# Caracterización del sesgo de selección en redes sociales en México a través de algunas características sociodemográficas de sus usuarios

# Characterization of Selection Bias in Social Networks in Mexico through Some Sociodemographic Characteristics of their Users

**Víctor Alfredo Bustos y de la Tijera, Abel Alejandro Coronado Iruegas, Silvia Laura Fraustro Velhagen, Gerardo Leyva Parra, Noemí López Delgado, Ricardo Antonio Olvera Navarro, Ana Miriam Romo Anaya y Víctor Silva Cuevas***

Se ha sugerido que los textos publicados en redes sociales representan una oportunidad para producir información oficial. Por esta razón, varias oficinas nacionales de estadística han comenzado a experimentar con su uso. Por lo tanto, es necesario indagar si las poblaciones de usuarios de dichas redes muestran diferencias importantes con la población general de México, según variables sociodemográficas relevantes; en otras palabras, si resultan representativas o no de acuerdo con dichas variables. La ausencia de representatividad introducirá sesgos en estimaciones. En esta nota nos proponemos llevar a cabo la cuantificación de tales discrepancias utilizando datos de la Encuesta Nacional sobre Disponibili-

It has been suggested that texts posted on social networks represent an opportunity to produce official information. For this reason, several national statistical offices have begun to experiment with their use. Therefore, it is necessary to investigate whether the populations of users of these networks show significant differences with the general population of Mexico, according to relevant sociodemographic variables; in other words, whether they are representative or not according to these variables. The absence of representativeness will introduce biases in estimations. In this note, we propose to carry out the quantification of such discrepancies using data from the National Survey on the Availabili-

* Instituto Nacional de Estadística y Geografía (INEGI), alfredo.bustos@inegi.org.mx, abel.coronado@inegi.org.mx, silvia.fraustro@inegi.org.mx, gerardo.leyva@inegi.org.mx, nohemi.delgado@inegi.org.mx, ricardo.olvera@inegi.org.mx,miriam.romo@inegi.org.mx y victor.silvac@inegi.org.mx, respectivamente.

dad y Uso de Tecnologías de la Información en los Hogares, ediciones 2017-2019. Se hace necesario vincular las respuestas registradas en el cuestionario con los textos publicados por los usuarios de redes que también son informantes. Se abriría, así, la posibilidad de entrenar algoritmos para etiquetar de manera sociodemográfica a los usuarios y a sus publicaciones con el fin de reponderarlos con vista en la reducción (o eliminación) de sesgos. Este es solo un primer paso en el uso de estrategias mixtas (encuestas-redes) en el estudio continuo de temas que interesan a la estadística oficial. Si tiene éxito, podremos considerar otros temas en el cuestionario de cualquier otra encuesta realizada por el Instituto Nacional de Estadística y Geografía para producir resultados representativos a partir de la información de las redes sociales.

**Palabras clave:** demografía; redes sociales; representatividad; sesgo de selección.

Recibido: 12 de abril de 2021.
Aceptado: 5 de agosto de 2021.

ty and Use of Information Technologies in Households 2017-2019 editions. It becomes necessary to link the responses recorded in the questionnaire with the texts published by network users who are also informants. This would open up the possibility of training algorithms for sociodemographic labeling of users and their publications in order to reweight them with a view to reducing (or eliminating) biases. This is only a first step in the use of mixed (survey-network) strategies in the ongoing study of topics of interest to official statistics. If successful, we may consider other topics in the questionnaire of any other survey conducted by the National Institute of Statistics and Geography to produce representative results from information from social network information.

**Key words:** demography; social networks; representativeness; selection bias.


Moscow, Russia, 18-02-2021: clubhouse app icon on smatphone screen surrounded by other social media apps and user run clubhouse. Clubhouse drop-in audio chat social media network. Shallow DOF / Victollio / iStock

## Introducción

La estadística oficial tiene como propósito principal el de proveer insumos de calidad a los tomadores de decisiones tanto en el ámbito privado como en el público; en este último se acuñó el término *policy driven*, que se interpreta como la necesidad de producir información para atender requerimientos identificados para el diseño, la instrumentación o el seguimiento de alguna política pública, cuya población objetivo queda, en general, claramente definida. Todo ejercicio de recolección de información debe buscar que esta resulte relevante a la mencionada población y al objetivo del estudio. Por ejemplo, en investigaciones por muestreo en hogares, el marco muestral debe representar de manera adecuada a la población objeto.

La proliferación de fuentes de información (consecuencia de la introducción de la telefonía celular, sensores y cámaras de vigilancia, de la disponibilidad de imágenes satelitales, así como del advenimiento de las redes sociales) es percibida como una gran oportunidad para complementar la producción de estadística oficial tradicional. Ello ha dado lugar a la necesidad de estudiar los retos y las oportunidades que tales tecnologías acarrean. Por ejemplo, la CBS holandesa está entre las primeras oficinas nacionales de estadística (ONE) en iniciar el estudio de estas fuentes alternativas (ver Struijs, 2014 y Struijs *et al*., 2014). A su vez, la Organización de las Naciones Unidas (ONU) creó, en el 2014, el Grupo Global de Trabajo (GWG, por sus siglas en inglés) para *Big Data* en la estadística oficial,[1 y 2] con participación de Australia, Bangladesh, Camerún, China, Colombia, Dinamarca, Egipto, Indonesia, Italia, México, Marruecos, Holanda, Omán, Pakistán, Filipinas, Tanzania, Emiratos Árabes Unidos y Estados Unidos de América, así como la United Nations Statistical Commission (UNSD), la United Nations Economic Commission for Europe (UNECE), la United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP), la UN Global Pulse, la Unión Internacional de Telecomunicaciones (ITU, por sus siglas en inglés), la Organización para la Cooperación y el Desarrollo Económicos (OCDE), el Banco Mundial, la Eurostat y la GCC-Stat (ver UNSD, 2015; Jansen, 2018 y Smith, 2018). De acuerdo con Snyder (2015), sus términos de referencia[3] asignan al GWG, entre otras, las tareas de "… aportar una visión estratégica, dirección y coordinación de un programa global de *Big Data* para la estadística oficial, incluyendo los indicadores para la *Agenda 2030 para el desarrollo sostenible*. También, promueve el uso práctico de fuentes *Big Data*, la promoción del desarrollo de capacidades, el entrenamiento y el intercambio de experiencias…". Durante su primera reunión, en octubre del 2014 en Beijing, el GWG estableció ocho equipos de trabajo:

1. Datos de redes sociales.
2. *Big Data* y los Objetivos de Desarrollo Sostenible (ODS).
3. Datos de telefonía móvil.
4. Temas transversales.
5. Mejorar acceso a fuentes *Big Data*.
6. Promoción y comunicación.
7. Capacitación, habilidades y fortalecimiento de capacidades.
8. Imágenes de satélite y datos geoespaciales.

Un ámbito de aplicación inmediato para los trabajos del GWG está dado por la *Agenda 2030 para el desarrollo sostenible*, la cual adoptó un marco global de monitoreo amplio abarcando 231 indicadores dentro de 17 objetivos. Por supuesto, dicho marco requiere datos que sean de alta calidad, accesibles, oportunos, confiables y desagregados por ingreso, sexo, edad, raza, etnicidad, estatus migratorio, discapacidad y localización geográfica, así como otras características relevantes dentro de los contextos nacionales. Una proporción no despreciable de los indicadores es de nueva creación, por lo que se estudia tanto su definición como las fuentes de datos que permitirán su cálculo. Ello ha conducido a dirigir una mirada atenta a las fuentes *Big Data*.

1 https://unstats.un.org/bigdata/

2 United Nations Global Working Group (GWG) on *Big Data* for Official Statistics (https://unstats.un.org/bigdata/) y sus seis International Conferences on *Big Data* for Official Statistics (ej., https://unstats.un.org/unsd/bigdata/conferences/2020/).

3 https://unstats.un.org/bigdata/documents/TOR%20-%20GWG%20-%202015.pdf

Sin embargo, dicha estrategia no está exenta de riesgos. Como señalan Lokanathan *et al*. (2017): "… aprovechar fuentes de datos nuevas y existentes (tanto del sector público como del privado) con el fin de monitorear el progreso hacia los ODS, así como para lograrlo, no está exento de desafíos…". Enfatizan que las diferencias en lo que denominan *datificación*[4] (en inglés, *datafication*) entre las economías desarrolladas y las emergentes impedirán que estas hagan un uso óptimo de la información disponible. El acceso a datos en manos del sector privado no será sencillo porque "… en industrias competitivas como el sector de las telecomunicaciones, compartir datos tendría implicaciones competitivas…". El acceso a su información dará lugar a modelos de negocio innovadores. Asimismo, destacan que la innovación requerirá el establecimiento de asociaciones entre una variedad de actores tanto del sector público como del académico y el privado.

Otros grupos de trabajo han sesionado y alcanzado diferentes avances. En Jansen (2019)[5] se hace un gran resumen de ellos para casi todos los creados por el GWG. Este autor señala que los tres grupos centrados en el uso de imágenes de satélite y de percepción remota, así como el de empleo de datos de telefonía celular en la estadística oficial y el de técnicas para la preservación de la privacidad han desarrollado manuales y han llevado a cabo talleres en diversas ciudades, así como la compilación de ejemplos en la utilización de dichos datos, entre otras actividades. El de uso de datos de escáner produjo algoritmos, con código fuente y documentos, para el cálculo del Índice de Precios al Consumidor. El de entrenamiento, competencias y desarrollo de capacidades realizó una evaluación global del grado institucional de preparación para incorporar *Big Data* en sus procesos, así como un análisis de los programas de entrenamiento en ciencia de datos.[6] Destaca, además, la creación de los nuevos equipos de trabajo en el uso de datos administrativos y sobre la biodiversidad y conservación del planeta. Cabe resaltar que no señala avances para los equipos sobre integración de datos, ni el que se refiere al empleo de información de redes sociales, al que declara en receso.[7]

Por otro lado, Lokanathan *et al*. (2017) señalan que "… es importante recordar que a pesar del gran acervo de literatura y aplicaciones que ya existen, el estado del arte en aplicaciones enfocadas en el desarrollo innovador de estas nuevas fuentes de datos aún se encuentra en sus etapas embrionarias…". Nuevamente, las diferencias en acceso a las tecnologías de información y a la *datificación* dificultarán la satisfacción del propósito de "… contar a los no contados…". Lo anterior, tiene repercusiones sobre el concepto estadístico de *representatividad* de estas nuevas fuentes de datos; es decir, con qué precisión reflejan a la población. De manera, adicional, mencionan que será necesario poner particular atención para "… abordar los dilemas éticos y de privacidad…" que surgirán de estas fuentes de datos.

Van Halderen *et al*. (2021) reportan algunos de los principales logros del Equipo de Trabajo sobre *Big Data* para los ODS. Destacan que uno de sus objetivos principales es "… proporcionar ejemplos concretos del uso potencial de *Big Data* para monitorear los indicadores asociados con los ODS…". Por ello el "… Equipo de Trabajo dirigió una encuesta global en 2015 para evaluar los macrodatos para las estadísticas oficiales, incluidos los ODS. La encuesta encontró que solo el 2 % de los países encuestados utilizaban macrodatos para los indicadores de los ODS. Por el contrario, casi el 30 % utilizaba macrodatos para las estadísticas de precios. El 60 % vio una necesidad urgente de orientación sobre el vínculo entre los macrodatos y los indicadores de los ODS…".

En Data-Pop Alliance (2016) se indica que "… destacan los riesgos y las oportunidades que *Big Data* presenta a las oficinas nacionales de estadís-

4 Tendencia tecnológica que convierte aspectos de nuestra vida en datos que, posteriormente, se transforman en información.

5 International Symposium on the Use of Big Data for Official Statistics, Hangzhou, China, October 16-18, 2019 (DE) http://www.stats.gov.cn/english/International Training/2019/202009/t20200930_1792523.html

6 Por ejemplo, Master in Official Statistics and Social and Economic Indicators, Complutense University of Madrid, Spain.

7 El último reporte del Social Media WG para el 2017 puede ser encontrado en https://unstats.un.org/unsd/bigdata/conferences/2017/gwg/GWG%20Task%20Team%20on%20Social%20Media%20Data%20-%202017%20report.pdf

tica en Latinoamérica en el contexto de los ODS…". Después de hacer una amplia revisión de los mayores retos y obstáculos para que las ONE aprovechen *Big Data* (barreras institucionales para la administración del cambio y la innovación, restricciones para el acceso y la completez de los datos, retos técnicos, brechas en capacidades humanas, retos metodológicos, riesgos éticos y políticos), se concluye desarrollando una hoja de ruta regional para el aprovechamiento de *Big Data* en la estadística oficial y en el seguimiento a los ODS. Se afirma que "… a pesar de los retos anteriores es posible desarrollar tendencias regionales importantes que, además de los ODS, faciliten un mayor uso y experimentación de *Big Data* a lo largo del ecosistema de datos latinoamericano…". Sobresalen cinco tendencias que se consideran propicias en la región: la experiencia latinoamericana en el movimiento de datos abiertos;[8] la aparición de asociaciones públicas y privadas sobre el tema de *Big Data*;[9] la presencia de comités, instituciones y grupos de trabajo fuertes y que abarcan a toda la región; el desarrollo de mejores prácticas adaptables; y la existencia de una red interdisciplinaria de innovación que involucra a las ONE y a otros actores.

Sobre esta base se desarrolla una hoja de ruta regional multipartita para *Big Data*, cuya premisa principal es la de construir sobre las fortalezas y oportunidades regionales existentes. Destacan tres ejes principales para este fin: creación de estructuras para alentar el desarrollo y la coordinación de proyectos sobre grandes volúmenes de datos tanto nuevos como ya existentes, movilizar la conciencia y la voluntad políticas para asegurar el establecimiento de políticas acerca de estos y desarrollo de mecanismos y herramientas para el uso de *Big Data* a través de la retroalimentación y del aprendizaje.

## El caso mexicano

Hacia principios del 2014 —y ante las constantes referencias que las autoridades del INEGI y las agencias internacionales hacían sobre el tema—, un grupo de entusiastas colaboradores del área de informática del INEGI se dio a la tarea de establecer las capacidades, de medir las limitaciones y de avanzar en el conocimiento de las plataformas tecnológicas relevantes, tanto de *hardware* como de *software*, en relación con el tema de los grandes volúmenes de datos. El propósito de este ejercicio inicial era, en esencia, didáctico. En consecuencia, fue necesario hacer uso de toda la creatividad de sus integrantes; por ejemplo, nueve computadoras personales fueron enlazadas para experimentar con el procesamiento paralelo y se instalaron algunas herramientas de *software* abierto —para trabajar bajo las mencionadas condiciones—, con lo que se inició una etapa de autocapacitación. Faltaba, sin embargo, la materia prima en la forma de información producida constantemente, disponible de forma no estructurada y, por supuesto, en grandes volúmenes. De manera adicional, acceder a ella tendría que resultar de bajo costo. Fue así como se llegó a la decisión de trabajar con la información de *Twitter* por la facilidad de acceso que concede la API pública, a través de la cual es posible descargar hasta 1 % de todos los tuits mundiales sin costo (ver esquema).

8 Iniciativa Latinoamericana por los Datos Abiertos (ILDA) (DE) https://idatosabiertos.org/acerca-de-nosotros/

9 Ver también DNP, 2017 y Dutra, 2018.

**Esquema**

De este modo, se tiene que alrededor de 2 millones de cuentas han publicado tuits de manera georreferenciada en territorio mexicano, entre enero del 2016 y octubre del 2019, alcanzándose un total de 143.3 millones de publicaciones descargadas.[10] Por otra parte, contra lo que ocurría antes del 2015, ahora sabemos por la Encuesta Nacional sobre Uso y Disponibilidad de Tecnologías de la Información en los Hogares (ENDUTIH)[11] que en el 2018 existían en México alrededor de 9.36 millones de cuentas activas,[12] lo cual contrasta con los 56.2 millones en *Facebook*.

Uno de los primeros proyectos que hicieron uso de la base de datos de tuits georreferenciados del INEGI es el denominado *Estado de ánimo de los tuiteros en México*.[13] Ya que no se disponía de información precisa sobre la representatividad de esta, se actuó con cautela al seleccionar el nombre con el que se publicarían sus resultados. En efecto, es difícil determinar si estos son representativos de la población mexicana, en uno de los extremos, o únicamente aplican a la de tuiteros que publican mensajes georreferenciados, en el otro. Tal duda dio lugar a la inclusión de preguntas sobre el uso y acceso a redes sociales en la ENDUTIH. Por ello, a partir del 2015, se cuenta, además, con información sociodemográfica de los usuarios activos, lo cual da lugar a los resultados que se presentan en este trabajo. Cabe aclarar que abarcan solo el periodo 2017-2019, pues la forma de preguntar ha venido cambiando, de modo que estos son los levantamientos más comparables a lo largo del tiempo. Asimismo, en adelante solo nos concentraremos en poblaciones con 15 años o más de edad en vista de las condiciones solicitadas por las redes. De otro modo, las comparaciones podrían resultar sesgadas o inválidas.

Los resultados que se presentan comparan estructuras porcentuales obtenidas de la Encuesta y correspondientes a seis subpoblaciones que se identifican como: *R*, residentes como una aproximación a lo que correspondería a la población total; *F*, usuarios de *Facebook*; *I*, de *Instagram*; *T*, de

10 Para salvaguardar su privacidad, el nombre del usuario es reemplazado en todos los mensajes por un código numérico.

11 La ENDUTIH y sus antecesores —los módulos Nacional de Computación (MONACO) 2001 y sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (MODUTIH), ediciones 2010-2014, así como la 2002— representan los esfuerzos realizados por el INEGI para medir la penetración de las tecnologías de información y comunicación (TIC) en los hogares mexicanos.

12 Activas dentro de los tres meses anteriores a la fecha de recolección de información.

13 https://www.inegi.org.mx/app/animotuitero/#/app/multiline

Cuadro 1

**Tamaños de diversas subpoblaciones en México, 2017-2019**

| | | 2017 | 2018 | 2019 |
|---|---|---|---|---|
| Población total | (*R*) | 123 430 703 | 124 664 007 | 125 781 270 |
| Población de 15 años de edad o mayor | | 91 698 197 | 93 078 513 | 94 890 564 |
| Población usuaria reciente de una o más redes | | 48 349 321 | 51 868 745 | 63 502 696[a] |
| 1. *Facebook* | (*F*) | 47 587 848 | 50 582 944 | 54 985 921 |
| 2. *Instagram* | (*I*) | 12 393 665 | 14 128 184 | 17 716 292 |
| 3. *Twitter* | (*T*) | 8 892 518 | 8 740 687 | 7 698 832 |
| 4. *Snapchat* | (*S*) | 4 016 848 | 3 829 825 | 3 381 436 |
| 5. *Linkedin* | (*L*) | 1 049 122 | 888 799 | 580 012 |

a En el levantamiento del 2019 se incluyó *Whatsapp* explícitamente entre las redes sociales; antes, quedaba incluida en la categoría *Otros*. Aparentemente, no era considerada una red social pues, al incluirla, el número de usuarios reciente de una o más redes crece en casi 12 millones en un año. Esta respuesta no es comparable en el tiempo, en estricto sentido, por lo que no será comentada en adelante.

**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

*Twitter*; *S*, de *Snapchat*; y *L*, de *Linkedin*. Las estructuras que se comparan se refieren a cuatro estratos sociodemográficos: bajo (1), medio-bajo (2), medio-alto (3) y alto (4); a 19 grupos quinquenales de edad; y a dos sexos; así como al cruce de variables. Ya que la población *R* representa la mejor aproximación que es posible obtener a partir de los datos de la Encuesta para la población total del país en cada momento, sus estructuras servirán como la base de comparación con las restantes subpoblaciones.

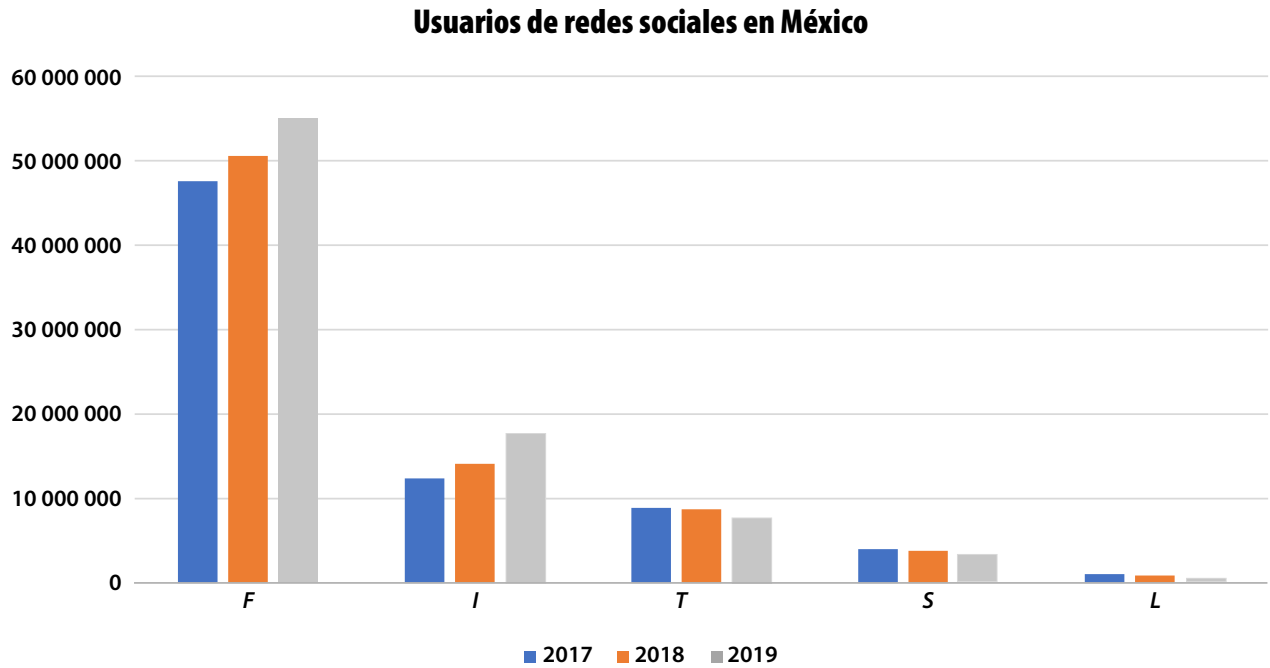
Tanto el cuadro como la gráfica 1 presentan un gran resumen para, entre otras, las subpoblaciones de usuarios con edades superiores a los 15 años. Lo primero que llama la atención es la desproporción entre los tamaños de las poblaciones de usuarios; por mucho, la de *Facebook* resulta ser siempre la mayor, seguida por la de *Instagram*. La tendencia decreciente del número de usuarios de *Twitter*, *Snapchat* y *Linkedin* contrasta con la de los de *Facebook* e *Instagram*; en particular, cabe destacar la alta tasa de crecimiento de esta última red en el periodo considerado.

## Sociodemografía de los usuarios de redes sociales en México

La gráfica 2 ejemplifica la manera en la que, con variaciones menores, se presentarán de forma visual los resultados a discutir. En este caso, exhibe seis conjuntos de barras, cada uno de los cuales representa una de las subpoblaciones mencionadas; las barras representan la proporción de hombres mayores de 15 años. La primera es el promedio de dicha proporción a lo largo de los tres años considerados. Con el fin de exhibir, en su caso, la presencia de posibles tendencias a través del tiempo se muestra, además, para cada uno de los años.

La población masculina de residentes exhibe un comportamiento estable apenas por debajo de 50 %, como es usual entre las concentraciones humanas. Salvo por dos de las redes consideradas, las poblaciones de usuarios son principalmente femeninas, mostrando marcadas diferencias con la de residentes; las excepciones son *Twitter* y *Linkedin*, cuyas desviaciones de la población de referencia (*R*) son, en consecuencia, aún mayores. Excepto

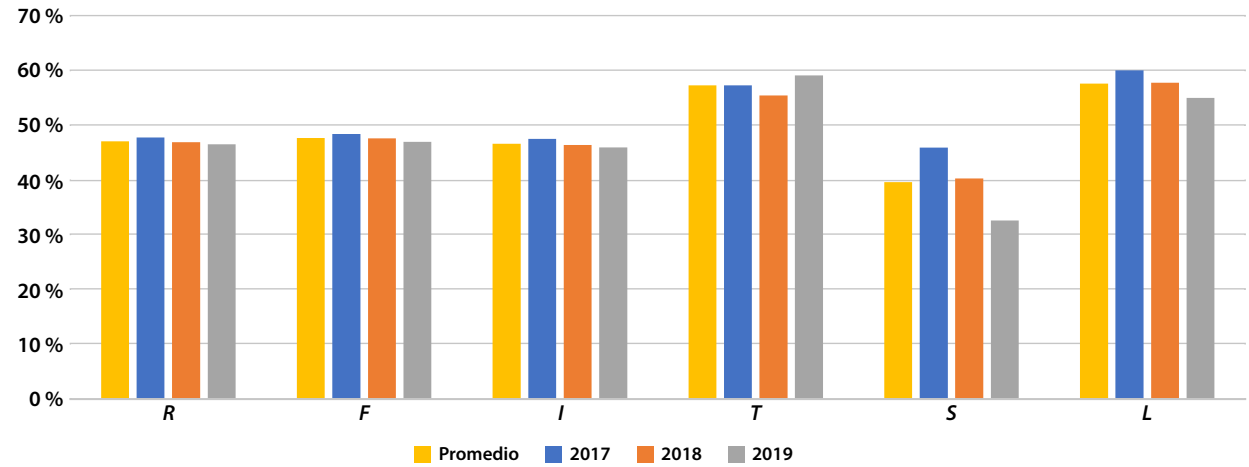
Gráfica 1




**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

Gráfica 2

**Proporción masculina para seis subpoblaciones, 2017-2019**



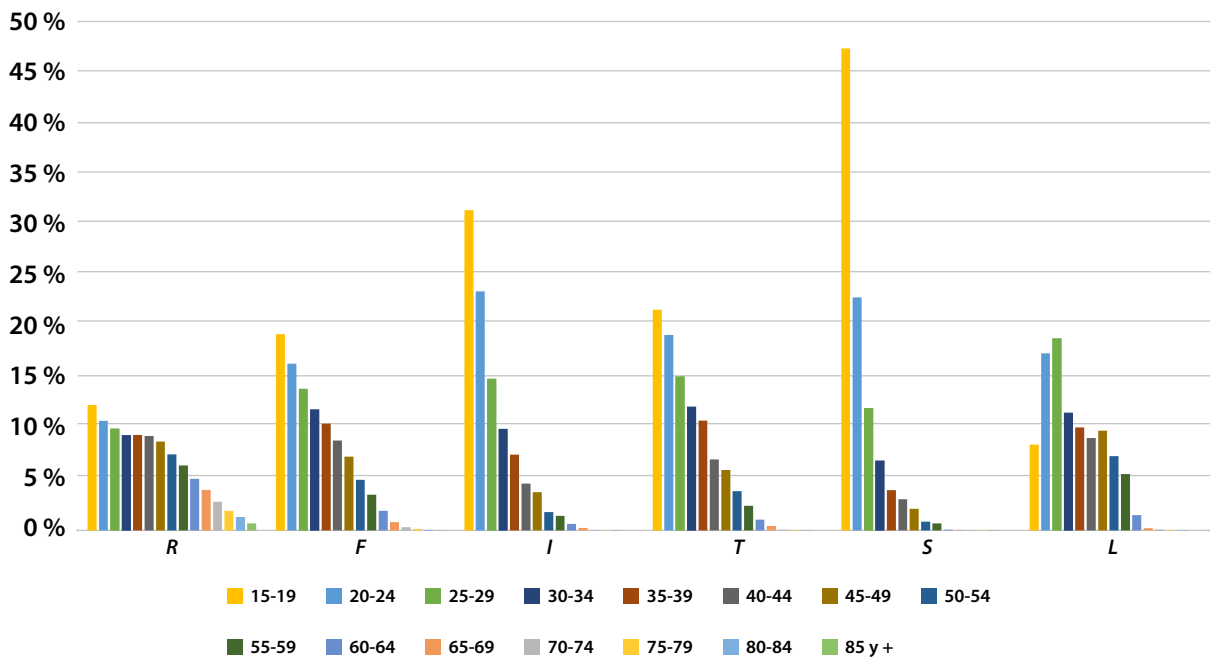

**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

por el caso de *Twitter*, se observan decrementos más o menos marcados para esta proporción entre el 2017 y 2019 y destaca la rápida caída de *Snapchat*. *Twitter* aparenta ser la única red que parece masculinizarse en el periodo.

En cuanto a la participación por grupos quinquenales de edad en las seis subpoblaciones, la gráfica 3 muestra contrastes marcados. En general, los usuarios son más jóvenes que en la de referencia. Por supuesto, aun entre las redes son aparen-

Gráfica 3

**Estructuras relativas por red social según grupos quinquenales de edad, promedio 2017-2019**




**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

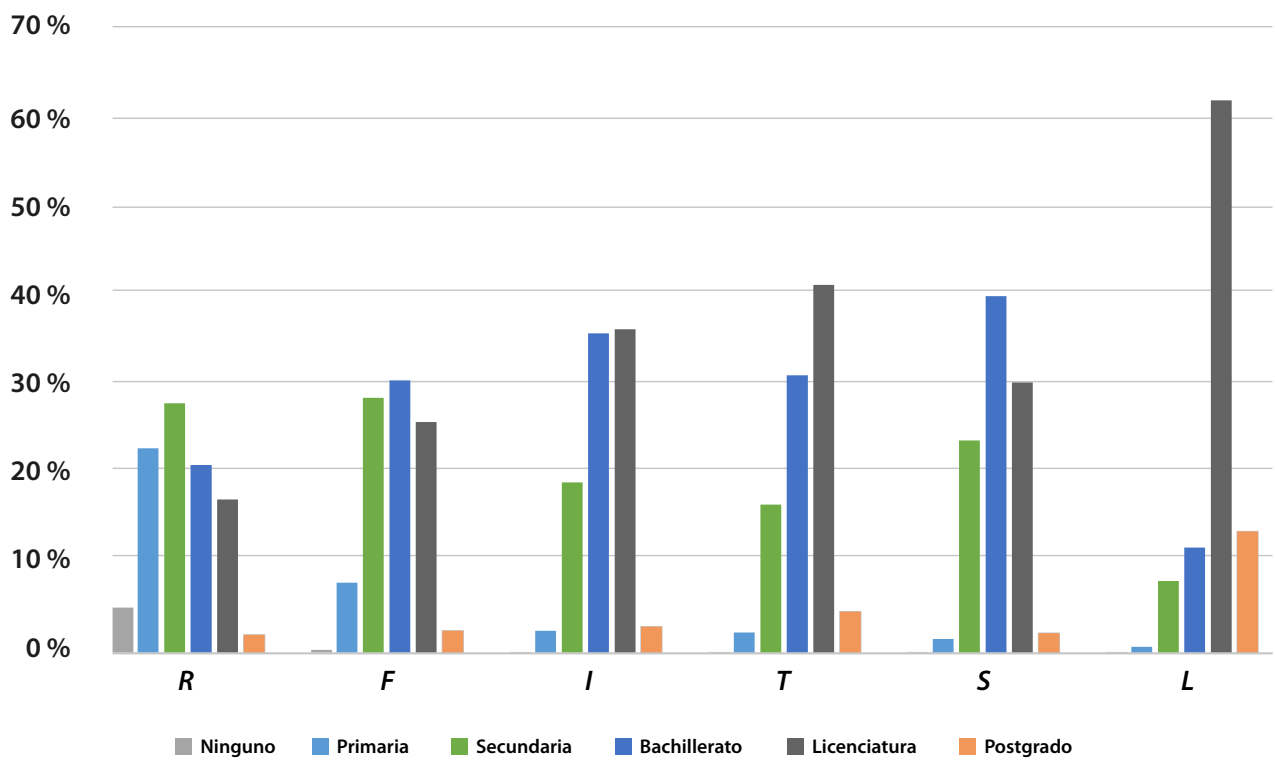
tes algunas diferencias marcadas, destacando *Instagram* y *Snapchat* como las más juveniles, con el grupo de 15-19 años de edad con el más numeroso en cada una de ellas; en términos porcentuales, la población en ese grupo es más grande que en la general por 2.5 veces que para *I* y por cuatro para *S*. En el otro extremo se ubica *Linkedin*, que no tiene a este como el grupo mayoritario. En ninguno de los casos, la edad parece distribuirse de manera semejante a lo que ocurre para *R*.

Veamos ahora qué ocurre al considerar el máximo grado escolar aprobado. La gráfica 4 muestra los seis niveles más relevantes. A simple vista se aprecia que la distribución de esta variable en cada una de las seis poblaciones es diferente. Por principio de cuentas, ninguna de las de usuarios de redes incluye una fracción significativa de personas mayores de 15 años que no han concluido algún grado educativo. Como se observa, las proporciones de los que solo cuentan con primaria completa son inferiores a las que se aprecian en la población general, pero a partir de la educación media superior se muestra lo opuesto, con una importante sobrerrepresentación. En estas condiciones destaca *Linkedin*, en la que poco más de 75 % de los usuarios cuentan con estudios superiores, lo que contrasta con 19 % para la población abierta.

De manera similar, podemos comparar el nivel sociodemográfico de los usuarios de redes con el de la población general. De nuevo, es clara la subrepresentación del estrato bajo, con la mitad o menos en todos los casos; para el del medio-bajo, esta se reduce en general. Para los dos restantes, en compensación, se da el caso opuesto. En particular, para el estrato alto se ratifica un claro sesgo favorable a este nivel, con casi dos veces el tamaño relativo de la población general para *I*, *T* y *S*, o más de cuatro veces para *L*. La red con la distribución más semejante a la población general es *Facebook*, pero aún con diferencias importantes (ver gráfica 5).

Gráfica 4

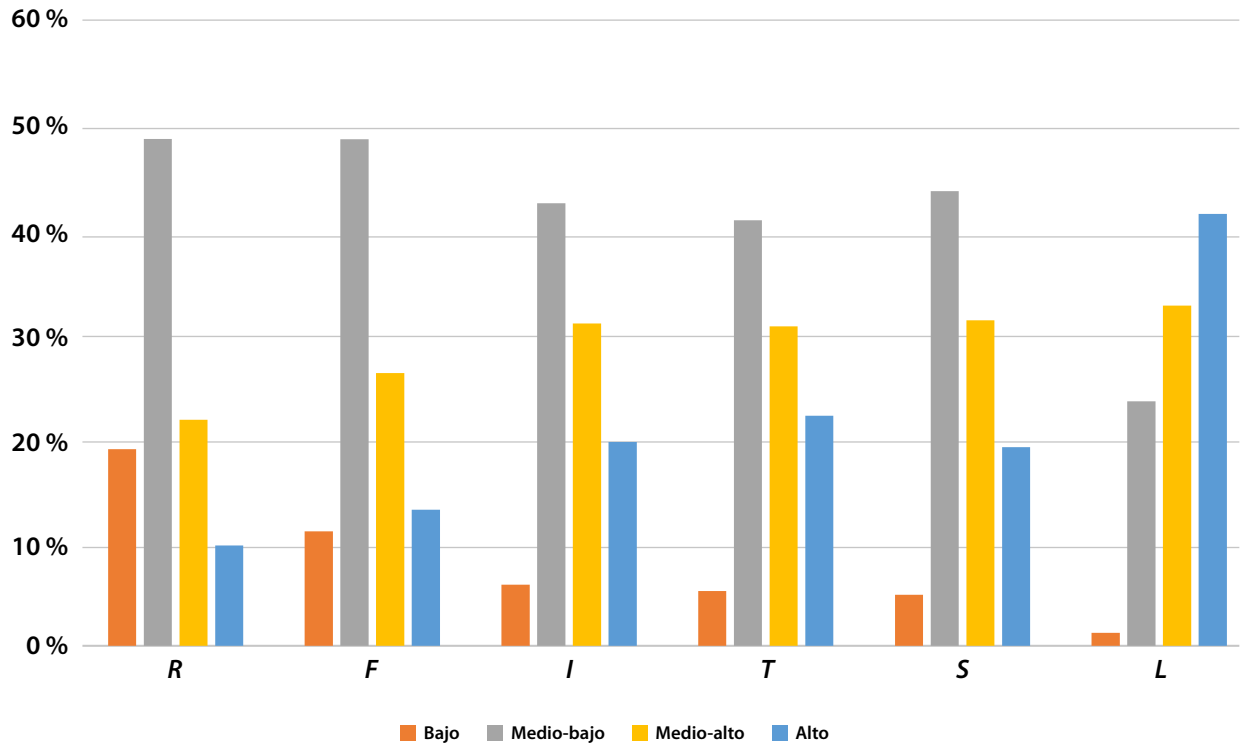
**Estructuras relativas por red social según niveles educativos seleccionados, promedio 2017-2019**




**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

Gráfica 5

**Estructuras relativas por red social según estrato sociodemográfico, promedio 2017-2019**




**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

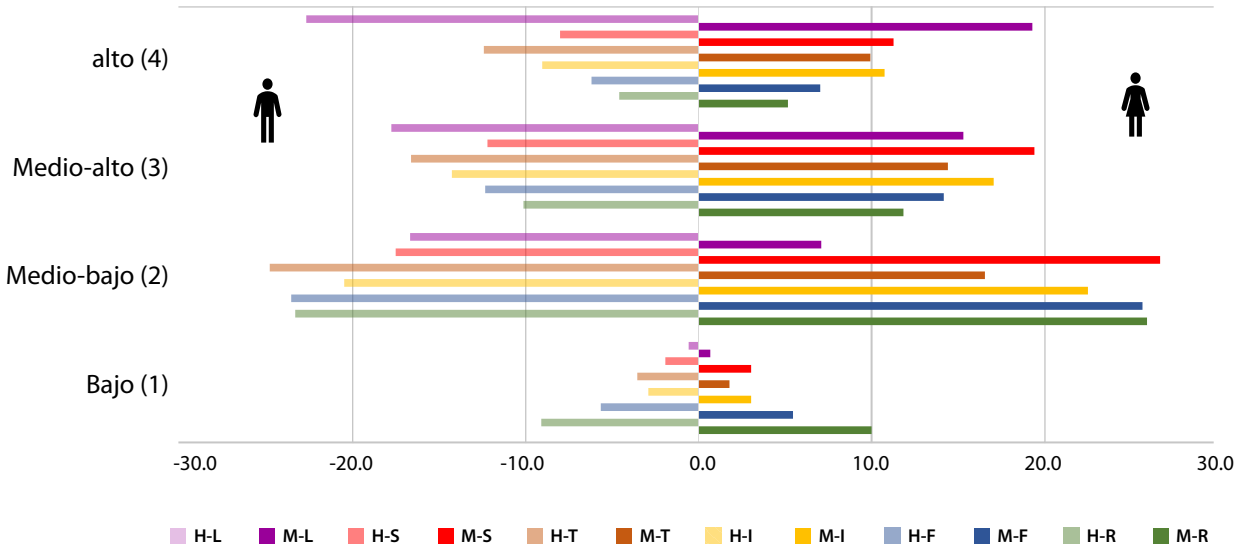
A partir de la breve discusión anterior, tenemos que es muy difícil que las distribuciones según sexo, edad, nivel educativo y estrato sociodemográfico en las poblaciones de usuarios de redes sociales coincidan con las correspondientes para la mexicana; en otras palabras, no son representativas de la población general, en términos de dichas variables.

En consecuencia, cabe preguntarse si, a causa de los sesgos exhibidos, tiene sentido hacer uso de los datos aportados por las redes en la producción de la estadística oficial para nuestro país. Si la explotación se lleva a cabo sin que la información considere los mencionados sesgos, la respuesta es no, salvo que la población bajo estudio sea la de los propios usuarios. Sin embargo, si se cuenta con datos sobre las mismas variables para cada usuario, se puede pensar que, a partir de los resultados anteriores, podemos modificar los pesos que se asignan a cada uno al calcular un indicador para lograr mejor representatividad nacional. El propósito, en este caso, se refiere a corregir la sub o sobrerrepresentación en las poblaciones de usuarios. Para ello, sin embargo, consideramos adecuado basar las modificaciones a las ponderaciones en estructuras relativas que consideran simultáneamente más de una de las variables. Se puede evitar, de este modo, el riesgo de ajustar de más o de menos algún subgrupo identificado por la combinación de niveles de estas.

Por ejemplo, los promedios de las estructuras porcentuales para el periodo, por sexo y estrato sociodemográfico, de las seis poblaciones mencionadas se muestran en la gráfica 6. En ella se aprecian discrepancias importantes entre la de residentes (en tonos verdes) y la de usuarios de *Twitter* (en café), quedando la de los de *Facebook* (en azules) entre las dos anteriores. Se tiene que los estratos

Gráfica 6

**Estructuras porcentuales según sexo y estrato sociodemográfico, promedio 2017-2019**




**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

sociodemográficos bajo y medio-bajo están subrepresentados entre los tuiteros mexicanos, en tanto que lo opuesto ocurre para los de medio-alto y alto, con lo que se favorecería a estas subpoblaciones en cualquier análisis basado en información de *Twitter*. Se aprecian, además, ligeros y diferentes sesgos según sexo en los estratos medio-bajo y alto.

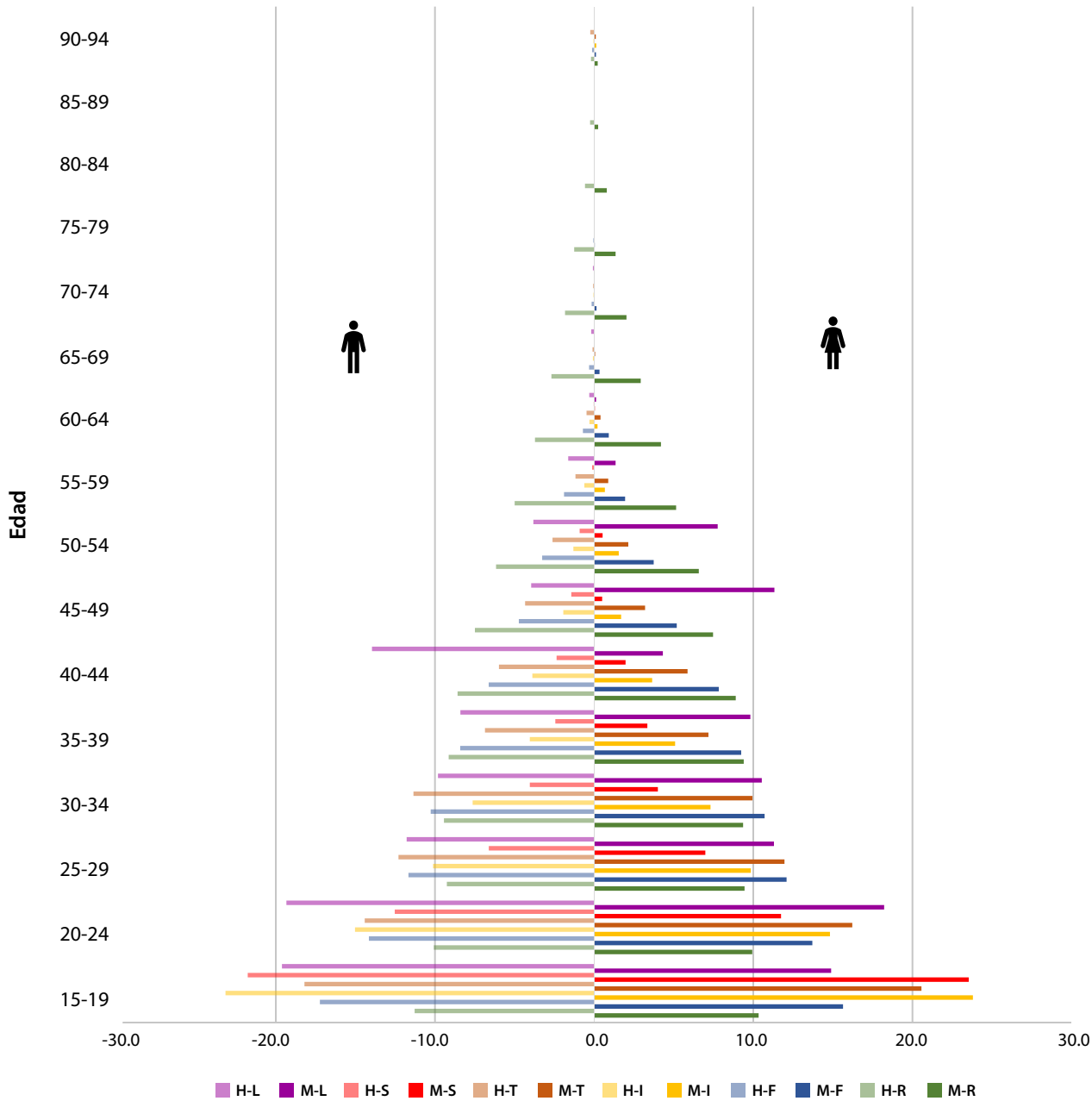
En la gráfica 7 se presentan las estructuras porcentuales por grupo de edad y sexo para cada una de las poblaciones consideradas, con un código de color semejante al del caso anterior. Salvo excepciones, no son apreciables sesgos importantes en favor de uno u otro sexo. Los usuarios de redes muestran una estructura etaria significativamente más joven que la que corresponde a la población de residentes, con *Twitter* mostrando un comportamiento más juvenil que *Facebook*, pero ambos rebasados por *Instagram* y *Snapchat*. Para algunos grupos de edad, en *Linkedin* es aparente un sesgo importante en favor de uno u otro sexo. Condiciones como esta justifican el uso de la distribución conjunta de esas variables, pues la corrección solo por una u otra marginal no corregiría la situación. En general, los grupos mayores a 40 años están subrepresentados entre los usuarios de redes sociales; lo opuesto ocurre entre las personas con edades entre los 15 y 39 años.

Según se aprecia en las gráficas 8, las distribuciones de edad pueden cambiar de forma importante a medida que nos desplazamos entre estratos sociodemográficos. En todos los casos es apreciable la reducción del tamaño relativo de las poblaciones jóvenes, así como la consecuente aparición de representantes de grupos de edades mayores. Aunque el cambio ocurre a ritmos menores, incluso entre la población abierta, tiene lugar el mencionado envejecimiento. Entre los usuarios de redes sociales, en el estrato bajo, la juventud de la estructura por edades es todavía más extrema. Esta disparidad se reduce de manera paulatina a medida que se avanza hacia los estratos altos. Por otro lado, cuando se asciende en el estrato sociodemográfico, también se

obtienen mayores promedios de edad. Todo ello concuerda con lo que se ha señalado para las gráficas 2 y 3. Nuevamente, el intento de corregir los sesgos a partir de las distribuciones marginales representaría una aproximación muy gruesa para la deseada corrección.

Gráfica 7

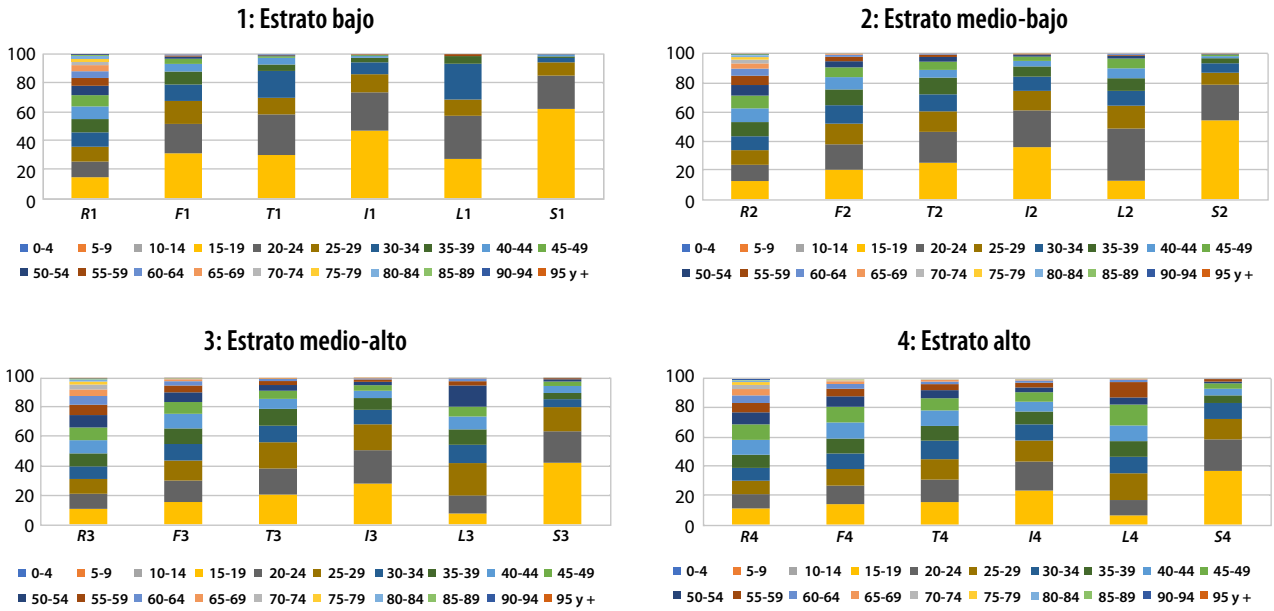
**Estructuras porcentuales por sexo y grupo quinquenal de edad, promedio 2017-2019**




**Fuente:** cálculos propios a partir de bases de datos públicas de la ENDUTIH.

Gráficas 8

**Estructura porcentual por estrato sociodemográfico y grupo de edad para seis poblaciones**




**Fuente:** elaboración propia a partir de la base de datos de la ENDUTIH 2018.

## ¿Sesgos en *Twitter*?

A manera de ejemplificación de las posibles consecuencias que tiene la explotación de información proveniente de redes sociales cuando no se realiza corrección alguna, en esta sección comentaremos dos temas que nos parecen relevantes: en el primer caso, se comparan resultados sobre movilidad de las personas en la Zona Metropolitana del Valle de México (ZMVM) a partir tanto de la encuesta que sobre el tema se levantó en el 2017 como de una colección de tuits georreferenciados que el INEGI ha venido recopilando; el segundo se refiere a la percepción sobre la intención de voto que podría derivarse de la misma colección de tuits y a su contraste con los resultados de la elección presidencial en México del 2018.

### Encuesta Origen Destino en Hogares de la Zona Metropolitana del Valle de México (EOD) 2017

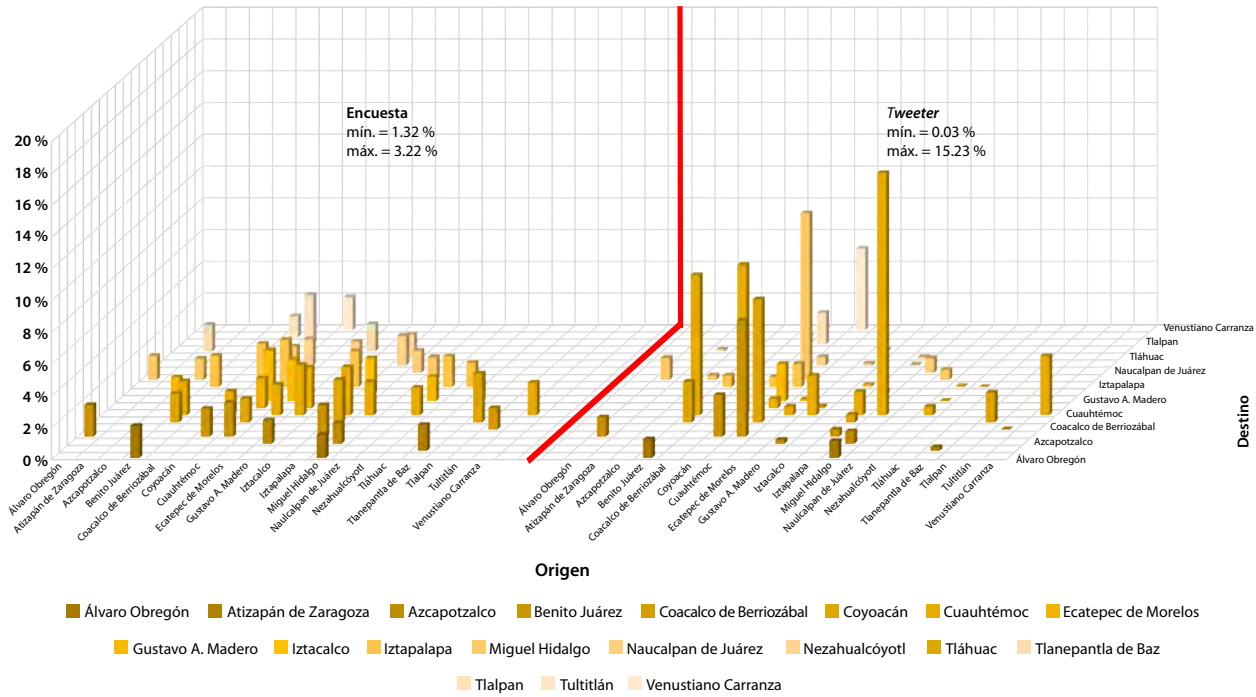
La comparación entre esta encuesta y el intento realizado para relacionarla con la captura de tuits georreferenciados es, también, elocuente en lo que toca al posible sesgo de selección en nuestra base de datos; esa es la razón por la que la discutiremos brevemente. En la gráfica 9 se exhiben dos conjuntos de frecuencias relativas estimadas para las 50 parejas de municipios[14] origen-destino más frecuentes. El primero (a la izquierda) proviene de los resultados de la EOD; el segundo, se obtiene aprovechando la georreferenciación de tuits. En este caso, contra lo que ocurre en la Encuesta, no es posible hacer un seguimiento por viaje individual. Por esta razón, para cada usuario se identificaron aquellos desde los que publica con mayor frecuencia durante un horario primordialmente nocturno y otro, sobre todo diurno.

A simple vista, ambos conjuntos muestran comportamientos diferentes. Para precisar la fuente de tales diferencias, se eligieron las cinco parejas de municipios origen-destino más frecuentes (ver cuadro 2.a) y las cinco menos habituales (ver cuadro 2.b), según *Twitter*. En ambas se incluye, también, una

14 Por simplicidad, nos referiremos como municipios a las actuales alcaldías o demarcaciones territoriales en la Ciudad de México.

Gráfica 9

## Frecuencia relativa entre 50 parejas origen-destino más frecuentes, según la EOD, 2017, por fuente




Álvaro Obregón · Atizapán de Zaragoza · Azcapotzalco · Benito Juárez · Coacalco de Berriozábal · Coyoacán · Cuauhtémoc · Ecatepec de Morelos · Gustavo A. Madero · Iztacalco · Iztapalapa · Miguel Hidalgo · Naucalpan de Juárez · Nezahualcóyotl · Tláhuac · Tlanepantla de Baz · Tlalpan · Tultitlán · Venustiano Carranza

**Fuente:** cálculos propios.

Cuadro 2.a

## Cinco parejas de municipios origen-destino más frecuentes según *Twitter*

| Origen | Destino | EOD | Twitter |
|---|---|---|---|
| Miguel Hidalgo | Cuauhtémoc | 2.11 % | 15.23 % |
| Cuauhtémoc | Miguel Hidalgo | 2.12 % | 10.46 % |
| Coyoacán | Cuauhtémoc | 1.54 % | 9.48 % |
| Benito Juárez | Cuauhtémoc | 2.18 % | 8.83 % |
| Cuauhtémoc | Coyoacán | 1.52 % | 7.77 % |

Cuadro 2.b

## Cinco parejas de municipios origen-destino menos frecuentes según *Twitter*

| Origen | Destino | EOD | *Twitter* |
|---|---|---|---|
| Iztapalapa | Nezahualcóyotl | 1.90 % | 0.08 % |
| Nezahualcóyotl | Iztapalapa | 1.95 % | 0.08 % |
| Nezahualcóyotl | Gustavo A. Madero | 1.56 % | 0.04 % |
| Tultitlán | Coacalco de Berriozábal | 1.38 % | 0.04 % |
| Tláhuac | Iztapalapa | 1.55 % | 0.03 % |

columna con las frecuencias según la EOD con fines de comparación, en tanto que los valores en las columnas EOD en los dos cuadros muestran cifras semejantes (entre 1.4 y 2.2 %), los de las de *Twitter* presentan una disparidad importante entre ambos cuadros (por arriba de 7 % y hasta 15 %, en el caso del cuadro 2.a, pero por debajo de 0.1 %, en el 2.b). Bajo el supuesto de que el diseño muestral de la Encuesta garantiza la representatividad de sus resultados, se tendría que *Twitter* exhibiría un sesgo favorable a las parejas de municipios en el cuadro 2.a.

Vale la pena destacar que, además, en el cuadro 2.a quedan incluidos solo municipios de la Ciudad de México, algunos de los cuales se encuentran entre los que exhiben los niveles socioeconómicos más altos del país, y que no son necesariamente los de mayor densidad poblacional. Lo contrario parece ocurrir, en cambio, con algunos de los del segundo conjunto; entre ellos se encuentran los que tienen mayor número de pobladores en el país, según el Censo de Población y Vivienda 2010, por lo que llama la atención que los desplazamientos entre ellos, según *Twitter*, aparezcan subrepresentados al ser comparados con resultados de la EOD 2017; las discrepancias entre los obtenidos mediante ambas fuentes pueden deberse, al menos en parte, a los ya comentados sesgos favorables a los niveles socioeconómicos altos entre los usuarios de *Twitter*.

## Elecciones federales 2018

Para la presidencial se presentaron cuatro candidatos a quienes identificaremos como Mead, Anaya, AMLO y Bronco. Las encuestas de preferencia electoral daban como favorito a AMLO, del Movimiento de Regeneración Nacional (MORENA).

Este evento nos brinda la oportunidad de estudiar el posible uso de la publicación de tuits como complemento de las encuestas de preferencias electorales e intención de voto previas a cada elección. Adicionalmente, por supuesto, contaríamos más tarde con el resultado de la elección misma de acuerdo con lo publicado por las autoridades electorales. Además de las encuestas difundidas por diversos medios de comunicación se contaba, en este caso, con información que permitía dar seguimiento a la evolución de estado de ánimo de los tuiteros a lo largo de las 12 semanas previas al evento, es decir, a partir de la designación de candidatos, de abril a junio.[15] Cabe señalar que, en este caso, se procedió a seleccionar tuits que fueron clasificados con contenido político. De manera adicional, se incluyeron solo mensajes que mencionaban a los candidatos (por nombre, apodo o algún otro identificador) o a las coaliciones que contendían. En todos los casos se evaluó la emoción del tuitero, pero se incluyeron en los resultados nada más aquellos clasificados como positivos por considerarlos *votos favorables*, eliminando, de este modo, los *negativos*, que carecen de sentido en el sistema electoral mexicano.

La gráfica 10 muestra (con las entidades ordenadas alfabéticamente), la acumulación de tuits favorables a cada uno de los candidatos durante los tres meses de campaña. A diferencia de lo consignado por diversas encuestas levantadas en el periodo, los tuiteros parecen favorecer al candidato Meade, cualquiera que sea el estado desde el que tuitean, de manera casi consistente. En general, Anaya y AMLO disputarían el segundo lugar en reñida competencia, variando entre una y otra entidad. La última posición, en cambio, correspondió siempre a Bronco, quien no parecía representar una seria competencia para los demás candidatos participantes.

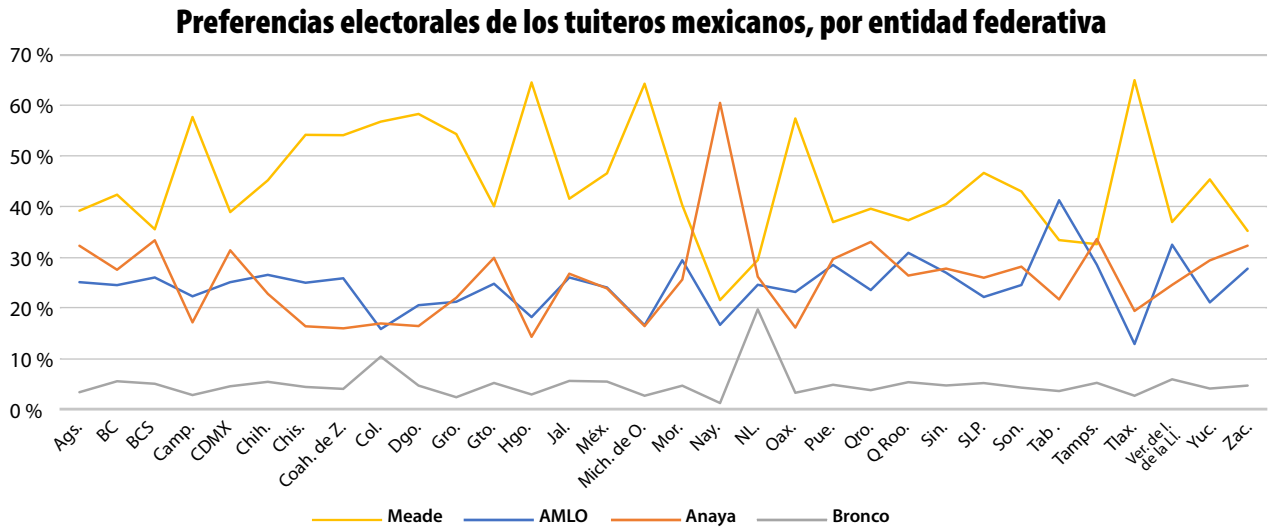
La gráfica 11 presenta la evolución nacional a lo largo de las semanas obtenida a partir de la base de datos desarrollada exprofeso. En ella se percibe nuevamente que el más favorecido por los tuiteros es Meade. Dicha preferencia muestra, sin embargo, una tendencia decreciente a lo largo del periodo considerado, pasando de 50 a 40 %, aunque nunca es alcanzado por ninguno de los otros candidatos; en segundo lugar, se tiene al candidato Anaya,

15 Algunas precisiones son necesarias para contextualizar los siguientes resultados: cada tuitero puede publicar más de un mensaje en cada agregado espacial o temporal. Tal vez sería útil asegurar que cada tuitero sea considerado solo una vez en cada agregado. Además, el anterior análisis no toma en cuenta la publicidad pagada en favor de algún candidato. No queda clara la ventaja de dar a la publicidad un origen geográfico definido, lo que lo haría aparecer entre los tuits georreferenciados; así pues, ni la publicidad pagada ni los bots fueron eliminados.

quien se mantiene con entre 25 y 30%; AMLO vio crecer su posición desde casi 20 % hasta más de 25 %; por último, Bronco se mantuvo siempre por debajo de 10 % de las preferencias. A esta gráfica se le añadió el resultado oficial de la elección.

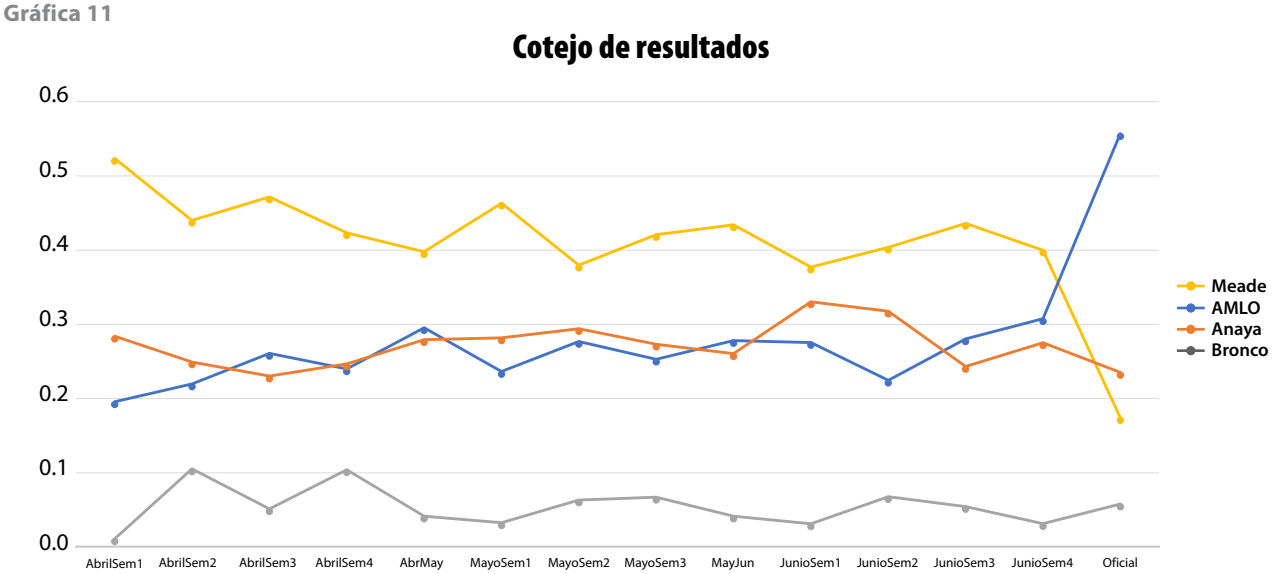
La ausencia de congruencia entre estos resultados y los mostrados por nuestro análisis a partir de *Twitter* (excepto para el candidato Bronco) parece evidenciar que la población que compone nuestra base de datos no representa a la que acudió a votar el 1 de julio de 2018, aunque aquella pueda estar contenida en esta. El día de la votación, el número de tuits capturados tuvo su máximo del segundo semestre del 2018. Sin embargo, el cociente de positividad para ese mismo día resultó en 1.57, el segundo valor más bajo del año, solo por detrás del 1.51 alcanzado el 29 de octubre, cuando se dio a conocer la decisión del nuevo gobierno de cancelar la construcción del Nuevo Aero-

Gráfica 10




**Fuente:** elaboración propia a partir de la base de datos de INEGI. *Estado de ánimo de los tuiteros.*

Gráfica 11




**Fuente:** elaboración propia a partir de la base de datos de INEGI. *Estado de ánimo de los tuiteros* e INE. *Numeralia proceso electoral 2017-2018.*

puerto Internacional de México. Este resultado parece reforzar lo señalado líneas arriba en el sentido de que, en el tema político, la población de tuiteros en nuestra base de datos no es representativa de la de votantes en México.

## Conclusiones

Aun cuando nuestro análisis sociodemográfico resulte en buenas noticias para el negocio de la publicidad, queda claro que, para las oficinas productoras de estadística oficial (como el INEGI), la información no puede ser aprovechada pues, al parecer, los usuarios de redes sociales solamente se representan a sí mismos. Sin embargo, la identificación y cuantificación de las discrepancias entre las poblaciones de usuarios y la objetivo en sus indagaciones abre la posibilidad ya comentada de modificar la importancia relativa de cada usuario, según sus características sociodemográficas, con el fin de que las mediciones realizadas en cada uno de los subgrupos se encuentren más cercanas a lo que ocurriría para la población abierta. Por supuesto, para ello debe realizarse un análisis a nivel de usuario y no de mensaje publicado pues, como ya se mencionó, el número de mensajes publicados por usuario puede resultar muy variable.

La experiencia acumulada mediante los anteriores ejercicios ha sido invaluable y contribuirá a situar al INEGI a la vanguardia en el ámbito de la explotación de información proveniente de redes sociales para la producción de estadística oficial. Particularmente importante es el hecho de que el grupo de técnicos y profesionales que laboran en el Instituto, y que se han capacitado en el uso de las técnicas relevantes, alcanza ya un número considerable. Asimismo, los planes para el fortalecimiento de la infraestructura muestran avances relevantes. Es preciso reconocer que, a lo largo de los diferentes trabajos reportados en este documento, se ha cumplido con el propósito didáctico que, como se mencionó, alentó los primeros esfuerzos.

Entre los temas de futura investigación se encuentran algunas aplicaciones que van más allá de la mera identificación de los sesgos incurridos al usar información de usuarios de redes sociales, y que se refieren al uso conjunto de datos provenientes de encuestas, por un lado, y por el otro, a los de redes sociales. Primero, se buscará establecer estrategias para la corrección de los sesgos identificados. Si informantes identificados en la ENDUTIH como usuarios de alguna red nos concedieran acceso a su cuenta, estaríamos en condiciones de asociar sus características sociodemográficas recogidas por la Encuesta con los textos de sus publicaciones en dicha red. De este modo, podríamos entrenar un algoritmo que nos permita predecir características sociodemográficas de los usuarios de redes que no formaron parte de la muestra, a partir de sus publicaciones. Dependiendo de la calidad de este resultado se estaría, ahora sí, en condiciones de reponderarlos. Ello nos permitiría, por ejemplo, pasar de los resultados del estado de ánimo de los tuiteros al de los mexicanos según *Twitter*. Es decir, de ser exitosa esta primera experiencia, estaremos en condiciones no solo de predecir el sexo, el grupo de edad o el logro académico de los usuarios de esta red, sino de mejorar la representatividad de los resultados obtenidos de la explotación de esa fuente de información.

## Fuentes


BBC. *Costo de datos móviles en América Latina: en qué países es más caro usar internet en el celular (y dónde cuesta menos)*. Redacción, BBC News Mundo, 5 de marzo de 2019 (DE) https://www.bbc.com/mundo/noticias-47455825, consultado el 21 de mayo de 2021.

Van den Brakel, J., E. Söhler, P. Daas y B. Buelens. *Social media as a data source for official statistics; the Dutch Consumer Confidence Index*. Survey Methodology. Vol. 43, No. 2, December 2017, pp. 183-210. Statistics Canada, Catalogue No. 12-001-X (DE) https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017002/article/54871-eng.pdf?st=LYTvhP20, consultado el 21 de mayo de 2021.

Data-Pop Alliance. *Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America*. Data-Pop Alliance, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute, November 2016.

Destatis. *Access to Big Data for statistical purposes* (Note by the Federal Statistical Office of Germany, Economic Commission for Europe). Paris, Conference of European Statisticians, 67th Plenary Session, 26-28 June

2019 (DE) https://undocs.org/ECE/CES/2019/20, consultado el 21 de mayo de 2021.

DNP. *Definición de la estrategia de Big Data para el estado colombiano y para el desarrollo de la industria de Big Data en Colombia, 2017-2018*: *Estado del arte y análisis comparativo de estrategias nacionales d*e Big Data. (DE) http://datapopalliance.org/wp-content/uploads/2018/09/Documento1_VersionFinal_DNP.pdf, consultado el 21 de mayo de 2021.

Dutra. *Las organizaciones deben implementar una estrategia centrada en los datos*. 2018 (DE) https://www.telefonica.com/es/web/public-policy/blog/articulo/-/blogs/las-organizaciones-deben-implementar-una-estrategia-centrada-en-los-datos, consultado el 21 de mayo de 2021.

Iacus, S., G. Porro, S. Salini y E. Siletti. "Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal", in: *Journal of Official Statistics*. Vol. 36, No. 2, 2020, pp. 315-338 (DE) http://dx.doi.org/10.2478/JOS-2020-0017, consultado el 21 de mayo de 2021.

Instituto Nacional Electoral. *Numeralia proceso electoral 2017-2018*. Final, 08/06/2018 (DE) https://www.ine.mx/wp-content/uploads/2018/08/1Numeralia01072018-SIJE08072018findocx-3.pdf, consultado el 21 de mayo de 2021.

Istat. *Experimental statistics new challenges for NSOs: Istat*. Geneva, Economic Commission for Europe, Conference of European Statisticians. Sixty-sixth Plenary Session, June 18-20, 2018 (DE) https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2018/CES_37_Sem_I_S2_Italy.pdf, consultado el 21 de mayo de 2021.

Jansen, R. *UN Global Working Group (GWG) on Big Data and its Task Teams*. Hangzhou, China, International Symposium on the Use of Big Data for Official Statistics, National Bureau of Statistics of China, Oct. 16-18, 2020 (DE) http://www.stats.gov.cn/english/pdf/202010/P020201012399997943871.pdf, consultado el 21 de mayo de 2021.

Letouzé, E. y J. Jütting. "Official Statistics, Big Data and Human Development", en: Data-Pop Alliance. *White Paper Series*. Data-Pop Alliance, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute and Paris21. March 2015 (DE) https://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf, consultado el 21 de mayo de 2021.

Lokanathan, S., T. Perera-Gomez y S. Zuhyle. *Mapping Big Data Solutions for the Sustainable Development Goals [Draft]*. LIRNEasia, 2017 (DE) https://lirneasia.net/2017/03/mapping-big-data-solutions-sustainable-development-goals/, consultado el 21 de mayo de 2021.

Mecinas, J. M. "The digital divide in Mexico: a mirror of poverty", in: *Mex. Law Rev*. Vol. 9, No. 1. Mexico, jul./dec. 2016, pp. 93-102 (DE) https://revistas.juridicas.unam.mx/index.php/mexican-law-review/article/view/10432/12508, consultado el 21 de mayo de 2021.

Moctezuma, D., M. Graff, S. Miranda-Jiménez, E. Sadit Tellez, A. Coronado and C. N. Sánchez. *A Genetic Programming Approach to Sentiment Analysis for Twitter*: *TASS '17*. 2017.

Shayaa *et al. Linking consumer confidence index and social media sentiment analysis*. Cogent Business & Management. 5: 1509424, 2018 (DE) https://pdfs.semanticscholar.org/a899/e7f0abbe336554706de7e2bb742f92d31f6a.pdf, consultado el 21 de mayo de 2021.

Snyder, N. *UN Global Working Group on Big Data*. UNECE Workshop on Statistical Data Collection, Washington, D. C., 29 April-1 May 2015.

Struijs, P., B. Braaksma and P. Daas. *Official statistics and Big Data, Big Data & Society*. April-June 2014, pp. 1-6, DOI: 10.1177/2053951714538417 (DE) https://journals.sagepub.com/doi/abs/10.1177/2053951714538417, consultado el 21 de mayo de 2021.

Struijs, P. "Official statistics and Big Data, XXIXª", en: Seminario Internacional de Estadística. EUSTAT (DE) https://en.eustat.eus/elementos/ele0018400/58-international-statistics-seminar/inf0018432_i.pdf, consultado el 21 de mayo de 2021.

Struijs, P. y P. Daas. "Quality approaches to Big Data in official statistics", in: *European Conference on Quality in Official Statistics, 2014* (DE) http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf, consultado el 21 de mayo de 2021.

Schiavoni, C., F. Palm, S. Smeekes & J. Van den Brakel. "A dynamic factor model approach to incorporate Big Data in state space models for official statistics", in: *J R Stat Soc*. Series A. 184. 2021, pp. 324-353 (DE) https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12626, consultado el 21 de mayo de 2021.

Stark, T. H. U"nderstanding the selection bias: Social network processes and the effect of prejudice on the avoidance of outgroup friends", in: *Social Psychology Quarterly*. 78(2), 2015, pp. 127-150 (DE) https://doi.org/10.1177/0190272514565252, consultado el 21 de mayo de 2021.

Van Halderen, G., I. Bernal, T. Sejersen, R. Jansen, N. Ploug y M. Truszczynski. *Big Data for the SDGs, Country examples in compiling SDG indicators using non-traditional data sources*. Working Paper Series. ESCAP Statistics Division, SD/WP/12/January 2021 (DE) https://www.unescap.org/sites/default/d8files/knowledge-products/SD_Working_Paper_no12_Jan2021_Big_data_for_SDG_indicators.pdf, consultado el 21 de mayo de 2021.

UNSD. *Report of the Global Working Group on Big Data for Official Statistics.* New York, Statistical Commission Forty Sixth Session, March 3-6, 2015 (DE) https://documents-dds-ny.un.org/doc/UNDOC/GEN/N14/692/71/PDF/N1469271.pdf?OpenElement, consultado el 21 de mayo de 2021.