existentes.

La mayoría de las veces el objetivo de un ejercicio de emparejamiento es ampliar el alcance de la información, pero también se han utilizado para alinear las estimaciones observadas en múltiples

[90] Reis, Fernando. "The Modernisation of European Social Statistics." Rome: Eurostat, 2012. https://www.destatis.de/EN/AboutUs/Events/DGINS/Document_PaperEUROSTAT.pdf?__blob=publicationFile.

encuestas y para mejorar la precisión de estas estimaciones mediante la integración con estudios más amplios. Existen dos enfoques respecto al emparejamiento estadístico:

1. El enfoque **macro** se refiere a la identificación de cualquier estructura que describa relaciones entre las variables no observadas conjuntamente en las fuentes de información que se desean integrar, tales como distribuciones conjuntas, distribuciones marginales o matrices de correlación (D'Orazio, 2006)[91].

2. El enfoque **micro** se refiere a la creación de un archivo completo de microdatos donde los datos de todas las variables están disponibles para cada unidad. Esto se logra mediante la generación de un nuevo conjunto de datos a partir de dos conjuntos de datos, usando para ello algunas variables comunes como variables bisagra.

La idea de la que parte el micro-emparejamiento estadístico es que, aunque las unidades de observación de las fuentes de datos provienen de la misma población, por lo general no se superponen. En ese sentido el micro-emparejamiento estadístico identifica y vincula registros de diferentes fuentes que corresponden a unidades similares. Esta es la diferencia básica en comparación con el Record Linkage, donde es necesario que las unidades incluidas en los ficheros datos se solapen y cuyo objetivo es enlazar la misma unidad en las distintas fuentes. **Por lo tanto, el record linkage trata con unidades idénticas, mientras que el statistical matching trata de unidades similares.**

En la práctica, el procedimiento de emparejamiento puede entenderse como un problema de imputación de las variables de una encuesta donante a una encuestas de receptores. Supongamos que las variables Y, Z se recogen a través de dos muestras diferentes tomadas de la misma población. Asimismo las variables X se recogen en ambas muestras y éstas están correlacionadas con Y y Z.

Esquema 9.- Microintegración de fuentes de datos

Fuente A (donante)	Fuente B (receptora)	Fuente B (ampliada)
X, Y	X, Z	X, $\widehat{Y}$, Z

La relación entre la variables comunes X con las variables observadas Y se utiliza para explorar la fuente de datos donante y determinar la imputación de las unidades del conjunto de datos receptor. Así se genera un conjunto de datos sintéticos con información completa sobre X, Y y Z.

[91] D'Orazio, Di Zio, and Scanu, *Statistical Matching: Theory and Practice*.

Como vemos, la microintegración es una estrategia estupenda para incorporar información de las fuentes Big Data a las fuentes tradicionales. Este tipo de ejercicio se ha realizado previamente con registros administrativos y encuestas, sin embargo aún no se conocen trabajos similares aplicados a las fuentes Big Data como fuentes donantes.

3.1.2. Uso como información de apoyo o complementaria

Un estudio de la literatura acerca de los datos administrativos revela sus diferentes usos por parte de las Oficinas Estadísticas[92]. En este apartado vamos a revisar algunos de esos usos y su trasposición de aplicabilidad a las fuentes Big Data:

A) Creación o complementación de registros estadísticos

Varias Oficinas de Estadística utilizan fuentes de datos administrativas para crear registros estadísticos. En estos registros se enumeran poblaciones completas de unidades de análisis tales como el directorio estadístico de empresas o el directorio de población y hogares; y en su alimentación también podrían utilizarse las fuentes Big Data.

Por ejemplo, para la alimentación del registro de establecimientos asociados al registro de empresas podría aprovecharse la información aportada por fuentes Big Data tales como Google Place, Facebook, FourSquare, OpenStreetMap, etc.

B) Diseño y planificación de encuestas

Las fuentes de Big Data se pueden utilizar en el diseño y planificación de encuestas, veamos en qué fases:

1. **Diseño muestral:** El primer uso puede ser el de proporcionar o complementar marcos de muestreo, pero además también pueden proporcionar información útil acerca de una población en estudio, información que puede ayudar en el diseño de muestras. Por ejemplo, pueden dar información relacionada con los estratos de la población: su tamaño, su composición (medias, desviaciones, las proporciones de interés), su extensión geográfica, etc.

[92] Eurostat, "Quality Assessment of Administrative Data for Statistical Purposes."

2. **Recogida de datos, incorporación de tecnología de Reality Mining a las encuestas tradicionales:** En una sociedad totalmente digital podría plantearse un paso más allá, de tal manera que la captura de datos de las encuestas podrían diseñarse utilizando tecnologías de Reality Mining[93]. En esta estrategia el diseño de la encuesta se realizaría mediante métodos de muestreo tradicional, mientras que en la fase de recogida de datos se utilizarán tecnologías de datificación reality mining. Por ejemplo, para un estudio de movilidad se selecciona una muestra de la población y se les solicita que activen una aplicación de seguimiento en su smartphone, como complemento a una encuesta tradicional

3. **Diseño de cuestionario:** Con el fin de profundizar en las relaciones entre las variables que pueden ser incorporadas en un cuestionario, se puede realizar previamente un análisis exploratorio de datos de las fuentes Big Data para mejorar el diseño del cuestionario.

4. **Planificación de trabajo de campo:** Además las fuentes Big Data también podrían utilizarse en la planificación de las encuestas, por ejemplo en la especificación del número requerido de entrevistadores o en la asignación del esfuerzo (áreas geográficas o cuotas) asignado a cada entrevistador.

C) Verificación o imputación de datos

Los datos recogidos en una encuesta se podrían verificar con los datos obtenidos para las mismas unidades de análisis y las mismas variables en fuentes Big Data. A su vez en caso de datos faltantes la fuente Big Data podría proporcionar un fichero para imputación Cold-Deck o de fichero donante[94].

D) Fuente auxiliar para estimaciones basadas en encuestas

Otro rol de las fuentes Big Data es su uso como fuente auxiliar, aprovechando sus características especiales, para la mejora de las estimaciones aportadas por encuestas. En este rol enumeramos dos usos potenciales:

1. **Nowcasting:** Elaboración de *estimaciones flash* para mejorar la puntualidad de los datos, aprovechando la rápida disponibilidad de datos en las fuentes Big Data. Por ejemplo la experiencia de Google Flu Trend u otras extensiones a campos como turismo, venta de

[93] https://en.wikipedia.org/wiki/Reality_mining
[94] https://en.wikipedia.org/wiki/Imputation_(statistics)

vehículos o empleo (Choi and Varian, 2012)[95]. El riesgo en este caso, tal como vimos en el Capítulo 2 para el caso de Google Flu Trends, es que las relaciones entre variables puedan verse afectadas por acontecimientos perturbadores extraordinarios; como por ejemplo una crisis económica grave, que es exactamente cuando las estadísticas son más importantes para guiar la intervención política.

2. **Estimaciones en dominios pequeños:** Las fuentes Big Data se pueden utilizar para la estimación en pequeños dominios, incorporandolas como información auxiliar para calibrar los resultados de las encuestas. Aunque, como vimos en el Capítulo 2, las fuentes Big Data a menudo no cubren completamente la población objetivo de una encuesta, una forma de transferir el poder de las fuentes Big Data es introducir algunas variables disponibles en dichas fuentes a las encuestas. Posteriormente los resultados de la encuesta se pueden calibrar con los totales y desgloses disponibles en la fuente Big Data; mejorando las estimaciones en dominios pequeños. Un ejemplo puede ser introducir preguntas de telefonía móvil en las encuestas a turistas, para luego mejorar las estimaciones en pequeñas áreas mediante calibrado a los datos ofrecidos por las operadoras. La ganancia de precisión en las estimaciones de una variable de enlace dependerá del grado de correlación entre los datos de la encuesta y la fuente Big Data.

   En esta dirección se han desarrollado proyectos de investigación sobre el uso de datos de la red de telefonía móvil para estimar las poblaciones en áreas pequeñas (Makita et al. 2013)[96] (Loibl et al. 2012)[97].

3.2. Riesgos del uso de fuentes Big Data en la estadística oficial

Una excelente revisión de los riesgos del uso de las fuentes Big Data en la estadística pública y de las soluciones para afrontar dichos riesgos se realiza en el documento de trabajo *Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a risk and quality framework* (Wirthmann et al, 2018)[98].

Un proceso de evaluación y gestión de riesgos puede dividirse en varias etapas en las que se incluyen al menos las siguientes: el establecimiento del contexto, la identificación de riesgos, el análisis de

[95]Choi and Varian, "Predicting the Present with Google Trends."
[96] Makita et al., "Can Mobile Phone Network Data Be Used to Estimate Small Area Population? A Comparison from Japan."
[97] Loibl and Peters-Anders, "Mobile Phone Data as Source to Discover Spatial Activity and Motion Patterns."
[98] Wirthmann, A., M. Karlberg, B. Kovachev, and F. Reis. "Structuring Risks and Solutions in the Use of Big Data Sources for Producing Official Statistics – Analysis Based on a Risk and Quality Framework." UNECE, April 24, 2015.

riesgos en términos de probabilidad e impacto, la evaluación de los riesgos y finalmente el tratamiento de los riesgos. En nuestro caso, simplificando el acercamiento realizado en el working paper anteriormente citado, vamos a centrarnos en los siguientes pasos:

1. **Identificación de riesgos:** En este paso deben identificarse eventos que puedan tener impacto en el logro de los objetivos definidos. La identificación debe incluir preguntas relacionadas con el tipo de riesgos, momento del evento, ubicación o cómo los eventos podrían prevenir, degradar, retrasar o mejorar el logro de los objetivos.

2. **Evaluación del riesgo:** El siguiente paso consiste en determinar los controles existentes y analizar los riesgos en términos de probabilidad, así como en términos de los posibles impactos. En la tabla resumen sólo recogemos el impacto de ocurrencia, que se mide utilizando una escala de 1 (insignificante) a 5 (extrema). El producto de la probabilidad y el impacto genera el indicador del nivel de riesgo.

3. **Reacción a los riesgos:** El paso final consiste en decisiones sobre cómo reaccionar ante los riesgos. Algunos riesgos que están por debajo de un nivel predefinido pueden ser ignorados o tolerados. En otros casos los costos de asumir los riesgos pueden ser tan altos que superan los beneficios potenciales. En este caso, la organización puede decidir abandonar la actividad relacionada o también podrían transferirse a terceros, tales como seguros que compensen los costos incurridos.

Tabla 5. Riesgos del uso de fuentes Big Data en la estadística pública

A) Riesgos relacionados con el acceso a los datos		
(R-1) Imposibilidad de acceso a los datos		
Descripción	**Impacto**	**Prevención**
Este riesgo consiste en que una operación estadística diseñada para usar fuentes Big Data no puede tener acceso a las mismas.	El impacto depende de la operación estadística y de la forma en que se utiliza la fuente Big Data. Si la fuente Big Data es indispensable entonces el impacto puede ser muy alto (4 = no es posible producir la operación estadística), mientras que podría ser menor si fuera posible producir la operación estadística (aunque con menor calidad) utilizando otras fuentes, dando como resultado un impacto en el rango de 2-3.	Para reducir este riesgo deben establecerse contactos previos con el proveedor de datos y establecer un acuerdo a largo plazo sobre el acceso a los datos. Además, debe realizarse un análisis jurídico exhaustivo sobre el uso de la fuentes Big Data en la operación estadística. También se deben evaluar las oportunidades de acceso a los datos mediante la legislación existente o futura.

(R-2) Pérdida de acceso a los datos		
Descripción	**Impacto**	**Prevención**
Este riesgo consiste en que una Oficina de Estadística pierde el acceso a una fuente Big Data utilizada en una operación estadística.	Si la operación estadísticas no se puede producir nos encontraríamos con un impacto muy alto (5). En otros casos, cuando la fuente Big Data es de naturaleza complementaria el efecto puede ser una pérdida de calidad, encontrándonos con un impacto en el rango de 2-3.	La estrategia de prevención es similar a la de la falta de acceso a los datos, pero con un mayor énfasis en la vigilancia constante. No hacer depender la operación de una única fuente podría ser una estrategia adecuada pero posiblemente más costosa.
B) Riesgos relacionados con el entorno jurídico		
(R-3) Incumplimiento de la legislación vigente		
Descripción	**Impacto**	**Prevención**
Este riesgo consiste en que una operación estadística basada en una fuente Big Data no ha tomado en consideración alguna norma jurídica de aplicación.	El impacto puede ser crítico (4), si la aplicación de la norma implica la paralización de la operación estadística en producción. Incluso podría ser extremo (5) por el impacto en la reputación de la Oficina Estadística.	Debe realizarse un análisis legal exhaustivo para cualquier operación estadística en todas sus fases. Este análisis puede conducir a una reingeniería de la operación para adaptarla a la normativa de aplicación.
(R-4) Cambios desfavorables en el entorno jurídico		
Descripción	**Impacto**	**Prevención**
Durante la producción de una operación estadística basada en fuentes Big Data podrían surgir cambios normativos incumplidos por la actividad.	El impacto suele ser crítico (4) pues la producción no conforme a una norma puede significar la detención la operación estadística.	Periódicamente debe supervisarse el desarrollo legislativo, con el fin de anticipar medidas correctivas y para defender el interés de las estadísticas oficiales en los foros pertinentes.
C) Riesgos relacionados con la manipulación o usos indebidos de los datos		
(R-5) Incumplimiento de la seguridad de los datos		
Descripción	**Impacto**	**Prevención**
Acceso no autorizado a los datos por parte de terceros, como consecuencia de una mala seguridad en la protección de los datos.	El daño potencial a la reputación de la Oficina Estadística puede ser grande (5). Cuando el incumplimiento de seguridad es del titular de la fuente Big Data el impacto en la reputación de la Oficina Estadística es mucho menor que cuando es producto de un error de la misma. A su vez, en este segundo caso surge un posible alto impacto negativo debido al daño en términos de confianza entre el proveedor y la oficina de estadística.	Se deben tener en cuenta los procedimientos de seguridad del titular de la fuente Big Data. Una manera directa de evitar que una brecha de seguridad en las instalaciones del propietario tenga un gran impacto para la Oficina de Estadística es usar múltiples fuentes para el mismo producto. Una manera de evitar que las brechas de seguridad en la Oficina de Estadística tenga un impacto negativo

		para el propietario de los datos originales es buscar métodos de trabajo que no impliquen la transferencia de datos potencialmente sensibles.

(R-6) Incumplimiento del deber de secreto estadístico

Descripción	Impacto	Prevención
Este riesgo consiste en la posibilidad de divulgar información individual de alguna persona física o jurídica por parte de la Oficina de Estadística.	El daño potencial a la reputación puede ser grande (5). Al igual que con el riesgo de violación de seguridad de datos, el incumplimiento del deber de secreto por parte de la Oficina de Estadística puede tener consecuencias negativas para el titular original de la fuente de datos y perjudicar la relación entre ambas partes. El incumplimiento del deber de confidencialidad de los datos individuales puede generar un alto impacto negativo respecto a la confianza de la población en la Oficina Estadística, perjudicando a su vez la colaboración o participación ciudadana en las encuestas o censos.	Desplegar procedimientos de control del secreto estadístico aplicados tanto a ficheros de microdatos como a tablas estadísticas.

(R-7) Manipulación de la fuente de datos

Descripción	Impacto	Prevención
Las fuentes de datos pueden ser manipuladas por los propietarios de las mismas, o en el caso de redes sociales por parte de terceros. Por ejemplo, se podrían generar muchos mensajes espurios en redes sociales para empujar un índice estadístico, derivado de estos datos, en una dirección determinada. Para los datos aportados voluntariamente puede ocurrir que los voluntarios formen parte de un grupo con el interés de dirigir algún índice estadísticos en alguna dirección.	El problema con las manipulaciones es que pueden durar mucho tiempo sin ser detectadas. Si una manipulación continúa durante mucho tiempo, el impacto en la calidad puede llegar a ser grande. Además seguramente también se vería afectada la credibilidad de la Oficina Estadística como proveedora de datos de calidad. Sin embargo, si una manipulación se descubre a tiempo y luego se divulga, posiblemente se reforzaría su credibilidad. Excepto en casos extraordinariamente malos, se podría imaginar un impacto máximo de (3).	La realización de ejercicios regulares de evaluación comparativa con fuentes alternativas es un posible enfoque preventivo. Estas fuentes alternativas podrían ser tradicionales o no. Basar la estadística en una combinación de fuentes podría evitar que las manipulaciones tuvieran un impacto significativo. En los casos en que se teman manipulaciones por parte de los proveedores, los marcos jurídicos deben ser un buen instrumento preventivo.

(R-8) Percepción pública adversa del gran uso de datos por las estadísticas oficiales

Descripción	Impacto	Prevención
La ciudadanía es muy sensible a las cuestiones de privacidad y al uso de datos personales procedentes de fuentes Big Data para fines no especificados inicialmente..	Como consecuencia de una percepción negativa de un uso inadecuado de una fuente Big Data, puede ocurrir que el titular de la fuente interrumpa los acuerdos de cesión, o que la calidad de	Las medidas preventivas podrían ser la definición de directrices éticas para el uso de fuentes Big Data en las estadísticas oficiales.

Especialmente sensible es el uso secundario de esos datos por parte de las Administraciones Públicas que adoptan medidas administrativas o legales contra los ciudadanos. Este riesgo consiste en la reducción de la confianza por parte de la ciudadanía sobre una organización productora de fuentes Big Data como consecuencia de su uso indebido o por cesiones no previamente informadas.	la misma se vea afectada por deserción de uso de la ciudadanía. El impacto será grave para una operación estadística que esté en producción, porque la actividad podría tener que interrumpirse. Se considera que el impacto va de 2 (menor) a 3 (mayor). Durante la fase de producción, el impacto podría aumentar a 4 (crítico).	Las directrices éticas deben estar fuertemente basadas en principios como el Código de Buenas Prácticas para las Estadísticas Europeas o los principios fundamentales de las estadísticas oficiales. Estas directrices deben acompañarse con una estrategia de comunicación adecuada, para prevenir la percepción adversa por parte de la ciudadanía.

(R-9) Pérdida de credibilidad

Descripción	**Impacto**	**Prevención**
Los usuarios de las estadísticas oficiales suelen tener una gran confianza en la precisión y validez de las estadísticas oficiales pues éstas se producen con un marco metodológico sólido públicamente disponible, así como con calidad contrastada y publicada. Además, la mayoría de los datos estadísticos se basan en la observación, es decir, se derivan de encuestas o censos, que establecen una relación fácilmente comprensible entre la observación y los datos estadísticos. El uso de fuentes Big Data que no se recopilan para el propósito principal de las estadísticas conlleva el riesgo de que esta relación se pierda y los usuarios pierdan la confianza en los datos de las estadísticas oficiales.	El impacto de la ocurrencia del riesgo dependería en gran medida de si las Oficinas Estadísticas pueden probar con éxito la exactitud y validez de los datos estadísticos. En caso de que esto no se pueda lograr, el impacto en términos de pérdida de confianza y credibilidad también podría extenderse a otros dominios estadísticos y no sólo a los vinculados con la fuente Big Data	Las acciones preventivas pasan por desarrollar y publicar una metodología sólida que sea reconocida por la comunidad científica, enriqueciendo los datos con metadatos de calidad, y ejecutando estrictos controles de calidad. Antes de difundir el producto estadístico éste podría publicarse como experimental y fomentado la discusión científica sobre el mismo.

D) Riesgos relacionados con las capacidades

(R-10) Falta de disponibilidad de expertos

Descripción	**Impacto**	**Prevención**
El análisis de las fuentes Big Data requiere de capacidades y herramientas que no son las comunes en las estadísticas oficiales. La falta de disponibilidad de expertos en la Oficina de Estadística puede dar lugar a que está no pueda procesar y analizar adecuadamente las fuentes Big Data.	La Oficina de Estadística que no puede procesar y analizar las fuentes de Big Data debido a la falta de habilidades de su personal puede tener dos posibles consecuencias negativas: 1) la fuente de datos no será explorada, al menos no en todo su potencial; 2) la fuente será utilizada incorrectamente.	Hay dos formas en que las Oficinas de Estadística pueden prevenir proactivamente este riesgo: 1) capacitación y 2) reclutamiento.

(R-11) Pérdida de expertos

Descripción	**Impacto**	**Prevención**

Este riesgo consiste en que las Oficinas de Estadística pierdan su personal después de que hayan adquirido habilidades relacionadas con Big Data.	El impacto de este riesgo sería el mismo que el riesgo de no tener personal con las habilidades adecuadas. Por lo tanto, el impacto sería crítico (4) como se argumentó anteriormente.	El principal instrumento de las Oficinas de Estadística para prevenir este riesgo es proporcionar condiciones de trabajo atractivas a su personal especializado.

Los riesgos también están relacionados con la calidad. La aplicación de un marco de calidad debe permitir el uso de diferentes fuentes y metodologías para obtener resultados con el nivel de calidad necesario para la satisfacción de las necesidades de los usuarios. Al igual que los riesgos, los niveles de calidad pueden derivarse del entorno institucional y el objetivo de cada organización. En este contexto, el entorno institucional define el nivel de riesgo global que una organización está dispuesta a soportar para alcanzar sus metas.

3.3. Marco de calidad para el uso de fuentes Big Data en la estadística pública

En la Resolución sobre los Principios Fundamentales de las Estadísticas Oficiales aprobada por la Asamblea General de NNUU el 29 de enero de 2014, indica que los datos para fines estadísticos pueden obtenerse de todo tipo de fuentes, ya sea encuestas estadísticas o registros administrativos. Sorprende que no haya mención explícita a las fuentes Big Data, siendo una resolución del año 2014, pero de la esencia del principio podríamos extraer que la intención es establecer que la estadística pública pueda realizarse no sólo a partir de encuestas, sino con cualquier tipo de fuente de datos útil para sus fines.

Esta propuesta de pluralismo de fuentes se ordena en el principio mencionado, indicando que éstas se deben seleccionar considerando: su calidad, oportunidad, costo y carga que impondrá a los encuestados. Los criterios de oportunidad, costo y carga a los encuestados son también considerados en el Código de Buenas Prácticas de las Estadísticas Europeas y son fácilmente comprensibles; sin embargo el criterio de calidad necesita ser mejor definido.

Al evaluar o describir la calidad de los datos o productos estadísticos las Oficinas de Estadística utilizan marcos de calidad de referencia. En la actualidad podemos encontrar diversos marcos de calidad en uso, tales como el del FMI[99], Sistema Estadístico Europeo[100], Statistics Canada[101] o el marco

[99] https://unstats.un.org/unsd/accsub/2010docs-CDQIO/Ses1-DQAF-IMF.pdf
[100] http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646
[101] http://www.statcan.gc.ca/pub/12-586-x/12-586-x2017001-eng.pdf

de calidad de la Australian Bureau of Statistics[102]. Estos marcos tienen diversas características en común pues entienden la calidad desde múltiples dimensiones y coinciden en muchas de las dimensiones consideradas.

Algunos marcos mencionados se desarrollaron inicialmente y principalmente como marcos de calidad para encuestas. Sin embargo, las Oficinas Estadísticas con mayor uso de fuentes administrativas amplían su visión con marcos específicos, tales como el marco calidad de Statistics Netherland para fuentes administrativas (Pieet Dass, 2009), o el de Australian Bureau of Statistics.

En ese sentido debemos referenciar una propuesta sobre marco de calidad para el uso de fuentes Big Data en la estadística pública, elaborada por UNECE Big Data Quality[103] e inspirada en el documento *Checklist for the Quality Evaluation of Administrative Data Sources*[104]. Este marco se estructura en tres hiperdimensiones, cada una con sus dimensiones de calidad, que a su vez se organizan en factores a considerar en el análisis.

3.3.1. Marco de calidad para fuentes Big Data

El documento de referencia sobre calidad de las fuentes Big Data para su uso en la estadística pública fue elaborado en 2014 por el UNECE Big Data Quality Task Team, bajo el título *"A Suggested Big Data Quality Framework"*[105] como entregable del proyecto *UNECE/HLG project - The Role of Big Data in the Modernisation of Statistical Production* en el que se presentó el Big Data Quality Framework (BDQF) que vamos a revisar en este apartado.

Este marco de calidad se fundamenta en los marcos de calidad de uso de fuentes administrativas para fines estadísticos, como por ejemplo el marco de calidad de Statistics Netherlands para datos administrativos[106] o el marco de control de calidad de productos estadísticos basados en fuentes administrativas[107] de la Australian Bureau of Statistics.

[102] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+Data+Quality+Framework
[103] UNECE Big Data Quality Task Team. "A Suggested Big Data Quality Framework." UNECE, December 2014.
[104] Piet Daas, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Tóth. "Checklist for the Quality Evaluation of Administrative Data Sources." Discussion Paper. The Hague/Heerlen: Statistics Netherlands, 2009.
[105] UNECE Big Data Quality Task Team. "A Suggested Big Data Quality Framework." UNECE, December 2014.
[106] Daas, Piet, Saskia Ossen, RJWM Vis-Visschers, and Judit Arends-Tóth. "Checklist for the Quality Evaluation of Administrative Data Sources." *Statistics Netherlands Discussion Paper* 9042 (2009).
[107] Pink, Brian. "Quality Management of Statistical Outputs Produced From Administrative Data." Information Paper. Australian Bureau of Statistics, 2011.

El marco proporciona una visión de la calidad desde la perspectiva de tres aspectos incluidos en el Generic Statistical Business Process Model (GSBPM): (1) Captura (proceso 5 del GSBPM), (2) Tratamiento (proceso 6 del GSBPM), (3) Análisis de resultados (proceso 7 del GSBPM); y se elaboró bajo tres principios vinculados con la calidad de los datos y de calidad del propio marco:

1. **Aptitud para el uso:** Es un principio fundamental que considera que la calidad de cualquier fuente de datos o producto sólo puede ser evaluada a la luz de su uso previsto. Este principio, que se utiliza ya en la aplicación de marcos actuales de calidad de datos estadísticos, también es relevante en la evaluación de la calidad de las fuentes Big Data y los productos estadísticos derivados de ellas.

2. **Genérico y flexible:** La intención es producir un marco genérico y flexible que pueda aplicarse en las fases de captura, tratamiento y difusión de resultados, permitiendo evaluar la calidad de diferentes tipos de fuentes Big Data en el contexto de usos distintos.

3. **Eficiencia:** La evaluación general de la aptitud de una fuente Big Data sólo pueden llevarse a cabo una vez evaluadas todas las dimensiones e indicadores de calidad. Con el fin de equilibrar el esfuerzo en la evaluación de la calidad y el valor añadido de la utilización de los datos, se identifican un conjunto de requisitos mínimos que deben cumplir la fuente Big Data.

El marco utiliza una estructura jerárquica compuesta de tres hiperdimensiones, siguiendo la misma organización desarrollada por Statistics Netherlands para fuentes administrativas, con dimensiones de calidad anidadas dentro de cada hiperdimensión. Estas tres hiperdimensiones son:

1. **Fuente:** Incluye dimensiones referidas a las entidades proveedoras de los datos, así como el marco de gestión y regulación de los mismos.

2. **Metadatos:** Aglutina las dimensiones que permiten describir los conceptos y variables utilizados en la fuente de datos, conocer la estructura de los ficheros, así como identificar los procesos que se le aplican.

3. **Datos:** Enumera la dimensiones propias de la calidad de los datos.

Estas hiperdimensiones se subdividen en dimensiones, y para cada una de ellas se enumeran diversos factores a considerar, de acuerdo con la Tabla 6. Estas dimensiones cubren tanto la etapa de captura de los datos (input) como la de divulgación de estadísticas basadas en fuentes Big Data (output).

Tabla 6. Estructura jerárquica del marco de calidad para el uso de fuentes Big Data

Hiperdimensión	**Dimensiones de calidad**	**Etapa**	**Factores a considerar**
Fuente	Entorno institucional	Input	Sostenibilidad de la entidad proveedora de datos Confiabilidad general de los datos Transparencia e interpretabilidad de la entidad proveedora y de los datos
		Output	Tipo de fuente de datos Acuerdos de acceso y control de calidad Rol de la fuente de datos en el producto final
	Privacidad y seguridad	Input	Legislación que afecta a los datos Restricciones de privacidad, seguridad y confidencialidad Percepción ciudadana sobre el uso de los datos
		Output	Legislación Limitaciones reales en el uso de datos Acciones emprendidas
Metadatos	Completitud	Input	Metadatos disponibles, interpretables y completos
	Complejidad	Input	Restricciones técnicas Datos estructurados, semiestructurados o no estructurados Legibilidad de los datos Presencia de jerarquías y anidamientos
		Output	Tratamiento de datos; limitaciones de los resultados de salida
	Accesibilidad y claridad	Output	Accesibilidad de datos y metadatos Definiciones claras, explicaciones Conformidad con los estándares
	Relevancia	Output	Grado en que los datos miden los conceptos que deben medirse para los usos previstos

	Usabilidad	Input	Recursos adicionales necesarios para el tratamiento de los datos Análisis de los riesgos
	Tiempo	Input	Oportunidad Periodicidad Cambios a través del tiempo
	Enlazamiento	Input	Presencia y calidad de variables de enlace Niveles al que se puede realizar enlazamiento
	Coherencia y consistencia	Input	Estandarización
	Validez	Input	Transparencia de métodos y procesos Solvencia de métodos y procesos
Datos	Exactitud y selectividad	Input Output	Error total de la muestra Datos de referencia con los que comparar Selectividad. Problemas de cobertura
	Tiempo	Output	Oportunidad Periodicidad
	Enlazamiento	Input Output	Calidad de las variables de enlace
	Coherencia y consistencia	Input Output	Coherencia entre los metadatos y los datos
	Validez	Input	Coherencia de los procesos y métodos con los datos observados

A) Hiperdimensión fuente

Esta hiperdimensión incluye dimensiones referidas a las entidades proveedoras de los datos, así como el marco de gestión y regulación de los mismos. Las dimensiones de esta hiperdimensión estudian el entorno institucional de la entidad proveedora de datos, así como el marco jurídico que afecta a los datos.

Entorno institucional: Esta dimensión estudia los factores institucionales y organizativos que pueden tener una influencia significativa en la eficacia y la credibilidad del organismo que elabora los datos.

La consideración del ambiente institucional asociado a un producto estadístico es importante, ya que permite la evaluación de factores que pueden influir en la validez, fiabilidad o idoneidad del producto estadístico final. Los factores a considerar en esta dimensión para la etapa de entrada de datos son los siguientes:

1. Sostenibilidad de la entidad proveedora de datos
2. Confiabilidad general de los datos
3. Transparencia e interpretabilidad de la entidad proveedora y de los datos

Asimismo, la información a incluir en la difusión de resultados es:

1) La naturaleza de la fuente Big Data (redes sociales, scrapeo web, datos de sensores, datos de satélites, etc).
2) Los acuerdo bajo los cuales los datos fueron transferidos a la Oficina Estadística
3) Los procesos de control de calidad aplicados a los datos entrantes
4) El rol de los datos en el producto final (por ejemplo, si se utilizó para evaluación comparativa, imputación, etc.)

Privacidad y seguridad: Esta dimensión analiza los factores jurídicos, tanto para la entidad proveedora de datos como para la Oficina Estadística, que puedan tener una influencia significativa en el uso previsto de los datos. Para ello se estudian las limitaciones legales y las restricciones organizacionales o las preocupaciones sobre confidencialidad y privacidad. Los factores a considerar en esta dimensión para la etapa de entrada de datos son los siguientes:

1. Legislación que afecta a los datos
2. Restricciones de privacidad, seguridad y confidencialidad
3. Percepción ciudadana sobre el uso de los datos

Asimismo, la información a incluir en la difusión de resultados es:

1) Legislación relacionada con la producción de los datos, su mantenimiento y acceso
2) Restricciones (privacidad, seguridad, confidencialidad) que limitan el uso de los datos
3) Acciones tomadas para mitigar las posibles percepciones negativas sobre el uso de datos de las partes interesadas

En los informes de calidad cualquier problema relacionado con la confidencialidad o protección de los datos que pueda imponer restricciones a la disponibilidad de los mismos con el nivel de detalle deseado debe ser informado. También debe hacerse referencia a la legislación pertinente e incluso enumerar cualquier paso (por ejemplo, un procedimiento legal específico) que permita la recopilación de los datos en el nivel de detalle deseado.

B) Hiperdimensión metadatos

Esta hiperdimensión aglutina las dimensiones que permiten describir los conceptos y variables utilizados en la fuente de datos, conocer la estructura de los ficheros, así como identificar los procesos que se le aplican.

Completitud: La completitud es el grado en que los metadatos están disponibles para facilitar la comprensión y uso adecuado de los datos. En ella se evalúa la exhaustividad de las descripciones disponibles tanto respecto al entorno institucional como respecto a los datos. Incluye descripciones referidas a las unidades observadas, variables recopiladas, tiempos de referencia, así como de los procedimientos aplicados para el tratamiento de datos y las medidas de calidad o de evaluación cualitativa de la calidad.

El acceso al diseño de registros puede considerarse como un **requisito mínimo** para poder usar una fuente de datos. Especialmente en el caso de datos complejos en el que la utilidad de los mismos dependerá en gran medida del tipo de información disponible acerca de su estructura y codificación. La ausencia de información relevante puede limitar drásticamente el uso potencial de los datos si esta información no puede deducirse o evaluarse a partir de los datos en sí.

En esta dimensión sólo hay que considerar un factor de calidad, que estudia si los metadatos están disponibles y si son interpretables y completos en los siguientes aspectos:

1. Procesos que llevaron a la recopilación de los datos
2. Procesos relacionados con el tratamiento de los datos
3. Descripción de los datos en sí

Dado que los posibles usos estadísticos de una fuente Big Data no se conocen de antemano, los informes de calidad deben incluir descripciones explicativas de los conceptos utilizados en la misma.

Complejidad: La complejidad se refiere a la falta de simplicidad y uniformidad en los datos. La complejidad de la fuente de datos puede ser evaluada en cuatro aspectos diferentes:

1. La estructura de datos: se refiere a la complejidad de las relaciones entre las diversas tablas de las fuentes de datos estructuradas, o la complejidad de los datos semiestructurados o no estructurados para su uso analítico.

2. El formato de los datos: se refiere al estudio de la complejidad de los formatos utilizados para la representación de los datos, por ejemplo los datos espaciales puedes tener diferentes formatos más o menos complejos.

3. Legibilidad de los datos: se refiere al grado de complejidad respecto al uso de códigos uniformes y conocidos, y a la correcta identificación de la causa de datos faltantes.

4. Jerarquías utilizadas en los datos: se refiere al estudio de la complejidad debido a la existencia de relaciones jerárquicas entre registros o entre variables.

Por lo tanto, los factores a considerar en esta dimensión son los siguientes:

1. Restricciones técnicas
2. Estructura
3. Legibilidad de los datos
4. Jerarquías y anidamiento

Asimismo, la información a incluir en la difusión de resultados es:

1) Tratamiento de datos: cómo se ha tratado la complejidad de los datos de entrada durante las etapas de entrada y de producción, con respecto a la estructura de datos, el formato y las jerarquías.

2) Limitaciones reales al uso de salidas estadísticas causadas por la complejidad de la fuente Big Data utilizada.

Accesibilidad y claridad: Se trata de una dimensión vinculada a los productos estadísticos y directamente ligada al Principio 15 del Código de Buenas Prácticas de las Estadísticas Europeas, que

especifica que *"Las estadísticas europeas se presentan de forma clara y comprensible, se difunden de forma adecuada y conveniente, su disponibilidad y acceso tienen carácter imparcial y van acompañadas de metadatos y orientación de apoyo"*. Los factores a considerar para el estudio de esta dimensión son los siguientes:

1. Accesibilidad de datos y metadatos
2. Definiciones claras
3. Conformidad con los estándares

El Código de Buenas Prácticas de las Estadísticas Europeas establece los siguientes indicadores de análisis:

1) Las estadísticas y los metadatos correspondientes se presentan y se archivan de tal forma que facilitan la interpretación adecuada y las comparaciones significativas.

2) Los servicios de difusión utilizan una tecnología moderna de información y comunicación y, si procede, copia impresa tradicional.

3) Cuando es posible, se suministran análisis a medida y se informa de ello al público.

4) El acceso a los microdatos está permitido con fines de investigación y está sujeto a normas o protocolos específicos.

5) Los metadatos están documentados con arreglo a sistemas de metadatos normalizados.

6) Se mantiene informados a los usuarios sobre la metodología de los procesos estadísticos, incluido el uso de datos administrativos[108].

7) Se mantiene informados a los usuarios sobre la calidad de la producción estadística con respecto a los criterios de calidad de las estadísticas europeas.

Relevancia: Se trata de una dimensión vinculada a los productos estadísticos y directamente ligada al Principio 15 del Código de Buenas Prácticas de las Estadísticas Europeas, que especifica que *"Las*

[108] En el caso que nos ocupa este indicador se extendería al uso de fuentes Big Data.

estadísticas europeas satisfacen las necesidades de los usuarios". El factor a considerar para el estudio de esta dimensión es el siguiente:

1. Grado en que los datos miden los conceptos que deben medirse para los usos previstos

El Código de Buenas Prácticas de las Estadísticas Europeas establece los siguientes indicadores de análisis:

1) Existen procedimientos para consultar a los usuarios, controlar la relevancia y la utilidad de las estadísticas existentes por lo que se refiere a sus necesidades y para considerar sus nuevas necesidades y prioridades.

2) Se satisfacen las necesidades prioritarias y se reflejan en el programa de trabajo.

3) Se realiza un control periódico y un seguimiento sistemático de la satisfacción de los usuarios.

Usabilidad: La usabilidad de un conjunto de datos identifica el grado en el que una Oficina Estadística será capaz de trabajar con ellos sin el empleo de recursos especializados o sin una carga significativa sobre los recursos existentes; así como el estudio de la facilidad con que se puede integrar con los sistemas y normas existentes. Los factores a considerar en el estudio de esta dimensión son los siguientes:

1. Recursos: ¿Cuáles son las habilidades necesarias para procesar y almacenar esta información? ¿Se requerirían inversiones adicionales en tecnología?

2. Análisis de riesgos: Identificar las posibles dificultades y ganancias para la Oficina Estadística si se requieren inversiones considerables para utilizar los datos.

Tiempo: Esta dimensión analiza la puntualidad de los datos y su periodicidad. La puntualidad y la frecuencia son dos aspectos importantes de la calidad las fuentes Big Data y de hecho en muchos casos son el valor añadido de las mismas. Los factores a considerar en el estudio de esta dimensión dentro de la hiperdimensión metadatos son los siguientes:

1. Puntualidad: Analiza tiempo requerido para el acceso a los datos desde el momento de su creación.

2. Oportunidad: Tiempo entre la recopilación de datos y el período de referencia al que se refieren los datos, que permita ofrecer datos oportunos para los usuarios.

3. Cambios en el tiempo: Estabilidad en el tiempo de la fuente de datos, lo que permite la elaboración de series homogéneas

A su vez, esta dimensión está ligada a la fase de difusión de resultados. El criterio revisa que las estadísticas se hacen públicas oportuna y puntualmente. Como ejemplo de indicadores tenemos los siguientes:

1) Tiempo transcurrido entre la recepción y la recopilaron los datos; el aumento en el tiempo es un indicador de menor calidad.

2) Tiempo entre la recopilación de datos y el período de referencia al que se refieren los mismos.

Enlazamiento: Esta dimensión se refiere a la facilidad con la que los datos se pueden vincular o fusionar con otros conjuntos de datos. Los factores a considerar en el estudio de esta dimensión en esta hiperdimensión son los siguientes:

1. Variables de enlace: Identificación de variables de enlace en las fuentes de datos.

2. Nivel de vinculación: Granularidad a la que pueden realizarse enlazamientos con otros ficheros.

Mientras que en la hiperdimensión datos sería el siguiente:

3. Calidad de las variables de enlace: Análisis de la calidad de las variables de enlace.

Coherencia y consistencia: Esta dimensión se refiere al grado en que la fuente Big Data utiliza conceptos estándar, además estudia la consistencia interna, la coherencia a lo largo del tiempo, y su consistencia con otras fuentes de datos. Esta dimensión está directamente ligada al Principio 14 del Código de Buenas Prácticas de las Estadísticas Europeas, que especifica que *"Las estadísticas europeas son consistentes internamente a lo largo del tiempo y comparables entre regiones y países; es posible combinar y utilizar conjuntamente datos relacionados procedentes de fuentes diferentes"*.

La coherencia de las estadísticas se refiere a las diferencias entre las cifras provisionales y finales, o a las discrepancias entre las cifras proporcionadas por la fuente Big Data y las cifras proporcionadas por otros conjuntos de datos que describen el mismo fenómeno. A su vez la comparabilidad se refiere a la capacidad de las estadísticas para ser comparadas en el tiempo, en el espacio y entre dominios. Los factores a considerar en el estudio de esta dimensión son los siguientes:

1. Conceptos estandarizados: el uso de estándares para variables clave.

Mientras que en la hiperdimensión datos sería el siguiente:

2. Coherencia con metadatos: el rango de valores encontrados en los datos puede ayudar a determinar la coherencia entre los metadatos y los datos reales.

Validez: La validez de un conjunto de datos indica el grado de coherencia entre lo medido y la medición. Considerando que las fuentes Big Data son fundamentalmente procesos de datificación ajenos a las Oficinas Estadísticas, esta dimensión de la calidad adquiere más relevancia que en encuestas o censos.

Los factores a considerar en el estudio de esta dimensión en esta hiperdimensión son los siguientes:

1. Transparencia de métodos y procesos
2. Solvencia de métodos y procesos

Mientras que en la hiperdimensión datos, en la fase de captura de datos, sería el siguiente:

3. Coherencia de los procesos y métodos con los datos observados

Asimismo, la información a incluir en la difusión de resultados es:

1) Validez convergente: Correlación de la métrica en estudio con otras métricas similares.

2) Utilidad conceptual: Grado en el que la medida es capaz de proporcionar una visión de los fenómenos del mundo real.

3) Validez metodológica: Indicaciones sobre el grado en el que los métodos subyacentes a la métrica son transparentes y teóricamente sólidos.

C) Hiperdimensión datos

Esta hiperdimensión enumera la dimensiones propias de la calidad de los datos.

Exactitud: La exactitud de la información estadística es el grado en el que la información describe correctamente los fenómenos que se diseñó para medir. Como vimos en el Capítulo 2, por lo general, se caracteriza en términos de error cuadrático medio y se descompone tradicionalmente en sesgo (error sistemático) y componentes de la varianza (error aleatorio). También puede ser descrito en términos de las principales fuentes de error que potencialmente causan inexactitud (por ejemplo, la cobertura, el muestreo, la falta de respuesta, etc). En ese sentido nos remitimos a lo especificado en el Apartado 2.2.4. respecto al Total Error Framework.

Como vimos en el Capítulo 2, una de las principales preocupaciones con muchas fuentes Big Data es su representatividad en términos de cobertura. Los factores a considerar en el estudio de esta dimensión, tanto en la captura de datos como en la difusión de resultados, son los siguientes:

1. Enfoque de Total Survey Error para analizar la precisión; incluyendo en particular: sobrecobertura, subcobertura, representatividad, datos faltantes (no observación y no respuesta), ajuste de los datos a los hechos y presencia de anomalías.

2. Datos de referencia, para el análisis de contraste e identificación de desviaciones de los datos o errores en los instrumentos de medida.

3. Representatividad, problemas de selectividad, autoselección.

Tiempo: Desde la perspectiva de la hiperdimensión datos hay dos dimensiones de calidad asociadas al tiempo y los productos estadísticos:

1. Oportunidad: La oportunidad generalmente se mide como el tiempo transcurrido entre el final del período de referencia y la fecha en que las estadísticas están disponibles para los usuarios.

2. Puntualidad: La puntualidad es la diferencia entre la fecha de publicación real y la fecha prevista.

Esta dimensión está ligada al Principio 13 del Código de la Buenas Prácticas de las Estadísticas Europeas, que especifica que *"Las estadísticas europeas se hacen públicas oportuna y puntualmente"*. En el Código se especifican los siguientes indicadores asociados a dicho principio:

1) El calendario de publicación de las estadísticas es conforme a las normas sobre comunicación europeas e internacionales.

2) Se hace pública una hora determinada del día para la publicación de estadísticas.

3) La periodicidad de las estadísticas tiene en cuenta los requisitos de los usuarios en la medida de lo posible.

4) Cuando no se cumple el calendario previsto de publicación, se notifica por adelantado, se dan explicaciones y se establece una nueva fecha.

5) Pueden hacerse públicos resultados preliminares con una precisión aceptable si se considera útil.

Enlazamiento: Esta dimensión se refiere a la facilidad con la que los datos se pueden vincular o fusionar con otros conjuntos de datos. Los factores a considerar en el estudio de esta dimensión, en la hiperdimensión, datos son los siguientes:

1. Variables de enlace: Identificación de variables de enlace en las fuentes de datos.

2. Nivel de vinculación: Granularidad a la que pueden realizarse enlazamientos con otros ficheros.

Mientras que en la hiperdimensión metadatos sería el siguiente:

3. Calidad de las variables de enlace: Análisis de la calidad de las variables de enlace.

Coherencia y consistencia: Esta dimensión se refiere al grado en que la fuente Big Data utiliza conceptos estándar, además estudia la consistencia interna, la coherencia a lo largo del tiempo, y su consistencia con otras fuentes de datos. Esta dimensión está directamente ligada al Principio 14 del Código de Buenas Prácticas de las Estadísticas Europeas, que especifica que *"Las estadísticas europeas son consistentes internamente a lo largo del tiempo y comparables entre regiones y países; es posible combinar y utilizar conjuntamente datos relacionados procedentes de fuentes diferentes"*. Los factores a considerar en el estudio de esta dimensión en esta hiperdimensión es el siguiente:

1. Coherencia con metadatos: el rango de valores encontrados en los datos puede ayudar a determinar la coherencia entre los metadatos y los datos reales.

Mientras que en la hiperdimensión metadatos sería el siguiente:

2. Conceptos estandarizados: el uso de estándares para variables clave.

3.3.2. Informes de calidad de fuentes Big Data y productos asociados

Un estudio de la literatura acerca de los datos administrativos revela los diferentes informes de calidad utilizados por las Oficinas Estadísticas[109]. En este apartado vamos a proponer algunos informes de calidad aplicables a las fuentes Big Data, siguiendo para ello las líneas propuestas para las fuentes administrativas:

A) Informes internos de calidad

En este apartado partimos de la idea de que el informe de calidad es para el personal estadístico usuario de las fuentes Big Data, que trabaja para una Oficina Estadística, y cuyo objetivo es producir estadísticas públicas.

Por tanto, el informe interno de calidad ha de permitir evaluar el grado en que los datos son apropiados para el propósito particular del usuario. El informe se abordará fundamentalmente para dirigirlo al personal técnico de la Oficina Estadística y puede ser muy detallado, ya que deberá recoger todas las dimensiones relacionadas con la captura y tratamiento de datos señalas en el apartado anterior.

Como hemos visto en el apartado de roles potenciales de la fuentes Big Data en la estadística pública, una misma fuentes puede tener diversos roles, por ello se necesitan **dos tipologías de informes**

[109] Eurostat, "Quality Assessment of Administrative Data for Statistical Purposes."

internos de calidad. Un tipo se referirá a las particulares de las fuentes Big Data y el otro se referirá a los productos estadísticos particulares basados en fuentes Big Data.

El **informe inicial de calidad de la fuente Big Data** evaluará la calidad de los datos, en general, sin mención a los productos específicos, a menos que la fuente contribuya a un solo producto. En todo caso se deben mencionar los productos estadísticos a los que aporta datos y se indicará al lector los informes de calidad sobre dichos productos.

En el **informe de calidad del producto** se evaluarán las características de las fuentes Big Data que sean pertinentes para el producto. En los mismos también se harán referencia a los informes iniciales de calidad de las fuentes Big Data utilizadas. Por lo tanto, los dos tipos de informes se superponen parcialmente, pero básicamente se complementan entre sí. Dado que los informes deben estar disponibles electrónicamente, se puede compilar fácilmente un informe interno completo sobre una fuente de datos particular usando su informe específico y las partes pertinentes de los informes específicos del producto.

Estos informes deberían realizarse una vez, con las revisiones necesarias cuando hayan modificaciones sustanciales. A su vez podría realizarse un informe ejecutivo para los proveedores de las fuentes de datos, incluyendo instrucciones o procedimientos para la mejora de la calidad de los datos. Estos informes se deben organizar y estructurar según las hiperdimensiones y dimensiones señaladas en el Marco de Calidad de las Fuentes Big Data:

Estructura del informe inicial de calidad de la fuente Big Data:

I. Calidad de la fuente de datos
    A. Entorno institucional
    B. Privacidad y seguridad
II. Calidad de los metadatos
    A. Completitud
    B. Complejidad
    C. Usabilidad
    D. Tiempo
    E. Enlazamiento
    F. Coherencia y consistencia
    G. Validez
III. Calidad de los datos

A. Exactitud y selectividad
B. Enlazamiento
C. Coherencia y consistencia
D. Validez

IV. Usos previstos para la fuente de datos

V. Referencias a los informes de calidad de los productos asociados

Estructura del informe de calidad del producto:

I. Calidad de la fuente de datos
A. Entorno institucional
B. Privacidad y seguridad

II. Calidad de los metadatos
A. Complejidad
B. Accesibilidad y claridad
C. Relevancia

VI. Calidad de los datos
A. Exactitud y selectividad
B. Tiempo
C. Enlazamiento
D. Coherencia y consistencia

VII. Referencias a los informes iniciales de calidad

VIII. Referencias a los informes de calidad de enlazamiento

B) Informes de calidad al combinar fuentes de datos

El uso eficiente de las diversas fuentes de datos, conduce a las Oficinas Estadísticas a utilizar combinaciones de las mismas. Ya hemos mencionado este problema en este capítulo y también lo hemos señalado como una de las dimensiones de calidad. El propósito de esta sección es brindar una presentación más completa del tema.

La situación es la siguiente: existen datos sobre una cierta población de interés en al menos dos fuentes de datos separadas. La Oficina de Estadística combina las fuentes separadas y crea un nuevo conjunto de datos, cada registro corresponde a una unidad de población y contiene toda la información sobre esta unidad previamente dispersa en múltiples fuentes.

La calidad de los datos combinados no es siempre la suma de calidad de las partes. Si la combinación se realiza de manera efectiva, la calidad resultante será superior a la de cada fuente considerada aisladamente; si, por otro lado, la combinación es pobre, la calidad resultante será inferior a la de las fuentes separadas.

En el informe de integración no hará falta incorporar información ya señalada en los informes de calidad de la fuente o del producto, basta con añadir referencias a los mismos. Los elementos que, por lo tanto, deben incluirse en un informe de calidad sobre el uso combinado de múltiples fuentes de datos son:

1. **Conceptos:** Una definición detallada de la población cubierta por cada fuente, las unidades de población en que se divide y su tiempo de referencia. También una definición de población de referencia, unidades de población y tiempo de referencia de los datos combinados. Demostración de cómo se identificaron las unidades de población y el tiempo de referencia de los datos combinados. Comparación de la población de referencia con la población objetivo (la que el Instituto de Estadística intenta cubrir con los datos combinados).

2. **Variables de identificación:** Presentación de la variable (o combinación de variables) que se utiliza para la identificación de registros en cada fuente. Presentación de las tasas de datos faltantes, grado de homogeneidad y grado de errores de estas variables, antes y después del tratamiento de datos aplicado para mejorar su calidad.

3. **Metodología de enlazamiento:** Una breve presentación del algoritmo de enlazamiento y referencia a documentos metodológicos relevantes o publicaciones científicas que lo describen en su totalidad.

4. **Tasas de error de enlazamiento:** Presentación de tasas de coincidencias falsas y tasas falsas de no coincidencia. El primero es el porcentaje de enlazamiento reportados que no corresponden a enlazamientos verdaderos, mientras que el último es el porcentaje de enlazamientos genuinos que no fueron reportados. El cálculo de las tasas requiere la aplicación de un piloto del método de enlazamiento de registros sobre archivos de datos para los cuales la Oficina de Estadística sabe qué pares de registros coinciden.

5. **Métodos alternativos:** Si el Instituto de Estadística lo desea, puede informar sobre los dos elementos anteriores para cualquier otro método de enlazamiento que se haya aplicado a los

conjuntos de datos de prueba. Esto ayudará a los terceros a juzgar si se ha elegido el método más apropiado. Los datos sobre el costo de los métodos en dinero, esfuerzo y tiempo serán útiles a este respecto.

6. **Superposición de fuentes:** Presentación del porcentaje de registros de cada fuente que se combinaron con las otras fuentes. Presentación del porcentaje correspondiente de la población objetivo cubierta por los datos combinados. Obviamente, si se usa una pequeña porción de los registros de cada fuente y se cubre un pequeño porcentaje de la población objetivo, la combinación de fuentes puede no ser beneficiosa.

7. **Calidad del conjunto de datos producido:** La calidad del conjunto de datos resultante de la combinación de las fuentes se informa como la de cualquier otro conjunto de datos. El Instituto de Estadística debe informar adicionalmente sobre el tratamiento aplicado a los datos al combinarlos. Aquí se pueden referenciar los informes de calidad del producto, enumerados en la sección anterior.

C) Informes de calidad para terceros

Esta sección, a diferencia de las anteriores, se refiere a la información que una Oficina de Estadística transmite a terceros sobre la calidad de los productos estadísticos que ha producido utilizando fuentes Big Data.

En este caso la estructura de los informes debe seguir la propuesta en el EURO-SDMX METADATA STRUCTURE (ESMS)[110] cuyo objetivo es documentar las metodologías, la calidad y los procesos de producción estadística en general. Para ello utiliza 21 conceptos de alto nivel, con un desglose de subtemas, derivados de los *Cross Domain Concepts* recogidos en el documento *SDMX Content Oriented Guidelines (2009)*[111] que en febrero de 2016 se integró dentro del *SDMX Glossary* (que a su vez reemplazó el *Metadata Common Vocabulary - MCV*). Para la difusión de indicadores se debería utilizar el ESMS-IP, que proporciona el conjunto de metadatos de referencia en la Unión Europea.

En el informe deberían incluirse al menos la información señalada en el marco de calidad :

1) La naturaleza de la fuente Big Data (redes sociales, scrapeo web, datos de sensores, datos de satélites, etc).

[110] http://ec.europa.eu/eurostat/data/metadata/metadata-structure
[111] https://sdmx.org/wp-content/uploads/01_sdmx_cog_annex_1_cdc_2009.pdf

2) Los acuerdo bajo los cuales los datos fueron transferidos a la Oficina Estadística

3) Los procesos de control de calidad aplicados a los datos entrantes

4) El rol de los datos en el producto final (por ejemplo, si se utilizó para evaluación comparativa, imputación, etc.)

5) Legislación relacionada con la producción de los datos, su mantenimiento y acceso

6) Restricciones (privacidad, seguridad, confidencialidad) que limitan el uso de los datos

7) Acciones tomadas para mitigar las posibles percepciones negativas sobre el uso de datos de las partes interesadas

8) Tratamiento de datos: cómo se ha tratado la complejidad de los datos de entrada durante las etapas de entrada y de producción, con respecto a la estructura de datos, el formato y las jerarquías.

9) Limitaciones reales al uso de salidas estadísticas causadas por la complejidad de la fuente Big Data utilizada.

10) Validez convergente: Correlación de las métricas en estudio con otras métricas similares.

11) Utilidad conceptual: Grado en el que las medidas son capaces de proporcionar una visión de los fenómenos del mundo real.

12) Validez metodológica: Indicaciones sobre el grado en el que los métodos subyacentes a la métrica son transparentes y teóricamente sólidos.

3.4. Inventario de casos de uso

El Equipo de Trabajo sobre Big Data, adscrito al High-Level Group for the Modernisation of Official Statistics, desarrolló en junio de 2013 una primera clasificación[112], posiblemente incompleta, de tipos de fuentes Big Data.

1. **Redes sociales, información de origen humano:** Se refiere al registro de información de experiencias humanas, que en la antigüedad se plasmaba en libros y obras de arte, y más tarde en fotografías, audio o video. Estos tipos de fuentes suelten tener información poco estructurada. En este grupo encontramos los siguientes subgrupos:

   1100. Redes sociales: Facebook, Twitter, Tumblr, etc.
   1200. Blogs y comentarios
   1300. Documentos personales
   1400. Fotos: Instagram, Flickr, Picasa, etc.
   1500. Videos: Youtube, etc.
   1600. Búsquedas en Internet
   1700. Contenido de datos móviles: mensajes de texto
   1800. Mapas generados por el usuario
   1900. Correo electrónico

2. **Sistemas empresariales tradicionales que recogen datos mediados por procesos.** Estos procesos registran y supervisan eventos comerciales de interés, como registrar a un cliente, fabricar un producto, tomar un pedido, etc. Los datos son altamente estructurados e incluyen transacciones, tablas de referencia y relaciones, así como metadatos que explican su contexto. En algunos casos estas fuentes de datos pueden incluirse dentro de la categoría de fuentes de datos administrativas. En este grupo encontramos los siguientes subgrupos:

   21. Datos producidos por agencias públicas
      2110. Registros médicos
   22. Datos producidos por las empresas
      2210. Transacciones comerciales
      2220. Registros bancarios / de acciones

[112] https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data

2230. e-Commerce

2240. Tarjetas de crédito

3. **Internet de las cosas, datos generados por máquinas:** A medida que los sensores proliferan y los volúmenes de datos crecen, se está convirtiendo en un componente cada vez más importante de la información almacenada y procesada por muchas empresas. Su naturaleza bien estructurada es adecuada para el procesamiento informático, pero su tamaño y velocidad van más allá de los enfoques tradicionales. En este grupo encontramos los siguientes subgrupos:

31. Datos de sensores

311. Sensores fijos

3111. Domótica

3112. Sensores del tiempo atmosférico / contaminación

3113. Sensores de tráfico / webcam

3114. Sensores científicos

3115. Videos / imágenes de seguridad / vigilancia

312. Sensores móviles (rastreo)

3121. Ubicación del teléfono móvil

3122. Coches

3123. Imágenes de satélite

32. Datos de sistemas informáticos

3210. Logs

3220. Web logs

3.4.1. Catálogos de NN.UU. de casos de uso

Tras el Meeting on the Management of Statistical Information Systems[113] de 2013, sobre la gestión de los sistemas de información estadística, se estableció un equipo de trabajo para resolver los problemas clave con el uso de Big Data para las estadísticas oficiales, identificar las acciones prioritarias y formular una propuesta de proyecto. Como parte de su labor, el equipo de trabajo reunió una cantidad significativa de información sobre el uso de Big Data en varias Oficinas Estadísticas nacionales e internacionales. El Equipo de Trabajo reconoció que esta información podría ser la base de un recurso útil, por lo que decidió desarrollar y publicar un **inventario de casos de uso de fuentes Big Data en**

[113] https://www.unece.org/index.php?id=31369

la estadística pública[114]. En 2014, el UNECE High-Level Group for the Modernisation of Official Statistics asumió la responsabilidad de mantener y actualizar el inventario, si bien en la actualidad dicho inventario se ha quedado obsoleto, aunque como se señala en el propio inventario no es ese su objetivo: *"No es realista suponer que el inventario pueda registrar toda la información posible relacionada con el uso de fuentes Big Data para las estadísticas oficiales (...). Por lo tanto, el inventario no pretende ser exhaustivo, sino que busca incluir al menos los recursos clave que sean más valiosos para la comunidad de estadísticos oficiales."*

Por otra parte, a partir del citado inventario, el Global Working Group (GWG) on Big Data for Official Statistics de la División Estadística de Naciones Unidas mantiene conjuntamente con el Banco Mundial otro inventario paralelo más ligero en metadatos y posiblemente más actualizado, bajo la marca **Big Data Project Inventory**[115] .

En la actualidad el inventario de la UNECE recoge información, según ficha normalizada, de veintiocho proyectos desarrollados por distintas Oficinas Estadísticas. A continuación se enumeran dichos proyectos, en el Anexo B se detallan los proyectos y se enlazan a sus correspondientes fichas normalizadas:

1	Australia (ABS) - Social Linked (semantic) Data Processing for Various Statistical Uses
2	Statistics Canada - Non-Residential Buildings Inventory: Feasibility Study
3	National Bureau of Statistics of China - Big Data Enterprise Statistical Indicator Ten-day Report
4	National Bureau of Statistics of China - Online Price Changes of Means of Production in Circulation Area in Shandong Zhuochuang
5	Statistics Finland - Traffic sensor data for commuting statistics
6	ESCAP - Developing a Curriculum and Training Modules on Using Big Data for Official Statistics
7	Eurostat - Feasibility study on the use of mobile positioning data for tourism statistics
8	Eurostat - Multi-purpose consumer price statistics, sub-project Scanner Data/Web Scraping
9	UNECE HLG Big Data Project 2014 - Sandbox Task Team Social Media - Sentiment Analysis
10	UNECE HLG Big Data Project 2014 - Sandbox Task Team Traffic Loops
11	UNECE HLG Big Data Project 2015 - Sandbox Task Team Enterprise Web sites
12	UNECE HLG Big Data Project - Sandbox Task Team Consumer Price Index (Scanner Data)

[114] https://statswiki.unece.org/display/BDI
[115] https://unstats.un.org/bigdata/inventor

13	UNECE HLG Big Data Project - Sandbox Task Team Consumer Price Index (Web scraped data)
14	UNECE HLG Big Data Project - Sandbox Task Team Job Vacancies
15	UNECE HLG Big Data Project - Sandbox Task Team Satellite Data
16	UNECE HLG Big Data Project - Sandbox Task Team Smart Meters
17	Italy (Istat) - Internet as a Data Source for ICT Usage by Enterprises and Public Institutions
18	Italy (Istat) - Persons and Places: Mobility Estimates based on Mobile Phone Data
19	Italy (Istat) - Specific purpose geographic basins and population statistics using mobile phone tracking data
20	Italy (Istat) - Use of scanner data for consumer price index
21	Mexico (INEGI) - Tweet Analysis
22	Central Statistical Office of Poland - Estimating demand on labour market by analysing job offer portals
23	Romania National Statistical Institute (INS) - Using scanner data
24	Statistics South Africa - Assessing use of scanner data for compiling the Consumer Price Index
25	Switzerland (FSO) - Price collection with scanner data
26	UK (ONS) Housing Website data to help improve address register
27	UK (ONS) Webscraped prices for price indices
28	United Kingdom (ONS) - Smartmeter type data for household structure/size and occupancy

3.4.2. El proyecto Sandbox

En 2014, el UNECE High-Level Group for the Modernisation of Official Statistics lanzó la iniciativa Sandbox para crear un entorno de colaboración alojado en el Irish Centre for High-End Computing para el estudio mediante proyectos de la capacidad de las fuentes Big Data para la estadística pública.

En esta iniciativa, más de 40 expertos de organizaciones estadísticas nacionales e internacionales trabajaron para identificar y abordar los principales desafíos del uso de fuentes de Big Data para las estadísticas oficiales. Los países involucrados fueron: Austria, Francia, Alemania, Hungría, Irlanda, Italia, México, Países Bajos, Polonia, Serbia, Rusia, Eslovenia, España, Suecia, Suiza, Turquía, Emiratos Árabes Unidos, Reino Unido y Estados Unidos de América. Las organizaciones internacionales fueron: Eurostat, UNECE (Comisión Económica de las Naciones Unidas para Europa), UNSD (División de Estadística de las Naciones Unidas) y OCDE (Organización para la Cooperación y

el Desarrollo Económicos). En 2015, la iniciativa Sandbox abordó, a través de equipos multinacionales, el estudio de cuarto fuentes de datos:

1. Visitas a ciertas entradas de la Wikipedia. En el Sandbox analizaron las visitas de lugares turísticos (sitios del patrimonio de la Unesco), descubriendo que esta fuente es un buen indicador de la popularidad de estos sitios.

2. Datos de comercio de la United Nations ComTrade Database. Esta fuente de datos global se utilizó principalmente para probar tecnología Big Data en una fuente de datos tradicional. Se utilizaron ocho paquetes de software diferentes, desde herramientas básicas de Hadoop para almacenar y limpiar datos, hasta herramientas estadísticas como R-Hadoop para el análisis de la asimetría bilateral así como librerías y paquetes de visualización de redes.

3. Datos de redes sociales: los datos de las redes sociales, donde los usuarios publican una cantidad considerable de información que de otro modo no estaría disponible, los convierte en una de las fuentes Big Data de mayor interés para las Oficinas Estadísticas. En 2015, se recopilaron datos de Twitter para México, Italia y el Reino Unido.

   Los datos mexicanos se usaron para muchos propósitos:

   - Estudiar la movilidad de las personas dentro del país (turismo interno) y desplazamientos fronterizos diarios/semanales entre EE. UU. y México.

   - Estudiar el uso de la infraestructura vial, utilizando los tweets generados mientras las personas viajan.

   - Estudiar la influencia de ciertas ciudades en los desplazamientos realizados por los habitantes de las localidades rurales y pequeños pueblos cercanos.

   - Analizar el sentimiento de las personas a nivel regional y temporal, para ser comparado con las encuestas de satisfacción del cliente tradicionales.

   Los datos del Reino Unido se utilizaron para estudiar la movilidad de las personas hacia los centros universitarios en diferentes épocas del año, utilizando algoritmos avanzados.

Finalmente, el grupo comenzó a recolectar tweets generados en la ciudad de Roma, para analizar los movimientos de los turistas durante el Jubileo Católico 2015-2016. Estos datos también se usaron para estudiar las reacciones de turistas y ciudadanos romanos durante y después de los ataques terroristas de París en noviembre de 2015.

4. Sitios web de empresas: este equipo trabajó para tratar de encontrar anuncios de vacantes de empleo en las páginas de los sitios web corporativos. Estudiaron cómo recopilar direcciones de Internet (URL) de las empresas, tratando de integrar diferentes fuentes, incluidas encuestas y datos administrativos.

   El equipo creó y probó un prototipo de herramienta de TI para identificar anuncios de trabajo y la metodología para crear posibles estadísticas. La herramienta de TI diseñada y probada se compone de cinco módulos (Spider, Downloader, Splitter, Determinator and Classifier) y, en principio, puede generalizarse y utilizarse para cualquier tipo de extracción de contenido de páginas web.

   Las primeras estadísticas creadas fueron prometedoras, pero se necesitan más pruebas y mejoras en la eficacia de la herramienta de TI. Estas actividades continuaron en un importante proyecto de la Unión Europea.

Los resultados del proyecto Sandbox aportan importantes indicaciones sobre el uso de Big Data para las estadísticas oficiales, resultados que pueden dirigir los esfuerzos futuros en este campo. El resultado más importante que surgió de los experimentos es que las estadísticas basadas en fuentes de Big Data serán diferentes de las que tenemos hoy.

Asimismo se destacó que los estadísticos oficiales que se ocupan de las fuentes Big Data deberían aprender a aceptar la inestabilidad general de las fuentes, incluso a corto o mediano plazo. Por ejemplo, las estadísticas de acceso de Wikipedia mostraron una caída general en el número total de accesos desde el momento en que se lanzó la versión móvil de Wikipedia. Del mismo modo, Twitter tuvo un número significativamente menor de tweets geolocalizados después de que Apple cambió las opciones predeterminadas para sus productos. La consistencia de la serie temporal se vería afectada por tales eventos, que deberían tratarse de maneras específicas y novedosas.

Por otra parte se señaló, tal como hemos visto en este Capítulo, que las características de las fuentes de Big Data afectan la calidad de las estadísticas oficiales, que está estrictamente relacionada con la

calidad de las fuentes. En el proyecto se concluye que producir estadísticas basadas en Big Data significaría aceptar diferentes nociones de calidad.

Otro resultado práctico importante se asoció con el acceso a las fuentes de Big Data. **Las altas expectativas iniciales sobre las oportunidades de Big Data tuvieron que enfrentarse a la complejidad de la realidad**.

El hecho de que los datos se produzcan en grandes cantidades no significa que estén disponibles inmediata y fácilmente para producir estadísticas. Las fuentes de calidad son de difícil acceso, independientemente del precio de las mismas, en ese sentido los datos de los teléfonos móviles representan un ejemplo notable. Por otro lado, las fuentes de datos de acceso público están limitadas en términos de calidad y, por lo tanto, requieren una cantidad significativa de procesamiento para su uso en la estadística pública.

Por otra parte, los problemas jurídicos pueden estar inesperadamente presentes incluso en fuentes de libre acceso. Por ejemplo, aunque los sitios web se pueden consultar sin límites a través de un navegador, a veces existen restricciones legales y técnicas para scrapear y descargar su contenido de forma automatizada. Otro ejemplo es que Twitter otorga acceso público a un subconjunto de tweets, pero tiene limitaciones precisas sobre la cantidad de datos que se pueden recopilar y el propósito del análisis.

Asimismo el proyecto llega a importantes conclusiones respecto al acceso a datos. En ese sentido sentencia que para alcanzar el siguiente nivel en el uso de fuentes Big Data para las estadísticas oficiales, se necesitan dos tipos de acciones:

1. En primer lugar, las negociaciones y los acuerdos con los proveedores, a quienes se debería alentar a compartir sus datos con las organizaciones estadísticas.

2. En segundo lugar, la intervención política a nivel legislativo, para facilitar el uso de las fuentes de Big Data con fines estadísticos y para superar los problemas de jurídicos.

En resumen, las fuentes de Big Data deben tratarse del mismo modo que las tradicionales en lo que respecta a su uso para las estadísticas públicas. **Las organizaciones estadísticas deberían recibir una tratamiento especial de acceso preferencial, en un marco de estrictas garantías de preservación de la privacidad.**

Los hallazgos relacionados con la tecnología mostraron posibles formas de mejorar el uso de las herramientas de TI en las organizaciones estadísticas, tanto para hacer frente a fuentes Big Data como para mejorar la eficacia general del tratamiento de datos. En particular, aunque las tecnologías Big Data se concibieron específicamente para manejar datos de gran tamaño, también se pueden usar para procesar datos de tamaño mediano de una manera más eficiente que la ofrecida por las herramientas tradicionales. Sin embargo, el aprendizaje y uso de estas herramientas requerirá una mayor colaboración entre el personal estadístico y el sector de tecnologías de la información.

3.4.3. La experiencia de la Oficina Estadística de UK

El Gobierno de Reino Unido ha creado el partenariado **Government Data Science Partnership** como una colaboración entre el Government Digital Service (GDS), Office for National Statistics (ONS) y el Government Office for Science con el fin de ayudar a los departamentos del gobierno a reconocer el potencial de la denominada ciencia de datos y para apoyar el desarrollo de habilidades y herramientas con el fin de impulsar una mayor utilización entre las Administraciones Públicas.

Entre sus actividades está por ejemplo el programa Data Science Accelerator que es un programa de formación pública sobre ciencia de datos a través de desafíos reales, que se desarrolla con el apoyo de tutores con experiencia. Además durante 2017 celebró su primera Conferencia Gubernamental de Ciencia de Datos, con el ánimo de crear un espacio de encuentro para compartir experiencias de uso.

En este contexto la Office for National Statistics (ONS) ha desarrollado dos líneas de actuación ejemplares para impulsar el uso de fuentes Big Data en la estadística pública, tal como señala en su artículo *Big Data at ONS*[116]. Por una parte ha constituido un grupo de trabajo interno sobre el uso de fuentes Big Data y por otra han lanzado el campus formativo *ONS Data Science Campus.*

Los resultados del grupo de trabajo sobre el uso de fuentes Big Data los publican en documentos metodológicos dentro de la serie *ONS methodology working paper series* y por otra parte sus desarrollos de software son de acceso público a través de su espacio en GitHub[117]. Este espacio GitHub es un excelente inventario de proyectos experimentales sobre el uso de fuentes Big Data para la estadística pública.

[116] https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theonsbigdataproject
[117] https://github.com/ONSBigData

En otro sentido,los objetivos del ONS Data Science Campus son tanto investigar el uso de nuevas las fuentes de datos para su uso público, como ayudar a construir capacidades en materia de ciencia de datos al servicio del Reino Unido, siendo responsable del programa Data Science Accelerator.

Por lo tanto, el ONS Data Science Campus cubre dos retos importantes en esta génesis del uso de fuentes Big Data por parte de las Administraciones Públicas. Por una parte concentra la actividad investigadora en materia de Big Data para bienes públicos, y por otra fomenta la formación tanto de científicos de datos como de empleaos públicos.

Bibliografía del Capítulo 3

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew Shapiro. "Using Social Media to Measure Labor Market Flows." Cambridge, MA: National Bureau of Economic Research, March 2014. http://www.nber.org/papers/w20010.pdf.

Bakker, BFM. "Micro-Integration." *Statistical Methods Paper (201108), The Hague/Heerlen: Statistics Netherlands*, 2011. https://www.cbs.nl/NR/rdonlyres/DE0239B4-39C6-4D88-A2BF-21DB3038B97C/0/2011x3708.pdf.

Choi, Hyunyoung, and Hal Varian. "Predicting the Present with Google Trends." Economic Record 88, no. s1 (2012): 2–9.

Daas, Piet, Saskia Ossen, RJWM Vis-Visschers, and Judit Arends-Tóth. "Checklist for the Quality Evaluation of Administrative Data Sources." *Statistics Netherlands Discussion Paper* 9042 (2009). http://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60.

D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu. *Statistical Matching: Theory and Practice*. John Wiley & Sons, 2006.

Eurostat. "QUALITY ASSESSMENT OF ADMINISTRATIVE DATA FOR STATISTICAL PURPOSES," 22. Luxembourg: Eurostat, n.d. http://ec.europa.eu/eurostat/documents/64157/4374310/36-QUALITY-ASSESSMENT-ADMINISTRATIVE-DATA-STATISTICAL-PURPOSES_2003.pdf/37373e67-d69c-4215-b727-5b036393b80f.

Florescu, Denisa, Martin Karlberg, Fernando Reis, Pilar Rey Del Castillo, Michail Skaliotis, and Albrecht Wirthmann. "Will 'big Data' Transform Official Statistics." In *Q2014–European Conference on Quality in Statistics*, 2014. http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.

UNECE Big Data Quality Task Team. “A Suggested Big Data Quality Framework.” UNECE, December 2014. https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2.

Karlberg, Martin, and Michail Skaliotis. “Big Data for Official Statistics—strategies and Some Initial European Applications.” In *UN Economic Commission for Europe: Conference on European Statisticians, September*, 25–27, 2013.

Leulescu, A, and M Agafitei. “Statistical Matching: A Model Based Approach for Data Integration.” *Eurostat-Methodologies and Working Papers*, 2013. http://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF.

Loibl, Wolfgang, and Jan Peters-Anders. “Mobile Phone Data as Source to Discover Spatial Activity and Motion Patterns.” *G1_Forum*, 2012, 524–533.

Makita, Naoki, Masakazu Kimura, Masayuki Terada, Motonari Kobayashi, and Yuki Oyabu. “Can Mobile Phone Network Data Be Used to Estimate Small Area Population? A Comparison from Japan.” *Statistical Journal of the IAOS* 29, no. 3 (2013): 223–232.

Pink, Brian. “Quality Management of Statistical Outputs Produced From Administrative Data.” Information Paper. Australian Bureau of Statistics, 2011. http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/67DAB50CE3A21F21CA25785A000DBA5E/$File/15220_Mar%202011.pdf.

Reis, Fernando. “The Modernisation of European Social Statistics.” Rome: Eurostat, 2012. https://www.destatis.de/EN/AboutUs/Events/DGINS/Document_PaperEUROSTAT.pdf?__blob=publicationFile.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe, and Frederick G. Conrad. “Social Media Analyses for Social Measurement.” *Public Opinion Quarterly* 80, no. 1 (2016): 180–211. doi:10.1093/poq/nfv048.

Wirthmann, A., M. Karlberg, B. Kovachev, and F. Reis. “Structuring Risks and Solutions in the Use of Big Data Sources for Producing Official Statistics – Analysis Based on a Risk and Quality Framework.” UNECE, April 24, 2015. https://ec.europa.eu/eurostat/cros/content/structuring-risks-and-solutions-use-big-data-sources-producing-official-statistics-%E2%80%93_en.

Capítulo 4. Conclusiones y propuestas de acción

4.1. Conclusiones

La sociedad de finales de Siglo XX y principios del Siglo XXI está cambiando rápidamente en muchos aspectos, entre ellos los vinculados al mundo de la información. Vivimos en la época SMAC (Social, Mobile, Analytics, Cloud) donde las personas, muchas de ellas denominadas nativas digitales, no conciben su vida sin un dispositivo móvil a través del que se relacionan con el mundo. Este estilo de vida, al que ya se llama digital, genera un tsunami de cambios y una verdadera montaña de datos en flujo constante.

A esto se suma lo que Kevin Ashton denominó Internet de las Cosas (IoT, por sus siglas en inglés), concepto que se refiere a la interconexión digital de objetos cotidianos con internet. La idea subyacente es que los objetos se equipan con sensores, que generan datos que se comunican por Internet. En esta línea nos encontramos con las Ciudades Inteligentes (Smart Cities), en las que los sistemas de iluminación, la señalización viaria y otros servicios públicos sensorizados son importantes generadores de datos públicos.

Este cambio de contexto en el mundo de la información tiene implicaciones directas en la Oficinas de Estadística. Entre ellas encontramos muchas cuestiones prácticas, pero también existe una importante y estratégica: **¿Qué posición quieren ocupar las Oficinas de Estadística en el futuro sociedad de la información?**

En 2001 Douglas Lane[118] propuso tres características que distinguían a lo que ahora denominamos Big Data: volumen, velocidad y variedad. Tradicionalmente las Oficinas de Estadísticas se han enfrentado a los problemas de volumen, pero en la actualidad aparecen dos elementos nuevos: la velocidad y la variedad. Siguiendo esta dirección, el primer documento (UNECE, 2013) que estudia el problema Big Data en la estadística pública *What Does 'Big Data' Mean for Official Statistics?*[119] lo define como una variante de la propuesta de Douglas Laney:

[118] http://www.gartner.com/analyst/40872/Douglas-Laney
[119] Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?" UNECE, March 10, 2013.

*"Big Data son las **fuentes de datos** que generalmente pueden ser descritas como de alto volumen, velocidad y variedad, que requieren formas rentables e innovadoras de procesamiento con el fin de mejorar los análisis y de apoyar las tomas de decisiones"*

Por lo tanto, para la estadística pública el problema Big Data se aborda como un problema de nuevas fuentes de datos. En esa dirección el problema se enfrenta considerando que estas fuentes de datos podrían complementar o sustituir las fuentes tradicionales utilizadas en la estadística pública, las encuestas y los registros administrativos, pero con algunas características peculiares:

1. La propiedad sobre las fuentes de datos generalmente no es pública, con los problemas derivados para el acceso, uso y mantenimiento de las fuentes.

2. La fuentes de datos no están pensadas para fines estadísticos con los problemas derivados de conceptualización y sesgos.

Asimismo, la **velocidad** es una característica disruptiva respecto a la historia de la datificación oficial fundamentada en censo y encuestas. En su momento el acceso a datos administrativos mejoró la oportunidad de la información ofrecida por las Oficinas Estadísticas, sin embargo con las fuentes Big Data la estadística pública se enfrenta a la información en tiempo real.

La posición de la estadística pública en un sistema de decisión se ha situado históricamente en los niveles analíticos o estratégicos y nunca en el operacional, que es donde se requiere información en tiempo real. Una incursión de la estadística oficial en el nivel operacional requeriría cambios tecnológicos y procedimentales. Desde la perspectiva de los cambios tecnológicos, se necesitarían evoluciones en las arquitecturas de datos de las Oficinas de Estadística. En la actualidad estas arquitecturas están dirigidas al procesamiento supervisado de datos por lotes, en cambio las arquitecturas Big Data (como la Kappa o la Lambda) están dirigidas al procesamiento automático de datos en modo batch o tiempo real. En ese sentido, el paso al procesamiento automático conlleva cambios procedimentales importantes en muchos de los procesos definidos en el Generic Statistical Business Process Model (GSBPM).

Otra de los retos que introducen las fuentes Big Data son la **variedad** de formatos y tipos de información disponible. Tradicionalmente la estadística pública ha extraído sus datos de formularios estructurados, sin embargo en las fuentes Big Data nos encontramos con recursos informativos de todo

tipo, asociados a formatos de imagen, voz y texto. Estos nuevos formatos aportan mucha riqueza informativa, pero su tratamiento requiere de nuevos procedimientos y tecnologías en las Oficinas Estadísticas.

Además en las fuentes Big Data nos tropezamos con los problemas derivados de los **sesgos por autoselección**. En las estadísticas, el sesgo de autoselección surge en cualquier situación en la que las unidades de observación se seleccionan a sí mismas para formar parte de la muestra, dando lugar a una muestra sesgada con muestreo no probabilístico. Con las fuentes Big Data nos enfrentamos a muestras politetápicas no probabilísticas con las consiguientes dificultades para corregir los problemas de sesgos. En algunos casos la corrección del problema puede resolverse en la fase de inferencia, para ello se podrían utilizar métodos de pseudo-estimación basada en diseños (con propensiones y calibrados clásicos en el muestreo no probabilístico), pero fundamentalmente métodos predictivos.

Como vimos en el Capítulo 2, las fuentes Big Data tienen importantes similitudes con las fuentes administrativas respecto a su capacidad informativa, incluso mejoran dicha capacidad. Estas características las convierten en fuentes atractivas para su inclusión en la parrilla de inputs de la estadística oficial. Sin embargo las fuentes Big Data aumentan los problemas metodológicos respecto a las fuentes administrativas:

1. Fuentes posiblemente sesgadas y con dificultades para conocer el sesgo
2. Incluyen tanto errores ajenos al muestreo como nuevas fuentes de error
3. La falta de persistencia dificulta la comparabilidad temporal
4. El uso de métodos estadísticos tradicionales no siempre es posible

Estos problemas plantean cuestiones importantes relativas a la idoneidad de las fuentes Big Data para su uso en las estadísticas oficiales, y en ese sentido algunos autores proponen que inicialmente su uso se circunscriba exclusivamente a la mejora de la calidad de las estimaciones dentro de los marcos metodológicos actuales, mientras que se estudian los niveles y causas de los errores de muestreo y no muestreo de estas fuentes.

Al evaluar o describir la calidad de los datos o productos estadísticos las Oficinas de Estadística utilizan marcos de calidad de referencia. En la actualidad podemos encontrar diversos marcos de calidad en uso, tales como el del FMI[120], Sistema Estadístico Europeo[121], Statistics Canada[122] o el

[120] https://unstats.un.org/unsd/accsub/2010docs-CDQIO/Ses1-DQAF-IMF.pdf
[121] http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646
[122] http://www.statcan.gc.ca/pub/12-586-x/12-586-x2017001-eng.pdf

marco de calidad de la Australian Bureau of Statistics[123]. Estos marcos tienen diversas características en común pues entienden la calidad desde múltiples dimensiones y coinciden en muchas de las dimensiones consideradas.

El documento de referencia sobre calidad de las fuentes Big Data para su uso en la estadística pública fue elaborado en 2014 por el UNECE Big Data Quality Task Team, bajo el título *"A Suggested Big Data Quality Framework"*[124] como entregable del proyecto *UNECE/HLG project - The Role of Big Data in the Modernisation of Statistical Production* en el que se presentó el Big Data Quality Framework (BDQF). El marco utiliza una estructura jerárquica compuesta de tres hiperdimensiones, siguiendo la misma organización desarrollada por Statistics Netherlands para fuentes administrativas, con dimensiones de calidad anidadas dentro de cada hiperdimensión. Estas tres hiperdimensiones son:

1. Fuente: Incluye dimensiones referidas a las entidades proveedoras de los datos, así como el marco de gestión y regulación de los mismos.

2. Metadatos: Aglutina las dimensiones que permiten describir los conceptos y variables utilizados en la fuente de datos, conocer la estructura de los ficheros, así como identificar los procesos que se le aplican.

3. Datos: Enumera la dimensiones propias de la calidad de los datos.

Por lo tanto el marco indica que la **estadística oficial debe evaluar las fuentes Big Data desde una perspectiva integral,** y no sólo desde el análisis de calidad de datos en sí mismo. En ese sentido el marco propone evaluar diversas dimensiones de calidad que permitan estudiar los riesgos del uso de una fuente Big Data, más allá de la bondad de sus datos. La evaluación de riesgos y la identificación de soluciones para el uso de fuentes Big Data en la estadística pública se revisan en el documento *Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a risk and quality framework* (Wirthmann et al, 2018).

El proyecto Sandbox ofreció importantes indicaciones sobre el uso de Big Data para las estadísticas oficiales, resultados que pueden dirigir los esfuerzos futuros en este campo. El resultado más importante que surgió de los experimentos es que las estadísticas basadas en fuentes de Big Data serán diferentes de las que tenemos hoy. Otro resultado práctico importante se asoció con el acceso a las

[123] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+Data+Quality+Framework
[124] UNECE Big Data Quality Task Team. "A Suggested Big Data Quality Framework." UNECE, December 2014.

fuentes de Big Data. **Las altas expectativas iniciales sobre las oportunidades de Big Data tuvieron que enfrentarse a la complejidad de la realidad:**

1. El hecho de que los datos se produzcan en grandes cantidades no significa que estén disponibles inmediata y fácilmente para producir estadísticas. Las fuentes de calidad son de difícil acceso, independientemente del precio de las mismas, en ese sentido los datos de los teléfonos móviles representan un ejemplo notable. Por otro lado, las fuentes de datos de acceso público están limitadas en términos de calidad y, por lo tanto, requieren una cantidad significativa de procesamiento para su uso en la estadística pública.

2. Asimismo el proyecto llega a importantes conclusiones respecto al acceso a datos. En ese sentido sentencia que para alcanzar el siguiente nivel en el uso de fuentes Big Data para las estadísticas oficiales, se necesitan dos tipos de acciones:

    a. En primer lugar, las negociaciones y los acuerdos con los proveedores, a quienes se debería alentar a compartir sus datos con las organizaciones estadísticas.

    b. En segundo lugar, la intervención política a nivel legislativo, para facilitar el uso de las fuentes de Big Data con fines estadísticos y para superar los problemas de jurídicos.

    En resumen, las fuentes de Big Data deben tratarse del mismo modo que las tradicionales en lo que respecta a su uso para las estadísticas públicas. **Las organizaciones estadísticas deberían recibir un tratamiento especial de acceso preferencial, en un marco de estrictas garantías de preservación de la privacidad.**

Es evidente que el desafío del uso de datos de fuentes Big Data dentro de la estadística pública significa necesariamente la modernización de las Oficinas Estadísticas. Ese desafío requiere al abordaje de diferentes retos, que sintéticamente podemos resumir en los siguientes puntos:

Estrategia: Es necesario definir cómo integrar las nuevas fuentes Big Data en la actividad de las Oficinas Estadísticas. Esta estrategia puede estar dirigida tanto a la integración de las nuevas fuentes en la producción habitual de las Oficinas, como en la identificación de nueva información estadística basada en dichas fuentes.

Acceso: Existe un debate intenso sobre los datos, su propiedad y el derecho de acceso para fines públicos. Si bien la legislación estadística puede obligar a facilitar el acceso a las Oficinas Estadísticas, esta capacidad tendrá que convivir en la tensión de intereses público-privado contrapuestos; tensión que necesitará de espacios de cooperación con los proveedores.

Privacidad: La datificación de buena parte de nuestras vidas genera actitudes diversas en la opinión pública sobre el derecho a la intimidad. Sin embargo cuando se trata de gran acumulación de datos por parte del Estado la confidencialidad, proporcionalidad y fin de los mismos pasan a ser una importante preocupación ciudadana. En ese sentido existe el peligro de que entre la sociedad se genere una visión de las Oficinas de Estadísticas como instituciones orwellianas.

Por otra parte, la generación de gran cantidad de datos a gran velocidad pone sobre la mesa nuevos retos tecnológicos para cumplir el mandato del deber de secreto estadístico, que impide que a través de la información publicada por las Oficinas Estadísticas se pueda identificar directa o indirectamente a las unidades de análisis.

Calidad: La dimensiones de evaluación de la calidad de las fuentes Big Data para su integración en la actividad de las Oficinas Estadísticas deben ser identificadas, especialmente debido a que son datos recopilados para fines no estadísticos.

Metodología: Con las fuentes Big Data nos encontramos ante la dificultad de datos recopilados para fines no estadísticos, por lo tanto estamos ante problemas similares a los planteados con los registros administrativos, al menos en lo que respecta a los conceptos usados en la recolección de datos y su relación con las definiciones internacionalmente armonizadas. Además algunas de las fuentes Big Data son muestras, con el problema añadido de ser muestras no probabilísticas y posiblemente sesgadas por el método o por las cuotas de mercado del agente recolector.

Tecnología: La incorporación de fuentes Big Data a la actividad estadística requerirá de la incorporación de tecnología Big Data a las Oficinas Estadísticas. Definir arquitecturas, hardware y software requeridos es uno de los retos que debe ser abordado.

Formación: El desarrollo de las capacidades y habilidades necesarias para explorar con eficacia los Big Data es esencial para su incorporación a la actividad de la Oficinas Estadísticas. Esto requiere esfuerzos sistemáticos, como cursos de formación adecuados y el establecimiento de comunidades de intercambio de experiencias y buenas prácticas.

4.2. Propuesta de acción

Los profesionales de las estadísticas oficiales son bastante conservadores y muy cautelosos en el uso de nuevas tipologías de datos. En ese sentido las fuentes Big Data no son una excepción. Sin embargo se han identificado usos potenciales de las fuentes Big Data en la estadística pública, por similitud con las fuentes administrativas, que categorizamos en dos bloques:

1. Uso como fuentes de estimación estadística
2. Uso como información de apoyo en el diseño y elaboración de estadísticas

La usabilidad de una fuente Big Data identifica el grado en el que una Oficina Estadística será capaz de trabajar con ella sin el empleo de recursos especializados o sin una carga significativa sobre los recursos existentes; así como el estudio de la facilidad con que se puede integrar con los sistemas y normas existentes. Los factores a considerar en el estudio de esta dimensión son los siguientes:

1. Recursos: ¿Cuáles son las habilidades necesarias para procesar y almacenar esta información? ¿Se requerirían inversiones adicionales en tecnología?

2. Análisis de riesgos: Identificar las posibles dificultades y ganancias para la Oficina Estadística si se requieren inversiones considerables para utilizar los datos.

Partiendo de estas premisas, las propuestas de acción para empezar a integrar fuentes Big Data en una Oficina Estadística deberían tener en cuenta los siguientes preceptos:

A) Uso incremental

Comenzar usando las fuentes Big Data como información de apoyo o complementaria, según los diferentes usos relacionados en el apartado 3.1.2., minimizando con ello los riesgos señalados en el apartado 3.2. Esta primera fase permite a la Oficina Estadística iniciar el estudio sobre la calidad, persistencia y complejidad de uso de cada una de las fuentes sin exponerse al riesgo de publicar datos basados en la misma.

Por ejemplo se pueden utilizar fuentes Big Data (Google Place, Facebook, Tripadvisor, FourSquare, Booking u otras) para complementar la elaboración, verificación o imputación de datos de directorios o encuestas de empresas y establecimientos.

En una segunda fase podría estudiarse la sustitución parcial de algunos campos de una encuesta por datos de una fuente Big Data, si previamente se ha identificado una buena correlación con los datos muestrales. Por ejemplo en encuestas a hoteles podría investigarse la sustitución de los datos de tarifas de alojamiento por información aportada en fuentes como Trivago o Booking.

En una tercera etapa se pasaría a analizar la sustitución de realizaciones completas de una encuesta por fuentes Big Data. Podría pensarse, por ejemplo, en recopilar datos de encuestas bimestrales en lugar de mensuales y realizar estimaciones con la fuente Big Data en los meses sin recogida de datos; pero sólo si las tendencias de la fuente demuestran una asociación suficiente con los resultados de la encuesta. Por ejemplo, podríamos pensar en la sustitución de recogida de precios para la elaboración de índices de precios o paridades de poder adquisitivo, por datos adquiridos mediante web-scraping.

En una cuarta fase podría abordarse la incorporación a una encuesta de variables de fuentes Big Data que sean de difícil o imposible recopilación mediante procedimientos tradicionales. Por ejemplo, en los censos de población el abordaje de la movilidad podría realizarse mediante el uso de datos de telefonía móvil, transporte público o aforos de carreteras.

Finalmente el uso de fuentes Big Data podría aplicarse en la sustitución completa de alguna operación estadística o en la elaboración de nuevas operaciones estadísticas. Para ello, de acuerdo con los criterios indicados en el apartado 3.3., la calidad de la fuente debe estar suficientemente acreditada.

B) Acceso fácil, persistente y de bajo coste

La fuente Big Data debe ser de fácil acceso por parte de la Oficina Estadística, bien sea debido a que la propiedad de la fuente es pública, o siendo privada existe la posibilidad de acceso abierto y persistente mediante scraping o APIs de datos. Además para el caso de fuentes privadas debe examinarse el coste-beneficio del acceso.

Como ejemplo de fuentes Big Data públicas podemos citar especialmente las que son producto de sensores tales como estaciones meteorológicas, estaciones de medición de la calidad del aire o de control de emisiones, aforos de carreteras, control de acceso a aparcamientos públicos, uso de transporte público, consumo de agua o control de la

generación de residuos sólidos. El desarrollo de la Internet de las Cosas en las Administraciones Públicas, a través de programas como las Smart Cities, abre un abanico considerable de acceso a fuentes Big Data de carácter público. Entre todas ellas es recomendable focalizar el interés en las que son de mayor cobertura y estabilidad pues algunas son de iniciativa municipal, dando lugar a que la cobertura o persistencia no sea siempre la deseable. En esta misma línea de fuente públicas también podemos señalar las vinculadas con la captación de imágenes, tales como ortofotos aéreas o cámaras de control de seguridad.

Una mención diferenciada requiere el acceso a datos de generación privada por adjudicación de servicios públicos o concesiones administrativas. En ese sentido se debe fomentar la incorporación de cláusulas en los pliegos de contratación o concesión que aseguren la datificación, titularidad y acceso por parte de las Administraciones Públicas.

En esta línea es importante que las Oficinas de Estadística acompañen los procesos de datificación de las Administraciones Públicas, asesorando y fomentando la reutilización de los datos para fines estadísticos, tal como ocurre con los registros administrativos.

Otra línea de acción a considerar es el desarrollo de captura de datos de Internet de la Cosas por la propia Oficina Estadística o la incorporación de tecnología de Reality Mining a las encuestas tradicionales. En una sociedad totalmente digital podría plantearse un paso más allá, de tal manera que la captura de datos de las encuestas podrían diseñarse utilizando tecnologías de Reality Mining[125]. En esta estrategia el diseño de la encuesta se realizaría mediante métodos de muestreo tradicional, mientras que en la fase de recogida de datos se utilizarán tecnologías de datificación reality mining. Por ejemplo, para un estudio de movilidad se selecciona una muestra de la población y se les solicita que activen una aplicación de seguimiento en su smartphone, como complemento a una encuesta tradicional.

C) Tratamiento no disruptivo

En un primer momento sería recomendable que la incorporación de una fuente Big Data sea lo menos disruptiva posible para cada una de las fases de los procesos definidos en el Generic Statistical Business Process Model (GSBPM)[126], facilitando la reutilización de tecnologías, métodos y procesos habitualmente utilizados por las Oficinas Estadísticas.

[125] https://en.wikipedia.org/wiki/Reality_mining

[126] https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

El uso incremental señalado anteriormente se debe acompañar con la incorporación gradual de capacidades, metodologías, procedimientos y tecnologías al servicio de dichos usos. Por lo tanto el desarrollo de estas características vendría dirigido por los casos de usos desarrollados por la Oficina Estadística, evitando la dispersión de las estrategias generalistas.

Esta propuesta de acción determina una estrategia viable que permita a las Oficinas Estadística, especialmente a aquellas con escasa capacidad para abordar o financiar grandes proyectos, la incorporación de fuentes Big Data en su parrilla de datos.

Anexos

Anexo A. Misión, principios y valores de la estadística pública

El primer Seminario interregional sobre organización estadística, patrocinado por las Naciones Unidas, se celebró en Ottawa en 1952 justo un año después de que la US Census Bureau comprara el UNIVAC I. Dos años después, en 1954, se publicó el *Manual de organización estadística*, y una segunda edición del Manual se hizo pública en 1980 con el título *Manual de organización estadística: Estudio sobre la organización de servicios nacionales de estadística y cuestiones conexas de administración*. Casi dos décadas más tarde, en un seminario sobre calidad de datos celebrado en 1999 con el patrocinio conjunto del Fondo Monetario Internacional y las Naciones Unidas, varios países solicitaron actualizar el Manual de 1980. En respuesta a esa petición, las Naciones Unidas preparó el *Manual de organización estadística, tercera edición: El funcionamiento y organización de una oficina de estadística*. En este manual se establecen, entre otras, las líneas directrices sobre la misión, principios y valores de una organización de estadística pública.

A.1. Misión

Naciones Unidas reconoce a las estadísticas oficiales como un elemento indispensable en el sistema de información de una sociedad democrática pues proporcionan a los gobiernos, a la economía y a la ciudadanía datos de la situación económica, demográfica, social y ambiental de un país o de una región. En ese sentido considera que la información estadística es esencial para el desarrollo en estas áreas, pero también para el conocimiento mutuo y el comercio entre los Estados y los pueblos del mundo. **Con este fin, NNUU indica que las Oficinas de Estadística han de compilar y facilitar, de forma imparcial, estadísticas oficiales de comprobada utilidad práctica para que los ciudadanos puedan ejercer su derecho a mantenerse informados.**

A.2. Principios y valores

Para que los ciudadanos confíen en las estadísticas oficiales, los organismos estadísticos deben contar con un conjunto de valores y principios fundamentales. De acuerdo con Naciones Unidas, los principios generales son la (1) independencia, (2) la pertinencia o relevancia, (3) la credibilidad, así

como (4) el respeto a los derechos de los informantes. Estos principios han sido desarrollados en los principios fundamentales de las estadísticas oficiales[127].

En coherencia con las líneas trazadas por Naciones Unidas, en el ámbito europeo, el Reglamento (CE) nº 223/2009, del Parlamento Europeo y del Consejo de 11 de marzo de 2009 relativo a la estadística europea, señala los siguientes principios en su artículo 2: (1) Independencia profesional, (2) imparcialidad, (3) fiabilidad, (4) secreto estadístico, (5) rentabilidad. Estos principios estadísticos se desarrollaron posteriormente en el Código de Buenas Prácticas de la Estadística Europea, que tiene por finalidad garantizar la confianza de la población en las estadísticas europeas mediante la determinación de la forma en que deben desarrollarse, elaborarse y difundirse las estadísticas con arreglo a los principios estadísticos europeos y a las mejores prácticas internacionales.

Las normas citadas desempeñan un papel vital en la obtención de la confianza en las estadísticas oficial. A su vez estas normas se refuerzan con los códigos éticos de los estadísticos, destacando la *Declaración sobre Ética Profesional* del Instituto Internacional de Estadística (ISI), que además se complementan con diferentes códigos éticos elaborados por los distintos sistemas estadísticos nacionales.

A.2.1. Principios enumerados por Naciones Unidas

Tal como señalamos anteriormente, NNUU enumeró en la tercera edición del Manual de Organización Estadística al menos cuatro valores o principios fundamentales para que los organismos estadísticos sean respetados por el público: la independencia, la pertinencia o relevancia, la credibilidad, así como el respeto a los derechos de los informantes. Posteriormente estos principios fueron ampliados con la Resolución aprobada por la Asamblea General el 29 de enero de 2014, sobre los Principios Fundamentales de las Estadísticas Oficiales:

1. *Relevancia, imparcialidad y acceso equitativo:* Las estadísticas oficiales constituyen un elemento indispensable en el sistema de información de una sociedad democrática y proporcionan al gobierno, a la economía y al público datos acerca de la situación económica, demográfica, social y ambiental. Con este fin, los organismos oficiales de estadística han de compilar y facilitar en forma imparcial estadísticas oficiales de comprobada utilidad práctica para que los ciudadanos puedan ejercer su derecho a la información pública.

[127] Documentos Oficiales del Consejo Económico y Social, 1994, suplemento Nº 9 (E/1994/29), cap. V. Para más información véase el apéndice II o consúltese el sitio web <http://unstats.un.org/unsd/statcom/doc94/s1994.htm>

2. *Patrones profesionales,principios científicos y ética:* Para mantener la confianza en las estadísticas oficiales, las Oficinas de Estadística han de decidir, con arreglo a consideraciones estrictamente profesionales, incluidos los principios científicos y la ética profesional, acerca de los métodos y procedimientos para la reunión, el procesamiento, el almacenamiento y la presentación de los datos estadísticos.

3. *Responsabilidad y transparencia:* Para facilitar una interpretación correcta de los datos, las Oficinas de Estadística han de presentar información conforme a normas científicas sobre las fuentes, métodos y procedimientos de la estadística.

4. *Prevención del mal uso:* las Oficinas de Estadística tienen derecho a formular observaciones sobre interpretaciones erróneas y la utilización indebida de las estadísticas.

5. *Fuentes de estadísticas oficiales:* Los datos para fines estadísticos pueden obtenerse de todo tipo de fuentes, ya sea encuestas estadísticas o registros administrativos. Las Oficinas de Estadística han de seleccionar la fuente con respecto a la calidad, la oportunidad, el costo y la carga que impondrá a los encuestados.

6. *Confidencialidad:* Los datos individuales que reúnan las Oficinas de Estadística para la compilación estadística, se refieran a personas naturales o jurídicas, deben ser estrictamente confidenciales y utilizarse exclusivamente para fines estadísticos.

7. *Legislación:* Se han de dar a conocer al público las leyes, reglamentos y medidas que rigen la operación de los sistemas estadísticos.

8. *Coordinación nacional:* La coordinación entre las Oficinas de Estadística a nivel nacional es indispensable para lograr la coherencia y eficiencia del sistema estadístico.

9. *Uso de patrones internacionales:* La utilización por las Oficinas de Estadística de cada país de conceptos, clasificaciones y métodos internacionales fomenta la coherencia y eficiencia de los sistemas estadísticos a nivel oficial.

10. *Cooperación internacional:* La cooperación bilateral y multilateral en la esfera de la estadística contribuye a mejorar los sistemas de estadísticas oficiales en todos los países.

De la revisión de estos principios se concluye que las Oficinas de Estadística son organizaciones de servicio. Los motivos de su existencia, crecimiento y la posibilidad de realizar una contribución palpable al gobierno y a la sociedad se basan en su capacidad de suministrar información útil para resolver problemas importantes. Sin embargo, las prioridades pueden cambiar con mayor rapidez que la capacidad del organismo para modificar sus actividades de producción. Por este motivo, es importante que los funcionarios superiores tengan la visión de conjunto y los contactos necesarios para poder detectar los problemas graves y diferenciarlos de los que pueden ser una mera moda pasajera.

Además, para que la organización de estadística pueda contar con la credibilidad de los usuarios y crear una relación de respeto y confianza mutua, es fundamental que tenga una sólida posición de independencia. Las actividades de recopilación, análisis y difusión de información estadística siempre deberán estar diferenciadas de las actividades de formulación de las políticas. Asimismo, una organización de estadística debe garantizar la confiabilidad del proceso de recolección y compilación de datos estadísticos, así como de los procedimientos internos. Para que la confiabilidad sea aceptada por la ciudadanía y estimule al personal de la organización, es necesario que se cumplan varias condiciones:

1. El proceso debe tener una buena base lógica.

2. Los mecanismos empleados deben ser sólidos.

3. la descripción de los mecanismos y de los procedimientos debe estar abierta a inspección y los resultados de esta sujetos al debate público.

4. Tanto los procedimientos como los mecanismos deben poder crecer y adaptarse a nuevas circunstancias y a nuevos entornos.

A menos que un organismo de estadística pueda garantizar que la información suministrada por los informantes sea estrictamente confidencial, no podrá confiar en la calidad de la información que reúne y se pondrá en riesgo su credibilidad.

INDEPENDENCIA

Para tener credibilidad y desempeñar su función es preciso que las Oficinas de Estadística tengan una posición de independencia ampliamente reconocida. Sin la credibilidad derivada de un alto grado de independencia, los usuarios perderán la confianza en la exactitud y la objetividad de la información del organismo y quienes le proporcionan los datos estarán menos dispuestos a cooperar con él.

En esencia, un organismo de estadística debería diferenciarse claramente de los sectores del gobierno encargados de las actividades de aplicación y de formulación de las políticas. Debería ser imparcial y evitar que se diera la impresión de que los procesos de recopilación, análisis e información de datos que realiza pudieran ser manipulados con fines políticos, o de que determinados datos, identificables en forma individual, pudieran ser cedidos con fines administrativos, regulatorios o de aplicación de la ley.

El Manual de Organización Estadística enumera algunas características relacionadas con la independencia:

1. Autoridad para adoptar decisiones de tipo profesional con respecto al ámbito de aplicación, el contenido y la frecuencia de los datos recopilados, analizados o publicados.

2. Autoridad para seleccionar y promover a los funcionarios profesionales, técnicos y operativos.

3. Reconocimiento, por parte de los funcionarios políticos ajenos al organismo de estadística, de su autoridad para publicar información estadística sin autorización previa.

4. Autoridad del jefe de estadística y de los funcionarios especializados para hablar sobre las estadísticas elaboradas por el organismo ante los funcionarios del gobierno y los organismos públicos.

5. Adhesión a calendarios predeterminados para la publicación de importantes indicadores económicos o de otro tipo a fin de evitar que se dé siquiera la impresión de una manipulación en las fechas de publicación con fines políticos.

6. Diferenciación clara entre la publicación de información estadística y la interpretación de dicha información por parte de los funcionarios superiores del gobierno.

7. Políticas de divulgación que alienten la presentación al público de los principales resultados obtenidos por los programas del organismo de estadística, a través de los medios de comunicación, Internet y otros.

RELEVANCIA / PERTINENCIA

Las Oficinas de Estadística deberían tratar de mejorar permanentemente para proporcionar información exacta, oportuna y relevante respecto a las necesidades cambiantes de las políticas públicas. Sin embargo, el problema es que los intereses en materia de políticas pueden modificarse a mayor velocidad de lo que puede adaptarse el sistema estadístico. Los asuntos de interés surgen en muy poco tiempo; primero son curiosidades, luego pasan a ser temas de debate y por último se convierten en cuestiones de suma importancia para los encargados de la formulación de políticas.

Habitualmente se tarda mucho menos en detectar un problema que en desplegar los medios necesarios para medir su magnitud y lograr que la medición sea comparable en el plano internacional. Dada esta disparidad, en un mundo cuyas prioridades se modifican con rapidez, un organismo de estadística que intente ser relevante de forma instantánea podría resultar permanentemente irrelevante.

Para una Oficina de Estadística no tiene mucho sentido tratar de abordar problemas que se perciban como transitorios pues para cuando se haya puesto en marcha un programa destinado a abordarlos, la agenda política se habrá modificado varias veces. De hecho, al examinar las distintas prioridades, el organismo de estadística deberá separar los problemas transitorios de los más permanentes.

Una vez que un organismo de estadística ha establecido una prioridad, no le resulta fácil modificarla con la misma rapidez con que parecen cambiar los temas de la agenda política. Por este motivo es fundamental establecer las prioridades con racionalidad y predecir con exactitud los cambios de orientación política. La planificación conlleva cuatro elementos importantes:

1. Concebir programas suficientemente generales como para poder adaptarlos con facilidad a pequeños cambios de orientación política.

2. Crear equipos que permitan abordar las situaciones imprevistas sin afectar el normal funcionamiento del organismo de estadística.

3. Desarrollar políticas de recursos humanos cuyo objeto sea lograr que el personal de las Oficinas de Estadística sea adaptable y pueda reorganizarse para enfrentar con éxito las modificaciones de los programas del organismo.

4. Compartir información técnica e ideas con otros Oficinas de Estadística. Este tipo de cooperación puede fomentar el desarrollo de métodos innovadores de recopilación, análisis y difusión de datos.

CREDIBILIDAD

Las estadísticas tienen una característica especial: los resultados de las actividades de un organismo de estadística deben ser reproducibles para ser creíbles, pero no es realista suponer que el usuario pueda reproducirlos. Este es el motivo de que las Oficinas de Estadística deban esforzarse en fortalecer su credibilidad, y de que exista una extremada susceptibilidad ante cualquier ataque a su credibilidad, o a la posibilidad de que el público pierda la fe en la confiabilidad de su producción.

Las Oficinas de Estadística deben ser sumamente rigurosas con respecto a las normas que deben satisfacer la obtención de datos, los métodos de procesamiento y la derivación de los resultados. Además, deben infundir al personal un espíritu de calidad acorde con el alto nivel de exigencia de las normas. De este modo se refuerza permanentemente la sensación de que lo que se produce es el resultado de la calidad de los insumos y de los métodos de producción y control.

RESPETO A LOS DERECHOS DE LOS INFORMANTES

La confidencialidad de la información individual es, probablemente, la mayor preocupación de los informantes. Los organismos que no logran persuadir a los informantes de que la información que aportan es totalmente confidencial no pueden confiar en la calidad de los datos que recopilan.

La potestad que confiere la legislación a las Oficinas de Estadística para recabar información no es de mayor utilidad a menos que todos los sectores de la sociedad estén dispuestos a cooperar. Las oficinas que han realizado los mayores esfuerzos para convencer a los encuestados de que la información que proporcionan es valiosa y que el tiempo que han dedicado a brindar esa información es reconocido y valorado suelen ser las que obtienen las mayores tasas de respuestas. Es preciso tener clara conciencia de que, en el trabajo estadístico, una baja tasa de respuestas es una falla tan importante como la falta de cuidado en la depuración y la difusión de las cifras.

A.2.2. Principios europeos desarrollados en el Código de Buenas Prácticas

El Código de Buenas Prácticas de las Estadísticas Europeas se basa en quince principios, que abarcan el entorno institucional, los procesos de elaboración de estadísticas y la producción estadística. Un grupo de indicadores de buenas prácticas para cada uno de los principios sirve de referencia para analizar la aplicación del Código. Los criterios de calidad de las estadísticas europeas se establecen en la Ley Estadística Europea aprobada por el Reglamento (CE) nº 223/2009, del Parlamento Europeo y del Consejo de 11 de marzo de 2009 relativo a la estadística europea que señala los siguientes principios en su artículo 2: (1) Independencia profesional, (2) imparcialidad, (3) fiabilidad, (4) secreto estadístico, (5) rentabilidad. Veamos cuales son los quince principios del Código:

ENTORNO INSTITUCIONAL

Los factores institucionales y organizativos tienen una influencia considerable en la eficacia y la credibilidad de una autoridad estadística que desarrolla, elabora y difunde estadísticas europeas. Los aspectos relevantes son la independencia profesional, el mandato de recogida de datos, la adecuación de los recursos, el compromiso de calidad, la confidencialidad estadística, así como la imparcialidad y la objetividad.

1. *Independencia profesional*. La independencia profesional de las autoridades estadísticas frente a otros departamentos y organismos políticos, reguladores o administrativos, y frente a los operadores del sector privado, garantiza la credibilidad de las estadísticas europeas.

2. *Mandato de recogida de datos*. Las autoridades estadísticas tienen un mandato jurídico claro para recoger información destinada a la elaboración de estadísticas europeas. A petición de las autoridades estadísticas, se puede obligar por ley a las administraciones, las empresas, los hogares y el público en general a permitir el acceso a los datos destinados a la elaboración de estadísticas europeas o a facilitar dichos datos.

3. *Adecuación de los recursos.* Los recursos a disposición de las autoridades estadísticas son suficientes para cumplir los requisitos de las estadísticas europeas.

4. *Compromiso de calidad.* Las autoridades estadísticas están comprometidas con la calidad; identifican sistemática y regularmente los puntos fuertes y débiles para mejorar continuamente la calidad del proceso y del producto.

5. *Confidencialidad estadística.* La privacidad de los informantes (hogares, empresas, administraciones y otros encuestados), la confidencialidad de la información que proporcionan y su uso exclusivo con fines estadísticos están totalmente garantizados.

6. *Imparcialidad y objetividad*. Las autoridades estadísticas desarrollan, elaboran y difunden estadísticas europeas respetando la independencia científica y de forma objetiva, profesional y transparente, de modo que todos los usuarios reciben el mismo trato.

PROCESOS ESTADÍSTICOS

Las normas, orientaciones y buenas prácticas, tanto europeas como internacionales, se respetan plenamente en los procesos utilizados por las autoridades estadísticas para organizar, recoger, elaborar y difundir las estadísticas europeas. La credibilidad de las estadísticas se ve reforzada por una reputación de buena gestión y eficacia. Los aspectos relevantes son una metodología sólida, unos procedimientos estadísticos adecuados, una carga no excesiva para los encuestados y la relación coste-eficacia.

7. *Metodología sólida*. Las estadísticas de calidad se apoyan en una metodología sólida, que requiere herramientas, procedimientos y conocimientos adecuados.

8. *Procedimientos estadísticos adecuados.* Las estadísticas de calidad se apoyan en procedimientos estadísticos adecuados, aplicados desde la recogida de los datos hasta la validación de los mismos.

9. *Carga no excesiva para los encuestados.* La carga de respuesta es proporcionada en relación con las necesidades de los usuarios y no es excesiva para los encuestados. Las autoridades estadísticas controlan la carga que supone responder a la encuesta y fijan objetivos para reducirla progresivamente.

10. *Relación coste/eficiencia.* Los recursos se utilizan eficientemente.

PRODUCCIÓN ESTADÍSTICA

Las estadísticas disponibles satisfacen las necesidades de los usuarios. Las estadísticas cumplen las normas de calidad europeas y responden a las necesidades de las instituciones europeas, los gobiernos, los centros de investigación, las empresas y el público en general. Los aspectos a tener en cuenta son la medida en que las estadísticas son relevantes, precisas y fiables, oportunas, coherentes y comparables entre regiones y países, y fácilmente accesibles para los usuarios.

11. *Relevancia.* Las estadísticas europeas satisfacen las necesidades de los usuarios.

12. *Precisión y fiabilidad.* Las estadísticas europeas reflejan la realidad de manera precisa y fiable.

13. *Oportunidad y puntualidad*. Las estadísticas europeas se hacen públicas oportuna y puntualmente.

14. *Coherencia y comparabilidad.* Las estadísticas europeas son consistentes internamente a lo largo del tiempo y comparables entre regiones y países; es posible combinar y utilizar conjuntamente datos relacionados procedentes de fuentes diferentes.

15. *Accesibilidad y claridad.* Las estadísticas europeas se presentan de forma clara y comprensible, se difunden de forma adecuada y conveniente, su disponibilidad y acceso tienen carácter imparcial y van acompañadas de metadatos y orientación de apoyo.

Anexo B. Proyectos Big Data en el inventario de UNECE

	Oficina estadística	**País**	**Tipo de fuente Big Data**	**Tema**
1	Australia (ABS) - Social Linked (semantic) Data Processing for Various Statistical Uses	Australia	Data from public administration (2100)	Education (1.3); Health (1.4); Income and consumption (1.5); Labour (1.2); Population and migration (1.1)
2	Statistics Canada - Non-Residential Buildings Inventory: Feasibility Study	Canada		Environment (3.1); Human settlements and housing (1.7); Population and migration (1.1); Prices (2.7)
3	National Bureau of Statistics of China - Big Data Enterprise Statistical Indicator Ten-day Report	China, People's Republic of		Economic accounts (2.2); Prices (2.7)
4	National Bureau of Statistics of China - Online Price Changes of Means of Production in Circulation Area in Shandong Zhuochuang	China, People's Republic of		Economic accounts (2.2); Prices (2.7)
5	Statistics Finland - Traffic sensor data for commuting statistics	Finland	Traffic sensors/webcam (3113)	Labour (1.2); Time use (1.11); Transport (2.4.4)

6	ESCAP - Developing a Curriculum and Training Modules on Using Big Data for Official Statistics	International		Economic accounts (2.2); Environment (3.1); Population and migration (1.1); Prices (2.7); Other
7	Eurostat - Feasibility study on the use of mobile positioning data for tourism statistics	International	Mobile phone location (3121)	Mobility (Transport 2.4.4); Tourism (2.4.5)
8	Eurostat - Multi-purpose consumer price statistics, sub-project Scanner Data/Web Scraping	International	Commercial transactions (2210)	Prices (2.7)
9	UNECE HLG Big Data Project 2014 - Sandbox Task Team Social Media - Sentiment Analysis	International	Social Networks: Facebook, Twitter, Tumblr etc. (1100)	Other
10	UNECE HLG Big Data Project 2014 - Sandbox Task Team Traffic Loops	International	Traffic sensors/webcam (3113)	Transport (2.4.4)
11	UNECE HLG Big Data Project 2015 - Sandbox Task Team Enterprise Web sites	International	Other	Labour (1.2)
12	UNECE HLG Big Data Project - Sandbox Task Team Consumer Price Index (Scanner Data)	International		

13	UNECE HLG Big Data Project - Sandbox Task Team Consumer Price Index (Web scraped data)	International		Prices (2.7)
14	UNECE HLG Big Data Project - Sandbox Task Team Job Vacancies	International		Labour (1.2)
15	UNECE HLG Big Data Project - Sandbox Task Team Satellite Data	International	Satellite images (3123)	Other
16	UNECE HLG Big Data Project - Sandbox Task Team Smart Meters	International		Energy (2.4.2)
17	Italy (Istat) - Internet as a Data Source for ICT Usage by Enterprises and Public Institutions	Italy		Information society (3.3.3)
18	Italy (Istat) - Persons and Places: Mobility Estimates based on Mobile Phone Data	Italy	Mobile phone: call/text times and positions (312.)	Mobility (Transport 2.4.4); Tourism (2.4.5)
19	Italy (Istat) - Specific purpose geographic basins and population statistics using mobile phone tracking data	Italy	Mobile phone: call/text times and positions (312.)	Population and migration (1.1)
20	Italy (Istat) - Use of scanner data for consumer price index	Italy		Prices (2.7)

21	Mexico (INEGI) - Tweet Analysis	Mexico	Social Networks: Facebook, Twitter, Tumblr etc. (1100)	Mobility (Transport 2.4.4); Population and migration (1.1); Tourism (2.4.5)
22	Central Statistical Office of Poland - Estimating demand on labour market by analysing job offer portals	Poland		Labour (1.2)
23	Romania National Statistical Institute (INS) - Using scanner data	Romania	Commercial transactions (2210)	Economic accounts (2.2); Prices (2.7)
24	Statistics South Africa - Assessing use of scanner data for compiling the Consumer Price Index	South Africa	Commercial transactions (2210)	Prices (2.7)
25	Switzerland (FSO) - Price collection with scanner data	Switzerland	Commercial transactions (2210)	Prices (2.7)
26	UK (ONS) Housing Website data to help improve address register	United Kingdom of Great Britain and Northern Ireland		Human settlements and housing (1.7)
27	UK (ONS) Webscraped prices for price indices	United Kingdom of Great Britain and Northern Ireland	E-commerce (2230)	Prices (2.7)

28	United Kingdom (ONS) - Smartmeter type data for household structure/size and occupancy	United Kingdom of Great Britain and Northern Ireland	Smart Energy Meters (311?)	Population and migration (1.1); Other