Vol. 13, Núm. 2, mayo-agosto, 2022.

ISSN 2007-2961

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

INTERNATIONAL JOURNAL OF STATISTICS AND GEOGRAPHY

REALITY, DATA AND SPACE

INEGI

Vol. 13, Núm. 2, mayo-agosto, 2022.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA
REALITY, DATA AND SPACE INTERNATIONAL JOURNAL OF STATISTICS AND GEOGRAPHY

Contenido

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Indizada en: Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal *Latindex Catálogo*; Citas Latinoamericanas en Ciencias Sociales y Humanidades (*CLASE*) y en la Red Iberoamericana de Innovación y Conocimiento (REDIB).

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA, Vol. 13, Núm. 2, mayo-agosto, 2022, es una publicación cuatrimestral editada por el Instituto Nacional de Estadística y Geografía, Avenida Héroe de Nacozari Sur 2301, Fraccionamiento Jardines del Parque, 20276, Aguascalientes, Aguascalientes, Aguascalientes, entre la calle INEGI, Avenida del Lago y Avenida Paseo de las Garzas, México. Teléfono 55 52781069. Toda correspondencia deberá dirigirse al correo: rde@inegi.org.mx

Editor responsable: Sergio Carrera Riva Palacio. Reserva de Derechos al Uso Exclusivo del Título Núm. 04-2012-121909394300-102, ISSN Núm. 2007-2961, ambos otorgados por el Instituto Nacional del Derecho de Autor. Certificado de Licitud de Título y Contenido Núm. 15099, otorgado por la Comisión Calificadora de Publicaciones y Revistas Ilustradas de la Secretaría de Gobernación. Domicilio de la publicación: Avenida Héroe de Nacozari Sur 2301, Fraccionamiento Jardines del Parque, 20276, Aguascalientes, Aguascalientes, Aguascalientes, entre la calle INEGI, Avenida del Lago y Avenida Paseo de las Garzas, México.

Versión electrónica: http://rde.inegi.org.mx
ISSN 2395-8537

Editorial

Este segundo número de 2022 de *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía* está conformado por los siguientes artículos. A continuación, se describe su contenido.

Modelo de estimación del impacto del INEGI en la investigación y divulgación científica (*Model for Estimating INEGI's Impact on Scientific Research and Dissemination*). Este trabajo propone una metodología de medición en el contexto científico que considera indicadores bibliométricos y altmétricos. Para ello, se realizó la identificación de fuentes de información científica, recuperación de esta en publicaciones y perfiles de autores en redes sociales, así como la estandarización de los datos obtenidos. Se validó la métrica empleando un subconjunto de datos representativos y se detectó que los resultados pueden ser equiparables; se aplicó al periodo 1984-2018, observándose notables incrementos en el peso de las consultas al Instituto Nacional de Estadística y Geografía (INEGI) en cada uno de los últimos tres lapsos sexenales (20, 32 y 40 %, respectivamente). Se considera que a futuro pueden generarse modelos predictivos del impacto por áreas del conocimiento que maneja el Instituto, como la económica y sociodemográfica; de gobierno, seguridad pública e impartición de justicia; geográfica; medio ambiente; así como de ordenamiento territorial y urbano.

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*). Desde unos años atrás se ha considerado que los textos publicados en los medios de comunicación en la web representan una oportunidad para producir información oficial, por lo que varias oficinas nacionales de estadística han experimentado con su uso. Dado que es necesario saber si los usuarios de estos muestran diferencias importantes con la población general de México, según variables sociodemográficas relevantes que garanticen la comparabilidad, en esta investigación se propone llevar a cabo la cuantificación de tales discrepancias utilizando datos de la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH), ediciones 2017-2019, del INEGI.

Contribución del sistema financiero al crecimiento económico de México: un análisis econométrico, 1997-2019 (*Contribution of the Financial System to Mexico's Economic Growth: An Econometric Assessment, 1997-2019*). Este trabajo contribuye a la discusión sobre el papel que tiene el conjunto de agentes económicos en el desarrollo nacional utilizando una novedosa aplicación empírica para el caso de México. El análisis de factores dinámicos permite concluir que los factores estimados representan de manera consistente, en sentido estadístico y económico, la evolución del sistema financiero en el país. Los resultados

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

apoyan también la hipótesis de que este es especialmente relevante en un contexto de recuperación económica.

Evaluación de técnicas de procesamiento de lenguaje natural y Machine Learning *para los procesos de codificación de encuestas en hogares* (*Evaluation of Natural Language Processing and Machine Learning Techniques for Household Survey Coding Processes*). El artículo tiene por objetivo valorar el uso e incorporación de técnicas de inteligencia artificial y aprendizaje automático (NLP y ML, por sus siglas en inglés, respectivamente) para incrementar el porcentaje de registros clasificados de manera automatizada. Para ello, se consideraron las variables de ocupación y actividad económica de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2018. Los resultados obtenidos muestran que sería posible trasladar 50 % de los registros que actualmente se codifican con asistencia humana hacia un proceso de codificación con algoritmos de NLP y ML.

Imputation of Non-Response in Height and Weight in the "Mexican Health and Aging Study" (*Imputación de no-respuesta en peso y talla en el Estudio Nacional de Salud y Envejecimiento*). El propósito de este trabajo es explicar las razones y el procedimiento de imputación de datos antropométricos autorreportados en las primeras cuatro rondas de ese programa estadístico del INEGI. Se destaca el efecto de esta frente a la eliminación de los casos con datos faltantes, comparando la distribución de dichos valores y sus efectos asociados con el Índice de Masa Corporal mediante un modelo de regresión. Se concluye que la incorporación de datos imputados ofrece resultados más sólidos en comparación con la omisión de situaciones con información faltante. De ahí la importancia de aplicar estos procedimientos estadísticos con un manejo adecuado de los datos y difundir la metodología aplicada para obtener los datos imputados desde la misma fuente de información, tal como se ofrece en dicho estudio.

Hoja de ruta para producir frecuentemente información estadística representativa mediante el uso conjunto de información de redes sociales y encuestas (*Roadmap for Frequently Producing Representative Statistical Information through the Joint Use of Social Networking and Survey Data*). Este documento muestra una propuesta metodológica para que las oficinas nacionales de estadística generen información característica sobre múltiples temas, con mayor regularidad, utilizando en conjunto datos de entrevistas en hogares y publicaciones en los medios de comunicación de la web. El estudio se basa en dar un nuevo rol a los datos como insumo para el entrenamiento de algoritmos de aprendizaje automático. Para que esto funcione, las respuestas de los usuarios-informantes deben ser vinculadas. Un futuro levantamiento de la ENDUTIH del INEGI se empleará para estudiar la viabilidad de la propuesta, puesto que esta ya investiga el uso de las redes sociales y recopila información sociodemográfica.

Propuesta de indicadores alternativos para medir la confianza del consumidor (*Proposed Alternative Indicators for Measuring Consumer Confidence*). En este trabajo se presentan sugerencias para mejorar la correlación del Indicador de

Confianza del Consumidor (ICC) con variables relevantes de la situación económica actual, como la actividad económica y el consumo privado. A partir de las 15 preguntas que produce la Encuesta Nacional sobre Confianza del Consumidor (ENCO) del INEGI y adicionalmente de indicadores de fuentes no tradicionales, específicamente de *Google Trends*, se estimaron ICC alternativos al actual.

Esperanza de vida sin limitaciones físicas ni mentales en México (*Life Expectancy without Physical or Mental Limitations in Mexico*). En esta investigación se presentan estimaciones producto de la aplicación del método de Sullivan en tablas de vida periodo construidas directamente con información censal y estadísticas de mortalidad. Es una propuesta original para el uso inmediato de los datos que produce y concentra el INEGI, que considera la diversidad nacional según sexo y entidad de residencia en el diseño de este indicador sintético de salud, que refleja la calidad de vida y no solo la expectativa de su duración. La autora sugiere que es necesario continuar aprovechando la información longitudinal que produce el Instituto para refinar los indicadores en esta materia.

Lecturas en lo que indican los indicadores (*Readings in What Economic Indicators Indicate*) es la reseña de una obra que actualiza y expande el material contenido en el libro *Lo que indican los indicadores. Cómo utilizar la información estadística para entender la realidad económica de México*, editado, publicado y difundido por el INEGI hace 10 años. Esta nueva publicación es el producto de la excelente coordinación del trabajo de 33 autores que abordan distintos temas y formas sobre cómo explicar y entender los indicadores para analizar y entender la coyuntura económica de nuestro país.

http://rde.inegi.org.mx

Modelo de estimación del impacto del INEGI **en la investigación y divulgación científica**

Model for Estimating INEGI's Impact **on Scientific Research and Dissemination**

José Román Herrera Morales,* Luis Francisco Barbosa Santillán, Jorge Rafael Gutiérrez Pulido,* María Andrade Aréchiga* y Sara Sandoval Carrillo***

En el ámbito de la investigación, el Instituto Nacional de Estadística y Geografía (INEGI) es reconocido como un proveedor confiable de información; sin embargo, no cuenta con un perfil de autor en sistemas especializados de información científica como *SCOPUS*, *ORCID*, *Google* Académico y otros. Esto significa que no es posible medir su impacto de forma directa haciendo uso de tales sistemas. Este trabajo propone una metodología para medirlo en el contexto científico, que considera no solo indicadores bibliométricos tradicionales sino, también, los altmétricos. Las tareas que se llevaron a cabo para tal fin son: identificación de fuentes de información científica en la web, recuperación de esta de las publicaciones y de perfiles de autores en redes sociales, así como la estandarización de los datos recuperados. Posteriormente, se procedió a la identificación de aquellos descriptores útiles para generar una métrica del impacto en la producción científica del INEGI; se analizó la contribución de cada uno de ellos aplicando técnicas de aprendizaje automático, como la regresión gradual con adición de

In the scientific context, INEGI is highly recognized as a reliable information provider. Despite this, it has not an author profile in specialized scientific information systems such as SCOPUS, ORCID, and Google Scholar among others. This means that its impact cannot be measured directly from them. This study proposes a methodology to measure such an impact, considering both traditional bibliometric and altmetric indicators. The tasks that were carried out for the creation of a reference database of publications that cite INEGI (BDREF) are: identification of sources of scientific information on the web, retrieval of this information from publications and authors' profiles in social networks, as well as the standardization of the data retrieved. Afterwards, we proceeded to identify those descriptors useful for generating an Impact Index to the scientific production of the INEGI (IIPCI). The significance of each of these descriptors was analyzed, applying machine learning techniques such as gradual regression with the addition of significant variables, for the construction of an opti-

* Universidad de Colima, rherrera@ucol.mx, jrgp@ucol.mx, mandrad@ucol.mx y sary@ucol.mx, respectivamente.
** Universidad Tecnológica de Puebla, luis.barbosa@utpuebla.edu.mx

variables significativas, para la construcción de un modelo de predicción optimizado; se validó la métrica empleando un subconjunto de datos representativos y se obtuvieron resultados que pueden ser equiparables; se aplicó dicha métrica al periodo 1984-2018, observándose notables incrementos del impacto científico del Instituto en cada uno de los últimos tres lapsos sexenales (20, 32 y 40 %, respectivamente). Asimismo, se tuvo la mayor tasa de crecimiento por año (26 %) del 2011 al 2012 y un máximo histórico de 6.9 unidades en el 2016.

Palabras clave: medición de impacto; indicadores bibliométricos; índices altmétricos; minería de datos.

Recibido: 12 de febrero de 2021.
Aceptado: 6 de mayo de 2021.

mized prediction model. The IIPCI was validated using a subset of representative data and comparable results were obtained. Applying the IIPCI in the period from 1984 to 2018, notable increases were observed in each of the last three six-year periods (20%, 32% and 40% respectively). Likewise, there was the highest growth rate per year (26%) in the period 2011-2012 and a historical maximum of 6.9 units in 2016.

Key words: impact measurement; bibliometric index; altmetrics; data mining.

Human hands using laptop for working his business having Data Analysis Diagram/ forrest9 / iStoc

Introducción

Para medir cualitativamente el grado de impacto que una institución, artículo, patente o libro ejerce en los trabajos que se producen dentro de la comunidad científica, se han propuesto diversas métricas; una de las más usadas son los indicadores bibliométricos. Dentro de estos destacan el número de citas, el factor de impacto (FI) y el Índice H (IH). El primero es el más utilizado en bibliometría de impacto debido a que cuantifica de forma directa la cantidad de veces que una investigación es usada para construir otras; además, es la base para obtener el FI y el IH (Velasco, 2012).

Esta propuesta que presentamos se enmarca en el proyecto 290379 aprobado en la convocatoria S0025-2016-02 del Fondo Sectorial Consejo Nacional de Ciencia y Tecnología (CONACYT)-Instituto Nacional de Estadística y Geografía (INEGI) donde, a petición del Laboratorio de Microdatos del INEGI, se busca conocer el impacto de la información que en este se genera, los servicios especializados que ofrece y responder preguntas como: ¿quiénes están utilizando fuentes de datos del Instituto para fundamentar sus investigaciones?, ¿qué tipos de publicaciones son las que más están citando al INEGI?, ¿de qué países e instituciones son los autores de estas publicaciones?, ¿cómo se ha comportado esta labor de apoyo a la investigación a lo largo del tiempo?, ¿qué tantas publicaciones científicas utilizan la información de Instituto para fundamentar su trabajo?, ¿cómo cuantificar el impacto científico que ha tenido el INEGI a través del tiempo con los datos que procesa y ofrece a la población? Las respuestas a estas interrogantes fueron plasmadas en el informe técnico de la primer etapa del proyecto (Herrera-Morales, 2018), fase que se enfocó en la identificación de referencias de uso a la información que el INEGI genera para, enseguida, rastrear menciones o citas que permitan medir su impacto no solo desde la perspectiva tradicional de la cantidad de citas que sus documentos reciben, sino empleando otros indicadores que dejan, también, conocer su impacto en otros contextos, como las redes sociales y la web. En las siguientes secciones se mencionan tanto los indicadores bibliométricos tradicionales como los altmétricos que fueron considerados como componentes de nuestra propuesta.

Principales métricas de impacto científico de publicaciones y revistas

El FI es una de las más importantes que se emplean en el contexto científico para expresar la calidad de trabajos de investigación; es una razón que se obtiene con la cantidad de citas recibidas por los artículos publicados en una revista, divididos entre el número de los publicados en un determinado periodo (Garfield, 1955), el cual es, típicamente, dos años. Sin embargo, esta métrica no expresa de forma directa la calidad de un artículo de investigación sino, más bien, el de la revista donde se publicó; así, una de mayor calidad es aquella que cuenta con un FI más elevado debido a que los estudios publicados en ella fueron más leídos y citados. Desafortunada o afortunadamente para los autores, al ser aplicado este indicador sobre un grupo de trabajos, da lugar a que algunos, sin ningún mérito en cuanto a citas, se cuelguen de la fama de otros más citados, dado que se trata de un promedio, en realidad (Seglen, 1997). Por ello, es importante enfatizar que este indicador hace referencia al impacto de la revista y no al de las publicaciones que cada autor tiene.

Hirsch (2005) estableció el IH, que evalúa la producción científica de un investigador, combinando su número de publicaciones y la cantidad de citas que ha recibido. Se puede usar, también, para dar a conocer la relevancia de una institución; asimismo, detecta los autores más destacados en un área de conocimiento. No obstante, resulta inadecuado comparar entre los de diferentes áreas, pues penaliza de manera injusta a los que publican poco, pero con una gran cantidad de citas recibidas en sus trabajos. Por estas deficiencias y la queja de la comunidad científica al ser evaluado su rendimiento y el de sus instituciones con estos indicadores, se desarrolló una serie de recomendaciones sobre la adecuada evaluación de la investigación (DORA, 2018) que establecen que el FI no debe usarse

como la única manera para evaluar personal, sino que también deberían utilizarse otras métricas basadas en el contenido científico que consideren diferentes aspectos (su visibilidad y la cantidad de citas recibidas, entre otros), como los indicadores altmétricos.

En fecha reciente se han publicado trabajos donde se intenta predecir el impacto que tendrá una publicación científica empleando sofisticados algoritmos que imitan el pensamiento humano, por ejemplo, las redes neuronales profundas (Abrishami, 2019) o considerando la cantidad de citas tempranas y el FI dentro de un modelo de regresión lineal (Abramo, 2019). Sin embargo, en estas investigaciones, la métrica importante sigue siendo la cantidad de citas recibidas por las publicaciones.

Indicadores altmétricos para medición de impacto científico

Como consecuencia del abuso de las bibliométricas para etiquetar qué institución o investigador merece ser premiado, se ha generado un estancamiento en el desarrollo científico debido a que se persigue a toda costa el publicar en revistas con alto FI abandonando algunas de las demás alternativas en la producción científica, como la participación en congresos o el desarrollo de proyectos complejos y, por ende, tardados, generando monocultivos científicos que conducen al empobrecimiento de la ciencia y la uniformización del pensamiento (Delgado-López-Cózar y Martín-Martín, 2019). Debido a estas problemáticas, a partir del 2010 surgieron nuevas propuestas sobre indicadores de impacto asociados a la web, denominadas *altmetrics* o métricas alternativas, y se presentan como un complemento a la evaluación de la actividad científica (Bornmann, 2014). La idea que subyace en ellas es que las menciones en blogs, noticieros, *facebook*, número de lecturas, tuits y descargas hechas en la web se consideren, también, como indicadores del impacto de las publicaciones científicas en contextos de mayor acceso y cobertura de la información para la sociedad en general (Torres-Salinas, 2013).

A pesar de lo complejo que resulta cuantificar este tipo de métricas alternativas, hoy en día, por fortuna, existen productos de empresas especializadas que ofrecen estos servicios. Entre los principales destacan: *Mendeley* de Elsevier, *Lagotto* de *PLOS*, *OurResearch* (antes *ImpactStory*), *Plum*X de *SCOPUS*, así como las plataformas *Crossref* y *Altmetric* (Ortega, 2020). La mayoría de estos proveedores emplean el *Document Object Identifier (DOI)* para registrar las interacciones de cada publicación en los diversos portales de internet desde donde se accede a estos recursos.

El uso de métricas alternativas (indicadores altmétricos) ha cobrado relevancia con el paso de los años; por ejemplo, Eysenbach (2011) demostró que, a partir de la divulgación de una publicación mediante tuits, se puede incrementar su número de citas. También, se ha evaluado la correlación que tienen los indicadores altmétricos con los bibliométricos, obteniéndose resultados variados, dependiendo del área de estudio. Nocera, Boyd, Boudreau, Hakim y Rais-Bahrami (2019) encontraron una muy baja correlación y Mullins *et al*. (2020), una alta. Sin embargo, todos los autores de estos trabajos coinciden en una idea en común: que las métricas alternativas captan el impacto sobre una dimensión diferente y deben ser usadas como un complemento de medición a los indicadores bibliométricos tradicionales (Bardus, 2020). En esta propuesta son consideradas como un elemento que permite medir el impacto con respecto a la contribución que realiza cada autor y publicación en los entornos de las redes sociales y en la divulgación científica en la web.

Proveedores de información científica

Se tiene una amplia oferta de fuentes y servicios especializados de esta: *ScienceDirect* de *SCOPUS* y *Mendeley* de *Elsevier*, *Web of Science* de *Clarivate Analytics* (antes Thompson-Reuters), *Plos-One* de *PLOS*, *Dimensions*, *PubMed* de la Biblioteca Nacional de Medicina de Estados Unidos de América, *EBSCO Host*; para el habla hispana, están *REDALYC* y *Scielo* y, en particular para México, el Repositorio

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Nacional del CONACYT. Además, están los servicios especializados de indexado y búsqueda, como *Google* Académico, *Microsoft Research Academic*, entre otros, que hacen uso de diversas herramientas y algoritmos especializados de recuperación de información científica y académica. Todos los anteriores fueron evaluados como posibles fuentes de datos y, con los seleccionados, se emplearon diversas técnicas para la recuperación de la información relevante de las publicaciones de interés para este proyecto.

Propuesta de medición de impacto para el INEGI

El Instituto es un organismo autónomo del Estado mexicano, cuyo objetivo es la generación y difusión de datos estadísticos y geográficos del país en aspectos como el territorio, el medio ambiente, la población, la economía y el gobierno. En el ámbito de la investigación, es reconocido como un proveedor confiable de información oficial y tiene presencia a través de las publicaciones de investigadores tanto nacionales como internacionales en las que hacen referencia al Instituto. Sin embargo, se identificó que no la tiene como autor en sistemas especializados de información científica; es decir, no cuenta con un perfil en *SCOPUS*, *ORCID*, *Google* Académico y otros, lo cual se traduce en que no se puede tener una determinación de cuáles son las publicaciones que ha generado y, por lo tanto, no se puede hacer una identificación, trazabilidad y medición de su impacto *cienciométrico*.

La idea de esta propuesta considera, primero, identificar qué publicaciones utilizan al INEGI como referencia en sus investigaciones y, enseguida, recuperar e integrar diversos indicadores (tanto bibliométricos tradicionales como altmétricos) que permitan medir el impacto de una forma integral, donde se consideran varios aspectos, no solo las citas. Esta información fue recopilada de varios sistemas especializados y, con ella, se formó la Base de Datos de Referencias Bibliográficas (BDREF). Nuestra metodología permite medir el impacto de las publicaciones del INEGI y hace uso de la BDREF para crear un modelo de estimación de su impacto con respecto a su uso en los ámbitos académico y científico.

En la siguiente sección se proporciona una descripción amplia de la metodología seguida en la construcción de indicadores del impacto relacionados con la producción científica. Asimismo, se presentan los resultados y las discusiones relacionadas con ellos. Finalmente, en la última sección se muestran las conclusiones de este trabajo.

Metodología

Esta consiste en dos etapas:

- Generación de la BDREF.
- Índice del Impacto en la Producción Científica del INEGI (IIPCI):
  - Identificación de predictores candidatos.
  - Construcción del IIPCI mediante la selección de predictores significativos.

A continuación, se detalla el proceso que se llevó a cabo en cada una de ellas.

Generación de la BDREF

Para su formación, fue necesaria, en primera instancia, la identificación de fuentes de información especializadas, como las mencionadas arriba en la subsección *Proveedores de información científica*. Enseguida, en una selección de estas fuentes de datos, se llevó a cabo la extracción de las etiquetas documentales o atributos descriptivos de las publicaciones, mismas que se clasificaron en dos grupos: las *citantes* (de terceros que utilizan al INEGI como referente) y las *citadas* (documentos generados por el Instituto que fueron usados como referencia) (Herrera-Morales, 2019). También, se realizó la integración y homogenización de los datos para asegurar su calidad, corrigiéndose varias inconsistencias, como: registros incompletos, duplicados, parecidos con ambigüedad de datos, con caracteres especiales y codificaciones de idiomas diferen-

tes. Además, la BDREF incluye los indicadores altmétricos recuperados de cada publicación *citante* y la información sobre los perfiles de los autores recuperados de redes sociales, que servirán como un insumo de datos para la construcción del IIPCI.

Índice del Impacto en la Producción Científica del INEGI

En esta subsección presentamos cada uno de los pasos que se siguieron para construirlo:

- Identificación de 12 predictores candidatos de la información en la BDREF.
- Obtención y normalización de la tabla de datos de cada predictor por año (periodo 1984-2018).
- Generación del Modelo Base de Impacto (MBI) con los 12 predictores, tomando una suma ponderada de estos.
- Optimización del Modelo mediante la técnica de regresión gradual con adición de variables.
- Reconstrucción del MBI.
- Comparación gráfica de los comportamientos temporales del modelo optimizado contra el MBI.

Identificación de predictores candidatos
De la BDREF se eligieron 12 indicadores (predictores) para estimar el impacto del Instituto: publicaciones INEGI *citadas*, publicaciones *citantes* al INEGI, citas de las publicaciones *citantes*, penalización por citas imprecisas (ver cuadro A1), búsquedas del término INEGI en *Google*, Índice Altmétrico de las Publicaciones *Citantes* (IAPC), publicaciones *citantes* con DOI, publicaciones *citantes* de acceso abierto, Índice Ponderado de Autores (IPA, *ResearchGate Score* e IH), Índice Ponderado de Redes de Colaboración de Autores de Publicaciones *Citantes* (IPRCAPC), proporción de publicaciones *citantes* entre publicaciones INEGI por año, proporción por año entre el número de citas de las publicaciones *citantes* y la cantidad de publicaciones citantes.

Cabe mencionar que los primeros predictores son indicadores directos que fueron recopilados y asociados a sus respectivas publicaciones, mientras que otros representan variables compuestas que llevaron un procesamiento adicional para tener un significado; este es el caso del IAPC, que contabiliza y pondera cada uno de los indicadores altmétricos recuperados para cada publicación *citante* que fue identificada (Martínez-Barajas *et al*., 2021) (ver cuadro A2). Situación similar es la del IPRCAPC, donde se hace un análisis de los coautores de cada publicación *citante* y se les asigna un peso de acuerdo con la cantidad de autores y sus respectivas nacionalidades (Mosqueda-González et al., 2021) (ver cuadro A3). Para una mayor descripción de ellos, ver el *Anexo A*.

Para obtener los valores observados por año de cada uno de los 12 predictores, se utilizaron expresiones SQL para extraer la información de la BDREF y se *normalizaron* los resultados para que todos los valores quedaran en un rango entre 0 y 1.

Generación del MBI con los 12 predictores
Se ha utilizado la información de los predictores mencionados considerando los valores normalizados para tener un MBI como el que se expresa en la fórmula 1:

$$MBI_i = X_1 + X_2 + X_3 - X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10} + X_{11} + X_{12} \quad (1)$$

donde:

i = impacto obtenido en cada uno de los años de interés (1984-2018).

En nuestro caso, hemos considerado la posible contribución de 12 predictores para construir el índice por lo que, a continuación, se hace un análisis de la relevancia de cada uno de estos.

Selección de los indicadores más significativos
Cuando se tienen múltiples predictores en un modelo se debe analizar que la información que aportan no esté cubierta por otro predictor, por lo tanto, es necesario optimizar este modelo. Existen diversos métodos cuantitativos para realizar este

tipo de análisis, como el de componentes principales (PCA) y selección de variables, entre otros, que se emplean cuando tenemos muchas variables explicativas y queremos saber cuáles son las que aportan algo nuevo. Para ello, primero se hace un análisis de dependencia de variables predictoras mediante el cálculo de la matriz de correlación de Pearson. Así, las m variables que presenten entre ellas valores altos de dependencia lineal se separarán de forma momentánea del modelo. Después, evaluamos la posible reincorporación de cada una de ellas, siguiendo la técnica de regresión gradual con adición de variables descrita en Guyon & Elisseeff (2003), la cual considera, en un inicio, todos los predictores disponibles en el MBI y, con un procedimiento iterativo de regresión, se procede a seleccionar el conjunto óptimo de estos, junto con cada uno de sus pesos asociados. A continuación, se describe de forma detallada la metodología.

En el primer paso, se dividen los 12 predictores (conjunto A) en dos subconjuntos, uno con los no correlacionados (conjunto B) y otro con el complemento de este, es decir, los correlacionados (conjunto C).

Después, formamos un subconjunto de entrenamiento $E(i)$ mediante la unión del conjunto B y el primer elemento del C. Entrenamos con el algoritmo gradiente descendente (Sebastian Ruder, 2017) el subconjunto en turno $E(i)$ para obtener un modelo de predicción candidato MRIi. En el siguiente paso, calculamos y almacenamos el estadístico *T-student* asociado al predictor tomado del conjunto C, siempre y cuando su valor-p sea menor a 0.05, es decir, que sea significativo para el modelo. Repetimos todo este procedimiento, pero ahora con el siguiente predictor del conjunto C y así, sucesivamente, hasta terminar con todos los elementos de este conjunto; es decir, cuando el número de iteración i sea igual al de elementos $n(C)$. Después, movemos del conjunto C al B el predictor $m(i)$ que presentó el mayor valor absoluto del estadístico T. Calculamos el error estándar del modelo entrenado con el conjunto actualizado B. Se verifica si se cumple la condición de paro, la cual establece que el error estándar sea menor o igual al umbral de 0.1; si no sucede así, se vuelve a iterar.

Al terminar todo este proceso —es decir, cuando se cumpla la condición de paro— obtenemos un modelo optimizado y reducido con $n(B)$ predictores, junto con sus pesos óptimos asociados (ver diagrama), para la predicción del impacto científico del INEGI, el cual presentamos en la sección de resultados.

Reconstrucción del MBI

El algoritmo de gradiente descendente tiene doble importancia en el método de medición del impacto, puesto que se utiliza para realizar el entrenamiento y, además, cuando ya se tiene el conjunto de predictores (conjunto D) optimizado, se usa también para la reconstrucción del MBI. El peso asignado en el algoritmo a cada variable explicativa se establece aleatoriamente en primera instancia. Este, al buscar ir reduciendo el error, va ajustando por sí mismo los pesos θ en cada repetición. El ajuste se da de manera gradual según la tasa de aprendizaje indicada; además, es guiado por la derivada o gradiente del error, de ahí su nombre. Por último, después de haberse repetido un número determinado de veces, alcanza el mínimo global de la función de costo, también conocida como función de error. Cuando el algoritmo converge al mínimo valor de error, los pesos óptimos del modelo de regresión son alcanzados. A continuación, se escribe matemáticamente la técnica utilizada. Sea:

$$J(\theta_{\emptyset}, \theta_1, \ldots, \theta_n) = \frac{1}{2} m \sum_{i=1}^{m} h_{\theta}(x_i - y_i)^2 \qquad (2)$$

donde:

$$h_{\theta}(x) = \theta_{\emptyset} X_{\emptyset} + \theta_1 X_1 + \ldots + \theta_n X_n \qquad (3)$$

θ_j = pesos de regresión.
J = función de error.
α = tasa de aprendizaje, con valor inicial de 0.03.
h = función de regresión.

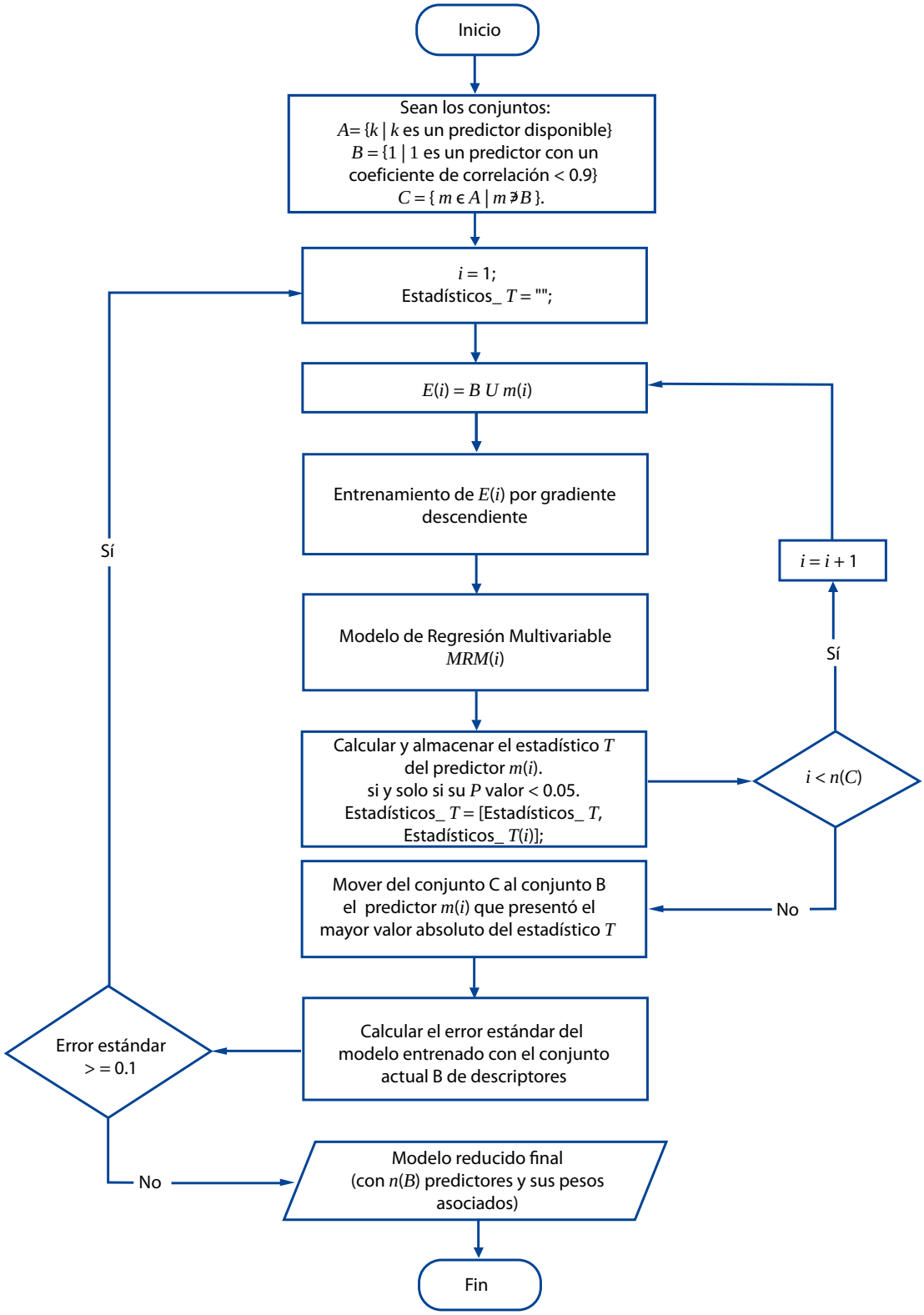
$$\text{Hacer } \{\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_j) \qquad (4)$$

simultáneamente actualizar $\theta_j\}$ repetir hasta que converja θ_j.

Entonces, considerando los pesos óptimos generados por este algoritmo junto con los predic-

Diagrama

Metodología empleada para obtener el modelo optimizado para cuantificar el impacto científico del INEGI



tores o variables asociadas, se obtiene el Modelo de Regresión Multivariable (MRM), referido en la metodología propuesta. Al final, con el modelo optimizado se podrán hacer comparaciones contra el MBI y evaluar si corresponde a una representación apropiada del original.

Resultados

En esta sección se describen los principales hallazgos obtenidos de las dos etapas propuestas en la metodología, la obtención de la base de datos de referencias y la identificación de predictores útiles para generar los modelos IIPCI.

Generación de la base de datos de referencias

Esta contiene la integración de las referencias de las publicaciones *citantes* al INEGI en el periodo 1984-2018 que fueron identificadas y recolectadas de las fuentes de datos seleccionadas *SCOPUS*, *REDALYC*, *Mendeley*, *Google* Académico, *Core API* y *Dimensions* (cf. arriba subsección *Generación de la BDREF* de *Metodología*). Para llevar a cabo su proceso de extracción, se desarrollaron herramientas de *software* que emplearon técnicas de obtención de datos para la recuperación de los atributos de las publicaciones, en particular web *scraping* o el empleo de sus respectivas API (Castrejón-Mejía *et al*., 2020, Alvarado-Villa *et al*., 2019, Madrigal-Martínez *et al*., 2018).

La BDREF (basada en el modelo de datos relacional) tiene una estructura *ad hoc* diseñada para almacenar la información descriptiva general sobre cada uno de estos documentos *citantes* y *citados*, entre los que destacan las etiquetas del título y la fecha de la publicación, el idioma, los autores, la revista y la editorial, así como algunos atributos de identificación estandarizados, ya sea para la publicación, la revista o los autores (como el DOI, ISSN, ISBN y *ORCID*, entre otros).

Entre los principales hallazgos que podemos mencionar sobre la BDREF se encuentran la identificación de 44 235 registros de publicaciones, de las cuales 28 984 son *citantes* (ver cuadro 1) y 15 251, *citadas*.

Considerando un análisis sobre el tipo de documento del conjunto de publicaciones *citantes*, se encontró que 82 % corresponde a artículos publicados en *journals*; el resto se dividió entre capítulos de libro (6.5 %), libros (2.4 %), tesis en diferentes grados (0.3 % de doctorado, 1.0 % de maestría y 1.7 % de pregrado) y otros formatos (cerca de 4 %).

Respecto al idioma predominante encontrado en la escritura de las publicaciones *citantes*, encontramos que el español tiene 53.1 % de los registros, seguido del inglés con 44.1 %; el resto lo cubren otros idiomas, como francés, portugués y alemán, con una participación menor a 1 % cada uno. Si se verifica que las publicaciones *citantes* cuentan con un identificador universal (DOI), resulta que 51 % sí lo tiene; el restante no fue identificado.

En cuanto a las publicaciones *citadas* (INEGI), se cuantificó un subconjunto de estas, considerando aquellas en las que se identificó con precisión su fecha de publicación, resultando 13 595, de las cuales el total cuenta con un *Universal Product Code (UPC)* asignado por el mismo Instituto. Acerca del DOI, ningún registro de publicaciones del INEGI lo tiene. Sin embargo, de este subconjunto de registros, se pudo precisar que 45.73 % está identificado con un código ISBN, mientras que 11.64 % cuenta con ISSN.

Cuadro 1

Publicaciones *citantes* recuperadas en cada fuente de datos seleccionada

Fuente de datos	*CORE*	*Dimensions*	*Google* Académico	*Mendeley*	*REDALYC*	*SCOPUS*
Cantidad	610	12 800	3 647	470	9 721	1 736

Referente a los perfiles académicos de los autores, se logró la identificación en el sistema *SCOPUS* de 15 054 de los 49 325 registrados en la BDREF, empleando para ello el *SCOPUS-ID*, que fue un atributo identificador que nos permitió hacer el *match* entre los investigadores que teníamos registrados en la base de datos. De los identificados se extrajeron los datos de su institución de adscripción e información relevante de su perfil académico. Después, teniendo certeza en los datos del nombre del autor y su adscripción, se recuperaron 4 764 perfiles utilizando el portal *ResearchGate*, que incluyen información complementaria de ellos, junto con el Índice de Impacto *RG Score*, que asigna la misma plataforma. Adicionalmente, se identificaron 1 184 perfiles en el portal de *Publons* y 223 en *Twitter*.

Modelado del Índice de Impacto INEGI

En esta subsección se evalúa la pertinencia de cada uno de los 12 predictores como posibles miembros del modelo de predicción optimizado (IIPCI), basado en algoritmos de aprendizaje automático y la técnica de regresión gradual con adición de variables. Se consideran como insumos de entrada los valores obtenidos para cada uno de los predictores extraídos de la BDREF, mismos que se muestran en el *Anexo B*.

Análisis de los predictores

De acuerdo con el proceso propuesto arriba en la subsección *Reconstrucción del MBI*, se obtienen los predictores que están correlacionados entre los 12 disponibles mediante la matriz de correlación de Pearson. Además, se validan visualmente con la gráfica 1a aquellos que presentaron correlación mayor a 0.9 y se aprecia cómo siguen una tendencia similar los siguientes predictores: IAPC (PubCitantes_Altmetric), publicaciones *citantes* con DOI (PubCitantes_DOI), publicaciones *citantes* de acceso abierto (PubCitantes_OpenAccess), IPA (AutoresCitantes_Score) y el IPRCAPC (AutoresCitantes_Colaboraciones).

En contraste, los que no presentaron una marcada correlación fueron: publicaciones INEGI *citadas* (INEGI_Citados), publicaciones *citantes* al INEGI (Pub_Citantes), citas de las publicaciones *citantes* (Citas_Pub_citantes), penalización por citas imprecisas (Cita_imprecisa) y búsquedas del término INEGI en *Google* (Tendencias_Google). Su comportamiento a través del tiempo se muestra en la gráfica 1b, donde se observa un incremento abrupto para el predictor Tendencias_Google en el 2004 debido a que fue cuando *Google* inició el registro de tendencias de búsqueda.

Asimismo, en la gráfica 1c se aprecia el comportamiento de los predictores compuestos, que son la proporción de publicaciones *citantes* entre publicaciones INEGI por año (INEGI_Citados) y la proporción por año entre el número de citas de las publicaciones *citantes* entre la cantidad de publicaciones *citantes* (Pub_Citantes); se percibe que no existe mayor correlación entre ellos.

Modelo matemático optimizado (IIPCI)

Teniendo el conjunto de 12 predictores candidatos (A) clasificados en dos conjuntos, los altamente correlacionados (C) y los no correlacionados (B), se aplicaron los pasos descritos en la metodología propuesta (ver el diagrama y subsección *Reconstrucción del MBI* de *Metodología*) para llegar a un modelo con ocho predictores significativos. El modelo matemático optimizado obtenido para IIPCI se muestra en la fórmula 5:

$$IIPCI_i = X_1 + 2.9X_2 + 1.4X_3 - 0.9X_4 + 0.9X_5 + 2.9X_8 + 0.7X_{11} + X_{12} \quad (5)$$

donde:

i = año de muestreo en el periodo 1984-2018.
$IIPCI$ = Índice de Impacto en la Producción Científica del INEGI.

Nótese cómo el cuarto término del modelo se resta de los demás, lo cual se debe a que representa los errores cometidos por el que hace las citas.

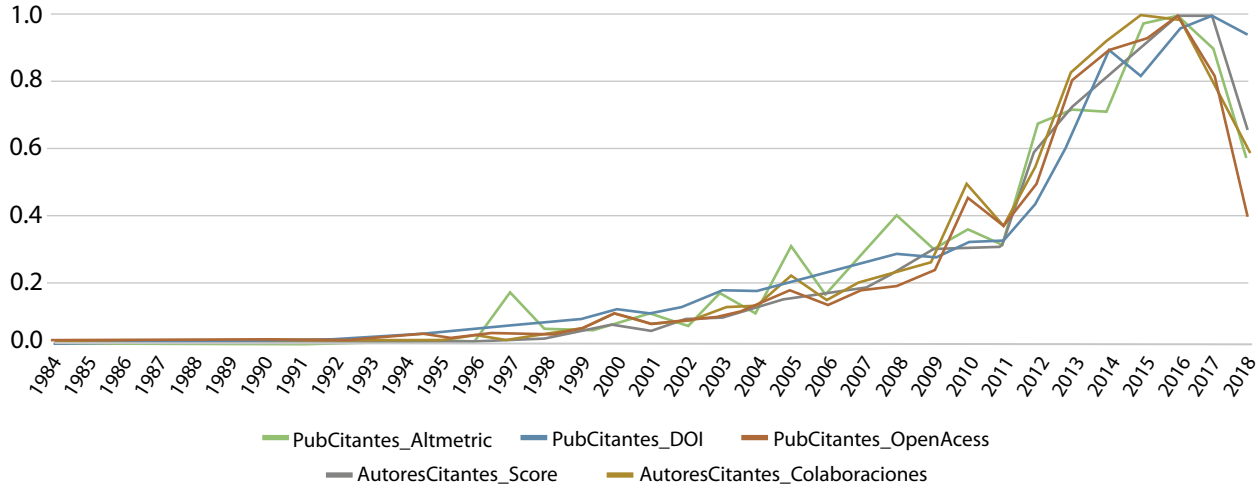
Reportamos, también, los resultados estadísticos del análisis de regresión obtenidos de cada

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

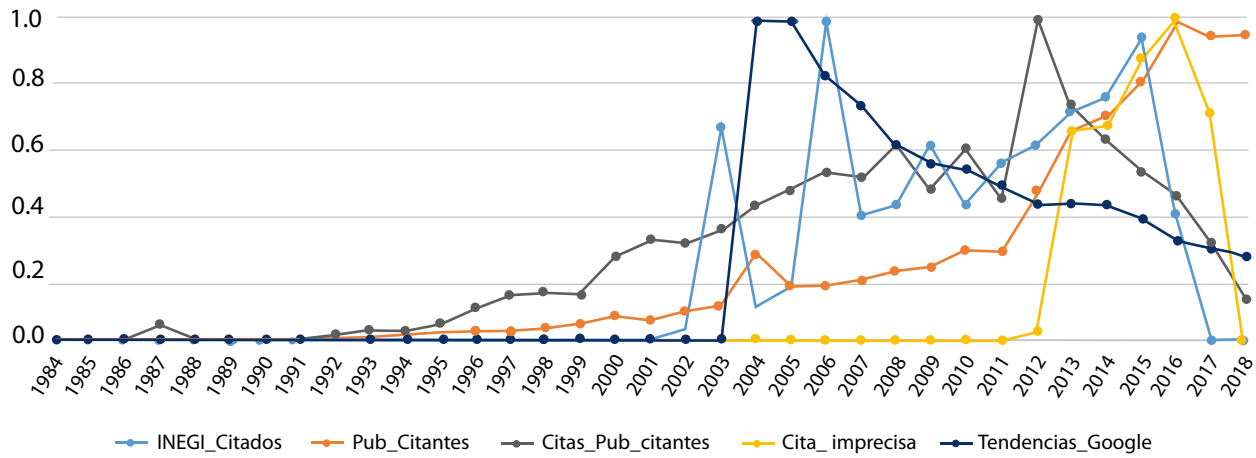
Gráficas 1

Tendencia de los valores normalizados de los predictores

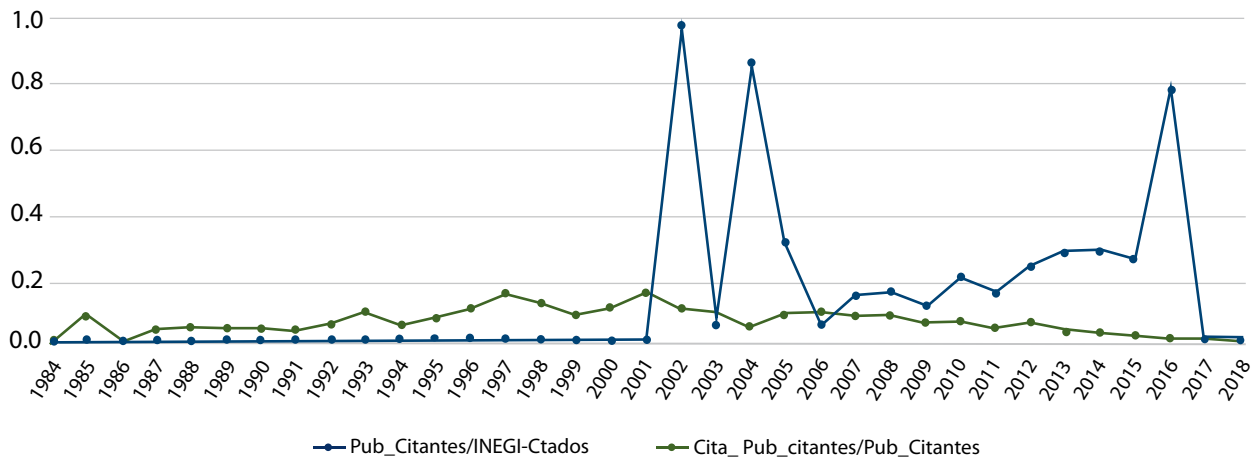
a) Comportamiento temporal de los cinco predictores correlacionados



b) Comportamiento temporal de predictores no correlacionados



c) Comportamiento temporal de los predictores compuestos



uno de los ocho predictores miembros del modelo (ver cuadro 2). En todos los casos, el valor-*p* (*p-value*) asociado fue menor a 0.05, lo cual garantiza una confianza mínima de 95 % e indica que el predictor es significante.

En el cuadro 2 se aprecia que no aparecen los valores para X_6, X_7, X_9, X_{10} debido a que no cumplieron con el criterio de regresión aplicado. También, se reporta gráficamente la evolución temporal que presenta año con año el índice IIPCI a partir de 1984 (ver gráfica 2).

Como se puede observar en la gráfica 2, se obtienen comportamientos equivalentes tanto con el MBI usando los 12 predictores como con el modelo

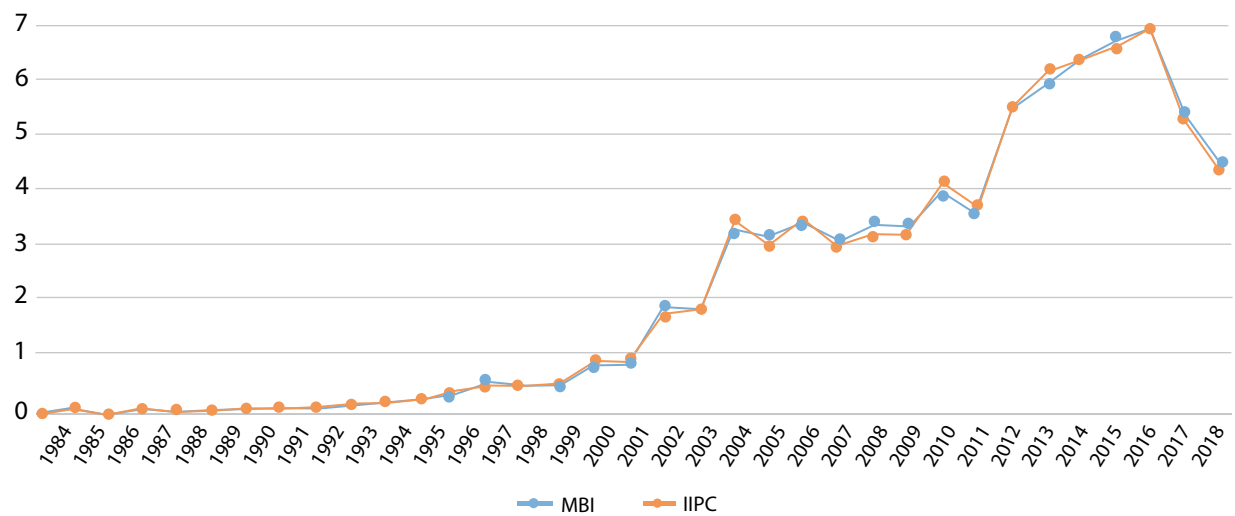
Cuadro 2

Resultados estadísticos del análisis de regresión para los predictores significativos del modelo IIPCI

Parámetro	Coeficiente estimado	Error estándar	*T-student*	*p-value*
Intercepción	-0.031	0.029	-1.056	0.30011034
INEGI_Citados (X_1)	0.998	0.105	9.541	2.6871E-10
Pub_Citantes (X_2)	2.952	0.168	17.522	1.2628E-16
Citas_Pub_citantes (X_3)	1.459	0.188	7.778	1.7959E-08
Cita_imprecisa (X_4)	-0.947	0.225	-4.203	0.00024315
Tendencias_Google (X_5)	0.916	0.086	10.633	2.4406E-11
PubCitantes_ OpenAccess (X_8)	2.965	0.340	8.717	1.8205E-09
Pub_Citantes/INEGI_ Citados (X_{11})	0.781	0.091	8.605	2.3782E-09
Citas_Pub_Citantes/ Pub_Citantes (X_{12})	1.032	0.115	8.953	1.0429E-09

Gráfica 2

Comparación del IIPCI considerando el MBI (con 12 predictores) vs. IIPCI optimizado (con ocho predictores)



REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

IIPC que solo emplea ocho de estos; prácticamente, las dos curvas siguen el mismo patrón de conducta, por lo tanto, resulta válido emplear el modelo optimizado IIPC para representar a través del tiempo el comportamiento del impacto científico del INEGI. Asimismo, debemos mencionar que no estamos comparando contra otros modelos de reducción de dimensionalidad —como PCA—, sino contra una optimización del mismo modelo base MBI.

Otra forma de apreciar el impacto del INEGI a través del tiempo es midiendo la zona bajo la curva del IIPCI, calculando el porcentaje del área que ocupa en cada uno de los sexenios recientes con respecto a la total. En la gráfica 3 se observa un evidente incremento de este por cada sexenio que transcurre; en particular, en el último (2012-2018) es donde se ha tenido el más alto con un poco más de 40 % del total estimado. Además, se tuvo la mayor pendiente positiva o tasa de crecimiento por año que corresponde a 26 % en el periodo 2011-2012 y un pico máximo histórico de 6.9 sobre 7.0 unidades en el 2016.

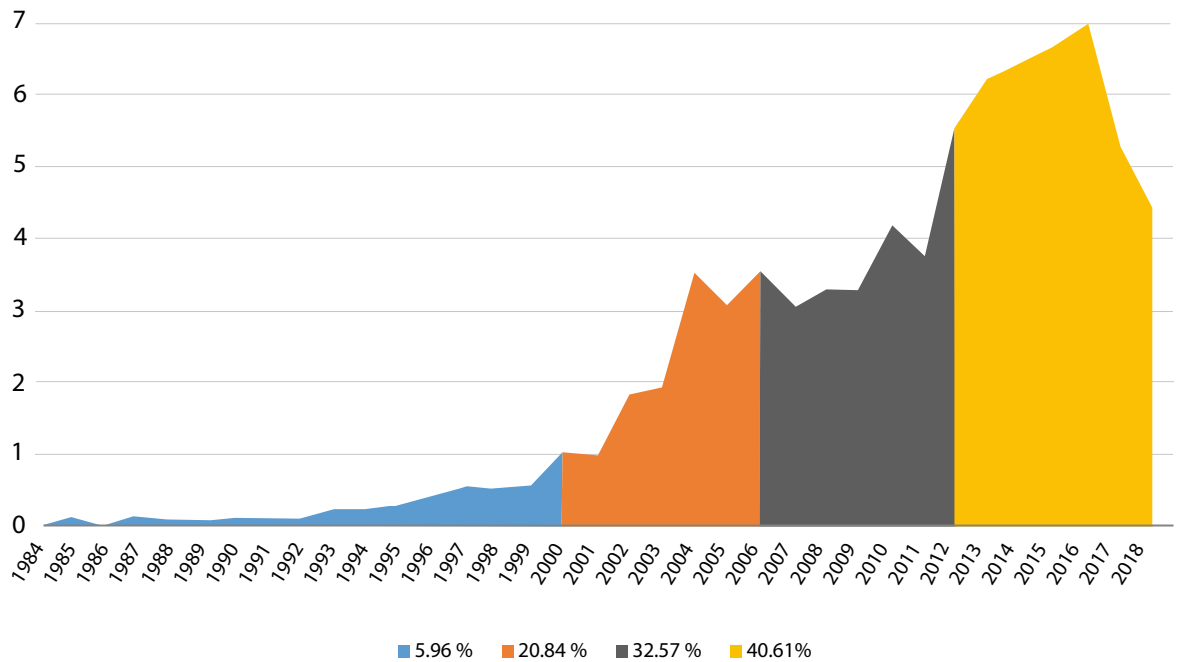
Conclusiones

En este trabajo se ha presentado una metodología de medición del impacto científico que ha tenido el INEGI en los ámbitos académicos y de divulgación científica, que consiste en un modelo de estimación basado en un índice compuesto de diversos indicadores y métricas relacionadas con el uso de información generada por el Instituto, como citas o referencias en publicaciones científicas, al cual denominamos Índice de Impacto en la Producción Científica del INEGI y no es comparable con otro tipo de métrica individual conocida o empleado en sistemas especializados de información científica.

Para llegar al diseño de este, se realizaron varias tareas, entre las que destacan la formación

Gráfica 3

IIPCI representado a lo largo del tiempo agrupado por los últimos periodos sexenales



de una base de datos de referencias de publicaciones *citantes* al INEGI (BDREF) mediante la identificación de repositorios digitales, minado de sus atributos referenciales y métricas asociadas a las publicaciones *citantes*, filtrado y depurado de registros duplicados e inconsistentes y la estandarización de datos.

El análisis de la información almacenada en la BDREF incluyó diversos patrones descriptivos evaluados a través del tiempo, como: cantidad de citas a publicaciones del INEGI, cantidad de publicaciones que citan al Instituto, cantidad de citas que obtuvieron las publicaciones *citantes* al INEGI, impacto ponderado de métricas altmétricas asociadas a publicaciones *citantes* a este organismo, entre las que se incluyen descargas, me gusta (*likes*), lecturas del resumen, menciones en *Twitter*, etc., así como penalizaciones aplicadas por falta de referencia formal, citas incorrectas o ambiguas al INEGI en las publicaciones citantes, cantidad de publicaciones *citantes* a esta institución que son de acceso abierto, cantidad de publicaciones *citantes* al INEGI que tienen un identificador universal DOI, impacto asociado a perfiles de autores con presencia en redes sociales académicas, Índice de Tendencias de Búsqueda en *Google Trends*, impacto de redes de colaboración entre coautores de publicaciones *citantes* al Instituto. Con esta serie de predictores se generó un modelo donde se analizó su respectiva contribución mediante la técnica de aprendizaje artificial denominada regresión gradual con adición de variables significativas.

Al tratarse el Índice de Impacto en la Producción Científica del INEGI de un modelo predictivo de estimación del impacto para el Instituto que fue entrenado y optimizado con 12 predictores extraídos mediante un proceso exhaustivo de minado de datos, se prevé que en los años siguientes se pueda tener dificultades para tener la disponibilidad de todos estos indicadores que se requieren para la alimentación del modelo, por lo que nos dimos a la tarea de buscar un subconjunto de datos de fácil disponibilidad y que pudiera emular de forma apropiada la entrada de información al modelo de estimación IIPCI con lo cual obtuvimos uno optimizado que utiliza menos predictores y que pudiera ser usado en un escenario más acotado, como el Laboratorio de Microdatos del INEGI.

Como contribución, este modelo puede ser replicable y aplicable para medir el impacto de diversas instituciones generadoras de contenido científico siguiendo la metodología descrita y requiriendo, para ello, recopilar los datos respectivos de cada uno de los indicadores considerados en el modelo. Con los resultados que se obtengan se podrían hacer comparaciones equiparables entre varios organismos.

A diferencia del modelo de regresión lineal que utilizan Abramo *et al*. (2019), el cual solo considera dos variables (número de citas tempranas y FI de la revista) para medir y predecir el impacto científico de una publicación, el nuestro resultó más completo e integral dado que emplea ocho variables para su estimación (ver *Anexo A*) considerando tanto predictores clásicos como altmétricos.

Los resultados derivados de este análisis asisten al INEGI en determinar qué información de sus programas estadísticos (censos, encuestas, etc.) es más utilizada y cuál es menos empleada en el campo científico. Esto puede contribuirle en la toma de decisiones estratégicas u operativas en el país al ubicar qué censos y encuestas debe fortalecer, así como orientar su periodicidad, entre otros aspectos. Aunque nuestra propuesta está pensada para el INEGI, puede ser aplicada a otras instituciones, incluyendo autores individuales que sean generadores de contenido con una problemática similar.

Finalmente, como trabajo futuro, se pueden generar modelos predictivos del impacto por áreas del conocimiento que maneja el INEGI, como económica y sociodemográfica; gobierno; seguridad pública e impartición de justicia; geográfica; medio ambiente; así como ordenamiento territorial y urbano. Para ello, se puede aplicar un análisis semántico latente de los resúmenes de los artículos c*itantes* del Instituto y, con ello, presentar resultados del impacto, detallados por área de conocimiento.

Fuentes

Abramo, G., C. A D'Angelo & G. Felici. "Predicting publication long-term impact through a combination of early citations and journal impact factor", en: *Journal of Informetrics*. 13(1). 2019, pp. 32-49.

Abrishami, A. & S. Aliakbary. "Predicting citation counts based on deep neural network learning techniques", en: *Journal of Informetrics*. 13(2). 2019, pp. 485-499.

Altmetric. *How is the Altmetric Attention Score calculated?* 2020 (DE) último acceso el 20 de junio de 2020 en https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-score-calculated-

Alvarado-Villa, D. A. & M. C. Santa Ana. *Acceso y recuperación de metadatos de publicaciones digitales científicas utilizando los servicios de* Dimensions. Tesis para obtener el grado de Ingeniero de *Software*. Facultad de Telemática, Universidad de Colima, 2019.

Alvarez, J. J., F. Almenárez Mendoza & M. Labrador. "An accurate way to cross reference users across Social Networks", en: *Southeast Con 2017*. Charlotte, NC, 2017, pp. 1-6 (DE) https://doi.org/10.1109/SECON.2017.7925366

Bardus, M., R. El Rassi, M. Chahrour, E. W. Akl, A. S. Raslan, L. I. Meho y E. A. Akl. "The Use of Social Media to Increase the Impact of Health Research: Systematic Review", en: *Journal of Medical Internet Research*. 22(7). e15607, 2020.

Bornmann, L. "Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics", en: *Journal of Informetrics*. 8(4). 2014, pp. 895-903 (DE) https://doi.org/10.1016/j.joi.2014.09.005

Castrejón-Mejía, O. E. *Aplicación de minería de datos basado en servicios de información de* SCOPUS. Tesis para obtener el grado de Ingeniero de Software. Facultad de Telemática, Universidad de Colima, 2020.

Delgado-López-Cózar, E. & A. Martín-Martín. "El factor de impacto de las revistas científicas sigue siendo ese número que devora la ciencia española: ¿hasta cuándo?", en: *Anuario ThinkEPI*, 13. 2019 (DE) https://doi.org/10.3145/thinkepi.2019.e13e09

DORA. *Declaración de San Francisco sobre la evaluación de la investigación*. Traducción por Beatriz Pardal-Peláez. 2018 (DE) https://sfdora.org/read/es

Eysenbach, G. "Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact", en: *Journal of Medical Internet Research*. 13(4). 2011, pp. 123 (DE) https://doi.org/10.2196/jmir.2012

Garfield, E. "Citation indexes for science; a new dimension in documentation through association of ideas", en: *Science*. 122(3159). 1955, pp. 108-111.

Guyon, I. & A. Elisseeff. "An introduction to variable and feature selection", en: *Journal of Machine Learning Research*. 3(Mar), 2003, pp. 1157-1182.

Herrera-Morales, J. R. *Primer informe técnico del proyecto 290379 del Fondo Sectorial CONACyT-INEGI: medición del uso de los servicios del Laboratorio de Microdatos y procesamiento remoto en investigaciones académicas y en la fundamentación de la definición, operación y evaluación de política pública*. México, 2019.

Hirsch, J. E. "An index to quantify an individual's scientific research output", en: *Proceedings of the National Academy of Sciences*. 102(46). 2005, pp. 16569-16572 (DE) https://doi.org/10.1073/pnas.0507655102

James, J. "Data Never Sleeps 7", en: *Domosphere*. 2019 (DE) último acceso el 20 de junio de 2020 en https://www.domo.com/learn/datanever-sleeps-7

Madrigal-Martínez, K. F., J. R. Herrera-Morales, A. R. Gallardo, S. Sandoval-Carrillo y P. C. Santana-Mancilla. "Aplicación de minería de datos para extracción de documentos académicos del repositorio de *REDALYC"*, en: *Coloquio de investigación multidisciplinaria*. ISSN 2007-8102, vol. 6(1). 2018, pp. 1822-1829.

Martínez-Barajas, M. A. *Recolección e integración de indicadores altmétricos asociados a publicaciones científicas que citan al INEGI*. Tesis para obtener el grado de Ingeniero de *Software*. Facultad de Telemática, Universidad de Colima, 2021.

Mosqueda-González, B. A. *Extracción de perfiles académicos y su impacto en redes sociales*. Tesis para obtener el grado de Ingeniero de *Software*. Facultad de Telemática, Universidad de Colima, 2021.

Mullins, Ch., C. J. Boyd & B. L. Corey. "Examining the Correlation Between Altmetric Score and Citations in the General Surgery Literature", en: *Journal of Surgical Research*. 248. 2020, pp. 159-164 (DE) https://doi.org/10.1016/j.jss.2019.11.008

Nocera, A. P., C. J. Boyd, H. Boudreau, O. Hakim & S. Rais-Bahrami. "Examining the Correlation Between Altmetric Score and Citations in the Urology Literature", en: *Urology*. 134. 2019, pp. 45-50 (DE) https://doi.org/10.1016/j.urology.2019.09.014

Ortega, J. L. "Altmetrics data providers: A meta-analysis review of the coverage of metrics and publication", en: *El Profesional de la Información*. 29(1). 2020 (DE) https://doi.org/10.3145/epi.2020.ene.07

Ruder, S. "An overview of gradient descent optimization algorithms", en: arXiv preprint arXiv:1609.04747v2. 2017.

Seglen, Per O. "Why the impact factor of journals should not be used for evaluating research", en: *British Medical Journal*. V. 314. 1997, pp. 498-502 (DE) https://doi.org/10.1136/bmj.314.7079.497

Torres-Salinas, D., Á. Cabezas-Clavijo & E. Jiménez-Contreras. "Altmetrics: New Indicators for Scientific Communication in Web 2.0", en: *Comunicar*. 21(41). 2013, pp. 53-60 (DE) https://doi.org/10.3916/c41-2013-05

Velasco, B., J. M. E. Bouza, J. M. Pinilla & J. A. San Román. "La utilización de los indicadores bibliométricos para evaluar la actividad investigadora", en: *Aula Abierta*. 40(2). 2012, pp. 75-84.

Anexo A

Predictores empleados en el IIPCI

X_1. INEGI_Citados. Se refiere al número de publicaciones que el mismo INEGI produce, derivado de los proyectos estadísticos que realiza periódicamente, como los censos y las encuestas.

X_2. Pub_Citantes. Representa la cantidad de publicaciones académicas por año que tomaron la información del INEGI como base para fundamentar sus trabajos.

X_3. Citas_Pub_citantes. Se refiere al número de citas que tienen las Pub_Citantes.

X_4. Cita_imprecisa. Son aquellas citas incluidas en Pub_Citantes que desafortunadamente no siguen las reglas convencionales para citar como referencia a una obra del INEGI. Se identificaron variantes incorrectas en la forma de citar (ver cuadro A1).

Cuadro A1

Penalización por tipos de imprecisión de la cita al INEGI de acuerdo con el grado de severidad de la falta

Tipo	Peso
Cita en el resumen	1
Cita en tabla o figura	2
Cita en el texto	3
Cita ambigua	4
No hay cita	5
No se identificó o la cita es correcta	0

Así, por cada publicación *citante* que incurra en alguna de las situaciones mostradas en el cuadro A1, se disminuye el índice de impacto, penalizando según el peso correspondiente.

X_5. Tendencias_Google. Este patrón se refiere al número de veces por año que la palabra INEGI ha sido consultada dentro de *Google*. Con el servicio de *Google Trends*, se puede explorar la relevancia o tendencia de cualquier expresión de búsqueda, obteniendo el número de veces que se ha empleado en un periodo y es manejado como porcentajes, considerando 100 % el día en que mayores incidencias se hayan realizado en el lapso seleccionado.

X_6. PubCitantes_Altmetric. Se refiere a un valor resultante que considera todos los indicadores altmétricos de cada una de las Pub_Citantes, de acuerdo con los pesos expresados en el cuadro A2:

$$X_6 = \sum_{k=1}^{m} \sum_{j=1}^{n} (W_{(j)} \, Alt_{(j)}) \qquad (1A)$$

donde:

$W_{(j)}$ = peso ponderado del altmétrico j.
$Alt_{(j)}$ = frecuencia del altmétrico j.
m = número total de publicaciones *citantes* en cada año.
k = índice de iteración de la publicación *citante* en la secuencia.
n = número total de indicadores altmétricos de la k-ésima publicación *citante*.
j = índice de iteración del indicador altmétrico en la secuencia.

Los pesos asignados a cada altmétrico se establecieron considerando las recomendaciones que da *Almetric Company* (Altmetric, 2020) y se detallan en el cuadro A2.

Cuadro A2

Pesos empleados para los indicadores altmétricos de las publicaciones

Tipo de indicador altmétrico	Peso
Lecturas en *Mendeley*	0.1
Lecturas del resumen	0.001
Mención en *Wikipedia*	3
Descargas	0.2
Likes	0.05
Menciones en blogs	5.5
Menciones en periódicos digitales	8-6
Vistas del texto completo	0.01
Tuits	1.0

Fuente: Altmetric, 2020.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

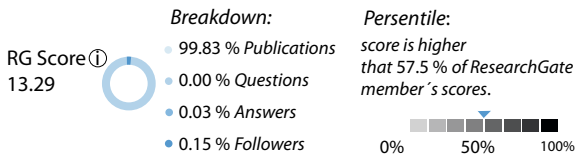
X_7. PubCitantes_DOI. Número de Pub_Citantes que tienen asignado un identificador único global (DOI).

X_8. PubCitantes_OpenAccess. Número de Pub_Citantes que son de acceso abierto (*Open Access*).

X_9. AutoresCitantes_Score. Este patrón es un indicador compuesto que hace referencia a la calidad del autor. Se obtiene sumando dos puntajes: el IH y el asignado por *ResearchGate* (*RG Score*) que toma en cuenta el número de publicaciones, preguntas, respuestas y número de seguidores que tiene el autor en su plataforma (ver gráfica A1).

Gráfica A1

Presentación del puntaje establecido por *ResearchGate* para calificar la productividad de un autor



X_{10}. AutoresCitantes_Colaboraciones. Se refiere a un patrón compuesto que sirve para representar el grado de colaboración académica entre coautores. Al considerar las redes de colaboración entre investigadores, se puede cuantificar el impacto del INEGI, asignando diferentes pesos de acuerdo con el tipo de colaboración que tengan los autores de una publicación. Por ejemplo, si una Pub_Citantes está hecha por un solo autor nacional, tendrá menor impacto que otra donde se tengan colaboraciones internacionales. Para obtener el grado de contribución que este patrón ejerce sobre el impacto del INEGI, se toman en cuenta los criterios para clasificación de las redes de colaboración de acuerdo con la adscripción de los coautores descritos en el cuadro A3.

Predictores compuestos. Combinar predictores para generar nuevos es válido, siempre y cuando tenga significado esta agrupación. Nos basamos en el concepto de factor de impacto, por ser una medida consolidada y ampliamente utilizada en el enfoque bibliométrico tradicional. Procedemos, entonces, a relacionar el número de citas que tuvieron las publicaciones, entre el número de ellas.

X_{11}. FI_INEGI_Citados. Se refiere a la proporción directa de la cantidad de Pub_Citantes entre el número de documentos INEGI_Citados en el mismo año.

X_{12}. FI_Pub_Citantes. Es la proporción de citas de los *citantes*, que se calcula con la cantidad de citas que obtuvieron las Citas_Pub_citantes entre el total de estas (Pub_Citantes).

Cuadro A3 **Continúa**

Clasificación de redes de colaboración de acuerdo con la cantidad de coautores que son de adscripción en instituciones mexicanas o extranjeras

Tipo	Descripción	Autores mexicanos	Autores extranjeros	Peso
RN1-Red Nacional A	Es cuando se tiene una publicación con un solo autor y es mexicano.	1	0	1
RN2-Red Nacional B	Hay dos o tres coautores, donde todos tienen adscripción en instituciones mexicanas.	{2..3}	0	2

Cuadro A3 Concluye

Clasificación de redes de colaboración de acuerdo con la cantidad de coautores que son de adscripción en instituciones mexicanas o extranjeras

Tipo	Descripción	Autores mexicanos	Autores extranjeros	Peso
RN3-Red Nacional C	Es de más de tres coautores, donde todos tienen adscripción en instituciones mexicanas.	> 3	0	3
RF1-Red Foránea A	Publicación con un solo autor de adscripción extranjera.	0	1	2
RF2-Red Foránea B	Publicación de dos o tres autores, donde todos tienen adscripción en instituciones extranjeras.	0	> 1	3
RF3-Red Foránea C	Publicación de más de tres coautores, donde todos tienen adscripción en instituciones extranjeras.	0	> 3	5
RI1-Red Internacional A	Tiene dos autores, uno nacional y otro con adscripción en el extranjero.	> 0	1	5
RI2-Red Internacional B	Se tienen tres coautores, donde al menos uno de ellos es nacional y los restantes con adscripción foránea diferente.	> 0	{1..2}	6
RI3-Red Internacional C	Se tienen cuatro autores, donde al menos uno es nacional y los otros deben ser al menos de tres diferentes adscripciones internacionales.	> 0	> 3	10

Anexo B

Continúa

Valores de los predictores para el IIPCI obtenidos para el periodo 1984-2018g

Periodo*	Valores normalizados de los predictores INEGI_Citados (X_1), Pub_Citantes (X_2), Citas_Pub_citantes (X_3), Cita_ imprecisa (X_4), Tendencias_Google (X_5), PubCitantes_ Altmetric (X_6), PubCitantes_DOI (X_7), PubCitantes_OpenAccess (X_8), AutoresCitantes_Score (X_9), AutoresCitantes_Colaboraciones (X_{10})									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
2018	0.00E+00	9.55E-01	1.25E-01	4.97E-03	2.62E-01	5.64E-01	9.38E-01	3.86E-01	6.51E-01	5.83E-01
2017	0.00E+00	9.53E-01	3.06E-01	7.09E-01	2.87E-01	8.98E-01	1.00E+00	8.19E-01	1.00E+00	7.73E-01
2016	3.88E-01	1.00E+00	4.50E-01	1.00E+00	3.09E-01	1.00E+00	9.54E-01	1.00E+00	9.94E-01	9.80E-01
2015	9.54E-01	8.12E-01	5.33E-01	8.83E-01	3.81E-01	9.75E-01	8.19E-01	9.29E-01	8.96E-01	1.00E+00
2014	7.62E-01	7.06E-01	6.30E-01	6.73E-01	4.21E-01	7.09E-01	8.91E-01	8.91E-01	8.02E-01	9.27E-01
2013	7.18E-01	6.60E-01	7.40E-01	6.51E-01	4.24E-01	7.16E-01	6.31E-01	8.03E-01	7.24E-01	8.29E-01
2012	6.09E-01	4.68E-01	1.00E+00	2.48E-02	4.27E-01	6.70E-01	4.23E-01	4.93E-01	5.82E-01	5.44E-01
2011	5.53E-01	2.78E-01	4.46E-01	0.00E+00	4.84E-01	3.00E-01	3.19E-01	3.63E-01	2.95E-01	3.58E-01
2010	4.22E-01	2.83E-01	5.99E-01	0.00E+00	5.34E-01	3.50E-01	3.05E-01	4.44E-01	2.97E-01	4.86E-01
2009	6.17E-01	2.27E-01	4.70E-01	0.00E+00	5.52E-01	2.87E-01	2.68E-01	2.18E-01	2.91E-01	2.49E-01
2008	4.22E-01	2.17E-01	6.10E-01	0.00E+00	6.13E-01	3.92E-01	2.76E-01	1.76E-01	2.17E-01	2.21E-01
2007	3.88E-01	1.85E-01	5.08E-01	0.00E+00	7.35E-01	2.77E-01	2.45E-01	1.64E-01	1.73E-01	1.85E-01
2006	1.00E+00	1.66E-01	5.26E-01	0.00E+00	8.28E-01	1.51E-01	2.18E-01	1.18E-01	1.57E-01	1.35E-01
2005	1.60E-01	1.64E-01	4.68E-01	0.00E+00	9.97E-01	2.91E-01	1.88E-01	1.63E-01	1.40E-01	2.10E-01
2004	9.47E-02	2.70E-01	4.22E-01	0.00E+00	1.00E+00	9.44E-02	1.60E-01	1.10E-01	1.11E-01	1.17E-01
2003	6.65E-01	1.06E-01	3.47E-01	0.00E+00	0.00E+00	1.51E-01	1.63E-01	9.13E-02	8.34E-02	1.06E-01
2002	0.00E+00	8.58E-02	3.01E-01	0.00E+00	0.00E+00	5.53E-02	1.20E-01	7.16E-02	7.53E-02	6.60E-02
2001	0.00E+00	6.07E-02	3.17E-01	0.00E+00	0.00E+00	9.00E-02	9.72E-02	5.43E-02	3.96E-02	6.36E-02
2000	0.00E+00	7.28E-02	2.62E-01	0.00E+00	0.00E+00	5.99E-02	1.04E-01	9.30E-02	5.81E-02	9.55E-02
1999	0.00E+00	4.88E-02	1.41E-01	0.00E+00	0.00E+00	4.17E-02	7.50E-02	4.45E-02	3.45E-02	4.04E-02
1998	0.00E+00	3.50E-02	1.46E-01	0.00E+00	0.00E+00	4.67E-02	6.99E-02	2.60E-02	1.65E-02	3.25E-02
1997	0.00E+00	2.66E-02	1.40E-01	0.00E+00	0.00E+00	1.57E-01	5.91E-02	3.12E-02	1.15E-02	1.39E-02
1996	0.00E+00	2.74E-02	9.96E-02	0.00E+00	0.00E+00	1.06E-02	5.17E-02	3.12E-02	8.15E-03	2.59E-02
1995	0.00E+00	1.90E-02	4.98E-02	0.00E+00	0.00E+00	1.06E-02	4.03E-02	2.54E-02	1.85E-02	1.82E-02
1994	0.00E+00	1.46E-02	2.91E-02	0.00E+00	0.00E+00	3.38E-03	2.84E-02	2.48E-02	7.17E-03	1.23E-02

Valores de los predictores para el IIPCI obtenidos para el periodo 1984-2018

Periodo*	Valores normalizados de los predictores INEGI_Citados (X_1), Pub_Citantes (X_2), Citas_Pub_citantes (X_3), Cita_ imprecisa (X_4), Tendencias_Google (X_5), PubCitantes_ Altmetric (X_6), PubCitantes_DOI (X_7), PubCitantes_OpenAccess (X_8), AutoresCitantes_Score (X_9), AutoresCitantes_Colaboraciones (X_{10})									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1993	0.00E+00	8.24E-03	2.92E-02	0.00E+00	0.00E+00	1.01E-02	1.65E-02	1.62E-02	3.99E-03	4.97E-03
1992	0.00E+00	6.78E-03	1.47E-02	0.00E+00	0.00E+00	2.02E-03	1.53E-02	9.82E-03	6.04E-03	2.98E-03
1991	0.00E+00	5.05E-03	6.66E-03	0.00E+00	0.00E+00	2.43E-04	1.25E-02	1.79E-02	5.68E-03	8.95E-03
1990	0.00E+00	4.12E-03	7.01E-03	0.00E+00	0.00E+00	1.51E-03	9.09E-03	1.68E-02	0.00E+00	1.06E-02
1989	0.00E+00	3.32E-03	6.52E-03	0.00E+00	0.00E+00	4.52E-04	7.95E-03	5.20E-03	0.00E+00	0.00E+00
1988	0.00E+00	9.30E-04	3.01E-03	0.00E+00	0.00E+00	1.54E-04	4.55E-03	5.78E-03	2.72E-03	3.31E-04
1987	0.00E+00	3.99E-04	0.00E+00	0.00E+00	0.00E+00	2.50E-03	3.41E-03	2.89E-03	0.00E+00	1.99E-03
1986	0.00E+00	2.66E-04	0.00E+00	0.00E+00	0.00E+00	4.01E-05	1.14E-03	4.04E-03	2.40E-03	1.33E-03
1985	0.00E+00	0.00E+00	2.87E-03	0.00E+00	0.00E+00	3.76E-05	0.00E+00	3.47E-03	0.00E+00	4.31E-03
1984	0.00E+00	7.97E-04	1.40E-04	0.00E+00	0.00E+00	1.14E-05	1.14E-03	4.04E-03	0.00E+00	4.31E-03

* No se incluyeron los años 2020 y 2019 debido a que no se cuenta con la información completa de todos los predictores, ya que el proceso de minado para integrar la BDREF culminó en mayo del 2019.

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.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

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.

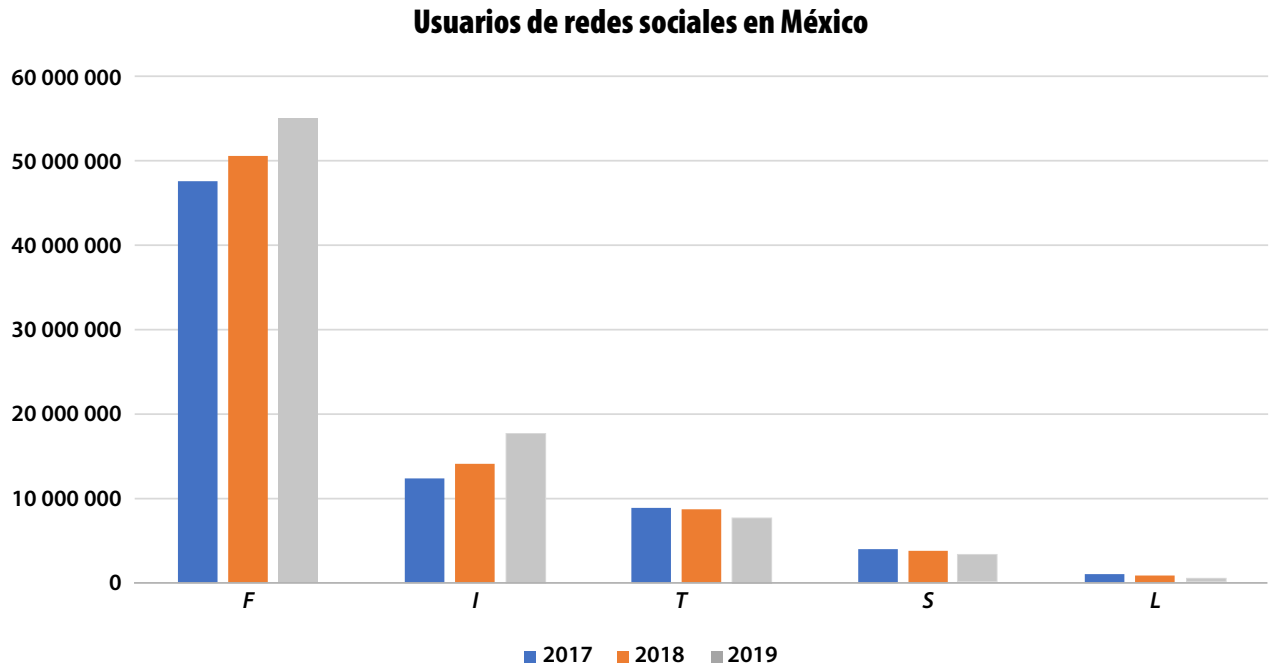
Tanto el cuadro como la gráfica 1 presentan un gran resumen para, entre otras, las subpoblaciones de usuarios con edades superiores a los 15 años. Lo primero que llama la atención es la desproporción entre los tamaños de las poblaciones de usuarios; por mucho, la de *Facebook* resulta ser siempre la 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.

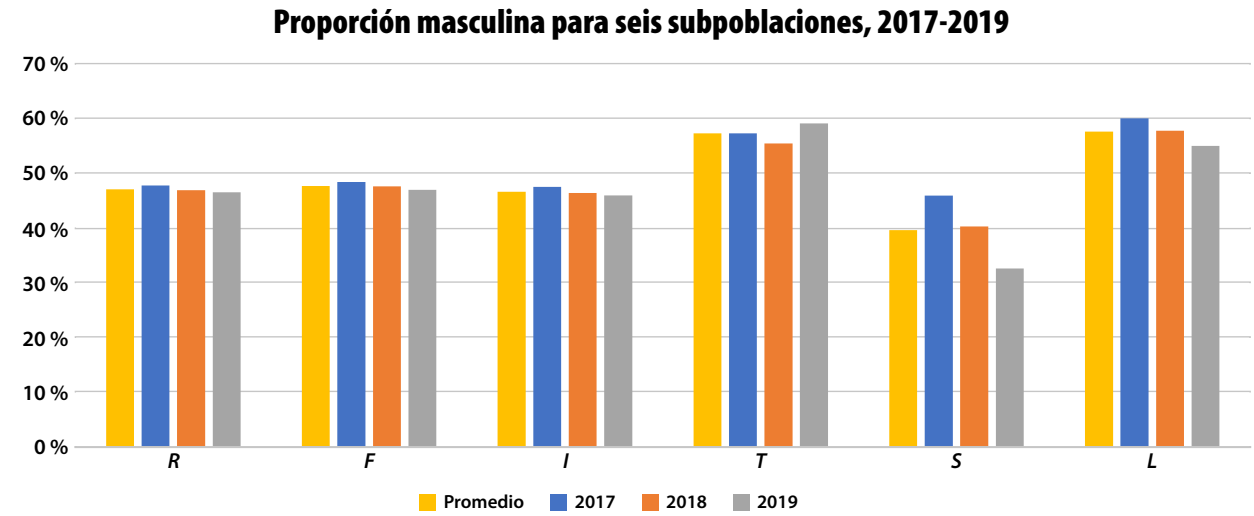
La población masculina de residentes exhibe un comportamiento estable apenas por debajo de 50 %, como es usual entre las concentraciones humanas. Salvo por dos de las redes consideradas, las poblaciones de usuarios son principalmente femeninas, mostrando marcadas diferencias con la de residentes; las 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

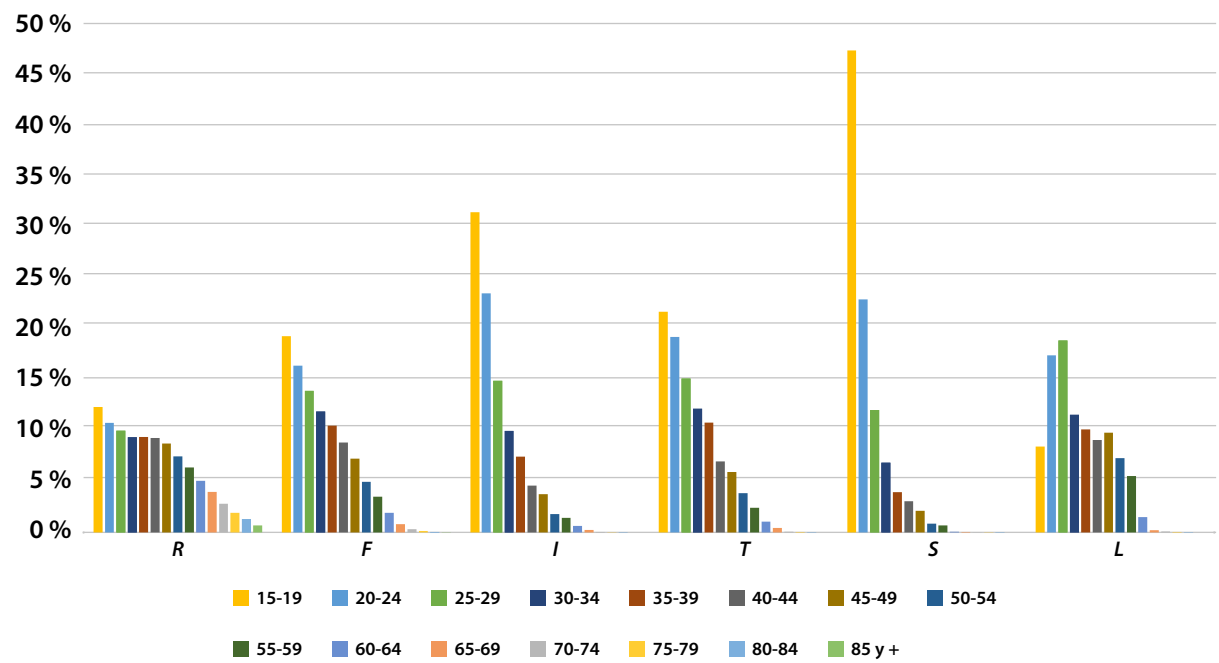


Fuente: cálculos propios a partir de bases de datos públicas de la ENDUTIH.

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



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

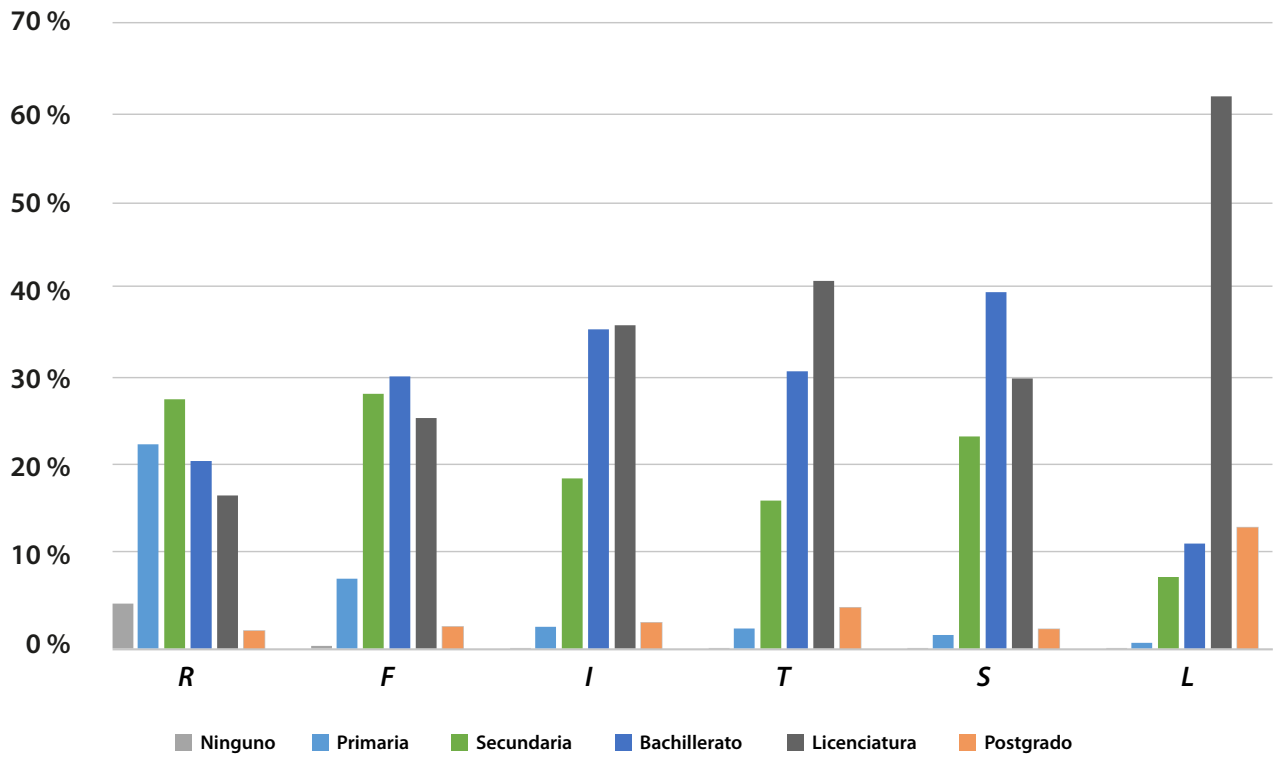
tes algunas diferencias marcadas, destacando *Instagram* y *Snapchat* como las más juveniles, con el grupo de 15-19 años de edad con el más numeroso en cada una de ellas; en términos porcentuales, la población en ese grupo es más grande que en la general por 2.5 veces que para *I* y por cuatro para *S*. 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

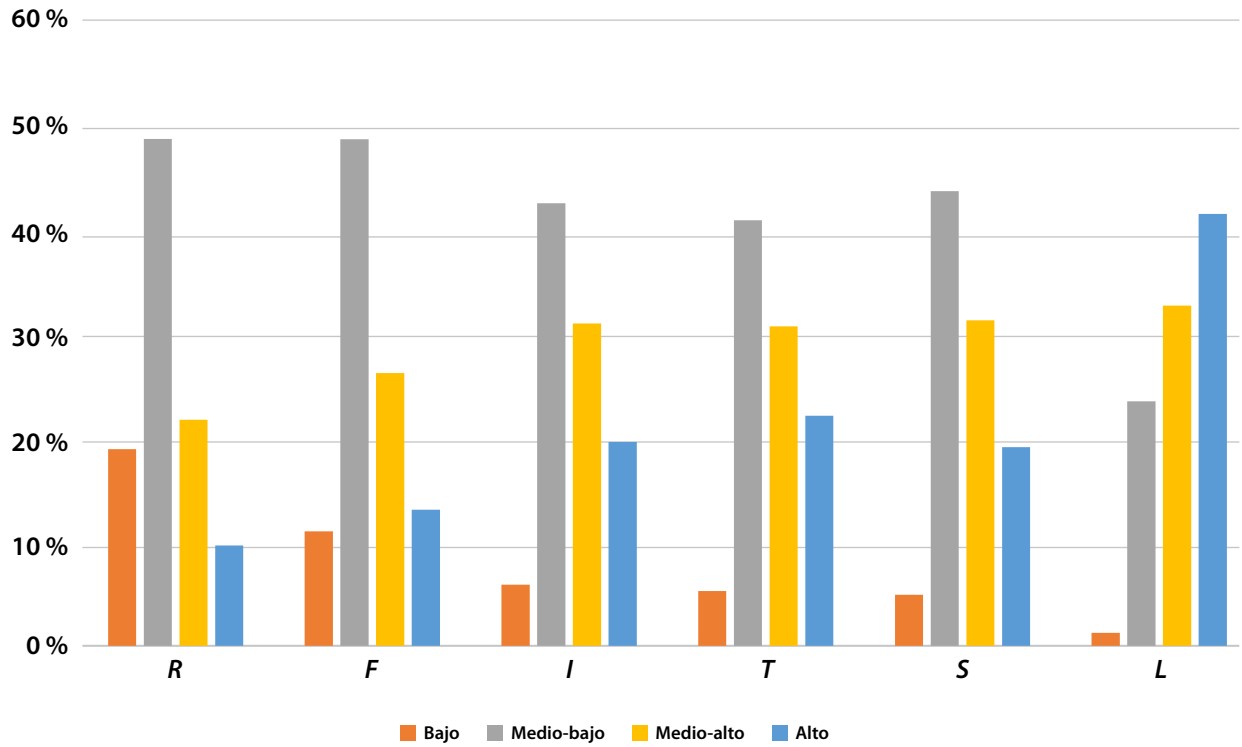
Estructuras relativas por red social según niveles educativos seleccionados, promedio 2017-2019



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.

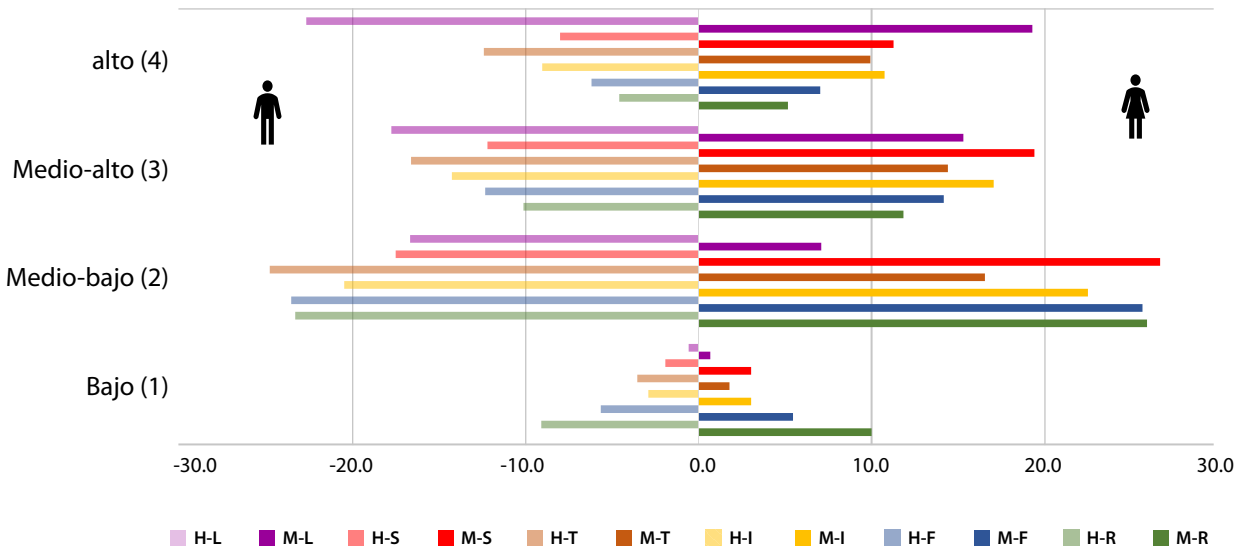
A partir de la breve discusión anterior, tenemos que es muy difícil que las distribuciones según sexo, edad, nivel educativo y estrato sociodemográfico en las poblaciones de usuarios de redes sociales coincidan con las correspondientes para la mexicana; en otras palabras, no son representativas de la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

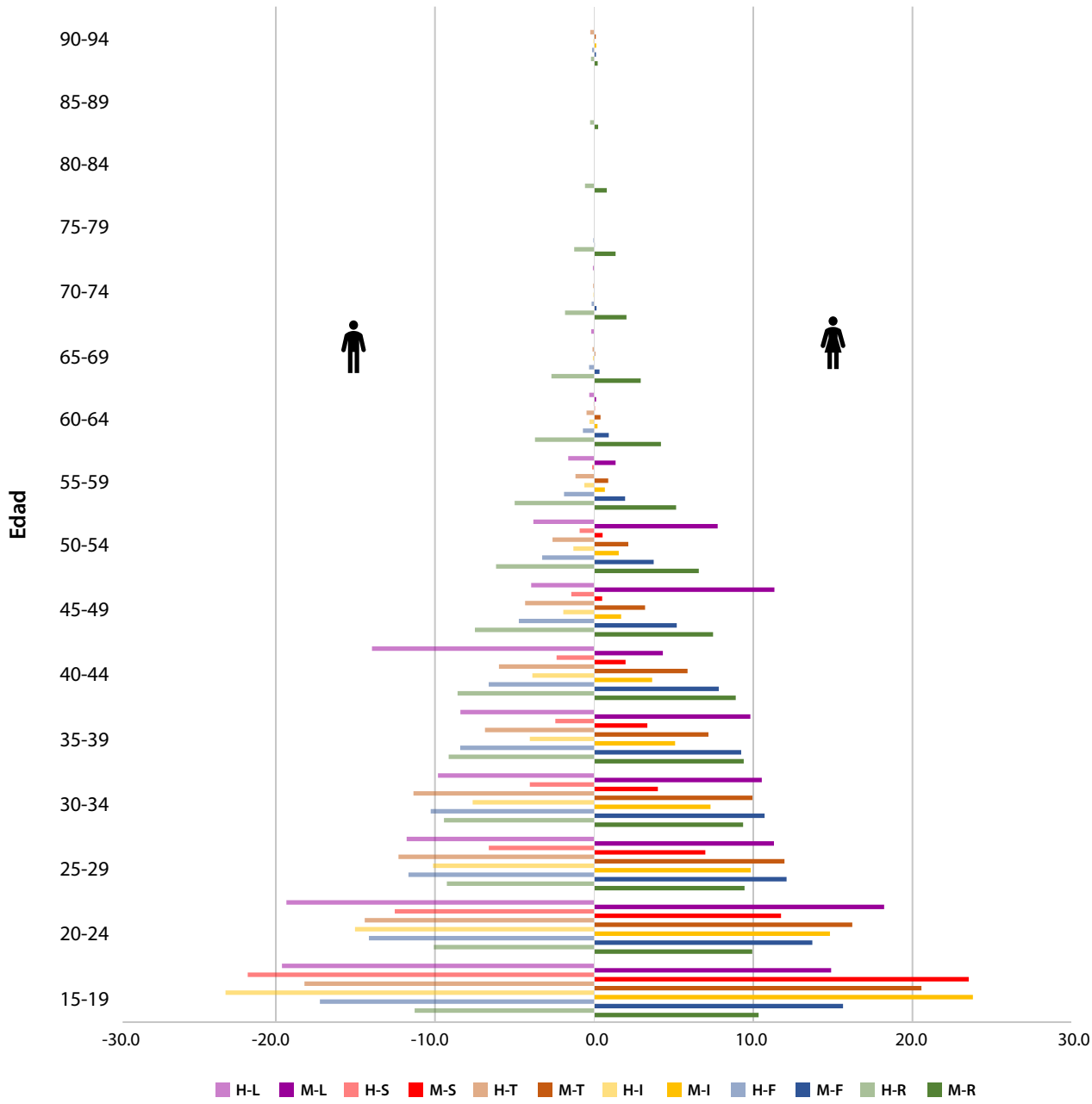
En la gráfica 7 se presentan las estructuras porcentuales por grupo de edad y sexo para cada una de las poblaciones consideradas, con un código de color semejante al del caso anterior. Salvo excepciones, no son apreciables sesgos importantes en favor de uno u otro sexo. Los usuarios de redes muestran 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

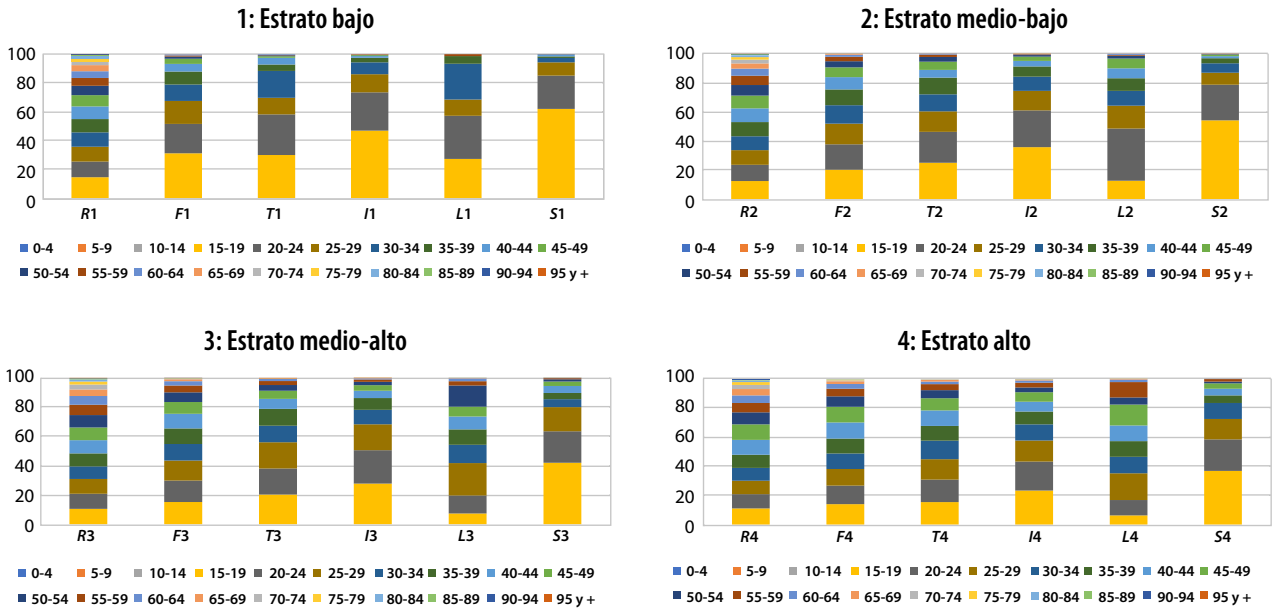


Fuente: cálculos propios a partir de bases de datos públicas de la ENDUTIH.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

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

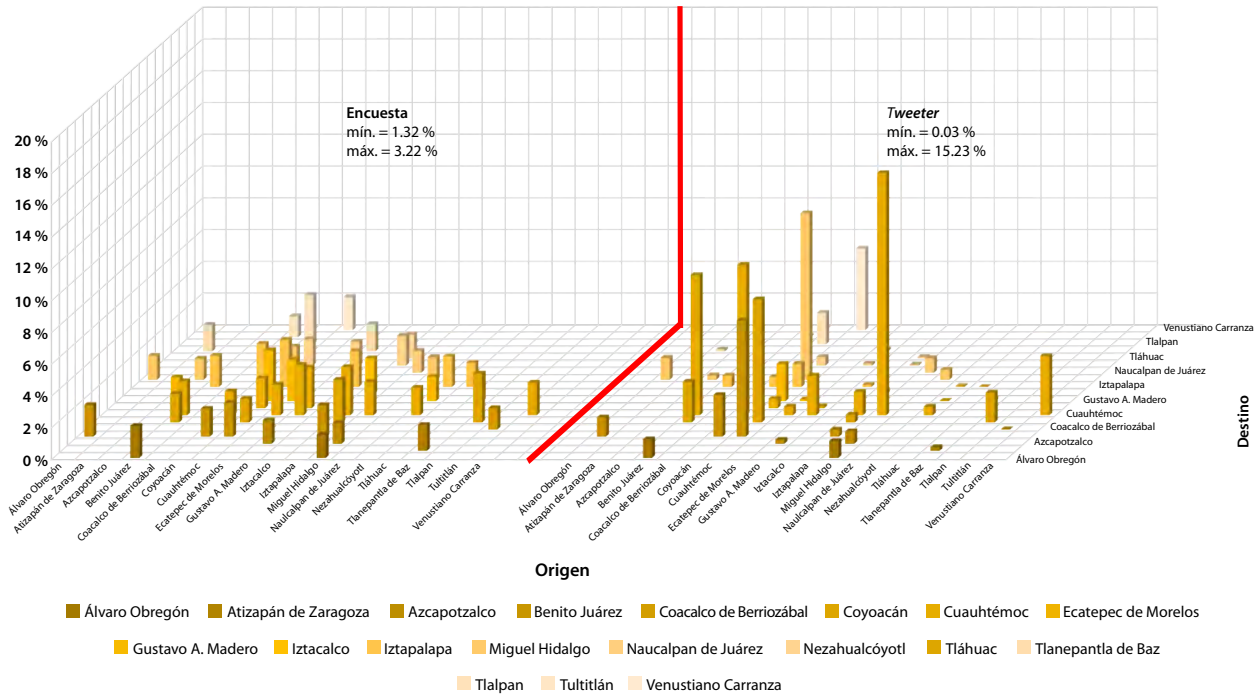
La comparación entre esta encuesta y el intento realizado para relacionarla con la captura de tuits georreferenciados es, también, elocuente en lo que toca al posible sesgo de selección en nuestra base de datos; esa es la razón por la que la discutiremos brevemente. En la gráfica 9 se exhiben dos conjuntos 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



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.

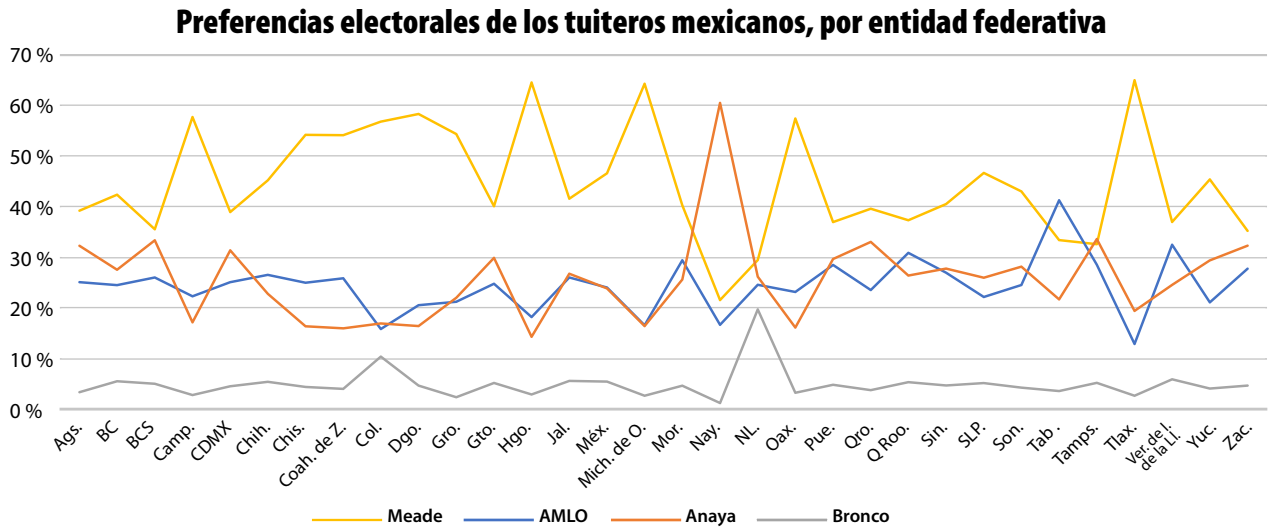
La gráfica 11 presenta la evolución nacional a lo largo de las semanas obtenida a partir de la base de datos desarrollada exprofeso. En ella se percibe nuevamente que el más favorecido por los tuiteros es Meade. Dicha preferencia muestra, sin embargo, una tendencia decreciente a lo largo del periodo 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.

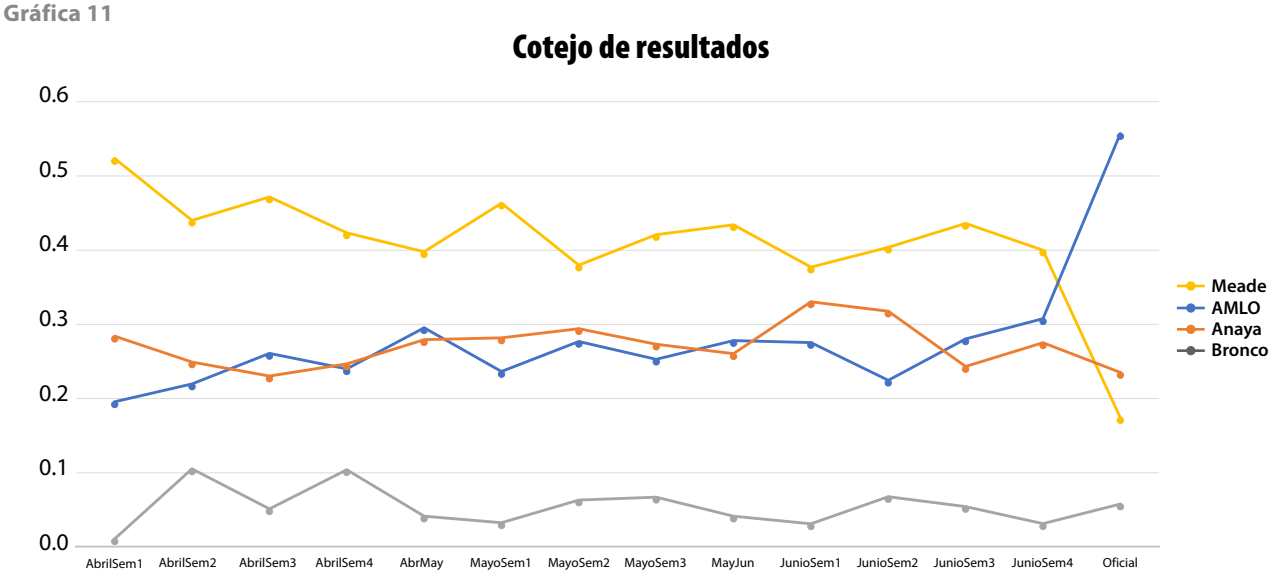
La ausencia de congruencia entre estos resultados y los mostrados por nuestro análisis a partir de *Twitter* (excepto para el candidato Bronco) parece evidenciar que la población que compone nuestra base de datos no representa a la que acudió a votar el 1 de julio de 2018, aunque aquella pueda estar 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 de* 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, XXIX[a]", 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.

Contribución del sistema financiero al ***crecimiento económico de México:*** un análisis econométrico, 1997-2019

Contribution of the Financial System ***to Mexico's Economic Growth:*** *An Econometric Assessment, 1997-2019*

Mauricio Montiel,* Francisco de Jesús Corona Villavicencio y Jesús López-Pérez****

Este trabajo contribuye a la discusión sobre si el sistema financiero tiene efectos positivos en el crecimiento económico, separando sus efectos de corto y largo plazo. Utilizando información para el periodo 1997-2019, se analiza la relación entre el Indicador Global de la Actividad Económica, como proxy del Producto Interno Bruto mensual, con los factores de producción, capital y trabajo, la inversión en capital humano y un factor para medir el desarrollo del sistema financiero. Este último se construye a través de modelos de factores dinámicos para incorporar diversas variables relacionadas con el desarrollo del sistema financiero. Enseguida, mediante la descomposición PT de Gonzalo y Granger (1995), se identifica que, ante un choque de corto plazo en el mercado financiero, la actividad económica reacciona positivamente, mientras que un sistema financiero sano contribuye también de manera significativa a la actividad económica en el largo plazo.

Palabras clave: actividad económica; Cobb-Douglas; corto plazo; descomposición PT; modelos de factores dinámicos.

This paper contributes to the academic discussion about whether finance has positive effects on economic growth, separating its short and long-term effects. By using information for the period 1997-2019, we analyze the relation between *Indicador Global de la Actividad Económica (IGAE)*, as a *proxy* of monthly GDP, with the production factors, capital and labor force, investment on human capital, and a factor that reflect the finance development. The latter is built upon dynamic factor models to incorporate several variables related to the financial system. Next, by using Gonzalo and Granger's (1995) PT decomposition, we identify that in the face of a short-term shock in the financial market, economic activity reacts positively, whilst a healthy financial system also contributes significantly to economic activity in the long term.

Key words: economic activity; Cobb-Douglas; Short-Run; PT decomposition; Dynamic Factor Models.

Recibido: 29 de junio de 2020.
Aceptado: 19 de octubre de 2021.

* El Colegio de la Frontera Norte, mmontiel.mea2018@colef.mx
** Instituto Nacional de Estadística y Geografía (INEGI), franciscoj.corona@inegi.org.mx y jesus.lopezp@inegi.org.mx, respectivamente.

3D rendering. Symbol made of digits / istock_onespirit / iStock

1. Introducción

El sistema financiero ha cobrado relevancia en el desarrollo contemporáneo de la economía mexicana. A partir de la reprivatización de la banca comercial en 1990, y posterior a la crisis de 1994, este se ha conformado acorde con las características y evolución del sector real de la economía, el cual ofrece una diversidad de productos financieros para inversionistas, ahorradores, empresas y emprendedores cada vez más amplia y sofisticada. Por ejemplo, durante la década del 2000 tuvo un gran auge la bursatilización de activos, principalmente hipotecas, que permitieron a los otorgantes de crédito ceder parte del riesgo crediticio. En fecha más reciente, la *Ley para Regular las Instituciones de Tecnología Financiera* (*Fintech*), decretada en el 2018, puso a México a la vanguardia en la regulación de empresas que, cada vez más rápido, han adoptado las tecnologías de la información para ofrecer productos y servicios financieros.

Por su parte, la más reciente Encuesta Nacional de Inclusión Financiera[1] señala que 68 % de la población cuenta con al menos un producto financiero. Aun en el contexto actual de la pandemia ocasionada por el COVID-19, el Banco de México (BANXICO) señala en su *Reporte de estabilidad financiera* (junio del 2021) que el sistema financiero mexicano ha mostrado resiliencia y una posición en general sólida, caracterizada por niveles de capital y liquidez por encima de los mínimos regulatorios aplicables.

1 Encuesta Nacional de Inclusión Financiera (ENIF) 2018, Instituto Nacional de Estadística y Geografía (INEGI).

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

No obstante, existe en la literatura en la materia un debate importante sobre el papel que puede tener el sistema financiero sobre el crecimiento económico de los países.[2] La discusión clásica versa en dos líneas, por un lado, autores como Lucas (1988) y Robinson (1952) argumentan que la importancia del primero en el segundo se encuentra severamente exagerada en la discusión académica; por otra parte, Schumpeter (1912), Gurley y Shaw (1955), Goldsmith (1969) y Miller (1998) concuerdan que para comprender el crecimiento económico es indispensable tomar en cuenta al sistema financiero.

El principal esfuerzo en probar las teorías antes mencionadas radica en identificar las variables apropiadas para incluirlas en un modelo de regresión lineal. Durlauf (2005) hizo una revisión de distintos estudios y encontró hasta 145 diferentes regresores, en los que la gran mayoría de las variables se han encontrado significativas, los cuales se agrupan en 43 *teorías de crecimiento*.

Para el caso de México, diversas investigaciones empíricas reconocen la importancia del desarrollo del sistema financiero para el crecimiento económico. Lustig (2001) y Moreno-Brid *et al.* (2005) coinciden al señalar que las problemáticas del primero generan barreras significativas al segundo. Bergoeing *et al*. (2002, 2007) identificaron que la colocación ineficiente del crédito durante el periodo 1982-1991 fue un factor importante del pobre desempeño económico en el país, mientras que Venegas *et al.* (2009) indican que el sistema financiero ejerce una influencia positiva, aunque pequeña,[3] sobre el crecimiento económico, y para ello evaluaron empíricamente el efecto de uno sobre el otro utilizando metodologías de series de tiempo para conocer los efectos de corto y largo plazo. Cermeño *et al.* (2012) hicieron un análisis para México y Estados Unidos de América (EE. UU.), donde encontraron que, para el caso de EE. UU., el sistema financiero promueve la tasa de crecimiento económico, pero para nuestra nación hallaron que la relación era inversa. En ese mismo sentido, Méndez-Heras *et al*. (2021) utilizaron un modelo de mínimos cuadrados ordinarios en dos etapas con datos de panel y uno dinámico de panel, y detectaron que había una relación positiva, estadísticamente significativa, entre el crédito de bancos comerciales y un Índice de Desarrollo Humano estatal elaborado por los autores.

En este orden de ideas, el principal objetivo de este trabajo es analizar de manera empírica la contribución que ha tenido el sistema financiero sobre el crecimiento económico en México. Para ello, utilizamos métodos de series de tiempo para relacionar el Indicador Global de la Actividad Económica (IGAE), como *proxy* del Producto Interno Bruto (PIB) mensual, con los factores de producción, capital y trabajo, con una variable para medir la inversión en capital humano y con otra para estimar el desarrollo del sistema financiero; el modelo permite separar los efectos de corto y largo plazo de las variables consideradas.

La presente investigación se distingue en introducir al modelo de crecimiento económico un amplio conjunto de variables relacionadas con el desarrollo del sistema financiero, propuestas originalmente por Svirydzenka (2016). En estudios empíricos, el grado de desarrollo del sistema financiero se ha incorporado en un modelo lineal de regresión probando una amplia variedad de variables relacionadas, como: indicadores del mercado de valores, participación de los bancos, dolarización, profundidad del sector bancario, competencia económica, represión financiera, sofisticación de productos bancarios, acceso al crédito, entre otros.

De esta forma, las aportaciones de este trabajo con respecto a la literatura existente son: i) aprovechar al máximo la información de un conjunto de series de tiempo financieras estimando variables latentes con periodicidad mensual de 1997 al 2019, con lo que se reduce la dimensionalidad

2 Para una revisión profunda de la literatura sobre la relación finanzas-crecimiento económico, se sugiere al lector ver, por ejemplo, Rousseau (2003), Levine (2005) y Beck (2008, 2012).

3 Los coeficientes estimados indican efectos modestos; ante un incremento de 10 % en la inversión total se tendría un aumento aproximado de 2.85 % en el PIB real, en el largo plazo, y ante un aumento de 10 % en el nivel de desarrollo financiero se reflejaría en un incremento de 0.8 % del PIB real en el largo plazo.

a través de un modelo de factores dinámico (DFM, por sus siglas en inglés); ii) estimar las posibles relaciones de largo plazo que comparte la actividad económica mensual con el factor trabajo, capital físico acumulado, capital humano y el sector financiero usando, junto con los factores estimados, la prueba de cointegración de Johansen (1991); y sujeto a estos resultados, iii) desentrañar los efectos de corto plazo a través del enfoque de Gonzalo y Granger (1995).

En particular, entender la relevancia del sistema financiero en México es importante, sobre todo en el contexto actual de recuperación económica postpandemia, toda vez que clarificar el entendimiento del rol de este en el crecimiento económico tendrá implicaciones de políticas públicas y creará investigaciones orientadas a la generación de aquellas apropiadas al ámbito nacional.

Lo que resta de este trabajo se distribuye como sigue: en la sección 2 se describe el marco teórico que sustenta las bases de esta investigación; en la 3 se expone la notación y la metodología econométrica; en la 4 se muestra la aplicación empírica y se describen los resultados; y, finalmente, en la 5 se presentan conclusiones.

2. Marco teórico

2.1. Relaciones teóricas entre el sistema financiero y el crecimiento económico

Los paradigmas de crecimiento económico suelen mencionar que este se encuentra supeditado al nivel de inversión en la economía, por ejemplo, y de acuerdo con el modelo neoclásico y a la teoría del modelo AK,[4] depende en específico de los niveles que se destinen en capitales físico y humano. Por otra parte, según el modelo schumpeteriano de innovación, lo que importa es la que se haga en tecnología y más específicamente, en el grado de inversión en investigación y desarrollo (I + D). Hasta este punto, tales enfoques de crecimiento no reparan en analizar las dificultades que las empresas pueden tener para financiar sus inversiones que lo impulsan, y la función de los bancos y otros intermediarios financieros para hacer frente a dichas restricciones.

De este modo, para comprender el rol que juegan los mercados e intermediarios financieros, es necesario relajar los supuestos de modelos idealizados de crecimiento económico. Ciertamente, el sector financiero es considerado como el lubricante que reduce las fricciones en el sistema económico y, por lo tanto, permite que la maquinaria funcione. Así, las principales funciones de los bancos son: i) distribución de riesgos, ii) promover el ahorro y iii) aliviar problemas de agencia. Aghion y Howitt (2008) introducen restricciones financieras a la teoría schumpeteriana, lo que les permite mostrar cómo los intermediarios financieros, al canalizar el ahorro, proveen financiamiento externo a los emprendedores que innovan y eso, a su vez, promueve el crecimiento económico.

2.2. Relaciones empíricas entre el sistema financiero y el crecimiento económico

La mayoría de la literatura en esta materia se relaciona con regresiones tipo panel entre países de la forma:

$$g_i = \beta_0 + \beta_1 DesFin_i + \beta_2 X_i + u_i$$

donde g_i es el promedio de crecimiento del país i durante el periodo de análisis, $DesFin_i$ es el nivel de desarrollo financiero del país, X_i es un vector de variables control u_i es un término de error. Los diferentes estudios difieren en: i) cómo tomar la variable de crecimiento de la economía, ya sea utilizar datos para diferentes países o bien para distintas industrias y regiones, o bien empleando datos a nivel de empresa; o ii) la forma en la que se mide $DesFin_i$.

4 Es un modelo de crecimiento endógeno que se emplea en la teoría del crecimiento económico; este utiliza un modelo lineal en el que la producción es una función lineal del capital $Y=AK^a L^{1-a}$, donde Y representa la producción total en una economía; A, la productividad total de los factores; K es capital; L, mano de obra; y el parámetro a mide la elasticidad de salida del capital. Para una presentación completa del modelo AK, ver por ejemplo Aghion y Howitt (2008).

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Entre las referencias más importantes se encuentra Goldsmith (1969), quien trató de evaluar si el sistema financiero ejerce una influencia en el crecimiento económico mostrando que existe una correlación positiva entre el tamaño del primero y el nivel de actividad económica. Por su parte, McKinnon (1973) interpretó la gran cantidad de evidencia que emerge de estudios de caso de países y sugirió que los sistemas financieros que funcionan mejor han estimulado de manera importante el crecimiento económico. King y Levine (1993) se basaron en el trabajo de Goldsmith (1969) y estudiaron un total de 77 naciones con el objetivo de determinar si el sistema financiero contribuye al crecimiento económico de largo plazo, a la acumulación de capital y a la productividad; estos autores concluyen que hay una fuerte relación entre el primero y el segundo.[5]

El grueso de evidencia empírica indica que el desarrollo financiero afecta al crecimiento económico de una forma positiva y monotónica, ya que los países con bancos y mercados más eficientes crecen más rápido, y cuando los sistemas financieros funcionan mejor alivian las restricciones de financiamiento externo que impiden la expansión industrial y de las empresas, mientras que los bancos proveen servicios de búsqueda de proyectos de inversión (*screening services*) y realizan monitoreo postpréstamo (Levine, 2005).

La evidencia más reciente ha sugerido que este no es necesariamente el caso de todos los tipos de actividad y en la totalidad de los niveles de desarrollo financiero (Popov, 2018). Por ejemplo, en un trabajo reciente, Ruiz (2018) utilizó una estimación de variables instrumentales y modelos de panel para controlar por endogeneidad al trabajar con el PIB per cápita inicial, y encontró evidencia que corrobora la existencia de una relación no lineal entre el desarrollo financiero y el crecimiento económico. Slesman *et al.* (2019) hallaron, en un estudio de 77 países emergentes y en vías de desarrollo, que al minimizar los riesgos políticos a través de mejoras en la calidad de las instituciones políticas es posible mejorar el crecimiento económico debido al desarrollo del sistema financiero. Maune *et al.* (2020) utilizaron un modelo de regresión múltiple para evaluar el impacto de la inclusión financiera en Zimbabue en el periodo del 2011 al 2017, y encontraron que la inclusión financiera tenía un impacto positivo en el crecimiento del país.

En el polo opuesto, hay quienes argumentan que en la práctica existen ciertas desventajas del desarrollo financiero. A través de la historia, las finanzas han sido señaladas como una actividad de búsqueda de rentas (*rent seeking*[6]). En consecuencia, terminan por aumentar la fragilidad de las instituciones financieras, la inestabilidad macroeconómica y la probabilidad de crisis financieras, por lo que se exacerba la caída de la tasa de crecimiento económico de largo plazo (Venegas *et al.*, 2009). Ibrahim y Alagidede (2018), en un panel de 29 países subsaharianos para el periodo 1980-2014, utilizando la técnica de estimación dinámica mediante el método generalizado de momentos (GMM, por sus siglas en inglés), encontraron que mientras el desarrollo financiero apoya al crecimiento económico, el grado en el que las finanzas lo ayudan depende de manera crítica del crecimiento simultáneo de los sectores real y financiero. Kadhuma y Kadhimb (2020) analizaron los efectos de la represión financiera en el crecimiento económico en Irak; sus resultados muestran que la liberalización financiera y de capitales tienen un efecto negativo y no significativo. Más recientemente, Cheng *et al.* (2021) usaron un panel de 72 naciones para el periodo 2000-2015 y, empleando GMM, hallaron que, sin importar el nivel de ingreso del país, el desarrollo financiero es siempre desfavorable para el crecimiento económico, pero este efecto es mayor en países de ingresos altos.

5 Las correlaciones estimadas entre los indicadores financieros de profundidad financiera (activos líquidos con respecto al PIB, créditos bancarios, activos respecto al crédito privado y crédito privado respecto al PIB) con el crecimiento del PIB per cápita son positivas del orden de 0.56, 0.44, 0.37 y 0.5, respectivamente, y son significativas a 1 por ciento.

6 Situación que se produce cuando un individuo, organización o empresa busca obtener ingresos captando renta económica mediante la manipulación o explotación del entorno político o económico, en lugar de obtener beneficios a través de transacciones económicas y producción de riqueza añadida.

Una parte importante de la literatura investiga la relación entre finanzas y crecimiento económico utilizando métodos de series de tiempo. Este enfoque permite emplear técnicas econométricas más poderosas para analizar naciones en particular a mayor profundidad (Levine, 2005). Estas investigaciones con frecuencia hacen uso de pruebas de causalidad de Granger, como lo realizan Blomstrom *et al.* (1996) para explicar el crecimiento económico con la formación de capital empleando rezagos de ambas series; de manera similar, Campos y Nugent (2002) investigaron la existencia y dirección de una relación causal entre inestabilidad sociopolítica y el crecimiento económico. Otro procedimiento utilizado ampliamente en la literatura es el de vectores autorregresivos (VAR), como en Dritsakis y Adamopoulos (2004), quienes analizaron el PIB trimestral en función de agregados monetarios y la apertura comercial.

Los modelos de series de tiempo también se han utilizado en la literatura de crecimiento endógeno para analizar cambios permanentes en variables que son afectadas potencialmente por políticas gubernamentales, las cuales conllevan a modificaciones permanentes en la tasa de crecimiento económico, tal es el caso de Jones (2005), quien evaluó modelos de crecimiento AK y de I + D, en los que modeló de manera explícita la inversión y el cambio tecnológico, y encontró que un incremento permanente en la tasa de crecimiento de la inversión afecta al crecimiento económico solamente en un corto periodo de ocho años.

Destacan, además, las restricciones gubernamentales que se han dado sobre el sistema financiero, como los límites establecidos sobre las tasas de interés, el encaje legal, entre otras, las cuales han distorsionado el proceso de desarrollo de este. De acuerdo con Rodríguez y López (2010), este efecto podría deberse a las imperfecciones que provocan tales restricciones en los mercados financieros, produciendo una asignación ineficiente de recursos.

De esta forma, en este trabajo nos centramos en cuantificar variables latentes financieras que puedan ser relacionadas con el crecimiento económico.

3. Metodología econométrica

En esta sección se describe esta y la notación empleada a lo largo de la investigación.

3.1. Modelos de factores dinámicos

En economía, estos fueron originalmente introducidos por Geweke (1977) y Sargent y Sims (1977), y son utilizados con frecuencia para representar la dinámica de un grupo de N series de tiempo correlacionadas a través de un pequeño número de factores comunes subyacentes ($r<N$). En la actualidad, dada la vasta cantidad de información recopilada a través de las décadas pasadas, el uso de los DFM de alta dimensionalidad se ha vuelto más atractivo debido a la flexibilidad que brinda el poder extraer en pocos factores la variabilidad asociada a un gran número de variables.

En este contexto, consideramos que la evolución común de un vector de series de tiempo que representan al mercado financiero, $X_t = (x_{1t},\ldots x_{Nt})'$, observadas desde $t = 1,\ldots,T$, son generadas por factores comunes no observados $F_t = (F_{1t},\ldots,F_{rt})'$ más ruidos idiosincráticos o comportamientos individuales, $\varepsilon_t = (\varepsilon_{1t},\ldots \varepsilon_{Nt})'$, donde estos dos últimos siguen su propia dinámica; en este caso, se asume como un modelo tipo VAR(1). En consecuencia, el DFM está representado como sigue:

$$X_t = PF_t + \varepsilon_t,$$

$$F_t = \Phi F_{t-1} + \eta_t,$$

$$\varepsilon_t = \Gamma \varepsilon_{t-1} + a_t, \qquad (4)$$

donde X_t, ε_t y $a_t = (a_{1t},\ldots a_{Nt})'$ son vectores de dimensión N x 1. La contribución de los factores sobre las observaciones está dada por la matriz de cargas, $P = (p_1,\ldots,p_N)'$, cuya dimensión es N x r, mientras que F_t y su término idiosincrático $\eta_t = (\eta_{1t},\ldots,\eta_{rt})'$ son vectores de dimensiones $r \times 1$. Se asume que a_t y η_t son ruidos blancos con matrices de covarianzas Σ_a y $\Sigma_n = diag(\sigma^2_{n_1},\ldots,\sigma^2_{n_r})$, respectivamente, en tanto que $\Phi = diag(\phi_1,\ldots\phi_r)$ y $\Gamma =$

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

$diag(\gamma_1, \ldots \gamma_r)$ y son matrices de dimensión $r \times r$ y $N \times N$, las cuales contienen los parámetros autorregresivos de los factores y de los componentes idiosincráticos, en ese mismo orden. Las diagonales principales de Φ pueden contener valores de 1 y, de esta manera, los factores comunes pueden ser no estacionarios. Por otra parte, y por conveniencia, se asume que la diagonal de Γ es tal que los errores idiosincráticos son estacionarios.

El procedimiento más popular para extraer factores en DFM de alta dimensionalidad es el de componentes principales (PC, por sus siglas en inglés) debido a la facilidad para implementarlo computacionalmente y que la teoría asintótica, dado los supuestos de las ecuaciones (4), es bien conocida. Su particularidad principal es que nos permite extraer los factores sin asumir nada en particular sobre la distribución de los errores; además, su implementación es computacionalmente simple, lo cual puede explicar por qué se usa con frecuencia en estudios que enfrentan una numerosa cantidad de variables y observaciones en forma de series de tiempo. La extracción de factores a través de PC separa el componente común, PF_t, del idiosincrático, ε_t, mediante promedios de sección cruzada sobre X_t de tal manera que cuando N y T tienden al infinito, el efecto del componente idiosincrático converge a 0, dejando solo los efectos asociados a los factores comunes; ver, por ejemplo, Corona *et al.* (2017). Se puede mostrar que la estimación de P obtenida por PC, $\hat{P}$, es equivalente a $\sqrt{N}$ veces los vectores propios correspondientes a los r valores propios más grandes de $X'X$, donde $X = (X_1, \ldots, X_T)$ es una matriz de dimensión $N \times T$. En consecuencia, la estimación de F con PC es, entonces:

$$\hat{F} = N^{-1} \hat{P}\, X. \tag{5}$$

La consistencia de $\hat{P}$ y $\hat{F}$ dependen de que el componente idiosincrático sea estacionario; por ende, se asume que $\varepsilon_t \sim I(0)$, ver Bai (2003, 2004), no obstante, en la práctica esto tiene que ser verificado (Bai y Ng, 2004).

Hasta este momento hemos asumido que r es conocido, pero en la práctica tiene que estimarse. En Corona *et al.* (2017) se realizan extensos experimentos Monte Carlo para diagnosticar qué criterio de estimación es el que funciona mejor bajo diferentes procesos generadores de datos, evaluando series de tiempo estacionarias y no estacionarias. En este trabajo se estudian los criterios de Bai y Ng (2002), Onatski (2010) y Ahn y Horenstein (2013), concluyendo que el de Onatski funciona mejor cuando las series de tiempo son incluso no estacionarias y las dependencias en el componente idiosincrático son relativamente fuertes, es decir, hay autocorrelación serial, heteroscedasticidad y correlación contemporánea entre los errores. También, en Corona *et al.* (2017) se muestran cómo estos criterios dependen de la magnitud de los valores propios de la matriz de covarianza de las observaciones, es decir, λ_j, para $j = 1, \ldots, N$; en este sentido, en este trabajo nos centramos en el tradicional criterio basado en la explicación de la varianza:

$$\hat{r} = min\left(r \left| \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{N} \lambda_j} > c \right.\right), \tag{6}$$

donde c es un número fijo predeterminado, conocido como el porcentaje de la explicación de la varianza, frecuentemente 0.8 o 0.9.

3.2. Relaciones dinámicas de corto y largo plazo: Gonzalo y Granger (1995) y Johansen (1991)

En un correcto orden de ideas, se asume que $Y_t = (y_t, l_t, k_t, h_t, \hat{F}_t)'$ es un vector $(4 \times r) \times 1$, donde y_t es el nivel de actividad económica; l_t es atribuible al factor trabajo y k_t, a la acumulación de capital físico; h_t mide al capital humano; y F_t son las variables latentes atribuibles al sistema financiero. Asumiendo que $Y_t \sim I(1)$, el vector de corrección de errores (VEC) es:

$$\begin{aligned} \Delta Y_t = {} & \Pi Y_{t-p} + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} \\ & + \cdots + \Gamma_{p-1} \Delta Y_{t-p+1} + u_t, \end{aligned} \tag{7}$$

donde se supone que u_t es ruido blanco. La prueba de cointegración de Johansen (1991) es ampliamente conocida en la literatura, y,

en resumen, se orienta en las raíces características de la matriz $\Pi = \alpha\beta'$ del modelo (7). En pocas palabras, el rango de Π especificará el número de vectores de cointegración que estén presentes en el sistema, es decir, el número de relaciones de largo plazo. Si $rango(\Pi)=0$, todos los elementos de Π son 0 y (7) puede reescribirse como un modelo VAR(p) en primeras diferencias. En este caso, Y_t tiene raíz unitaria y ninguna de las variables está cointegrada. Si, por el contrario, $rango(\Pi) = 4 \times r$, el vector Y_t es estacionario e igualmente no cointegrado, es decir, el modelo (7) es un VAR(p). En los casos intermedios donde $1 < rango(\Pi) < 4 \times r$, quiere decir que existen múltiples vectores de cointegración y, si suponemos que $rango(\Pi) = m$, tal que el número de vectores de cointegración independientes sea m, entonces solo esas m relaciones lineales de las variables conducirán a un resultado estacionario. En otras palabras, hay m vectores de cointegración, siendo estas $\beta' Y_t$.

En este último caso, Stock y Watson (1988) muestran que series de tiempo cointegradas permiten una representación de factores similar de la siguiente manera:

$$Y_t = A_1 f_t + \tilde{Y}_t, \tag{8}$$

donde la dimensión de f_t es $k=N-m$ y $\tilde{Y}t \sim I(0)$. En palabras, $Z_t \sim I(1)$ son las tendencias comunes de Y_t. Gonzalo y Granger (1995) otorgan las condiciones para estimar f_t; primero establecen que $f_t = BY_t$, es decir, es una combinación lineal de las variables originales, de tal forma que $A1Z_t$ puede ser asociado al componente permanente de Y_t y $\tilde{Y}$, al transitorio. Esta descomposición es conocida como descomposición PT y, de acuerdo con Blanchard y Quah (1989), es necesario que se cumpla lo siguiente para que esta sea válida:

1. $Y_t = P_t + T_t$.
2. ΔP_t y T_t son estacionarias.
3. Para el modelo $H(L)\ Y_t = u_t$:

   a. $\lim_{s\to\infty} \frac{\partial E_t(Y_{t+s})}{\partial u_{P_t}} \neq 0.$

   b. $\lim_{s\to\infty} \frac{\partial E_t(Y_{t+s})}{\partial u_{T_t}} = 0.$

Es decir, solo los choques permanentes tienen efecto en el largo plazo en las observaciones, mientras que los transitorios no. De esta manera, sujeto a los resultados de cointegración, la idea intuitiva de Gonzalo y Granger (1995) es que si $\beta' Y_t \sim I(0)$, se puede expresar una descomposición que cumpla con las condiciones anteriores tal que $Y_t = A_1 f_t + A_2 Z_t$, donde $Z_t = \beta' Y_t$. Es demostrable que dicha combinación es, finalmente:

$$Y_t = A_1 f_t + A_2 Z_t = P_t + T_t, \tag{9}$$

donde $A_1 = \beta_\perp (\alpha_\perp' \beta_\perp)^{-1}$, $f_t = \alpha_\perp' Y_t$ y $A_2 = \alpha(\beta'\alpha)^{-1}$. Gonzalo y Granger (1995) dan las pautas para estimar A_1 y A_2, que se basa en estimar el complemento ortogonal, $\alpha_\perp$, lo cual es basado en máxima verosimilitud usando un procedimiento similar al proceso de estimación de a y β en Johansen (1991). En este trabajo, por interpretabilidad, para analizar el corto plazo nos centraremos en A_2, que contiene los efectos de los choques transitorios, mientras que, para el largo plazo, nos enfocaremos en las tradicionales ecuaciones de cointegración, es decir, $\beta' Y_t$. Nótese que, en consecuencia, A_1 es la contribución de las tendencias comunes sobre las observaciones.

4. Aplicación empírica

En esta sección detallamos los datos utilizados en este trabajo y los resultados de aplicar la metodología descrita en el apartado anterior.

4.1. Datos

Las fuentes de información empleadas en este trabajo tienen periodicidad mensual, desde julio de 1997 hasta octubre del 2019. Estas se describen a continuación:

- Series financieras:
  - Cartera vigente (*cv*): como proporción del PIB (fuente: BANXICO e INEGI).
  - CETES a 28 días (*cetes28*): en términos reales, utilizando la inflación calculada con el Índice Nacional de Precios al Consumidor

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

(INPC) base 2013 = 100 (fuente: BANXICO e INEGI).
- CETES a 91 días (*cetes91*): en término reales deflactado con el INPC base 2013 = 100 (fuente: BANXICO).
- CETES a 128 días (*cetes128*): en términos reales deflactado con el INPC base 2013 = 100 (fuente: BANXICO).
- CETES a 364 días (*cetes364*): en términos reales deflactado con el INPC base 2013 = 100 (fuente: BANXICO).
- Índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores (*ipcg*): indicador del mercado accionario en su conjunto (fuente: BANXICO).
- Índice del Tipo de Cambio Real (*itcr*): con respecto a 49 países (fuente: BANXICO).
- Agregado monetario M1 (*m*1): compuesto por instrumentos altamente líquidos en poder de los sectores residentes tenedores de dinero; en particular, incluye billetes y monedas emitidos por el Banco de México, así como depósitos de exigibilidad inmediata en bancos y entidades de ahorro y crédito popular; serie deflactada con el INPC base 2013 = 100, desestacionalizada y como proporción del PIB (fuente: BANXICO e INEGI).
- Agregado monetario M2 (*m*2): instrumentos monetarios a plazo en poder de los sectores residentes tenedores de dinero; en particular, incluye la captación con un plazo residual de hasta cinco años en bancos, entidades de ahorro y crédito popular, así como uniones de crédito, las acciones de los fondos de inversión de deuda y los acreedores por reporto de valores; serie deflactada con el INPC base 2013 = 100, desestacionalizada y como proporción del PIB (fuente: BANXICO).
- Agregado monetario M3 (*m3*): valores públicos en poder de los sectores residentes tenedores de dinero y que fueron emitidos por el Gobierno Federal, BANXICO (BREMS) y el Instituto para la Protección al Ahorro Bancario (IPAB); serie deflactada con el INPC base 2013 = 100, desestacionalizada y como proporción del PIB (fuente: BANXICO).
- Agregado monetario M4 (*m*4): tenencia por parte de no residentes de todos los instrumentos incluidos en M3; serie deflactada con el INPC base 2013 = 100, desestacionalizada y como proporción del PIB (fuente: BANXICO).

• Series económicas:
- PIB mensual (*igae*): PIB real mensualizado a través de la técnica de Denton-Cholette (Denton, 1971) utilizando como variable preliminar al IGAE; serie desestacionalizada (fuente: INEGI).
- Asegurados permanentes y eventuales del IMSS (*imss*): número de trabajadores asegurados, permanentes y eventuales inscritos en el Seguro Social; serie desestacionalizada (fuente: IMSS).
- Inversión fija bruta, Índice de Volumen Físico Acumulado (*ifb*): total, incluye construcción y maquinaria y equipo, base 2013 = 100; se desestacionaliza (fuente: INEGI).
- Presupuesto asignado al Consejo Nacional de Ciencia y Tecnología (*conacyt*): serie desestacionalizada y como proporción del PIB, base 2013 = 100 (fuente: SHCP).

La desestacionalización de las variables no desestacionalizadas previamente se realizó con el paquete estadístico *R* con ayuda de la librería *seasonal*, que incluye los métodos del paquete *X13-ARIMA-SEATS*.

Finalmente, las series se expresan en logaritmos naturales por dos motivos: i) para eliminar la posible dependencia multiplicativa en la varianza de las series y ii) para realizar el ejercicio de cointegración donde los coeficientes representen las elasticidades.

Se realizaron pruebas Dickey-Fuller aumentadas (ADF, por sus siglas en inglés), donde se concluye que todas las series de tiempo son $I(1)$.

4.2. Obteniendo indicadores financieros

El primer paso consiste en estimar factores subyacentes del grupo de variables financieras a través de

DFM de alta dimensionalidad usando los métodos descritos en la subsección 3.1. Para lo anterior, un primer paso descriptivo consiste en analizar la estructura de correlación entre las series de tiempo. Es claro que, mientras más correlacionadas estén las variables, más sentido tendrá aplicar la reducción de la dimensionalidad. En este caso, la dependencia efectiva muestral[7] es de 0.960, por lo cual existe una clara dependencia lineal multivariada.

Estimando la expresión (6) nos otorga un $\hat{r} = 3$, lo cual nos dictamina que, de las 12 series de tiempo financieras, tres variables latentes explican al menos 90 % de la variabilidad total observada. La estimación de la matriz $\hat{P}$ se puede observar en el cuadro 1.

Claramente, el primer factor es dominado negativamente por las tasas de interés y de manera positiva, por el resto de las variables con excepción del *m*2, aunque su contribución es menor. El segundo carga positivamente para el resto de las variables, donde sobresalen las contribuciones de *m*2, *itcr* y *m*4. Por último, el tercero es dominado de forma negativa por casi todas las variables financieras. La intuición económica de estos factores es que, si el primero, el cual tiene mayor grado de explicación, carga de manera positiva respecto a la actividad económica, se verá beneficiado cuando las tasas de interés son bajas y el resto de las variables altas. Para el segundo, se podría argumentar lo contrario, que la economía se verá beneficiada por tasas de interés altas, mientras que el tercer factor tiene una interpretación poco más compleja, pero en definitiva indica que la economía se ve beneficiada cuando el tipo de cambio real y las tasas de interés decrecen.

La validación económica de los factores se presenta en la siguiente subsección, no obstante, econométricamente, la estimación de estos es consistente siempre y cuando se pueda corroborar que el componente idiosincrático es estacionario. Para estos fines, se realiza la prueba PANIC de Bai y Ng (2004) dada por el siguiente estadístico:

$$Ø = - \frac{2\sum_{i=1}^{N} log\, \varphi_i - 2N}{\sqrt{4N}}, \qquad (10)$$

donde φ_i son los *p*-valores individuales de las pruebas de raíces unitarias individuales ADF para los

7 En inglés, *Sample Effective Dependence*, se calcula como $SED=1-|R|^{1/(N-1)}$, donde $|R|$ es el determinante de la matriz de correlaciones. Este número está acotado entre 0 y 1, donde mayor significa mayor dependencia lineal multivariada.

Cuadro 1

Matriz de pesos del DFM aplicado a las series financieras

Variable	Factor 1	Factor 2	Factor 3
cv	1.00	1.13	-0.55
*cetes*28	-1.14	0.42	-0.72
*cetes*91	-1.13	0.48	-0.76
*cetes*128	-1.15	0.40	-0.68
*cetes*364	-1.13	0.47	-0.75
ipcg	1.14	0.14	0.04
itcr	0.60	1.20	-2.22
*m*1	1.15	0.04	-0.61
*m*2	-0.05	2.12	1.57
*m*3	0.91	-1.23	-0.90
*m*4	0.99	1.21	0.13

Fuente: elaboración propia a partir de análisis econométrico.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

elementos contenidos en $\hat{\varepsilon} = Y\text{-}\hat{P}\hat{F}$. La hipótesis nula indica que hay una raíz unitaria múltiple, es decir, los errores son no estacionarios, mientras que la hipótesis alternativa, lo contrario. De esta manera, se obtiene un Ø =9.964 que genera un *p*-valor de 0.00, por lo cual podemos concluir que los errores idiosincráticos son estacionarios y, por ende, los factores estimados son consistentes estadísticamente.

4.2.1. Validación económica

Para validar en un sentido estructural los factores subyacentes estimados, se propone compararlos y analizarlos con indicadores ya publicados por instituciones internacionales y validados de manera empírica, pero con periodicidad mayor (en este caso anual) y su cobertura de tiempo es menor, motivo por el cual no pueden ser directamente utilizados en este trabajo. De esta forma, los tres factores subyacentes son comparados con los indicadores publicados en el Fondo Monetario Internacional (FMI) por Svirydzenka (2016), los cuales se pueden apreciar en el cuadro 2.

Para esta comparación, los factores subyacentes se analizan promediando los valores mensuales para obtener el anual y se comparan con los indicadores propuestos previamente descritos. Las gráficas 1 muestran los comportamientos para el primer factor estimado.

Cuadro 2

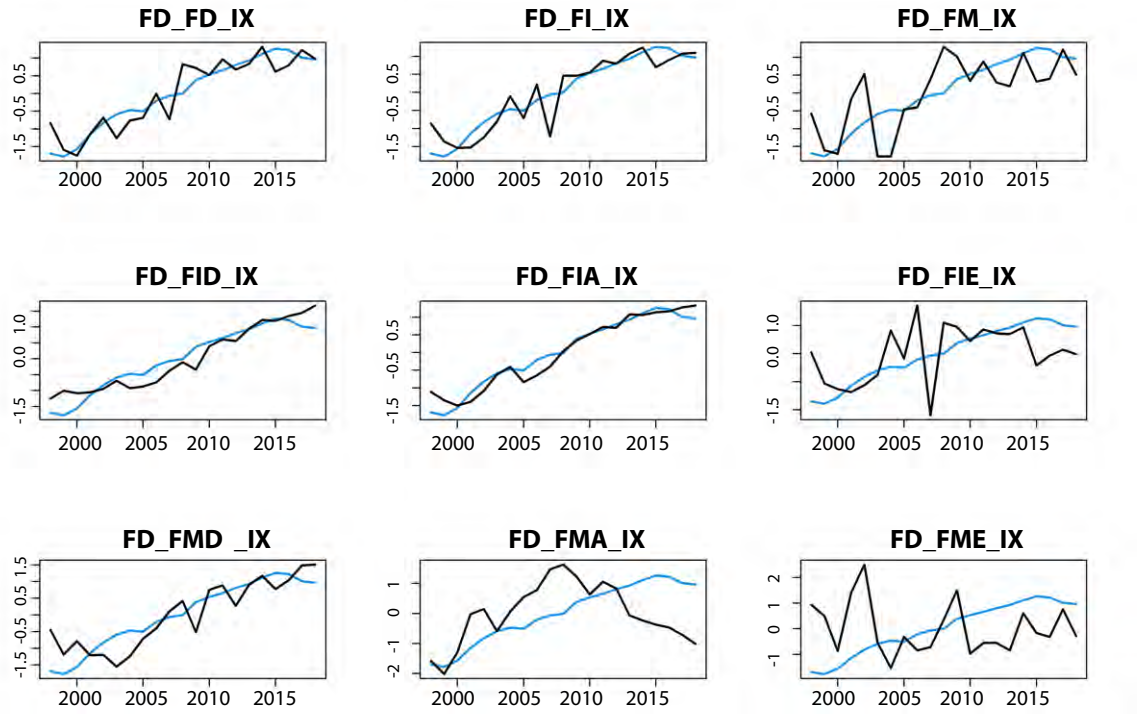
Variables de Svirydzenka (2016) incluidas en el estudio

Variable	Definición
Índice de Desarrollo Financiero Agregado (FD_FD_IX)	Calculado con base en los subíndices (FD_FI_IX) y (FD_FM_IX).
Índice Agregado de Desarrollo de las Instituciones Financieras (FD_FI_IX)	Subíndice que mide el grado de desarrollo general de las instituciones financieras calculado con base en los subíndices de profundidad, acceso y eficiencia de las instituciones financieras.
Índice Agregado de Desarrollo de los Mercados Financieros (FD_FM_IX)	Subíndice que mide el grado de desarrollo general de los mercados financieros calculado con base en los subíndices de profundidad, acceso y eficiencia de los mercados financieros.
Profundidad de las Instituciones Financieras (FD_FID_IX)	Subíndice calculado con base en los indicadores de crédito al sector privado (% PIB), activos de los fondos de pensiones (% PIB), activos de los fondos de inversión (% PIB) y las primas de seguro de vida y no vida (% PIB).
Acceso a las Instituciones Financieras (FD_FIA_IX)	Subíndice calculado con base en los indicadores de sucursales (bancos comerciales) por cada 100 mil adultos y el número de cajeros por cada 100 mil adultos.
Eficiencia de las Instituciones Financieras (FD_FIE_IX)	Subíndice calculado con base en los indicadores de margen de interés neto, distribución de los depósitos de préstamo, los ingresos no vinculados a intereses en relación con los ingresos totales, gastos generales sobre el total de activos, rendimiento de los activos, rendimiento de los fondos propios.
Profundidad de los Mercados Financieros (FD_FMD_IX)	Subíndice calculado con base en los indicadores de capitalización del mercado de valores en relación con el PIB, existencias comercializadas en relación con el PIB, títulos de deuda internacional del gobierno (% PIB), total de títulos de deuda de las sociedades no financieras (% PIB) y el total de títulos de deuda de las sociedades financieras (% PIB).
Acceso a los Mercados Financieros (FD_FMA_IX)	Subíndice calculado con base en el porcentaje de capitalización del mercado al margen de las 10 empresas más grandes y el número total de emisores de deuda (interna y externa, empresas no financieras y empresas financieras).
Eficiencia de los Mercados Financieros (FD_FME_IX)	Subíndice calculado con base en la relación de rotación del mercado bursátil (acciones comercializadas/ capitalización).

Fuente: series definidas con base en Svirydzenka (2016), y descargadas de https://data.imf.org/?sk=F8032E80-B36C-43B1-AC26-493C5B1CD33B

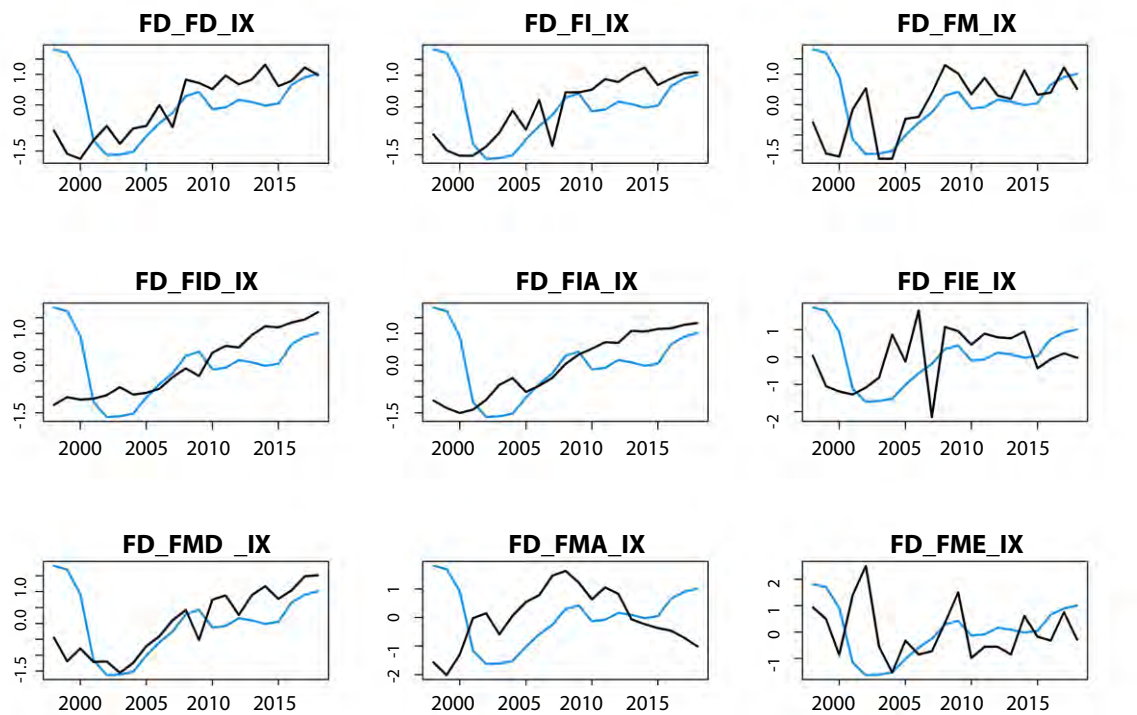
Gráficas 1

Primer factor común (color azul) respecto a indicadores financieros de Svirydzenka (2016) (color negro)



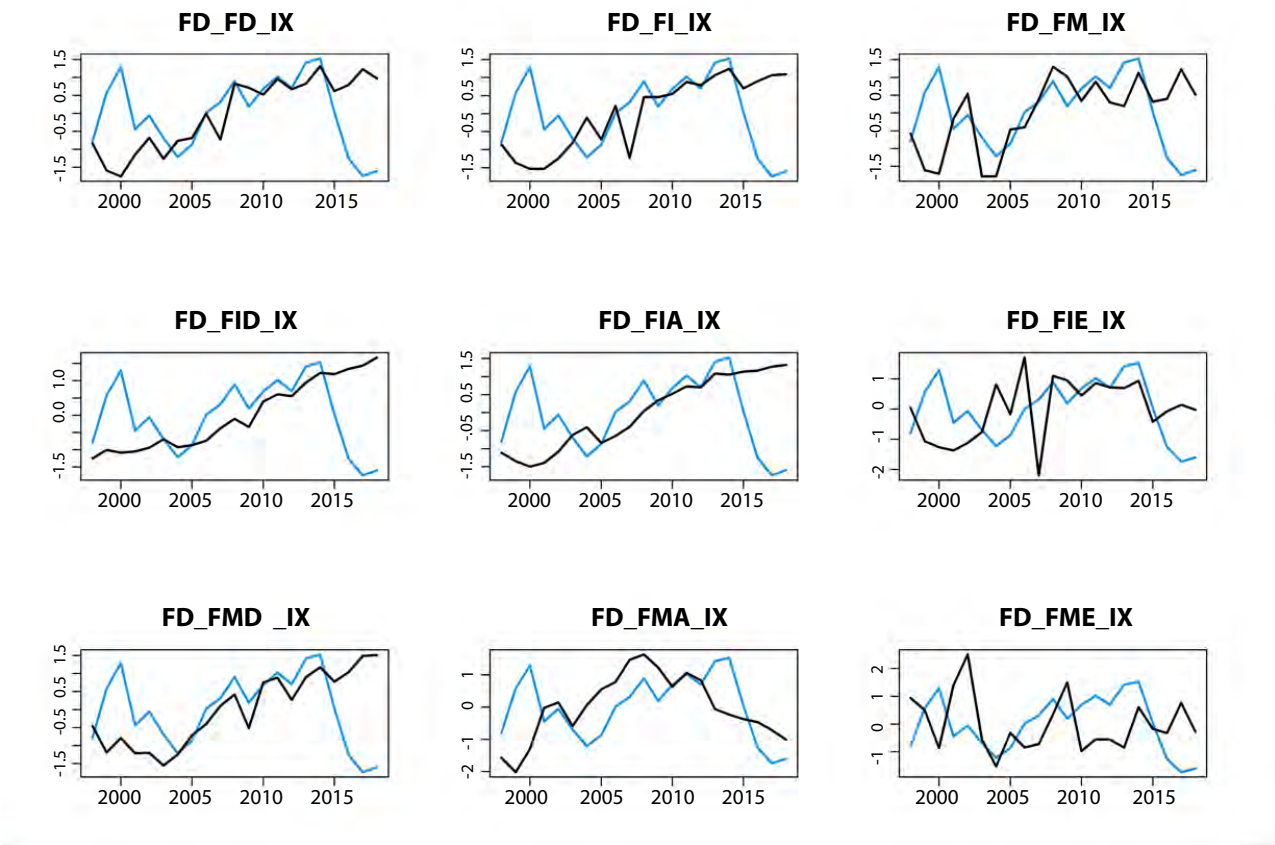
Gráficas 2

Segundo factor común (color azul) respecto a indicadores financieros de Svirydzenka (2016) (color negro)



Gráficas 3

Tercer factor común (color azul) respecto a indicadores financieros de Svirydzenka (2016) (color negro)



Se puede apreciar que el factor común tiene una tendencia positiva durante todo el periodo, cuyas correlaciones son positivas para todos los casos, excepto para FD_FME_IX, donde el valor estimado es de -0.22. Las correlaciones superiores a 0.90 se obtienen con FD_FIA_IX, FD_FID_IX, FD_FD_IX y FD_FI_IX, con valores de 0.97, 0.92, 0.91 y 0.90, respectivamente.

Asimismo, vale la pena comentar que estas altas correlaciones no son espurias bajo la concepción de Engle y Granger (1987), ya que se puede argumentar que las series están cointegradas para estos pares. Para este primer factor y a un nivel de significancia de 10 %, las correlaciones espurias se obtienen para FD_FM_IX, FD_FIE_IX y FD_FMA_IX, aunque los valores son 0.69, 0.46 y 0.35, respectivamente; es decir, en estos casos no se puede establecer estadísticamente que las series están cointegradas. Las gráficas 2 muestran lo mismo que las 1, pero haciendo referencia al segundo factor.

Es posible observar que el segundo factor común tiene una caída fuerte en los primeros periodos y después presenta una tendencia positiva en el resto, cuyas correlaciones son positivas para todos los casos, excepto para FD_FMA_IX, donde el valor estimado es de -0.46. En este caso no existen correlaciones altas, siendo la mayor con FD_FMD_IX con un valor de 0.42. Estas correlaciones, aunque bajas, no son espurias, es decir, para este se-

gundo factor y a un nivel de significancia de 10 %, se puede establecer estadísticamente que todos los pares de series están cointegrados. Finalmente, respecto a este análisis, las gráficas 3 presentan las relaciones del tercer factor respecto a cada una de las series financieras.

En este caso, el tercer factor común no parece presentar ninguna tendencia a largo plazo, mostrando algunas caídas y recuperaciones importantes durante todo el periodo; sin embargo, las correlaciones son positivas para casi todos los casos, excepto para FD_FID_IX, FD_FIA_IX y FD_FME_IX, donde el valor estimado es de -0.036, -0.038 y -0.091, respectivamente. En general, todas las correlaciones positivas están por debajo de 0.30, siendo la más alta la de FD_FID_IX con un valor de 0.27. Aun con estas correlaciones bajas, las relaciones no son espurias, ya que se puede argumentar que las series están cointegradas para estos pares; es decir, para este tercer factor y a un nivel de significancia de 10 %, se puede establecer estadísticamente que todos los pares de series están cointegrados.

4.2.2. Experimento Monte Carlo

Una vez verificada la consistencia estadística de la estimación de los factores a través de la prueba de Bai y Ng (2004) y también en un sentido estructural al correlacionar los factores subyacentes con los indicadores financieros de Svirydzenka (2016), se realiza además un experimento Monte Carlo con el fin de validar en un sentido muestral que las estimaciones de los factores, bajo las condiciones específicas en tamaño de muestra y no estacionariedad de las series utilizadas en este trabajo, se pueden extraer los que serían los verdaderos factores.

La justificación de verificar la consistencia obedece a que, aunque la estimación de los factores por PC genera buenos resultados para tamaños de muestra relativamente pequeños usando diferentes procesos generadores de datos (Corona *et al.*, 2018), las condiciones pueden variar según el conjunto de datos utilizado.

Ante este panorama, se genera un experimento Monte Carlo usando M=500 réplicas, T=300, N=12 y r=3. Los parámetros del DFM son los siguientes:

$$\phi=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\sum_n=\begin{pmatrix}2&0&0\\0&0.5&0\\0&0&0.25\end{pmatrix},\Gamma=\begin{pmatrix}0.9&0&0\\0&0.5&0\\0&0&0.9\end{pmatrix}.$$

Asimismo, Σ_a es simulada tal que permite débil correlación cruzada. Nótese que este experimento indica que los factores son no estacionarios, por ende, las series son no estacionarias; el primero es más fuerte que el segundo y este, que el tercero. Finalmente, aunque los errores idiosincráticos son estacionarios, estos tienen alta correlación serial y débil cruzada. En otras palabras, son condiciones similares a las cuales se estimaron los factores financieros.

En cada réplica computamos el promedio de las bondades de ajustes (R^2) al realizar las regresiones $F = a + b\hat{F} + e$, es decir, tratamos de explicar la variación en los factores simulados (verdaderos) con los estimados en la sección anterior. Los resultados indican que para los cuantiles 5, 50 y 95 % las $R2$ son 0.94, 0.98 y 0.99, y con ello podemos concluir que al hacer la estimación de los factores a través de PC se obtienen precisiones muy cercanas a lo que serían los verdaderos factores. De esta manera, el experimento Monte Carlo arroja evidencia de que, en muestras finitas, las estimaciones de los factores son confiables y robustas para el número de series consideradas y las dinámicas de cada uno de los componentes del DFM.

4.3. Relaciones dinámicas: corto y largo plazo

Los resultados de cointegración usando la prueba de Johansen (1991), específicamente la del máximo valor propio a 5 % de significancia, nos indican que no rechazamos la hipótesis nula para $m\leq3$, cuyo valor crítico es de 28.14, donde el estadístico de la prueba es de 26.41. De esta manera, existen hasta tres combinaciones lineales de series de tiempo no estacionarias que sí lo son. Centrándo-

nos en la primera ecuación de cointegración normalizada respecto a la actividad económica, la cual es más informativa en sentido estructural, los coeficientes de largo plazo están dados por la siguiente ecuación (valores *p* entre corchetes):

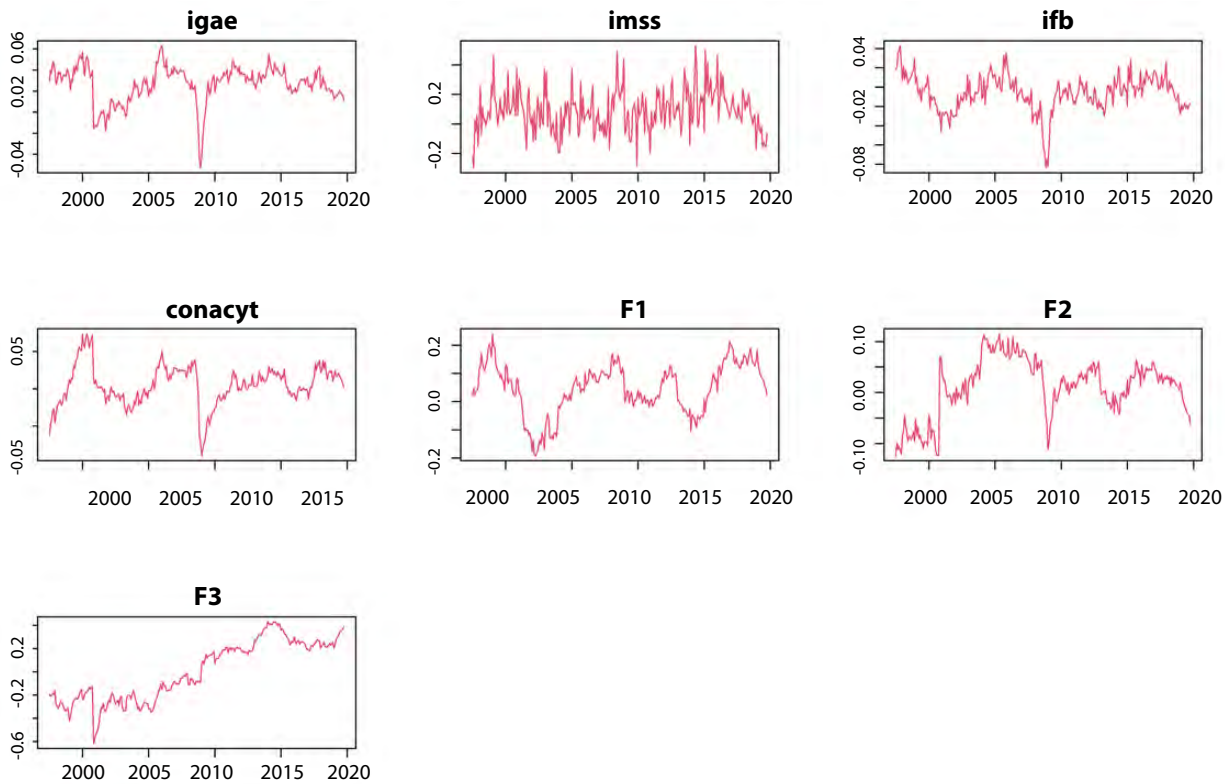
$$igae_t = \underset{[0.00]}{6.84} + \underset{[0.00]}{054imss} + \underset{[0.00]}{0.21ifb} + \underset{[0.03]}{0.01conacyt_t} + \underset{[0.02]}{0.01\hat{F}_{1t}} - \underset{[0.00]}{0.2\hat{F}_{2t}} - \underset{[0.00]}{0.03\hat{F}_{3t}} + \hat{e}_t. \quad (11)$$

Esta ecuación tiene varias consideraciones en términos económicos, primero nótese que la suma de los parámetros asociados a los factores trabajo, capital y capital humano es de 0.76, menor a 1; más aún, la suma de todos los coeficientes de la ecuación (11) es de apenas 0.81, lo cual nos puede dar indicios de que las variables *proxy* de la función Cobb-Douglas no capturan el crecimiento en el largo plazo como función de los factores de la producción y las variables de desarrollo financiero, es decir, nuestro método genera estimaciones sesgadas hacia abajo. Esto puede mostrar indicios de omisión de variables, para lo cual se podrían considerar en particular aquellas que permitan capturar de mejor manera la aportación del capital humano, ya que la contribución de esta variable es muy pequeña. En Durlauf (2005) se sugieren, por ejemplo, nivel de estudios, nivel de estudio en las mujeres, grado de cumplimiento de nivel básico de educación, proporción de estudiantes de ingeniería, entre otras.

No obstante, con respecto a los factores comunes, nótese que estos no son interpretables como variables en sí, sino como factores subyacentes que son producto de una estimación proveniente de variables observables. En ese sentido, lo más relevante es que todas las series de tiempo son significativas para explicar los niveles de la actividad económica en el largo plazo. En las gráficas 4 se muestra el comportamiento de las relaciones de cointegración normalizando respecto a cada una de las variables.

Gráficas 4

Relaciones de cointegración



Es de interés denotar el comportamiento estacionario de cada una de estas, lo que verifica que existen combinaciones lineales de series de tiempo no estacionarias que sí lo son. Nótese que el tercer factor, aunque presenta tendencia, es estacionario una vez eliminado el efecto de una tendencia determinística, obteniendo un valor p en la prueba ADF de 0.02. Enfocándonos en la actividad económica, la relación de cointegración nos indica la desviación que existe respecto a los efectos de largo plazo, llamando la atención la caída de la crisis financiera del 2009.

Para validar estadísticamente el modelo, nos centramos en evaluar los supuestos de no autocorrelación serial en los errores, $\hat{u}_t$, en la ecuación (7), para lo cual estimamos pruebas Ljung-Box con hasta siete rezagos, ecuación por ecuación y estimando la mediana de los p-valores para cada una de estas. En este sentido, probamos que ninguna de las ecuaciones tiene problemas de autocorrelación, por lo que puede concluirse que los coeficientes asociados al VEC se estiman de manera eficiente. Otra prueba es la normalidad residual, y aunque algunos residuales no se distribuyen normalmente, como las ecuaciones del *ifb* y de los factores financieros, nótese que, asintóticamente, podemos relajar este supuesto sin repercusiones en las propiedades de los estimadores asociados al VEC expresado en (7).

Los efectos de corto plazo son estimados a partir de la matriz de la ecuación (9), es decir, las ponderaciones lineales que no tienen efectos de largo plazo. La estimación normalizada para el *igae* se expresa como sigue:

$$igae_t = 0.49\mathrm{Z}(imss_t)_t + 2.48Z(ifb_t)_t + 0.51\mathrm{Z}(conacyt_t)_t + 10.35\mathrm{Z}(\hat{F}_{1t})_t - 1.31Z(\hat{F}_{2t})_t - 7.57\mathrm{Z}(\hat{F}_{3t})_t. \quad (12)$$

En este caso, podemos apreciar que en el largo plazo se mantienen los mismos signos que en el corto, aunque la magnitud asociada a los factores financieros es mayor que las variables consideradas como estructurales, que son los factores clásicos de la producción. Esto debe tomarse con cautela dado que las series que dan forma a estos son estandarizadas, por lo que, como hemos comentado, los coeficientes no deben interpretarse directamente como propensiones o elasticidades, no así el signo, el cual nos indica la contribución de los factores subyacentes, en este caso, las relaciones de cointegración sobre las observaciones.

Una interpretación natural del primer factor común es la siguiente:

1. Largo plazo: cuando las tasas de interés disminuyen (cambio marginal negativo) y el resto de las variables financieras incrementan (positivo), esto beneficia a la actividad económica en el largo plazo de manera significativa, donde sobresale el efecto del Índice de Precios y Cotizaciones, la liquidez y el efecto de la cartera vigente, es decir, un sistema financiero sano.
2. Corto plazo: cuando hay un choque de corto plazo en el mercado financiero, producto de un incremento de la actividad financiera respecto a su comportamiento de largo plazo, la actividad económica reacciona positivamente.

En la siguiente subsección se presenta una discusión detallada de los resultados en términos económicos.

4.4. Discusión de los resultados

Hemos indicado el debate existente acerca del papel que puede tener el sistema financiero sobre el crecimiento, y nuestros resultados coinciden con el pensamiento de diversos economistas teóricos que están de acuerdo en que para comprender el crecimiento económico es indispensable tomar en cuenta al sistema financiero.

Es en ese sentido que esta investigación se orienta en observar el papel que juegan las finanzas en el proceso de crecimiento económico en el corto y largo plazo. Se utilizan variables de frecuencia mensual desde 1997 hasta el 2019, lo que permite realizar un análisis más detallado y

robusto en comparación con la literatura previa. En todos los cálculos se verifica la consistencia estadística, económica y empírica. Se determinaron tres variables latentes que representen al sistema financiero mexicano, las cuales son estimadas consistentemente de acuerdo con la prueba PANIC de Bai y Ng (2004). Además, en un sentido estructural, se valida que dichos factores tienen alta relación lineal con indicadores financieros publicados por organismos internacionales que miden la profundidad, el acceso y la eficiencia de este. También, mediante un experimento Monte Carlo, se computa el promedio de las bondades de ajustes entre los factores simulados y los estimados. Los resultados indican que para los cuantiles 5, 50 y 95 % las R^2 son 0.94, 0.98 y 0.99, es decir, la estimación obtiene precisiones muy cercanas a lo que serían los verdaderos factores y, por lo tanto, estos son muestralmente consistentes.

Los factores estimados se utilizan para realizar pruebas de cointegración en conjunto con variables económicas clave del crecimiento. En este sentido, la prueba de Johansen (1991) indica la existencia de hasta tres vectores de cointegración. Centrándonos en la primera ecuación de cointegración normalizada respecto a la actividad económica, los coeficientes de largo plazo de los factores de trabajo, capital y capital humano (0.54, 0.21, 0.01, respectivamente) son todos positivos, pero suman menos de la unidad al igual que la suma de todos los coeficientes de la ecuación, lo cual nos puede dar indicios de que las variables *proxy* de la función Cobb-Douglas no capturan en su totalidad la contabilidad del crecimiento económico a través de los factores de la producción. Por otra parte, esto también nos puede indicar que empíricamente existen otras variables que no están siendo consideradas.

Las estimaciones aquí realizadas están basadas en el modelo con restricciones financieras a la teoría de crecimiento schumpeteriana. Uno de los resultados más relevantes es que todas las series son significativas para explicar los niveles de la actividad económica en el largo plazo; además, los signos de los factores de producción y el primer factor financiero son positivos, de acuerdo con lo esperado.

En específico, esto implica que, cuando las tasas de interés reales presentan un cambio marginal negativo y el resto de las variables financieras muestran uno positivo se beneficia a la actividad económica en el largo plazo de manera significativa, donde sobresale el efecto del Índice de Precios y Cotizaciones, la liquidez y el de la cartera vigente. Con respecto al corto plazo, calculado con la metodología Gonzalo y Granger (1995), los coeficientes para trabajo, capital y capital humano son 0.49, 2.48 y 0.51, respectivamente, y con respecto a los tres factores financieros sus coeficientes son 10.35, -1.31 y -7.57, esto quiere decir que cuando hay un choque en el mercado financiero, producto de un incremento de la actividad financiera respecto a su comportamiento de largo plazo, la actividad económica reacciona positivamente. Se debe tener en cuenta que los tres factores comunes no son interpretables como variables en sí, sino como factores subyacentes que son producto de una estimación proveniente de variables observables, por lo que la interpretación no es directa, no así el signo, el cual nos indica su contribución, en este caso, las relaciones de cointegración a largo plazo sobre las observaciones.

Finalmente, se realiza un ejercicio de verificación en términos estructurales desagregando el indicador financiero FD_FD_IX con ayuda del primer factor subyacente estimado, dado que ambos están altamente correlacionados y también están cointegrados. Se utiliza la técnica de Denton-Cholette para desagregar el indicador financiero y se hace una estimación estilo Cobb-Douglas ampliada con la variable desagregada obteniendo los siguientes resultados (valores p entre corchetes):

$$igae_t = \underset{[0.00]}{5.84} + \underset{[0.00]}{0.60} imss_t + \underset{[0.00]}{0.17} ifb_t + \underset{[0.00]}{0.01} conacyt_t + \underset{[0.00]}{0.19}\, \tilde{F}_t + \hat{w}_t . \quad (13)$$

Para verificar evidencia de cointegración desde la perspectiva de Engle y Granger (1987), se realiza la prueba ADF sobre $\hat{w}_t$ y se concluye que esta es estacionaria, por lo cual también se concluye que hay evidencia de cointegración, es decir, las va-

riables consideradas se relacionan en el largo plazo, en este caso, con la actividad económica. Llama la atención con respecto a la ecuación (11) que los coeficientes asociados a los factores estructurales son similares, no obstante, el valor de $\tilde{F}_t$ es mayor que los valores sumados asociados a los efectos de los factores subyacentes estimados previamente. En conclusión, este ejercicio verifica que desde la perspectiva de Engle y Granger también hay evidencia en favor de la contribución del sistema financiero en el crecimiento económico en el largo plazo.

5. Conclusiones

Este trabajo contribuye a la discusión sobre el papel que tiene el sistema financiero en el crecimiento económico utilizando una novedosa aplicación empírica para el caso de México. En este sentido, el análisis de factores dinámicos permite concluir que los factores estimados representan consistentemente, en sentido estadístico y económico, la evolución del sistema financiero en el país.

Los factores se relacionan de manera significativa con el crecimiento económico en el largo plazo en el sentido de que cuando las tasas de interés disminuyen y el resto de las variables financieras incluidas en el modelo incrementan se beneficia la actividad económica, sobresaliendo el efecto del Índice de Precios y Cotizaciones, la liquidez y el de la cartera vigente, es decir, un sistema financiero sano. En el corto plazo, cuando hay un choque en el mercado financiero producto de un incremento de la actividad financiera respecto al comportamiento de largo plazo, la actividad económica reacciona positivamente.

Lo anterior es relevante, por ejemplo, para la conducción de política monetaria, pues la literatura en la materia indica que la fijación de controles en las tasas de interés por debajo de los niveles de mercado o la existencia de mecanismos que induzcan este comportamiento conducen, por lo general, a una asignación de crédito por parte del sistema financiero formal, combinada con un floreciente mercado informal formado por consumidores o productores que se prestan entre sí a tasas de mercado. En ese sentido, son importantes iniciativas planteadas por las autoridades financieras, como la Tasa de Interés Interbancaria de Equilibrio (TIIE) de fondeo que busca establecer una tasa de referencia que esté basada en transacciones observadas en el mercado y que el mercado al que hacen referencia exista y sea representativo, y evitar así abusos o manipulaciones como los ocurridos con la tasa LIBOR en el Reino Unido.

Los resultados apoyan también la hipótesis de que el sector financiero es especialmente relevante en un contexto de recuperación económica. Por ejemplo, en el *Reporte de estabilidad financiera* del Banco de México (2021, pp. 7-9) se encuentra que en las economías emergentes el otorgamiento del crédito es más importante para que la recuperación sea robusta y rápida, y se indica la importancia de mejorar el entorno y fomentar las condiciones de certidumbre para la inversión y la actividad productiva.

Finalmente, para futuras investigaciones, sería conveniente probar otras variables para capturar el efecto de la variable asociada al capital humano, ya que en la estimación de la función Cobb-Douglas, para explicar la contabilidad del crecimiento económico, se encontró que la suma de los coeficientes es menor a la unidad, en particular el coeficiente asociado con dicha variable es pequeño.

Fuentes

Aghion, P. y P. W. Howitt. *The economics of growth*. MIT press, 2008.

Ahn, S. y A. Horenstein. "Eigenvalue ratio test for the number of factors", en: *Econometrica*. 81(3), 2013, pp. 1203-1227.

Bai, J. "Inferential theory for factor models of large dimensions", en: *Econometrica*. 71(1), 2003, pp. 135–171.

_____ "Estimating cross-section common stochastic trends in non-stationary panel data", en: *Journal of Econometrics*. 122(1), 2004, pp.137-183.

Bai, J. y S. Ng. "Determining the number of factors in approximate factor models", en: *Econometrica*. 70(1), 2002, pp. 191-221.

_____"A panic attack on unit roots and cointegration", en: *Econometrica*. 72(4), 2004, pp. 1127-1177.

Banco de México (BANXICO). *Sistema de Información Económica*. 2020.

_____*Reporte de estabilidad financiera, primer semestre del 2021*. Banco de México, 2021 (DE) https://www.banxico.org.mx/publicaciones-y-prensa/reportes-sobre-el-sistema-financiero/%7BFBAA9B09-D285-B1CC-33A8-84E35347B970%7D.pdf

Beck, T. *The econometrics of finance and growth*. Policy Research Working Paper Series 4608. The World Bank, 2008.

_____ *Finance and growth: Lessons from the literature and the recent crisis.* Prepared for the LSE Growth Commission, 2012.

Bergoeing, R., P. J. Kehoe, T. J. Kehoe y R. Soto. "A Decade Lost and Found: Mexico and Chile in the 1980s", en: *Review of Economic Dynamics*. 5(1), 2002, pp. 166-205.

_____ "A Decade Lost and Found: Mexico and Chile in the 1980s", en: Kehoe, T. J. and Edward C. Prescott. *Great Depressions of the Twentieth Century*. Minneapolis, Federal Reserve Bank of Minneapolis, 2007, pp. 217-256.

Blanchard, J. y D. Quah. "The Dynamic Effects of Aggregate and Supply Disturbance", en: *The American Economic Review*. 79(4), 1989, pp. 653-673.

Blomström, M., R. E. Lipsey y M. Zejan. "Is fixed investment the key to economic growth?", en: *The Quarterly Journal of Economics.* 111(1), 1996, pp. 269-276.

Campos, N. y J. Nugent. "Who is afraid of political instability?", en: *Journal of Development Economics.* 67, 2002, pp. 157-172.

Cermeño, R., J. García y C. González. "Desarrollo financiero y la volatilidad del crecimiento: evidencia de series de tiempo para México y Estados Unidos", en: *Monetaria.* 38(2), 2012, pp. 209-250.

Cheng, C. Y., M. S. Chien y C. C. Lee. "ICT diffusion, financial development, and economic growth: An international cross-country analysis", en: *Economic Modelling*. 94, 2021, pp. 662-671.

Corona, F., P. Poncela y E. Ruiz. "Determining the Number of Factors After Stationary Univariate Transformations", en: *Empirical Economics*. 53(1), 2017, pp. 351-372.

_____ "Estimating non-stationary common factors: implications for risk sharing", en: *Computational Economics*. 55(1), 2020, pp. 37-60.

Denton, F. "Adjustment of Monthly or Quarterly Series to Annual Totals an Approach Based on Quadratic Minimization", en: *Journal of the American Statistical Association*. 66(333), 1971, pp. 99-102.

Dritsakis, N. y A. Adamopoulos. "Financial Development and Economic Growth in Greece: An Empirical Investigation with Granger Causality Analysis", en: *International Economic Journal.* 18(4), 2004, pp. 547-559.

Durlauf, S. "Growth econometrics", en: *Handbook of Economic Growth*. 1, 2005, pp. 555-677.

Engle, R. y C. Granger. "Co-Integration and Error Correction: Representation, Estimation, and Testing", en: *Econometrica*. 55(2), 1987, pp. 251-276.

Geweke, J. *Latent Variables in Socio-Economic Models*. 1977.

Goldsmith, W. *Financial Structure and Development*. 1969.

Gonzalo, J. y C. W. J. Granger. "Estimation of common long-memory components in cointegrated systems", en: *Journal of Business & Economic Statistics*. 13(1), 1995, pp. 27-35.

Gurley, J. y E. Shaw. "Financial Aspects of Economic Development", en: *The American Economic Review*. 45(4), 1995, pp. 515-538.

Ibrahim, M. y P. Alagidede. "Effect of financial development on economic growth in sub-Saharan Africa", en: *Journal of Policy Modeling*. 40(6), 2018, pp. 1104-1125.

Instituto Mexicano del Seguro Social (IMSS). *Consulta dinámica*. 2020.

Instituto Nacional de Estadística y Geografía (INEGI). *Banco de Información Económica*. México, INEGI, 2020.

Johansen, S. "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models", en: *Econometrica*. 59(6), 1991, pp. 1551-1580.

Jones, C. I. "Time series tests of endogenous growth models", en: *The Quarterly Journal of Economics*. 110(2), 1995, 495-525.

Kadhuma, T. A. A. y F. A. Kadhimb. "The Impact of Liberalisation of the Financial Sector on Economic Growth in Iraq for the Period 2004-2018", en: *International Journal of Innovation, Creativity and Change*. 14(1), 2020, pp. 709-726.

King, R. y R Levine. "Finance, entrepreneurship and growth", en: *Journal of Monetary Economics*. 32(3), 1993, pp. 513-542.

Levine, R. "Finance and growth: theory and evidence", en: *Handbook of Economic Growth*. 1, 2005, pp. 865-934.

Lucas Jr, R. E. "On the mechanics of economic development", en: *Journal of Monetary Economics*. 22(1), 1988, pp. 3-42.

Lustig, N. "Life Is Not Easy: Mexico's Quest for Stability and Growth", en: *Journal of Economic Perspectives*. 15(1), 2001, pp. 85-106.

Maune, A., E. Matanda y J. Mundonde. "Does financial inclusion cause economic growth in Zimbabwe? An empirical investigation", en: *Acta Universitatis Danubius. Economica.* 16(1), 2020, pp. 195-215.

McKinnon, R. *Money and Capital in Economic Development.* 1973.

Méndez-Heras, L. B., F. Venegas-Martínez y D. E. Linthon-Delgado. "El impacto del crédito bancario sobre el desarrollo humano en México: un análisis de datos panel a nivel estatal, 2004-2016", en: *Ensayos Revista de Economía.* 40(1), 2021, pp. 1-28.

Miller, M. "Financial Markets and Economic Growth", en: *Journal of Applied Corporate Finance.* 11(3), 1998, pp. 8-13.

Moreno-Brid, J. C., J. Santamaria y J. C. Rivas. "Industrialization and Economic Growth in Mexico after NAFTA: The Road Travelled", en: *Development and Change.* 36(6), 2005, pp. 1095–1119.

Onatski, A. "Determining the number of factors from empirical distribution of eigenvalues", en: *The Review of Economics and Statistics.* 92(04), 2010, pp. 1004-1016.

Popov, A. "Evidence on finance and economic growth", en: *Handbook of Finance and Development.* 2018.

Robinson, J. "The Generalisation of the General Theory", en: *The Generalisation of the General Theory and Other Essays.* London, Palgrave Macmillan, 1952, pp. 1-76.

Rodríguez Benavides, D. y F. López Herrera. "Desarrollo financiero y crecimiento económico en México", en: *Problemas del Desarrollo. Revista Latinoamericana de Economía.* 40(159), 2010, pp. 41-60.

Rousseau, P. "Historical perspectives on financial development and economic growth", en: *Review, Federal Reserve Bank of St. Louis.* 2003, pp. 81-106.

Ruiz, J. L. "Financial development, institutional investors, and economic growth", en: *International Review of Economics & Finance*. 54, 2018, pp. 218-224.

Sargent T.J. y C.A. Sims. "Business cycle modeling without pretending to have too much a priory economic theory", en: *New methods in business cycle research.* Federal Reserve Bank of Minneapolis, Minneapolis, 1977.

Schumpeter, J. *Theorie der wirtschaftlichen Entwicklung (The Theory of Economic Development*, traducida por R. Opie. Cambridge, MA, Harvard University Press, 1934). Leipzig, Dunker & Humblot, 1912.

Secretaría de Hacienda y Crédito Público (SHCP). *Estadísticas oportunas de finanzas públicas.* 2020.

Slesman, L., A. Z. Baharumshah, y W. N. W. Azman-Saini. "Political institutions and finance-growth nexus in emerging markets and developing countries: a tale of one threshold", en: *The Quarterly Review of Economics and Finance*. 72, 2019, pp. 80-100.

Stock, J. H. y M. W. Watson. "Testing for common trends", en: *Journal of the American Statistical.* 83(444), 1988, p. 1097-1107.

Svirydzenka, K. *Introducing a New Broad-based Index of Financial Development.* International Monetary Fund, 2016.

Venegas, F, M. Tinoco y V. Torres. "Desregulación financiera, desarrollo del sistema financiero y crecimiento económico en México: efectos de largo plazo y causalidad", en: *Estudios Económicos.* 24(2), 2009, pp. 249-283.

Evaluación de técnicas de procesamiento de lenguaje natural y *Machine Learning* **para los procesos de codificación de encuestas en hogares**

Evaluation of Natural Language Processing and Machine Learning Techniques **for Household Survey Coding Processes**

José Alejandro Ruiz Sánchez,* Jael Pérez Sánchez y Adrián Pastor López Monroy*****

* Instituto Nacional de Estadística y Geografía (INEGI), jose.ruizs@inegi.org.mx
** INEGI, jael.perez@inegi.org.mx
*** Centro de Investigación en Matemáticas, A. C., pastor.lopez@cimat.mx

Artificial intelligence (AI), machine learning and modern computer technologies concepts. Business, Technology, Internet and network concept./ putilich/ iStock

De los múltiples procesos productivos llevados a cabo dentro de las oficinas nacionales de estadística se encuentra el de codificación, el cual consiste en la asignación automática o manual de claves alfanuméricas a un registro u observación. Este mapeo a un conjunto de categorías predefinidas permite agrupar registros bajo una misma descripción, lo cual facilita su manejo y análisis. Un porcentaje importante de estas tareas de codificación se realizan con ayuda de algoritmos determinísticos basados en reglas de decisión; sin embargo, otros procesos utilizan en mayor medida la asistencia de expertos humanos. El trabajo que a continuación presentamos tiene por objetivo valorar el uso e incorporación de técnicas de procesamiento de lenguaje natural (PLN) y de *Machine Learning (ML)* para incrementar el porcentaje de registros clasificados de manera automática. Para ello, tomamos las variables de *ocupación* y *actividad económica* de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2018. Los resultados obtenidos muestran que sería posible trasladar 50 % de los registros que actualmente se codifican con asistencia humana hacia un proceso de codificación automatizada con algoritmos de PLN y *ML*.

Palabras clave: procesamiento de lenguaje natural; *Machine Learning;* codificación; ENIGH.

Recibido: 26 de julio de 2021.
Aceptado: 27 de octubre de 2021.

National Statistic Offices carry out multiple production processes, coding being one of them. Coding is referred to the assignment of alphanumeric keys to a particular observational unit. The coding process can be either automatic or manual and it is based on certain additional information. This mapping to a set of predefined categories allows grouping records under the same description, which facilitates its management and analysis. Currently a great percentage of the coding tasks are made by deterministic algorithms or decision rules. However, there are processes where human intervention to code is still largely required. The paper we present assesses the use and incorporation of Natural Language Process (NLP) and Machine Learning (ML) to increase the percentage of automatically coded records. We evaluate the process on two variables from the National Survey of Household Income and Expenditure (ENIGH by its acronym in Spanish) 2018: *occupation* and *economic activity*. Our results show it could be possible to transfer 50% of the records coded by humans to be automatically coded by NLP and ML.

Key words: Natural Language Process; Machine Learning; Coding; ENIGH.

Introducción

En todo proyecto de generación de información estadística por lo general existen preguntas abiertas que dan como resultado un conjunto de descripciones (usualmente en forma de texto) sobre la temática que se está levantando. Al proceso de asignarle una clave alfanumérica a estas con fines de explotación de la información se le llama codificación. En los censos de población y vivienda, así como en las encuestas en hogares, hay diferentes variables que entran al proceso de codificación, por ejemplo: parentesco, lengua indígena, religión, lugar de nacimiento, lugar de residencia, ocupación de la persona o la actividad económica en la cual trabaja.

Actualmente, el área del Instituto Nacional de Estadística y Geografía (INEGI) a cargo de la codificación de encuestas en hogares, como la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), realiza el proceso en dos etapas: la primera se le llama automática porque se hace con un conjunto de

reglas determinísticas que son aplicadas a través de algoritmos computacionales, donde se conjuga con un preprocesamiento de los textos para lograr una mayor cantidad y calidad de registros codificados de manera correcta; la segunda se denomina codificación asistida, pues es realizada por humanos a través de un sistema que facilita dicho trabajo. Tanto para la codificación automática como para la asistida, se garantiza un nivel de calidad mínimo aceptable que está en función de la complejidad de cada variable. Para la mayor parte de las variables, la calidad de la codificación se encuentra establecida en un nivel mínimo de 98 %, no así para *ocupación y actividad económica*, donde clasificar las descripciones resulta mucho más complejo que para el resto de variables.

Para las variables de *ocupación* y *actividad económica*, la calidad oscila entre 90 y 94 %, dependiendo del proyecto que se trate, sea censo o encuestas en hogares, con un porcentaje de codificación automática entre 70 y 76; el complemento se hace de manera asistida. La verificación de la calidad de la codificación se hace de forma implícita en el proceso; para la automática se emplea un muestreo probabilístico para cada lote de información y para cada estrategia de codificación, mientras para la asistida, se emplea un muestreo de aceptación por lotes (tal cual se hace en la industria) donde, para cada carga de trabajo hecha por cada codificador se obtiene una muestra; si esta no pasa el umbral de calidad mínimo aceptable, se vuelve a codificar la carga completa (ver *Anexo*).

El trabajo que a continuación presentamos tiene por objetivo valorar el uso e incorporación de técnicas de procesamiento de lenguaje natural (PLN) y de *Machine Learning* (*ML*) para incrementar el porcentaje de registros clasificados de manera automática. Para ello, tomamos las variables de *ocupación* y *actividad económica* de la ENIGH 2018.

El documento está dividido en dos secciones: en la primera reportamos el proceso metodológico que seguimos, así como los principales resultados, y en la segunda proponemos la operacionalización del proceso metodológico en conjunción con los actuales procesos productivos.

1. Procesamiento de lenguaje natural para codificación automática

A través de los años, el INEGI ha desarrollado procesos de codificación para distintas variables, lo cual ha resultado en una gran cantidad de información etiquetada por personal especialmente capacitado para tales tareas. Sin embargo, para algunas de las variables, el proceso de codificación sigue representando un reto en términos de tiempo, recursos humanos y mejora en calidad. Los avances recientes en el campo del PLN y *ML*, junto con los insumos acumulados por las oficinas nacionales de estadística (ONE), ofrecen la posibilidad de incorporar nuevas metodologías y así obtener mejores resultados en términos de calidad y eficiencia. Por PLN nos referimos al conjunto de técnicas y algoritmos empleados para el tratamiento automático del lenguaje (Jurafsky and Martin, 2008) y en nuestro caso, al procesamiento, modelación y representación del texto contenido en la base de datos en cuestión. Para la parte de *ML*, nos enfocamos en clasificadores automáticos basados en procesos de optimización, que parametrizan modelos mediante el insumo de datos para descubrir relaciones y patrones relevantes entre ellos (Svensén and Bishop, 2007).

En esta sección elaboramos una primera aproximación a las técnicas de PLN aplicadas a las tareas de clasificación automática para las variables *ocupación* y *actividad económica* que reportan los informantes de la ENIGH. Nuestro objetivo es explorar la viabilidad del uso de estos métodos dentro de alguna de las fases del proceso de codificación.

1.1 Datos

El ejercicio lo realizamos sobre la ENIGH 2018, cuyas variables de *ocupación* y *actividad económica* ya están codificadas según los procesos actuales del INEGI. La base de datos tiene 158 568 registros etiquetados. Los campos descriptivos con base en los cuales se realizó la codificación son: *Nombre de la ocupación, Tareas o funciones, Activi-*

dad económica de la empresa o institución y *Nombre de la empresa*. Adicionalmente, se integran las siguientes características de apoyo: *Lugar donde realizó las actividades*, *Clasificación de la empresa*, *Si tiene personal a su cargo*, *Nivel académico*, *Grado académico*, *Nivel académico aprobado*, *Grado aprobado*, *Tamaño de la empresa*, *Trabaja dentro del país* y *Pregunta complementaria a nombre de la empresa*.

1.2 Procesamiento y resultados

En términos generales, el desarrollo metodológico que hemos seguido para la aplicación de las técnicas de PLN y *ML* consta de tres fases: 1) preprocesamiento de los datos, 2) vectorización numérica de los textos y 3) clasificación a través de algoritmos de aprendizaje supervisado[1] de *ML*.

1.2.1 Preprocesamiento

Como se comentó en la sección anterior, el proceso actual de codificación que se realiza en la ENIGH se compone de dos etapas, aquella basada en algoritmos determinísticos de reglas de decisión y otra donde interviene personal especializado. Para la primera, es especialmente importante contar con un texto estandarizado; para ello, se realizan correcciones ortográficas, lematizaciones y truncado de palabras. Este mismo proceso es el que hemos seguido como paso inicial para el tratamiento de los datos para el uso de PLN y *ML*. Una vez estandarizado el texto, generamos conjuntos de *n*-gramas (partición de una palabra o frase en subconjuntos de igual cardinalidad; así, existen *n*-gramas a nivel carácter y palabra. Este último consiste en la agrupación de palabras secuenciales; así, un *n*-grama a nivel palabra de cardinalidad o longitud 2 resulta de la unión de dos palabras secuenciales), los cuales serán el insumo para la etapa de vectorización. Los *n*-gramas usualmente benefician la efectividad de clasificación debido a que capturan información del contexto o conceptos compuestos por *n* palabras.

1 Los métodos de *ML* supervisado establecen relaciones entre un conjunto de covariables con otra variable objetivo, la cual es de interés predecir o estimar.

1.2.2 Vectorización

En términos generales, esta etapa consiste en representar numéricamente un texto, el cual puede estar compuesto por un conjunto de caracteres alfanuméricos (*tokens*) y, por lo tanto, contener *n*-gramas o palabras completas.

Uno de los métodos tradicionales en PLN para vectorización consiste en contabilizar la repetición de cada uno de los *tokens* que aparece en un documento (bolsa de palabras). Estos pueden ser palabras, *n*-gramas o cualquier otra unidad simbólica que se determine de manera previa. Algunas de sus variantes surgen de agregar ponderaciones a este conteo, tal es el caso de la técnica *Term frequency-Inverse document frequency* (*TF-IDF*).

Consideremos N, número de documentos (o registros), cada uno de los cuales está compuesto por una cantidad finita de términos o *tokens*. Cada término i puede ser representado por un vector numérico de dimensión menor o igual a N, de acuerdo con la siguiente fórmula:[2]

$$w_{i,j} = (1 + log\ (tf_{i,j})) * log\ log\ \left(1 + \frac{N}{(1 + df_i)}\right),$$

donde:

$tf_{i,j}$ = número de veces que aparece el término i en el documento j.
df_i = número de documentos que contienen el término i.
N = número de documentos.

Por ejemplo (ver ilustración), consideremos los siguientes tres ($N = 3$) documentos (textos) obtenidos de la ENIGH. En este caso la palabra *FRUTA* aparece en ocho ocasiones y solo en uno de los textos ($df_i = 1$).

2 Para el procesamiento de la técnica *TF-IDF,* usamos el paquete *Keras* en *Python*. La fórmula que se muestra fue tomada de la siguiente liga: https://github.com/keras-team/keras-preprocessing/blob/master/keras_preprocessing/text.py

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Ilustración

	FRUTA	ALBAÑIL	DEPARTAMENTO	PISO	AGRICULTURA	…
1. "ALBAÑIL PEGAR PISO HACER ACABAR COMO REPELLAR PEGAR PISO *ETC*. DEPARTAMNETO NO DEPARTAMENTO"	0	1	2	2	0	…
2. "AYUDANTE DE ALBAÑIL BATIR MEZCLA ACARREAR LADRILLO QUEHACER COSA DE LADRILLO"	0	1	0	0	0	…
3. "AGRICULTOR CHAPEAR SU AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA"	8	0	0	0	2	…

↓

	FRUTA	ALBAÑIL	DEPARTAMENTO	PISO	AGRICULTURA	…
1. "ALBAÑIL PEGAR PISO HACER ACABAR COMO REPELLAR PEGAR PISO ETC. DEPARTAMENTO NO DEPARTAMENTO"	0	1**log*(3/2)	2**log*(3/1)	2**log*(3/1)	0	…
2. "AYUDANTE DE ALBAÑIL BATIR MEZCLA ACARREAR LADRILLO QUEHACER COSA DE LADRILLO"	0	1**log*(3/2)	0	0	0	…
3. "AGRICULTOR CHAPEAR SU AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA"	8**log*(3/1)	0	0	0	2	…

De esta manera, cada uno de los documentos (renglones) en la ilustración es representado por un vector numérico de una dimensión finita. De forma intuitiva, la técnica castiga términos que aparecen en todos los documentos y, por lo tanto, suelen ser un tanto discriminativos. Naturalmente, este proceso puede ser fácilmente extendido cuando se incluyen *n*-gramas a nivel carácter y/o palabra.

Una vez que obtenemos una representación vectorial para cada documento (en nuestro caso, un documento está conformado por la concatenación de las respuestas textuales de un informante de la ENIGH convertidas en *n*-gramas), los resultados pasan a la siguiente etapa, que es la de clasificación automática a través de algoritmos de *Machine Learning*.

1.2.3 Resultados de la codificación automática

Los resultados de los distintos ejercicios que se presentan a continuación tienen como finalidad aportar información sobre las variaciones metodológicas y parametrizaciones que resultaron favorables para nuestros problemas de clasificación (*ocupación* y *actividad económica* para la ENIGH).

De los 158 568 registros originales en nuestra base de datos, excluimos 58 pertenecientes a clases (códigos) con menos de cuatro registros. De los datos restantes, usamos 75 % de ellos como conjunto de entrenamiento (manteniendo las proporciones de las categorías) y el restante 25 % como conjunto de prueba. Los ejercicios realizados fueron para clasificar, por separado, la variable *ocupa-*

ción conformada por 461 clases, y la de *actividad económica* que cuenta con 157.

Con respecto a los algoritmos de *ML*, para este proyecto empleamos los tradicionales de aprendizaje supervisado (donde existe una variable a predecir y un conjunto de predictores), como *Support Vector Machine* (*SVM*), regresión logística, *Radom Forest*, *K Nearest Neighbors* (*K-NN*), *Naive Bayes*, entre otros (Chih-Chung y Chih-Jen, 2019; Platt, 1999; Hsiang-Fu *et al*., 2011; Breiman, 2001; Zhang, 2004). De estos, el que obtiene en general resultados más sobresalientes es el *SVM*, el cual es un algoritmo de aprendizaje automático que tiene el objetivo de encontrar un hiperplano óptimo que separe las instancias que pertenecen a dos diferentes clases (Svensén and Bishop, 2007). De manera específica, y sea $\{\boldsymbol{X}_i, y_i\}$ el conjunto de pares instancia-categoría de ejemplos de entrenamiento, donde $\boldsymbol{X}_i \in \mathbb{R}^d$ $y \in \{-1, +1\}$ con d dimensionalidad del problema (e.g., el tamaño del vocabulario). El *SMV* trata de determinar un mapeo de los ejemplos de entrenamiento a las categorías objetivo por medio de la utilización de la siguiente función lineal:

$$f(x) = sign(\sum_i \alpha_i y_i k(x_i, \boldsymbol{x}) - b)$$

donde α_i y y_i son los pesos y etiqueta del ejemplo de entrenamiento i. Para mapear los vectores de entrada (x_i, x_j) al espacio de características, se utiliza la función de kernel $k(x_i, \boldsymbol{x})$. De manera intuitiva, $k(x_i, \boldsymbol{x})$ mide la similitud entre las instancias x_i y x_j. Los parámetros α y b son aprendidos por medio de optimización, y la selección de la función de kernel es de vital importancia (Svensén and Bishop, 2007).

De entre los algoritmos evaluados, el que mejores resultados arrojó fue el *SVM*. En próximas iteraciones de este proyecto incluiremos algoritmos en el estado del arte para valorar su factibilidad.

Excepto cuando se especifique lo contrario, en los ejercicios usamos las mismas variables predictoras como insumo para la clasificación, las cuales se enlistan en el cuadro 1.

El primer conjunto de ejercicios tuvo como finalidad probar la utilidad de segmentar una palabra, combinar segmentos de palabras o hacer uso de palabras completas en las variables de texto; es decir, experimentamos con diversas combinaciones de *n*-gramas, tanto en longitud como uniones entre ellos. En primer lugar, cada uno de los registros de texto los seccionamos a nivel carácter en *n*-gramas de longitud 6 (en diversos ejercicios probamos distintas longitudes y esta fue la que mejores resultados arrojó). Después, se escogieron los 30 mil *n*-gramas más frecuentes (en diversos ejercicios

Cuadro 1

Variables (características o *features*) usadas para la clasificación

Principales (texto)	Auxiliares (categóricas)
Nombre de la ocupación	Lugar donde realizó las actividades
Tareas o funciones	Clasificación de la empresa
Actividad económica de la empresa o institución	Si tiene personal a su cargo
Nombre de la empresa	Nivel académico
	Grado académico
	Nivel académico aprobado
	Grado aprobado
	Tamaño de la empresa
	Trabaja dentro del país
	Pregunta complementaria a nombre de la empresa

probamos con distintas dimensiones y esta fue la que mejores resultados arrojó) para conformar una matriz de 30 mil columnas a través del proceso *TD-IDF*, cuyo resultado lo empleamos dentro de un algoritmo de *SVM* para clasificar automáticamente los registros del conjunto de prueba de la ENIGH (las métricas reportadas a lo largo de esta sección son: *accuracy*, precisión, *recall* y F1; sin embargo, los ejercicios fueron guiados tomando en consideración el resultado del *accuracy*, el cual representa la proporción de registros cuyo código asignado por el proceso actual del INEGI coincide con aquel código asignado por el algoritmo de *ML*). El resultado de este acercamiento inicial se muestra en el primer renglón del cuadro 2. El segundo ejercicio incorpora tanto *n*-gramas a nivel carácter de longitud 6 como de 10; para cada uno de estos conjuntos, generamos una matriz *TF-IDF* de 30 mil columnas, que al concatenarlas obtenemos una representación vectorial de 60 mil para cada registro de nuestra base de datos. El tercer y último ejercicio que reportamos en el cuadro 2 resulta de generar una matriz *TF-IDF* con *n*-gramas a nivel carácter de longitud 6 y *n*-gramas a nivel palabra de cardinalidad 2 (es decir, un *n*-grama está conformado por la agrupación de secuencia de palabras completas; en este caso, se crean pares de palabras). Posterior a la aplicación de la técnica *TF-IDF*, y antes de la implementación de los algoritmos de *ML*, realizamos un proceso de normalización con la norma L2 para las matrices resultantes *TF-IDF* (el proceso de normalización se hace sobre matrices concatenadas).

Como se muestra en el cuadro 2, este último ejercicio fue el que mejor resultado arrojó en términos de *accuracy* (0.8793).

Con el segundo conjunto de ejercicios (ver cuadro 3) evaluamos la conveniencia de que la asignación de un código a una determinada observación esté basada no solo en el resultado de un algoritmo, sino en la combinación de resultados de distintos algoritmos (ensamble). Para nuestras tareas de codificación, los mejores resultados (en términos del *accuracy*) los obtuvimos al clasificar los registros de la ENIGH a través de un ensamble con los siguientes métodos: *SVM*, regresión logística, *Random Forest*, redes neuronales, *XGBoost, K-NN*, *Naive Bayes* y árboles de decisión. También, encontramos que la mejora es significativa usando palabras completas.

Cuando obtenemos los resultados individuales de cada algoritmo de clasificación, observamos que algunos producen clasificaciones sustancialmente más certeras, por lo que sería razonable en uno de ensamble otorgar mayor peso a aquellos que en lo individual lo hacen mejor. Haciendo la analogía con un sistema de votación, vamos a dar más votos a la clasificación que nos indican aque-

Cuadro 2

Accuracy para distintas representaciones del texto. Clasificador _SVM_

	Actividad económica	Ocupación
6-gramas (carácter)	0.8782	0.8204
6-gramas (carácter), 10-gramas (carácter)	0.8781	0.8189
6-gramas (carácter), 2-gramas (palabra)	0.8793	0.8188

Cuadro 3

Accuracy para distintas representaciones del texto. Clasificador ensamble

	Actividad económica	Ocupación
6-gramas (carácter)	0.8849	0.8474
6-gramas (carácter), 10-gramas (carácter)	0.8825	0.8467
6-gramas (carácter), 2-gramas (palabra)	0.8905	0.8505

llos algoritmos que tuvieron un desempeño mejor. Dado un registro por clasificar, aquel código que tenga más votos será el que sea asignado a este. Siguiendo esta idea, encontramos una mejora en el *accuracy* si otorgamos un número de votos distinto a cada algoritmo del ensamble. El número de votos asignado a cada algoritmo fue obtenido guiándonos por el desempeño y los resultados obtenidos: cuatro a *SVM*, dos a regresión logística, uno a *Random Forest*, tres a redes neuronales, dos a *XGBoost*, uno a *K-NN*, uno a *Naive Bayes* y tres a árboles de decisión. Las mejoras son notorias en métricas como *Recall* y F1 (ver cuadro 4).

Para la variable de *ocupación* (ver cuadro 5), los pesos que mejores resultados arrojaron fueron los siguientes: cuatro a *SVM*, dos a regresión logística, dos a *Random Forest*, cuatro a redes neuronales, tres a *XGBoost*, uno a *K-NN*, uno a *Naive Bayes* y tres a árboles de decisión.

Todos los resultados anteriores fueron calculados considerando una sola fase en el proceso de clasificación (como es común); es decir, independientemente del número de clases, todos los registros se intentan clasificar en alguna de ellas y en una sola etapa. Sin embargo, una de las características de los códigos de *ocupación* y *actividad económica* es que son jerárquicos. Por ejemplo, para el caso de *actividad económica*, cada código está compuesto por cuatro dígitos: los dos primeros corresponden al sector al que pertenece el código, de tal manera que varios de estos pueden pertenecer al mismo sector. La variable de *actividad económica* está conformada por 157 clases que pertenecen a 25 sectores y *ocupación* tiene 461 agrupadas en 52 sectores. Esta estructura en los códigos puede ser aprovechada para intentar aumentar el poder predictivo del algoritmo.

Para probar esta idea, realizamos el siguiente experimento: al conjunto de entrenamiento adicionamos un grupo de variables dicotómicas dentro de las predictivas; cada una de las dicotómicas es marcada con 1 si el registro pertenece a ese sector y 0 si es lo contrario. Este nuevo grupo de variables adicionales, junto con las del cuadro 2, las usamos para entrenar y obtener un algoritmo integral de clasificación (*modelo clase*) basado en los registros del conjunto de entrenamiento. Para el conjunto de prueba, y dado que no debemos incorporar directamente el sector al que pertenece cada registro, adicionamos una etapa para clasificar y pronosticar el sector, esto lo logramos entrenando un modelo basado en registros del conjunto de entrenamiento y con las variables del cuadro 2 (*modelo sector*); de esta manera, para el conjunto de prueba obtenemos, en primer lugar, un sector pronosticado que transformamos en variable dicotómica; después,

Cuadro 4

Resultados usando n-gramas de seis caracteres y de dos palabras. Actividad económica

	Accuracy	Precisión	*Recall*	F1
Ensamble con mismos pesos	0.8905	0.6925	0.6149	0.6365
Ensamble con pesos diferenciados	0.8921	0.6767	0.6420	0.6512

Cuadro 5

Resultados usando n-gramas de seis y 10 caracteres. Ocupación

	Accuracy	Precisión	*Recall*	F1
Ensamble con mismos pesos	0.8447	0.6441	0.5384	0.5639
Ensamble con pesos diferenciados	0.8505	0.6437	0.5637	0.5831

usamos esta clasificación sectorial para ordenar las clases (códigos) basado en el *modelo clase*.

Los resultados de este ejercicio se muestran en los cuadros 6 y 7. Para comparar, en el primer renglón replicamos los resultados mostrados en el cuadro 2, donde empleamos *SVM* y matrices *TF-IDF* con *n*-gramas a niveles carácter de longitud 6 y palabra de cardinalidad 2. Usando *SVM* y con la misma arquitectura en la matriz *TF-IDF*, observamos que el ejercicio basado en dos etapas (primero sector y luego clase) para la variable de *actividad económica* mejora solo en la métrica de precisión cuando clasificamos *actividad económica* (ver cuadro 6); sin embargo, para *ocupación* parece haber una mejora importante tanto en *accuracy* como en precisión (ver cuadro 7).

2. Incorporación de algoritmos de *ML* dentro del proceso de codificación en el INEGI

Un tema central para las ONE es asegurar la correcta incorporación de nuevos procedimientos a la producción corriente. En esta sección presentamos una propuesta para adecuar el proceso actual de codificación en las encuestas en hogares que realiza el INEGI e incorporar técnicas de PLN que ayuden a aminorar la carga en la codificación asistida.

Como referencia, si solo empleáramos PLN para codificar el 100 % de los registros, obtendríamos un *accuracy* de 89.2 % para la variable de *actividad económica* y de 85 % para la de *ocupación* (recordemos que el *accuracy* se refiere a la proporción de observaciones o registros cuyo código original coincide con aquel asignado por el algoritmo de *ML*). Sin embargo, otra opción es codificar con el algoritmo de *ML* solo aquellos registros que pertenecen a clases cuyo *accuracy* por clase (*accuracy* interno) es superior a determinado umbral (por ejemplo, mayor a 95 %). Este ejercicio lo resumimos en la gráfica 1. Como se observa, si deseamos un *accuracy* de 95 %, el algoritmo lograría clasificar cerca de 65 % del total de registros en la base de datos de la ENIGH, al excluir aquellas clases cuyo *accuracy* individual no pasó el umbral.

El escenario anterior no considera la coexistencia de los procesos actuales de codificación con una metodología de PLN. En la sección 1 comentamos

Cuadro 6

Modelo de clasificación jerárquica usando *SVM*. Actividad económica

	Accuracy	Precisión	*Recall*	F1
Una sola etapa (6-gramas carácter, 2-gramas palabras)	0.8793	0.6372	0.6760	0.6511
Dos etapas (6-gramas carácter, 2-gramas palabras)	0.8774	0.6600	0.6452	0.6443

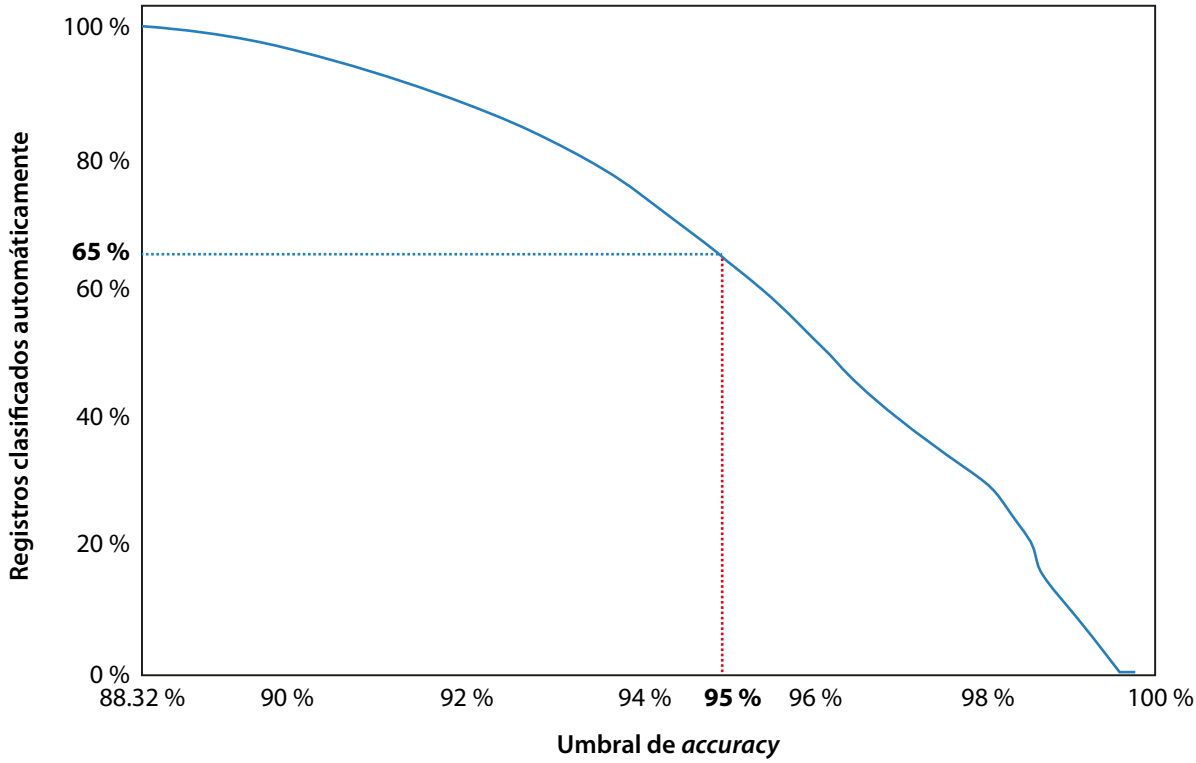
Cuadro 7

Modelo de clasificación jerárquica usando *SVM*. Ocupación

	Accuracy	Precisión	*Recall*	F1
Una sola etapa (6-gramas carácter, 2-gramas palabras)	0.8188	0.5353	0.5918	0.5531
Dos etapas (6-gramas carácter, 2-gramas palabras)	0.8312	0.5786	0.5730	0.5648

Gráfica 1

Umbral de *accuracy* por clase vs. porcentaje de registros clasificados automáticamente

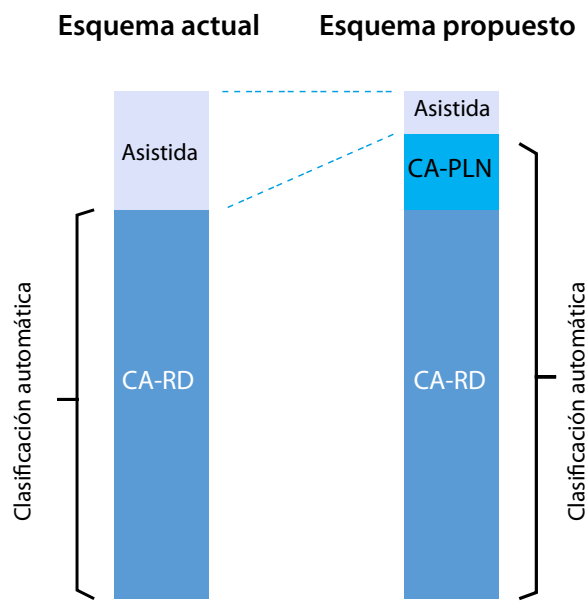


sobre el proceso actual, el cual está dividido en dos etapas: en la primera se emplea un algoritmo de clasificación automática con reglas determinísticas (CA-RD) y en la segunda se utilizan codificadores humanos (codificación asistida) para aquellos registros que no se logró captar en la etapa inicial. Nuestra propuesta consiste en la incorporación, como fase intermedia, de la metodología desarrollada en secciones previas; es decir, una vez aplicado CA-RD, usamos el algoritmo de PLN (CA-PLN) para un subconjunto de registros que no pudieron ser clasificados empleando CA-RD. Así, la clasificación automática sería la unión de CA-RD y CA-PLN. Los registros restantes se codificarían por humanos, reduciendo así, la carga de trabajo (ver figura).

Ahora bien, ¿cuáles registros que anteriormente eran clasificados por humanos se tratarían ya usando *ML*? La primera posibilidad es confiar en el algoritmo de *ML* para sustituir en su totalidad la codificación asistida; de esta forma, el 100 % de los

Figura

Propuesta operativa para el uso de PLN para las tareas de codificación



REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

registros serían clasificados automáticamente (CA-RD + CA-*ML*). El costo de esta decisión recaería en la calidad de la clasificación (coincidencia entre la clasificación asistida y la clasificación del algoritmo de *ML*, es decir, el *accuracy*): si tomamos como *ground-truth* la clasificación de la codificación asistida y esta la comparamos con el código que sería asignado usando CA-*ML*, el porcentaje de coincidencia sería solo de 68.6.

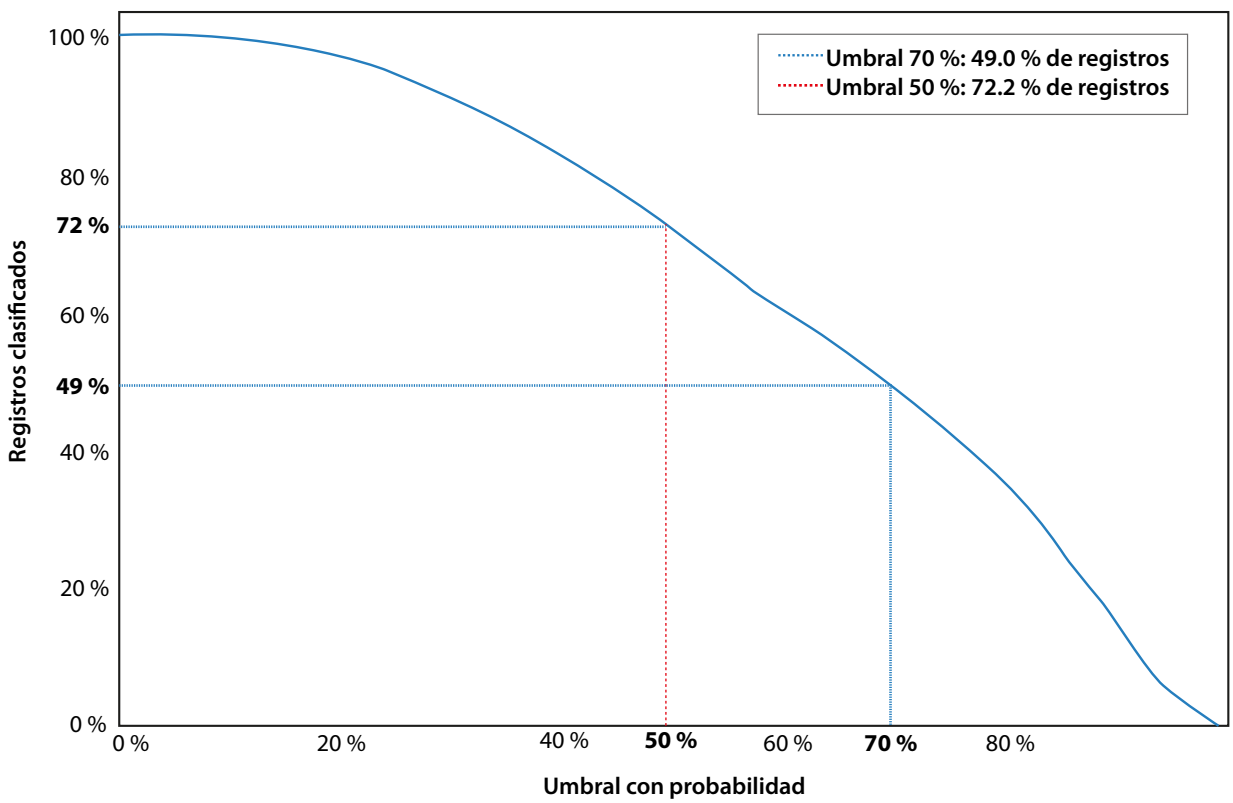
Otra posibilidad es considerar solo un subconjunto de los registros clasificados de forma asistida. Para ello, proponemos emplear una métrica de certidumbre (similar a la probabilidad) que viene asociada con los algoritmos de *ML*. En la implementación de estos se puede obtener el siguiente par de valores para cada registro: código asignado y un grado de certidumbre asociado a ese código. La idea intuitiva para su empleo recae en su correlación positiva con la coincidencia entre el algoritmo de *ML* y la parte asistida: a mayores valores de la métrica de certidumbre, mayor será la probabilidad de que coincidan. Esta métrica esta normalizada y toma valores entre 0 y 1.

El proceso que se propone para los registros que no pueden ser clasificados con reglas determinísticas es el siguiente: a) seleccionar un umbral de certidumbre (por ejemplo, 0.7), b) asignar el código pronosticado por el algoritmo de *ML* solo a los registros que tengan un valor de certidumbre igual o mayor a ese umbral y c) los registros restantes serían clasificados de forma asistida. De esta forma, tendremos un esquema como el de la figura arriba presentada.

Entre mayor es el valor del umbral elegido, mayor es el porcentaje de coincidencias entre lo que clasifican los humanos y lo que dice el algoritmo de *ML*; sin embargo, entre mayor es el umbral, menor es el número de registros que pueden ser clasificados con el algoritmo de *ML*. Por ello, el umbral deberá estar en función del nivel mínimo aceptado de coincidencias (*accuracy*). Este *trade-off* se presenta en la gráfica 2.

Gráfica 2

Trade-off entre los registros codificados por _ML_ y el umbral de certidumbre



REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Así, para el caso de actividad económica, si seleccionamos un umbral de 0.7, el porcentaje de coincidencias con respecto a la codificación asistida es de 87.7 y el de registros que pasan el umbral respecto al total de registros asistidos, de 49; en otras palabras, la carga de trabajo para los humanos se reduciría casi a la mitad, y el porcentaje de coincidencia entre el código que ellos asignarían con el que asignaría el algoritmo sería de 87.7. Si el umbral se incrementa a 0.8, el valor de coincidencias aumenta a 91.7 %, pero la carga de trabajo para los humanos se reduce a 34.7 por ciento.

Otro posible uso de los algoritmos de *ML* dentro de los procesos actuales de codificación es para identificar potenciales errores humanos. Para ello, se aprovecharía la distribución de probabilidades que un algoritmo de *ML* genera para alertar de posibles discrepancias entre lo que pronostica un algoritmo con alta probabilidad y la etiqueta que otorga un codificador humano. Este proceso puede ser especialmente de utilidad en proyectos donde se contratan de forma eventual codificadores humanos y que, por lo general, tienen poca experiencia, por ejemplo, en los censos de población y vivienda.

3. Conclusiones

Las oficinas nacionales de estadística llevan a cabo tareas diarias cuyas características las hace propensas al aprovechamiento de los avances metodológicos y tecnológicos, tal es el caso de los procesos de codificación, donde, a partir de un conjunto de covariables auxiliares, se determina uno de los códigos que le será asignado a un registro en una base de datos. Estas covariables pueden ser de texto, categóricas o numéricas.

El objetivo de este documento es evaluar la factibilidad de incorporación de técnicas de procesamiento de lenguaje natural dentro de los actuales procesos productivos que se siguen para la codificación de variables de encuestas en hogares realizadas por el INEGI. Una de ellas es la ENIGH, y entre las variables que se codifican para esta se encuentran las de *ocupación* y *actividad económica*. Actualmente, estas variables requieren de una cantidad considerable de recursos humanos debido a la dificultad de automatización con técnicas convencionales.

La propuesta elaborada en este trabajo permitiría reducir la carga a los codificadores humanos al incorporar PLN para automatizar la asignación de códigos para las variables de *ocupación* y *actividad económica*. De acuerdo con las pruebas que llevamos a cabo, se podrían codificar automáticamente 50 % de los registros que en la actualidad se hacen de forma manual y obtener porcentajes de coincidencia (*accuracy*) cercanos a 90 por ciento.

Estos resultados, a su vez, constituyen un punto de referencia para potenciales mejoras en los algoritmos evaluados, toda vez que los avances en esta área están en constante desarrollo y evolución.

Fuentes

Breiman, L. "Random Forest", en: *Machine Learning*. 45, 2001, pp. 5-32.

Chih-Chung, C. y L. Chih-Jen. *LIBSVM: A Library for Support Vector Machine*s. 2019.

Hastie, T. *et al*. *The Elements of Statistical Learning*. Springer, 2009.

Hsiang-Fu, Y., H. Fang-Lan y L Chih-Jen. "Dual Coordinate Descent Methods for Logistic Regression and Maximum Entropy Models", en: *Machine Learning*. 85, 2011, pp. 41-75 (DE) https://doi.org/10.1007/s10994-010-5221-8, consultado el 1/08/2020.

Jurafsky, D. y J. H. Martin. *Speech and Language Processing*: An *Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Second Edition. Pearson, 2008.

Measure, A. *Automated Coding of Worker Injury Narratives*. U.S. Bureau of Labor Statistics, 2014.

Platt, J. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. Advances in Large Margin Classifiers, 1999.

Svensén, M. y C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2007.

Zhang, H. *The Optimality of Naive Bayes*. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004. 2, 2004.

Anexo

Para cada proyecto de codificación de encuestas o censos de población y vivienda, se tiene la necesidad de diseñar un conjunto de estrategias y materiales de codificación de acuerdo con las variables principales a codificar, así como sus preguntas de apoyo; para cada uno, se tiene que diseñar un *traje a la medida* para el proceso automático (ver esquema). Para codificar las variables de *ocupación* y *actividad económica* en la ENIGH, bajo el proceso tradicional (algoritmos determinísticos), se utilizan las variables que se muestran en el cuadro 8.

La información se codifica al máximo nivel de desagregación de las clasificaciones utilizadas; para ocupación, se empleó el Sistema Nacional de Clasificación de Ocupación (SINCO) 2011, que tiene 468 grupos unitarios (claves) diferentes y para actividad económica, el Sistema de Clasificación Industrial de América del Norte (SCIAN) 2008 versión hogares, que cuenta con 176 distintos subsectores (ver cuadro 9).

Cuadro 8

Variable	Tipo de variable
Nombre de la ocupación	Texto
Tareas o funciones	Texto
Actividad económica de la empresa o institución	Texto
Nombre de la empresa	Texto
Lugar donde realizó las actividades	Código alfanumérico
Clasificación de la empresa	Código alfanumérico
Trabajo dentro del país	Código alfanumérico
Tamaño de la empresa	Código alfanumérico
Nivel de instrucción	Código alfanumérico
Grados aprobados	Código alfanumérico

Cuadro 9 Continúa

Fragmento de una base de datos codificada

var_texto_1	var_texto_2	var_texto_3	var_auxiliar_1	var_auxiliar_2	var_auxiliar_5	Código asignado
auxiliar de almacén	capturar información atender proveedores.	elaboración de carne y productos cárnicos	1	3	1	3132
ayudante de panadería	atender a clientes, acomodar, hacer puestos de pan, preparar el pan relleno, freir donas.	a la elaboración y venta de pan en un local comercial.	1	12	1	4211

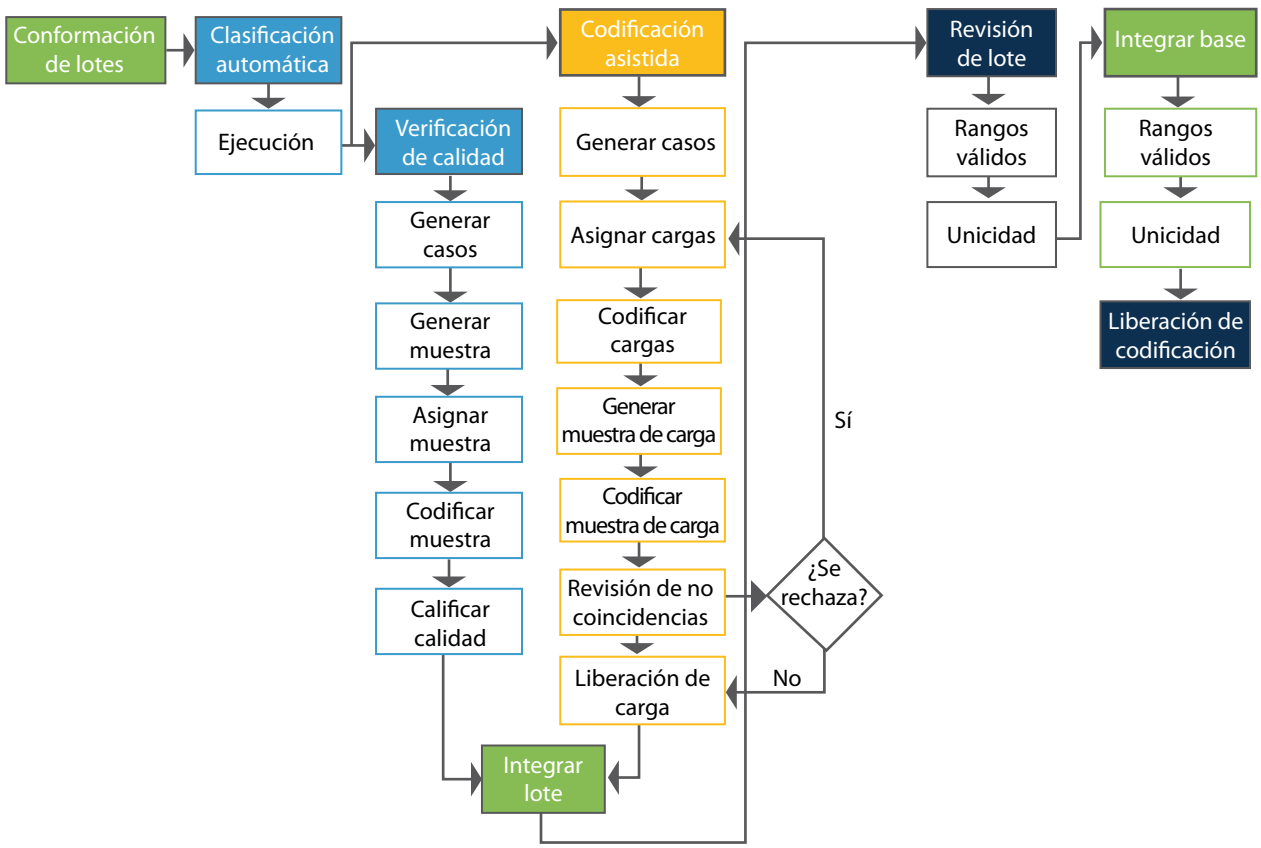
Cuadro 9 Concluye

Fragmento de una base de datos codificada

var_texto_1	var_texto_2	var_texto_3	var_auxiliar_1	var_auxiliar_2	var_auxiliar_5	Código asignado
ayudante de albañil	preparar mezcla, pasar y pegar blocs, limpiar la herramienta y preparar cimbra.		2	11	1	9221

Esquema

Proceso de codificación actual para encuestas en hogares del INEGI



Imputation of Non-Response in Height and Weight in the **"Mexican Health and Aging Study"**

Imputación de no-respuesta en peso y talla **en el Estudio Nacional de Salud y Envejecimiento**

Matthew Miller,* Alejandra Michaels-Obregón, Karina Orozco Rocha,*** Rebeca Wong******

The way missing data in population surveys are treated can influence research results. Therefore, the aim of this paper is to explain the reasons and procedure for imputing anthropometric data such as height and weight self-reported by individuals in the first four waves of the Mexican Health & Aging Study (MHAS). We highlight the effect of the imputation versus the exclusion of the cases with missing data, by comparing the distribution of these values and their associated effects on the Body Mass Index using a regression model. We conclude that the incorporation of imputed data offers more solid results as opposed to eliminating the cases with missing data. Hence the importance of applying these statistical procedures, with appropriate treatment of the data, making the methodology and the imputed data available to the users by the same source of information, as offered in the MHAS.

Key words: MHAS; imputation; height; weight; BMI.

El manejo de los datos faltantes en entrevistas por encuestas puede influenciar los resultados de una investigación. Por ello, el objetivo de este trabajo es explicar las razones y el procedimiento de imputación de datos antropométricos, como la altura y el peso, autorreportados en las primeras cuatro rondas del Estudio Nacional sobre Salud y Envejecimiento en México (ENASEM). Destacamos el efecto de la imputación *versus* la eliminación de los casos con datos faltantes, comparando la distribución de dichos valores y sus efectos asociados con el Índice de Masa Corporal mediante un modelo de regresión. Se concluye que la incorporación de datos imputados ofrece resultados más sólidos en comparación con la eliminación de los casos con datos faltantes. De ahí la importancia de aplicar estos procedimientos estadísticos con un manejo adecuado de los datos y difundir la metodología aplicada para obtener los datos imputados desde la misma fuente de información, tal como se ofrece en el ENASEM.

Palabras claves: ENASEM; imputación; altura; peso; IMC.

Recibido: 28 de junio de 2021.
Aceptado: 5 de noviembre de 2021.

* mrmiller@utmb.edu
** almichae@utmb.edu
*** korozco9@ucol.mx
**** rewong@utmb.edu

Body Mass Index Abstract / Hailshadow / iStock

Introduction

Missing data are a common problem in statistical information collected through population surveys, and an inadequate treatment in the processing and analysis of the information can generate biases and inaccuracies in the results obtained (Abellana & Farran, 2015; Kontopantelis et al., 2017). Missing data in the Mexican Health and Aging Study (MHAS) are no exception, since they are present in a variety of variables including social, economic, and health dimensions. The source of missing data tends to be that the respondent has no knowledge or refuses to disclose the information to the interviewer. In the variables on income and assets, the fraction of missing data is around 10% (Wong et al., 2017a), while in anthropometric variables, such as self-reported height and weight, it is close to 20% (Montevarde & Novak, 2008). In MHAS, the advantage in the economic variables is that the study includes bracket questions as follow-up after a non-response, in order to recover some of the missing data. However, the self-report of anthropometric variables such as height and weight do not use this strategy.

Regarding these two types of variables, there has been more documentation on the mechanisms or techniques to impute missing data in economic variables, such as earned-income variables in the National Survey of Occupation and Employment, ENOE (Durán, 2019), household-income variables in the National Survey of Household Income and Expenditure, ENIGH (Vargas & Valdés, 2018) or economic indicators in National Economic Surveys, EEN (Corona, et al. 2019). These data are collected by the Mexican National Institute of Statistics and Geography (Instituto Nacional de Estadística y Geografía, INEGI). We know less about the mechanisms to impute missing data in the anthropometric var-

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

iables, hence the importance of documenting the procedure performed for the MHAS.

The anthropometric variables of weight and height are used to calculate quite an important indicator for health and aging research: body mass index (BMI), providing an assessment for level of underweight, normal weight, overweight or the obesity of a person. This indicator is critical and used by multiple studies related to a variety of health dimensions of older adults. Palloni et al. (2015) research the effects of overweight and obesity on the incidence of type 2 diabetes and older adult mortality; or research such as Kumar et al. (2015) that analyze longitudinally the effects of BMI on disability and mortality over an 11-year follow up among Mexicans aged 50 years and older who are non-disabled at baseline in 2001. Now we know that obesity is also a risk factor for severe Covid-19 infection (Satter et al., 2020; González et al., 2021). Indeed, it is estimated that the prevalence of obesity has been rising over the last decade, with 45% of adults 50 years of age and older being overweight and 23% obese in Mexico in 2015 (Rodriguez & Wong, 2019).

This paper aims to provide the rationale and explain the procedure of imputation of the missing data in height and weight self-reported by the individuals in the MHAS. To highlight the effect of imputation versus deletion of observations with missing data, we compared the distributions of these variables among three groups: cases where the data were observed (non-imputed cases), cases where the data were imputed (imputed cases), and all cases (non-imputed plus imputed). Finally, we constructed a database containing the means and standard deviations of height, weight, and BMI of each individual in each wave, along with dummy variables indicating whether height and weight were imputed. These variables are shared with users in an MHAS data file along with the proper documentation.

This work has five sections. First, we present conceptual aspects about missing data and imputation. In the second section, we describe the anthropometric data for weight and height in the Mexican Health and Aging Study for the four waves. Next, we present how we prepared the data for imputation, the procedure for imputation, and the creation of final datasets for end-users. In the fourth section, we present results highlighting the differences between imputed and non-imputed weights and heights, and their effect on the calculated BMI. Finally, we present the conclusions about the importance of imputation in anthropometric data.

1. Conceptual aspects of imputation

There is a variety of ways to handle missing data, such as case deletion or imputation. The selection of the proper mechanism depends on how the missing data are considered: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Kontopantelis et al., 2017).

There are two types of case deletion. The first excludes from the analysis all the cases with missing data in the variables of interest (listwise)—that is, working only with the cases with complete information for all variables. This implies a reduction of the analytical sample size and, depending on the proportion of missing data, the statistical power of hypothesis tests and standard errors may be affected. This method assumes that missing data are MCAR, meaning that the likelihood that data are missing is totally independent of all observed or missing data. The other alternative is pairwise deletion (or available case analysis), which eliminates those cases with missing data in a specific variable in each analysis. But they are included in other analyses using variables with complete information. This means working with different sample sizes in different parts of the analysis. Like the previous method, this one assumes that missing data are MCAR (Abellana & Farran, 2015).

There are different alternatives for the imputation procedure, such as simple or multiple im-

putation. Imputation seeks to replace missing data with plausible values of each incomplete variable. Plausible values are simulated by estimating relationships between imputed variables and those with no missing values. Imputation adds a layer of uncertainty to results derived from imputed data, as it is not definitively known that the missing values would equal the imputed values if they were observed. Therefore, We recommend creating multiple sets of imputed data using a process that involves a degree of randomness. Such a procedure is called "multiple imputation," and we use it here to compute plausible heights and weights for subjects in each wave of the MHAS for whom such data are missing.

Multiple imputation typically assumes that missing data are MAR, meaning that the likelihood that data are missing is independent of the missing values themselves, given the observed values. Although it is difficult to tell whether our data are MAR from the observed data alone, we believe that assuming as much is reasonable, considering how many variables contributed to our imputations (Rässler, Rubin, & Zell, 2012; van Buuren, Boshuizen, & Knook, 1999).

We identified other associated variables, which can contribute to imputation of height and weight in our data, and we needed to impute any missing values in those other variables too. Height, weight, and other variables that contributed to imputation of height and weight had a non-monotonic pattern of missingness, so we employed the multivariate imputation using chained equations (MICE) or a fully conditional specification (FCS) algorithm because it provides the flexibility that we need (van Buuren, 2007).

2. Data

The Mexican Health and Aging Study (MHAS) is longitudinal and representative of adults aged 50 and over living in rural and urban areas of Mexico. The study is also known by its name in Spanish (Estudio Nacional de Salud y Envejecimiento en México, ENASEM**)**. The goal is to study aging with a broad health, economic and sociodemographic perspective. Furthermore, this study is highly comparable to the U. S. Health and Retirement Study (HRS). The baseline sample was surveyed in 2001. It included households with at least one resident aged 50 years or older (born no later than 1951) and his/her spouse or partner, regardless of age (Wong et al, 2017b). The follow-up surveys were successfully fielded in 2003, 2012, 2015, and 2018. In 2012 and 2018, the MHAS cohort was supplemented with representative samples of adults born between 1952 and 1961 and of those born between 1962 and 1967, respectively.

For this research, the first four waves are used. The MHAS questionnaire is made up of various sections such as: demographic, non-resident children, health, health care services, cognition, help and children, employment, housing, pension, income, and assets. Within the health section, various aspects of self-reporting are asked, such as the diagnosis of chronic diseases as well as weight and height. The latter information is captured with the following questions: "How much do you weigh now?", the answers to which are coded in kilos; and "How tall are you without shoes?", the answers to which are coded in meters and centimeters.

3. Methods

a) Preparing data for imputation

The variables that we seek to impute are self-reported height and weight. The first step is to prepare the data so that the values to be imputed in each variable are identified.

In the raw dataset, numeric variables contain values that although they appear as real numbers are intended to denote observations where those variables were unobserved for some known reason (usually "refused to answer" or "don't know"). These are values such as 888, or 999 in a 3-field variable. Stata, the software used to perform all imputations and analyses described in this document, regards such values as observed and valid, so these values

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

need to be replaced with explicitly missing values. The MHAS codebooks for each wave list such values for each variable (MHAS, 2001–2015). Stata has 27 different missing values: ".," ".*a*," ".*b*," …, and ".*z*." Because only the soft missing value "." can be imputed in STATA, we assign a soft missing value (.) to the values in every variable that will be imputed.

MHAS selected a subsample in each wave to obtain objective anthropometric measurements, including height and weight, which contributed to the imputation of self-reported heights and weights for those observations selected for the subsample in each wave. Some recorded values of self-reported heights and weights differed so greatly from measured values that the accuracy of the recorded self-reported value is suspect. Therefore, for the imputation exercise, self-reported heights and weights that differed from observed measured values in the same survey participant by more than 10% of the measured value were replaced with soft missing values. Table 1 shows the numbers of self-reported heights and weights in each wave that are missing for this reason.

Furthermore, if a height reported in 2003, 2012, or 2015 differed from the height reported by the same respondent in at least one prior wave by more than 10% of the height in the prior wave, the height in the later wave was also assumed to be inaccurate and replaced with a soft missing value ($N_{2003} = 313, N_{2012} = 477, N_{2015} = 696$). This is because heights in the target population of the MHAS should not change significantly over time.

The process of preparing or "cleaning" the data for imputation in this way is outlined in Figure 1, and the proportions of observations in each wave with missing height and missing weight after the data were cleaned are shown in Figure 2.

b) Imputation Procedure

As previously mentioned, height, weight, and other variables that contributed to the imputation thereof were imputed with the MICE technique. MICE involves random draws from posterior predictive distributions. Thus, for the sake of reproducibility, the seed for pseudorandom-number generation was set to 101 each time that the command "*mi impute chained*" was called in Stata. The covariates for imputation of self-reported height and weight included sex, age, locality size, and years of education. MICE requires that any variable X involved in imputation of another variable Y also be imputed if X has missing values. Table 2 shows the numbers of observations in which each of those variables was imputed.

In addition to these covariates, measured heights and weights contributed to imputation of self-reported heights and weights within the subsam-

Table 1

Numbers of Self-Reported heights and Weights that Differed from Measured Values by more than 10%

Wave	2001	2003	2012	2015
Total MHAS sample size	15,186	13,704	15,723	14,779
Anthropometric subsample size	2,944	2,641	2,086	2,054
Cases with different height	43	53	67	70
Cases with different weight	317	263	252	270

Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

Figure 1

Flowchart of Process of Preparing Data for Imputation in Each Wave

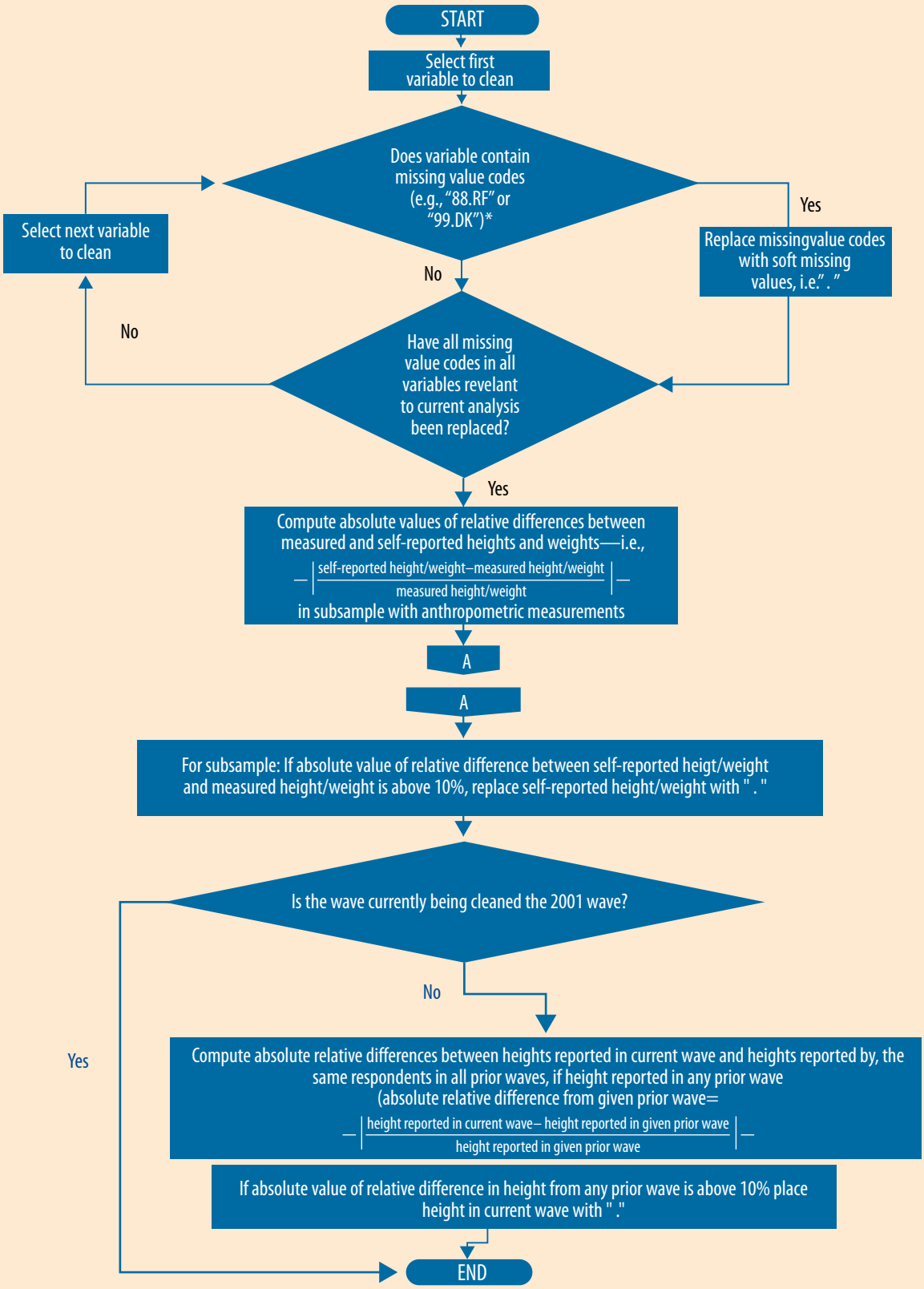
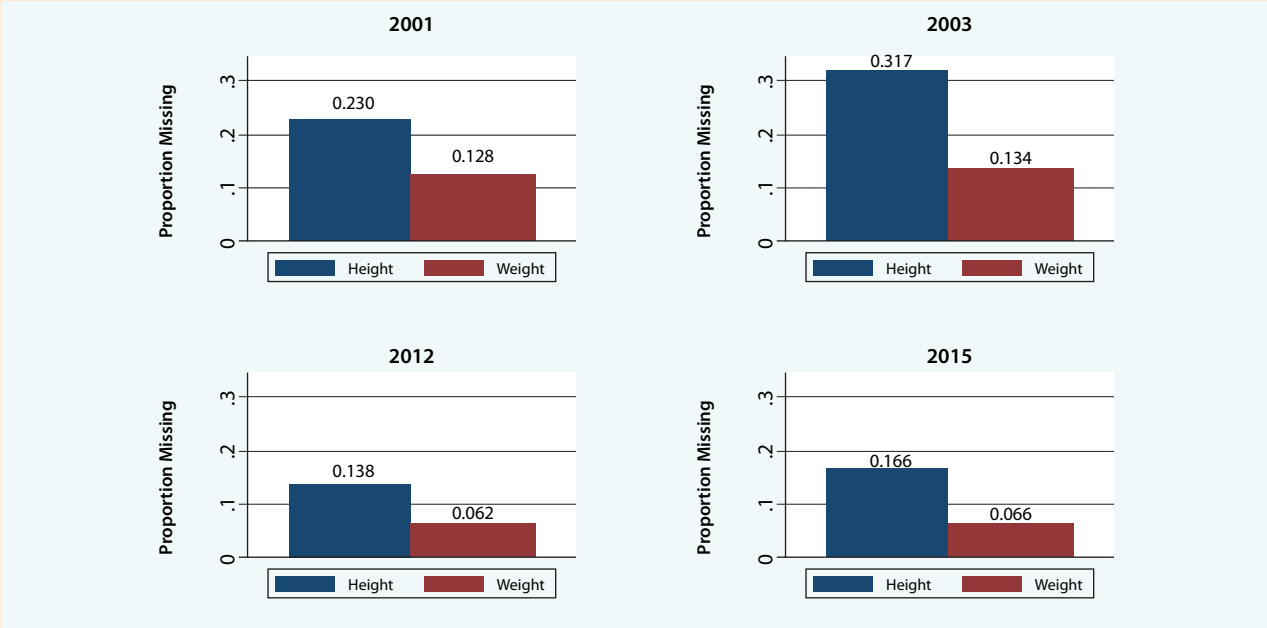


Figure 2

Proportion of Individuals Missing Self-Reported Height and Self-Reported Weight in each Wave



Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

ples selected for anthropometric measurements. We substituted zeroes for measured heights and weights outside of the subsample to allow Stata to perform the MICE algorithm. MICE sequentially performs a univariate imputation on each variable with missing values, in our case predictive mean matching (PMM) for all such variables. More detailed justifications for these choices can be found in the document "Imputation of Height and Weight in the Mexican Health and Aging Study" found on the MHAS webpage.

The following table shows which variables were imputed in each wave and the univariate imputation method used to impute them (Table 3).

All variables with missing values were imputed using the univariate method predictive mean matching (PMM)—called "regression switching" by van Buuren, Boshuizen, and Knook (1999). For each observation with a missing value of the imputed variable, the PMM algorithm finds a predetermined number of observations that are "closest" to the observation with a missing value, according to a certain measure of distance, among all observations with non-missing values of the imputed variable. One of those observations is selected at random, and the observed value from the selected observation is assigned for the missing value. In each wave, each imputed value was selected from one of the five closest observations with non-missing values.

In 2001, missing values of education, self-reported height and weight, and measured height and weight were imputed; sex, age, and locality size had no missing values in 2001. The length of the burn-in period —the number of times PMM was performed before settling on an imputed value— in 2001 was set at 450 iterations.

In 2003, missing values of self-reported height and weight, age, education, and measured height and weight were imputed, and the length of the burn-in period was set at 350 iterations.

In 2012, missing values of age, education, and self-reported height and weight were imputed for the entire sample, and the averages of two meas-

Table 2

Number of missing values among covariates used in the imputations of height and weight

Wave (total sample size)	2001 (15,186)	2003 (13,704)	2012 (15,723)	2015 (14,779)
Sex	0	0	0	0
Age	0	11	25	4
Locality size	0	0	0	0
Education	19	90	68	175

Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

urements each of height and weight were imputed for the subsample selected for anthropometric measurements. The length of the burn-in period was set at 300 iterations for this wave.

In 2015, missing values of age, education, and self-reported height and weight were imputed for the entire sample, and averages of two measurements each of height and weight were imputed for the subsample selected for anthropometric measurements. The length of the burn-in period was set at 300 iterations for this wave.

c) Post-Imputation

After imputation of height and weight, BMI was generated as a "passive variable," a function of one or more imputed variables, in each wave. To examine how imputing missing values can affect results versus entirely excluding observations with missing values from analysis, three linear regression models of the natural logarithm of BMI were estimated in each wave using both imputed and non-imputed data. Each model had one independent variable at a time: diabetic status, years of education, or locality size; and similar

Table 3

Variables imputed and imputation method in each MHAS wave

Variable	Wave			
	2001	2003	2012	2015
Self-reported height	PMM			
Self-reported weight	PMM			
Self-reported body mass index (BMI)	Calculated from self-reported height and weight after imputation			
Measured height (subsample only) [1/]	PMM			
Measured weight (subsample only) [1/]	PMM			
Years of education	PMM			
Age	Complete	PMM		
Locality size[2/]	Complete			
Gender	Complete			

Notes:
Complete indicate that the variable had no missing values and, thus, was not imputed in that wave.
[1/] Average of two measurements each— for waves 2012 and 2015.
[2/] 2003 data from 2001.w
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

models were constructed using only non-imputed data. The models that included imputed data were pooled across 10 imputations, and the standard errors of estimated coefficients were adjusted to account for the added variability introduced by such pooling.

Finally, for each wave we calculated the means and standard deviations of height, weight, and BMI across 10 imputations for each subject. These are the imputed variables that are provided in the MHAS website (http://www.mhasweb.org/). In cases where such values are observed, the imputed values are the same as the observed values. For each case in each wave, two separate dummy variables are included which indicates if the values for height and for weight were imputed. The goal is to provide as much information as possible to the MHAS data user, who can decide whether or not to use the imputed variables.

4. Results

Tables 4 and 5 show great similarity between the distributions of self-reported height and weight

Table 4

Continue

Self-Reported Heights (cm) and Weights (kg) by Percentile in Non-Imputed Cases and by Imputation Status, 2001–2015

	Self-Reported Heights			Self-Reported Weights		
2001	Non-Imputed	Imputed	All*	Non-Imputed	Imputed*	All*
Percentile	*n*= 11,677	3,509	15,186	13,225	1,961	15,186
1st	135.0	144.7	139.0	40.0	46.7	41.1
5th	146.0	148.5	147.0	50.0	51.9	50.0
10th	150.0	150.1	150.0	53.0	55.0	53.0
25th	155.0	153.2	154.0	60.0	60.5	60.0
50th	160.0	156.4	160.0	70.0	66.3	68.8
75th	168.0	161.7	166.0	78.0	72.1	78.0
90th	173.0	166.1	172.0	87.0	78.0	86.0
95th	177.0	168.2	175.0	94.0	82.4	93.0
99th	185.0	171.2	183.0	109.0	94.1	108.0
2003	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	*n*= 9,278	4,426	13,704	11,765	1,939	13,704
1st	140.0	145.1	140.0	40.0	51.0	40.0
5th	146.0	148.1	147.0	49.0	55.6	50.0
10th	150.0	150.2	150.0	53.0	57.7	54.0
25th	155.0	153.0	153.7	60.0	62.1	60.0
50th	160.0	156.4	160.0	69.0	67.1	68.2
75th	168.0	161.5	165.0	78.0	73.3	78.0

Table 4 Concludes

Self-Reported Heights (cm) and Weights (kg) by Percentile in Non-Imputed Cases and by Imputation Status, 2001–2015

Self-Reported Heights				Self-Reported Weights		
2003	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	*n*= 9,278	4,426	13,704	11,765	1,939	13,704
90th	173.0	166.1	170.0	87.0	80.5	86.0
95th	176.0	168.1	175.0	94.0	85.7	93.0
99th	183.0	172.0	182.0	108.0	92.9	106.0
2012	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	13,622	2,101	15,723	14 ,746	977	15,723
1st	140.0	144.7	140.0	42.0	44.6	42.0
5th	145.0	147.9	145.8	50.0	49.7	50.0
10th	149.0	149.7	149.4	53.0	53.9	53.0
25th	153.0	152.4	153.0	60.0	60.3	60.0
50th	160.0	155.5	160.0	69.0	66.3	69.0
75th	167.0	160.5	165.0	78.0	72.5	78.0
90th	172.0	165.4	171.0	88.0	79.0	87.0
95th	175.0	167.6	175.0	95.0	84.7	95.0
99th	182.0	171.3	180.0	108.0	103.5	108.0
2015	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	12,386	2,393	14,779	13,807	972	14 ,779
1st	140.0	145.0	140.0	41.0	43.7	41.0
5th	145.0	147.7	146.0	49.0	49.1	49.0
10th	149.0	149.2	149.0	53.0	51.4	53.0
25th	153.0	151.8	152.4	60.0	58.8	60.0
50th	160.0	155.0	159.0	69.0	65.1	68.0
75th	166.0	160.4	165.0	78.0	71.3	78.0
90th	172.0	165.6	170.0	88.0	78.5	87.0
95th	175.0	168.0	175.0	95.0	83.3	95.0
99th	182.0	172.5	180.0	109.0	95.6	109.0

* Imputed values in each observation averaged across 10 imputations.
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

among the non-imputed cases and among all cases (combining imputed and non-imputed). For additional analysis, we include box plots of BMI in 2012 that control for locality size and diabetic status (Figures 3 and 4, respectively), showing similar results. The results showing similar distributions between all cases and non-imputed cases are expected, as imputation of missing values should not distort the distribution of the data used to perform imputations.

Histograms of self-reported height and weight in 2012 among imputed cases showed more centralized distributions than histograms among non-imputed cases (see Figure 5). The values for imputed cases were averaged across 10 imputations; this

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Table 5

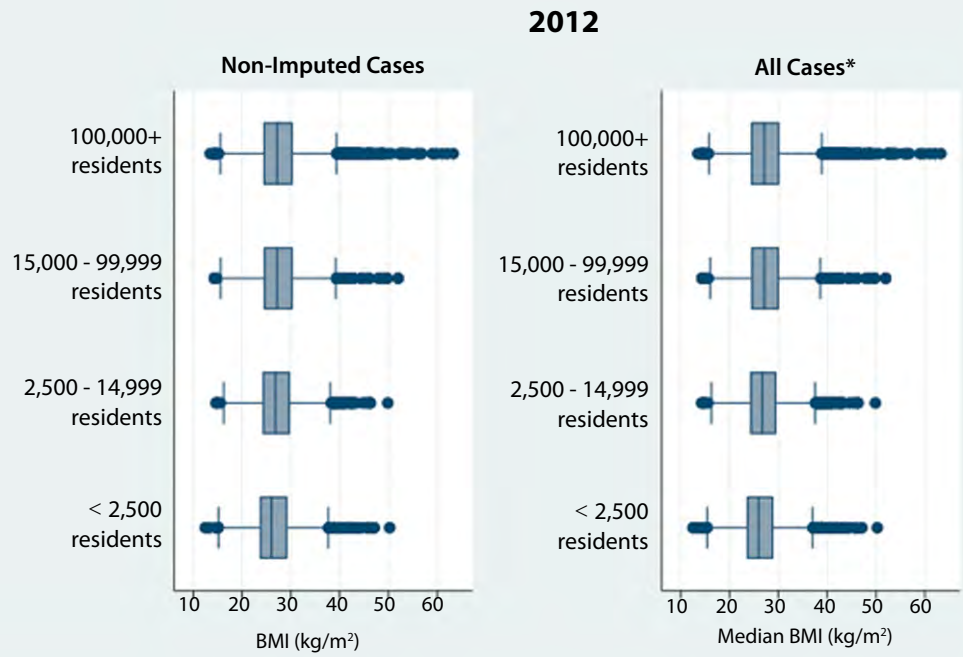
Summary Statistics of Self-Reported Heights and Weights by Imputation Status, 2001–2015

	Number of Missing Values (% of Sample)	Non-Imputed Cases		Imputed Cases*		All Cases*	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
2001 (*N* = 15,186)							
Self-Reported Height (cm)	3,493 (23.0)	160.94	9.99	157.20	6.07	160.08	9.37
Self-Reported Weight (kg)	1,937 (12.8)	69.93	13.98	66.52	9.62	69.49	13.55
2003 (*N* = 13,704)							
Self-Reported Height (cm)	4,347 (31.7)	161.05	9.62	157.33	6.18	159.87	8.85
Self-Reported Weight (kg)	1,833 (13.4)	69.80	14.11	66.49	9.98	69.36	13.68
2012 (*N* = 15,723)							
Self-Reported Height (cm)	2,170 (13.8)	159.99	9.43	156.63	6.06	159.53	9.12
Self-Reported Weight (kg)	977 (6.2)	70.00	13.99	66.81	10.90	69.80	13.84
2015 (*N* = 14,779)							
Self-Reported Height (cm)	2,452 (16.6)	159.85	9.33	156.39	6.34	159.28	9.00
Self-Reported Weight (kg)	971 (6.6)	69.77	14.19	65.72	10.79	69.50	14.03

* Imputed values in each observation averaged across 10 imputations.
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

Figure 3

Box Plots of BMI by Locality Size Among Non-Imputed Cases and Among All Cases (2012)

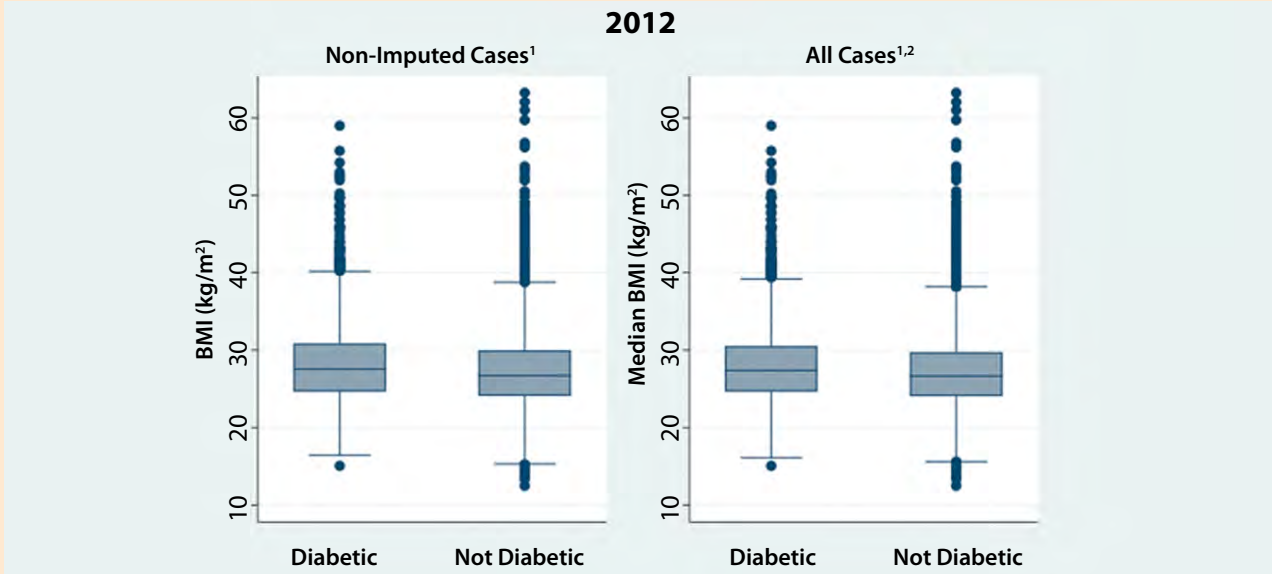


* BMI´s for imputed cases are medians among 10 imputations.

Source: Own calculation using data from the Mexican Health and Aging Study 2012.

Figure 4

Box Plots of BMI by Diabetic Status Among Non-Imputed Cases and Among All Cases (2012)

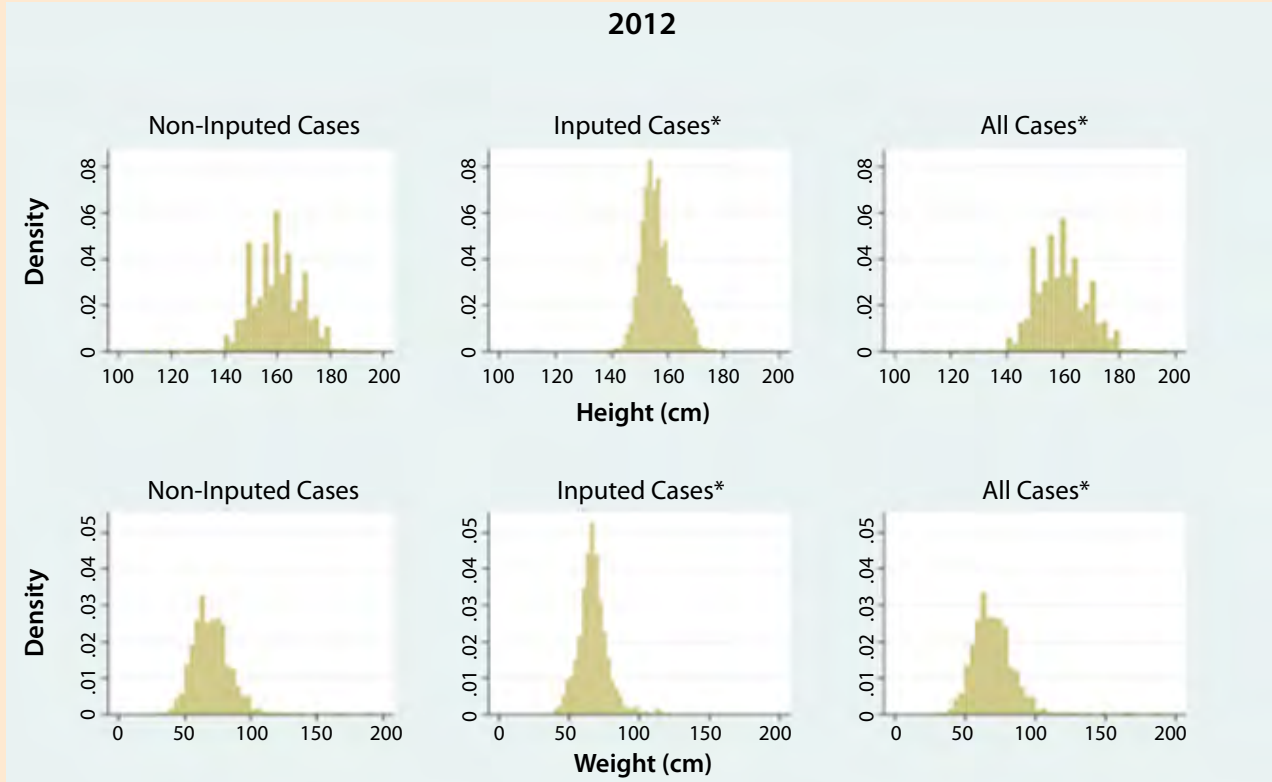


[1] Excludes observations where diabetic status is missing.
[2] BMI´s for imputed cases are medians among 10 imputations.

Source: Own calculation using data from the Mexican Health and Aging Study 2012.

Figure 5

Histograms of Self-Reported Heights and Weights (2012 Wave)



* Imputed values in each observation are averaged across 10 imputations.

could explain the differences between the two distributions because sample means are less variable than the data from which they are computed.

Table 6 presents the regression coefficients of the aforementioned models of log-BMI along with *p*-values, and shows how outright excluding observations with missing data can bias results. For example, in 2001 and 2003, using all (imputed and non-imputed) data showed a statistically significant positive association between log-BMI and education. On the other hand, in the models that excluded observations in which either BMI or education was missing, those associations were estimated to be smaller in magnitude and not statistically significant. Also, in every wave the models with education and locality size as independent variables had smaller coefficients when missing data were excluded than when their imputed values were included. Although the differences varied in magnitude, the fact that such differences were consistently evident across waves implies that the impact of deleting observations with missing data on analysis of these data may be meaningful.

Table 6 Continue

Regression Parameter Estimates of Log-Transformed BMI on Years of Education, Locality Size[1], or Diabetic Status

	Non-Imputed Cases Only			All Cases[2]		
	N[3]	β	*p*-value	*N*[3]	β	*p*-value
2001						
Years of Education	11,107	3.269×10^{-4}	NS	15,186	1.3979×10^{-3}	***
Locality Size (2,500–14,999)		0.026	**	15,186	0.038	***
Locality Size (15,000–99,999)	11,117	0.034	***		0.043	***
Locality Size (100,000+)		0.044	***		0.052	***
Diabetes (Yes)	10 ,830	0.016	***	14,721	0.016	**
2003						
Years of Education	8,875	6.377×10^{-4}	NS	13,704	2.8981×10^{-3}	***
Locality Size in 2001 (2,500–14,999)		0.031	***		0.041	***
Locality Size in 2001 (15,000–99,999)	8,929	0.040	***	13,704	0.052	***
Locality Size in 2001 (100,000+)		0.045	***		0.062	***
Diabetes (Yes)	8,914	0.024	***	13,650	0.025	***
2012						
Years of Education	13, 042	2.07×10^{-3}	***	15,723	2.5986×10^{-3}	***
Locality Size (2,500–14,999)		0.025	***		0.028	***
Locality Size (15,000–99,999)	13,104	0.037	***	15,723	0.039	***
Locality Size (100,000+)		0.037	***		0.041	***
Diabetes (Yes)	13,081	0.034	***	15,689	0.031	***
2015						
Years of Education	11,746	2.4337×10^{-3}	***	14,779	3.5923×10^{-3}	***
Locality Size (2,500–14,999)		0.025	***		0.033	***

Table 6 Concludes

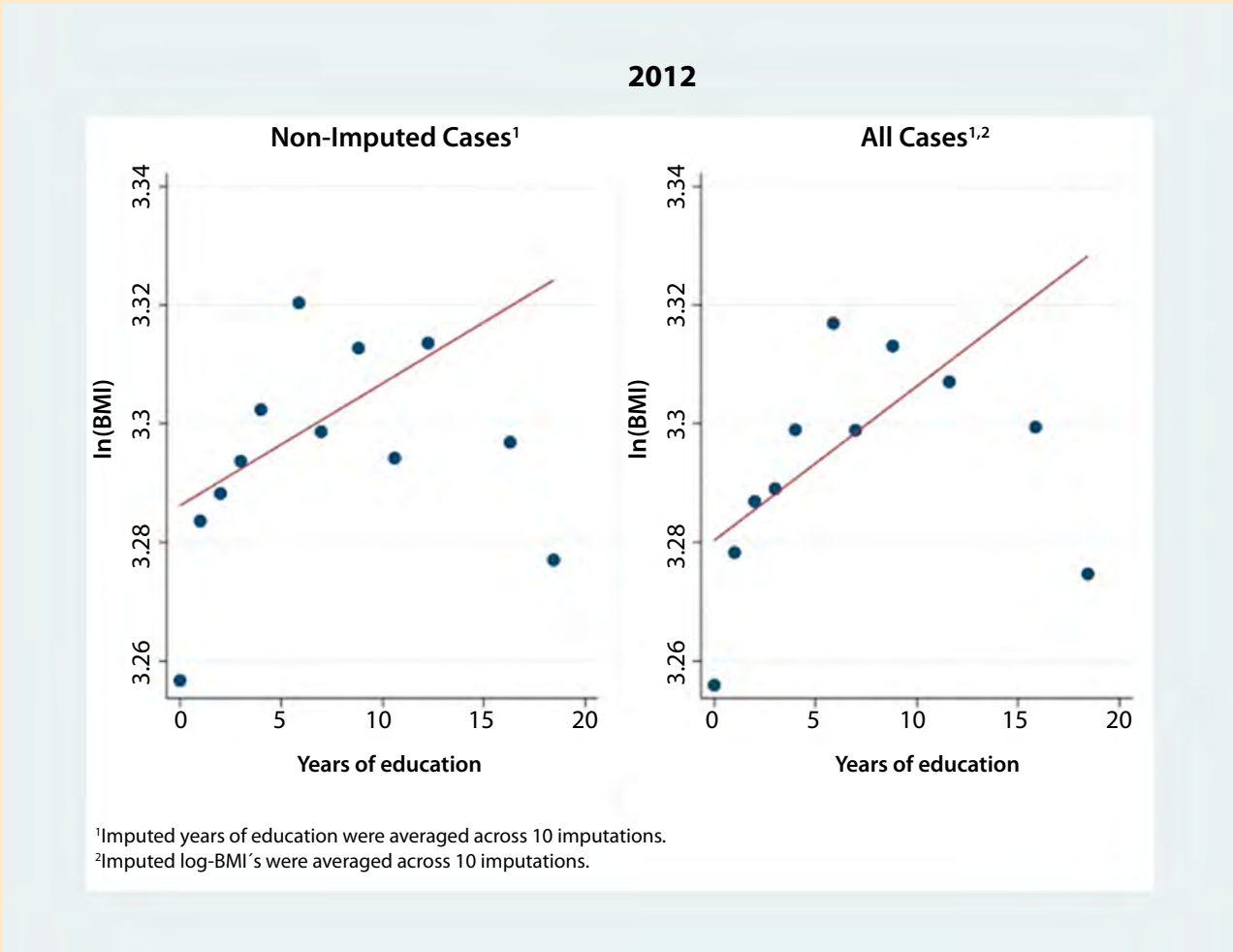
Regression Parameter Estimates of Log-Transformed BMI on Years of Education, Locality Size[1], or Diabetic Status

	Non-Imputed Cases Only			All Cases[2]		
	N[3]	*β*	*p*-value	*N*[3]	*β*	*p*-value
Locality Size (15,000–99,999)	11,909	0.023	***	14,779	0.028	***
Locality Size (100,000+)		0.032	***		0.043	***
Diabetes (Yes)	11,898	0.034	***	14,759	0.032	***

Notes: [1] Versuslocalitysize<2,500.
[2] Models pooled across 10 imputations.
[3] Number of observations in which neither log-transformed BMI nor independent variable is missing.
*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$. NS ($p>.05$).
Each model had one independent variable at a time: years of education, locality size, or diabetes. Models were constructed using only non-imputed cases and all cases.
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

Figure 6

Binned Scatter Plots of Mean Log-BMI by Years of Education among Non-Imputed Cases and All Cases (2012 Wave)



[1]Imputed years of education were averaged across 10 imputations.
[2]Imputed log-BMI´s were averaged across 10 imputations.

Source: Own calculation using data from the Mexican Health and Aging Study 2012.

Comparison between case deletion and multiple imputation with respect to the estimated association between log-BMI and diabetes is more complicated, however. In 2003 multiple imputation showed a stronger association than pairwise deletion showed, as with education and locality size, but in 2012 and 2015 the opposite was true.

Table 6 and Figure 6 show that –although exclusion of cases with missing values biases the slope of the linear association between log-BMI and each of education, locality size, and diabetic status towards zero– this effect is less pronounced in the 2012 and 2015 waves than in the 2001 and 2003 ones. This result may be explained because the last two waves had smaller fractions of missing height and weight than the earlier two waves.

5. Conclusion

We provided a rationale and explained the procedure for imputation of non-response across MHAS waves. Multiple imputation produced more powerful results than case deletion did, without significantly distorting the distributions of height, weight, and body mass index (BMI) computed from these heights and weights. Therefore, we recommend imputing missing data and/or using the imputed values that we have generated here when analyzing data that includes self-reported height and weight from MHAS 2001, 2003, 2012, and/or 2015. More generally, when working with data with missing values, we recommend that users consider multiply imputing missing data whenever possible.

Our results justify the strategy of providing imputed values for the MHAS users, in particular because BMI is a critical variable for many studies of health of mid- and old-age Mexican adults. Our strategy is to provide users with an alternative to excluding the cases with missing values in height or weight, which could bias their results in a meaningful manner. We believe that our imputed variables provide a robust alternative for most users, and that researchers should not need to perform their own imputations. Even though the extent of bias when excluding cases with missing values may vary depending on the specific research and analyses performed, the researchers may at least now be able to test the sensitivity of their results when the cases with missing values are excluded.

As previously stated, the imputations described in this document used data from the 2001, 2003, 2012, and 2015 MHAS waves. Raw data from another wave, fielded in 2018, is now publicly available. Next, we will use the process described above to impute self-reported heights, weights, and BMI's in 2018.

References

Abellana, R., & Farran, A. (2015). "The identification, impact and management of missing values and outlier data in nutritional epidemiology", in: *Nutrición Hospitalaria*. 31(3), 189–195 (DE) https://doi.org/10.3305/nh.2015.31.sup3.8766

Corona, F., López-Pérez, J., & Muriel, N. (2019). "Funcionamiento en muestras finitas de técnicas de imputación y retropolación: caso de las series de encuestas económicas nacionales del INEGI", in: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. 10(3), 100–116.

Durán, B. (2019). "Comparación de metodologías de imputación aplicadas a ingresos laborales de la ENOE", in: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. 10(3), 4-27.

González-González, C., Orozco-Rocha, K., Samper-Ternent, R., & Wong, R. (2021). "Adultos mayores en riesgo de COVID-19 y sus vulnerabilidades socioeconómicas y familiares: un análisis con el ENASEM", in: *Papeles de Población*. 27(107), 141-165. Epub 06 de diciembre de 2021 (DE) https://doi.org/10.22185/24487147.2021.107.06

Kontopantelis, E., Parisi, R., Springate, D. A., & Reeves, D. (2017). "Longitudinal multiple imputation approaches for body mass index or other variables with very low individual-level variability: the mibmi command in Stata", in: *BMC Research Notes*. 10(1), 1–21 (DE) https://doi.org/10.1186/s13104-016-2365-z

Kumar, A., Karmarkar, A., Tan, A., Graham, J., Arceri, C., Ottenbacher, K., & Al Snih, S. (2015). "The effect of obesity on incidence of disability and mortality in Mexicans aged 50 years and older", in: *Salud Publica Mex*. 57(1), s31–s38.

MHAS Mexican Health and Aging Study, (2001–2015). Data Files and Documentation (public use): Mexican Health and Aging Study, (Data File Codebooks). Retrieved from www.MHASweb.org on September 9, 2020.

Monteverde, M., & Novak, B. (2008). "Obesidad y esperanza de vida en México", in: *Población y Salud Mesoamérica.* 6(1), 1–13 (DE) https://doi.org/10.1038/jid.2014.371

Palloni, A., Beltrán-Sánchez, H., Novak, B., Pinto, G., & Wong, R. (2015). "Adult obesity, disease and longevity in Mexico", in: *Salud Pública de México.* 57(1), s22–s30.

Rässler, S., Rubin, D. B., & Zell, E. R. (2012). "Imputation", in: *WIREs Computational Statistics.* 5(1), 20–29. doi: 10.1002/wics.1240

Rodriguez, M., & Wong, R. (2019). "Envejecimiento en México: Obesidad", in: *Boletín Informativo del ENASEM*. 19(1), 1–2 (DE) http://www.enasem.org/MHAS_AgingInMexico.pdf

Sattar, N., McInnes, I. B., & McMurray, J. J. V. (2020). "Obesity is a risk factor for severe COVID-19 infection: Multiple potential mechanisms", in: *Circulation,* 4–6 (DE) https://doi.org/10.1161/CIRCULATIONAHA.120.047659

Van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification", in : *Statistical Methods in Medical Research*. 16(3), 219–242. doi: 10.1177/0962280206074463

Van Buuren, S., Boshuizen, H.C., & Knook, D.L. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis", in: *Statistics in Medicine*. 18(6), 681–694. doi: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

Vargas Chanes, D., & Valdés Cruz, S. (2018). "Ajuste estadístico a la distribución del ingreso en el Módulo de Condiciones Socioeconómicas 2015 mediante imputaciones múltiples", in: *Realidad, Datos y Espacio, Revista Internacional de Estadística y Geografía.* 9(Número especial), 155–175.

Wong, R., Michaels-Obregon, A., & Palloni, A. (2017b). "Cohort Profile: The Mexican Health and Aging Study (MHAS)", in: *International Journal of Epidemiology.* 46(2), 1–10 (DE) https://doi.org/10.1093/ije/dyu263

Wong, R., Orozco, K., Zhang, D., & Michaels, A. (2017a). "Imputation of non-response on economic variables in the Mexican Health and Aging Study (MHAS / ENASEM) 2015", in: *Aging. University of Texas Medical Branch*.

Hoja de ruta para producir frecuentemente información estadística representativa

mediante el uso conjunto de información de redes sociales y encuestas

Roadmap for Frequently Producing Representative Statistical Information

through the Joint Use of Social Networking and Survey Data

Víctor Alfredo Bustos y de la Tijera, Silvia Laura Fraustro Velhagen, Noemí López Delgado y Ricardo Antonio Olvera Navarro*

Desarrollamos una propuesta metodológica para que las oficinas nacionales de estadística produzcan información representativa sobre múltiples temas, con mayor frecuencia, utilizando en conjunto datos de encuestas de hogares y publicaciones en redes sociales. La propuesta se basa en dar un nuevo rol a los datos como insumo para el entrenamiento de algoritmos de aprendizaje automático (ML, por sus siglas en inglés). Comenzamos clasificando a los encuestados según sus datos registrados en el cuestionario. Las publicaciones en las redes sociales de estos, si las hubiera, heredan sus etiquetas de clase. Utilizándolas como entrada, se entrenan algoritmos de ML. Para el seguimiento, las recientes en el momento de las nuevas recopilaciones de la encuesta se etiquetan y se entrenan de nuevo los algoritmos. En cualquier caso, cuando se considera apropiado el resultado de entrenar un algoritmo, se utiliza para etiquetar automáticamente grandes volúmenes de publicaciones actuales y futuras de usuarios no incluidos en la encuesta. El seguimiento futuro se lleva a cabo

We developed a methodological proposal for National Statistical Offices (NSOs) to produce representative information on multiple topics, with greater frequency, using household survey data and social media posts together. The proposal is based on giving a new role to the data as input for training machine learning (ML) algorithms. We begin by classifying respondents according to their data recorded in the questionnaire. Their social media posts, if any, inherit their class tags. Using them as input, ML algorithms are trained. For follow-up, recent ones at the time of new survey collections are tagged and algorithms are trained again. In either case, when the result of training an algorithm is deemed appropriate, it is used to automatically tag large volumes of current and future posts from users not included in the survey. Future tracking is carried out through tweets posted between survey rounds. The above procedure also has application in selection bias mitigation. In this case, a minimal set of sociodemographic (SD) variables collected through surveys can be used to develop a da-

* Instituto Nacional de Estadística y Geografía (INEGI), alfredo.bustos@inegi.org.mx, silvia.fraustro@inegi.org.mx, nohemi.delgado@inegi.org.mx y ricardo.olvera@inegi.org.mx, respectivamente.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

a través de tuits publicados entre rondas de encuestas. El procedimiento anterior también tiene aplicación en la mitigación del sesgo de selección. En este caso, se puede usar un conjunto mínimo de variables sociodemográficas (SD) recopiladas a través de encuestas para desarrollar una base de datos de autores etiquetada según SD. Se hará referencia a esta durante los estudios temáticos para mitigar la falta de representatividad de la población de usuarios. Para que todo lo anterior funcione, las respuestas a las encuestas y las publicaciones en redes de los usuarios-informantes deben ser vinculadas. Proponemos una forma de conseguirlo. Un futuro levantamiento de la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares del Instituto Nacional de Estadística y Geografía se empleará para estudiar la viabilidad de la propuesta, puesto que ya investiga el uso de las redes sociales y recopila información sociodemográfica.

Palabras clave: redes sociales; representatividad; sesgo de selección; etiquetado.

Recibido: 22 de septiembre de 2021.
Aceptado: 5 de noviembre de 2021.

tabase of authors labelled according to SD. This will be referenced during the thematic studies to mitigate the lack of representativeness of the user population. For all of the above to work, survey responses and user-informant network postings must be linked. We propose a way to achieve this. A future survey of the National Survey on Availability and Use of Information Technologies in Households (ENDUTIH in Spanish) of the National Institute of Statistics and Geography (INEGI) will be used to study the feasibility of the proposal, since it already investigates the use of social networks and collects sociodemographic information.

Key words: official statistics; machine learning; social networks; tagging; representativeness; selection bias.

Young Girl Using Smart Phone Social Media Concept/ bombuscreative/ iStock

Introducción

La sola posibilidad de complementar las fuentes tradicionales de información mediante la incorporación de los así llamados grandes datos o *Big Data* ha dado lugar a una amplia bibliografía en años recientes. Numerosos organismos nacionales e internacionales han llevado a cabo diversas acciones para estudiar su uso en sus actividades cotidianas. Por ejemplo, la CBS holandesa está entre las primeras oficinas nacionales de estadística (ONE) en iniciar el estudio de estas fuentes alternativas, así como de sus implicaciones para la estadística oficial (ver Struijs *et al*., 2014 y Struijs y Daas, 2014). A su vez, la Organización de 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[1,2] (ver UNSD, 2015; Jansen, 2020; 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* de grandes datos, la promoción del desarrollo de capacidades, el entrenamiento y el intercambio de experiencias…". Durante su primera reunión en octubre 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.

Estos han sesionado y alcanzado diferentes avances. En Jansen (2020)[4] se hace un gran resumen de dichos avances para casi todos los grupos creados por el GWG. Cabe resaltar que no están señalados aquellos para los equipos sobre integración de datos ni el que se refiere al uso de información de redes sociales, al que declara en receso.[5]

Un ámbito de aplicación inmediato para los trabajos del GWG está dado por la *Agenda 2030*, la cual adoptó un marco global de monitoreo amplio abarcando 231 indicadores dentro de 17 ODS (ver Van Halderen *et al*., 2021). 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*[6] entre las economías desarrolladas y las que están en vías de serlo impedirán que estas hagan un uso óptimo de la información disponible. Los mismos autores indican 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…"; adicionalmente, exponen que será necesario poner particular atención para "… abordar los dilemas éticos y de privacidad…" que surgirán de estas fuentes de datos.

Para América Latina, en Data-Pop Alliance (2016) se indica que "… destacan los riesgos y las oportunidades que *Big Data* presenta a las oficinas nacionales de estadística en Latinoamérica en el contexto de los ODS…". Entre los primeros, incluye barreras institucionales para la administración del cambio y la innovación, restricciones para el acceso y la completez de los datos, retos técnicos y

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

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

5 El reporte más reciente del Social Media WG para 2017 se puede encontrar en https://unstats.un.org/unsd/bigdata/conferences/2017/gwg/GWG%20Task%20Team%20on%20Social%20Media%20Data%20-%202017%20report.pdf

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

metodológicos, brechas en capacidades humanas, así como riesgos éticos y políticos. Por otro lado, sobresalen cinco tendencias que considera propicias en la región: la experiencia latinoamericana en el movimiento de datos abiertos;[7] la aparición de asociaciones públicas y privadas sobre el tema de *Big Data*;[8] 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 tanto a las ONE como a otros actores. 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.

Aun considerando legislar para que las oficinas nacionales de estadística accedan a las fuentes *Big Data*, como se sugiere en Destatis (2019), es claro que no se puede garantizar que las empresas de redes sociales sigan funcionando. Hay experiencias de estas que han visto reducido el número de sus usuarios o hasta han desaparecido, por ejemplo, *Sixdegrees*, *Friendster* y *My Space*. La producción de estadística oficial habrá de establecer con adecuada anticipación las estrategias a instrumentar para garantizar la continuidad y comparabilidad de los resultados.

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

8 Ver también DNP (2017) y Dutra (2018).

Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos

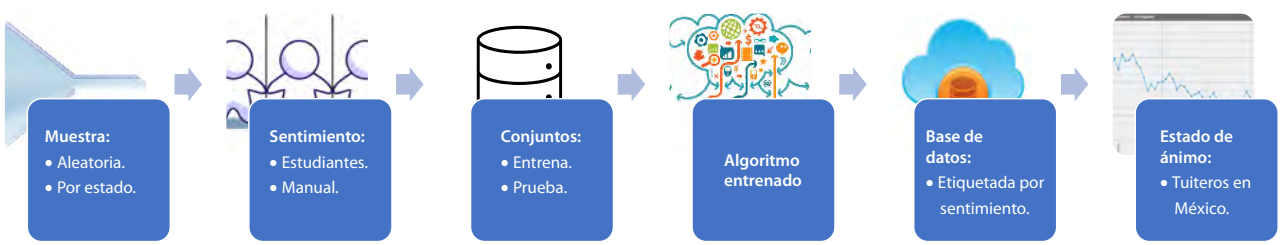
Este proyecto[9] se encuentra entre los usos más destacados que el Instituto Nacional de Estadística y Geografía (INEGI) ha dado al aprendizaje automatizado, así como a textos provenientes de *Twitter* (INEGI, 2017). En sus inicios, la única certeza era que los usuarios de esta red que publicaban textos georreferenciados se representaban a sí mismos; es decir, que la inferencia que resultara del análisis de sus textos aplicaría a ese segmento de habitantes, pero no necesariamente al resto de la población mexicana. De ahí el nombre que se dio al proyecto.

El procedimiento seguido en aquel momento para llegar a la publicación se resume en la figura 1. Entre los mensajes capturados se obtuvo una muestra aleatoria. Cada mensaje en esta fue etiquetado manualmente por uno o más estudiantes según su interpretación del sentimiento (positivo, neutro o negativo) del autor al publicarlo. Se obtuvo, así, un conjunto de tuits etiquetados que fueron separados para formar conjuntos de entrenamiento y prueba. Con este insumo se entrenaron y evaluaron diversos algoritmos para elegir la combinación más adecuada. A esta parte del proceso se la denominó *Pío análisis*. A partir de ella fue posible etiquetar de manera automática millo-

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

Figura 1

Proceso del estado de ánimo de los tuiteros en México



Fuente: elaboración propia.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

nes de mensajes almacenados en nuestra base de datos, así como los que cotidianamente son capturados con el fin de dar seguimiento diario al estado de ánimo de los tuiteros.

Limitaciones por resolver

Sesgo de selección

Al igual que ocurre en el caso de otros estudios observacionales, es necesario preguntarnos si la muestra disponible de usuarios de una red social es representativa de la población para la cual se quiere inferir. Es claro que, mientras más se aleje la primera de la segunda en términos de variables relevantes a la materia de estudio, se corre el riesgo de que los resultados que se produzcan apliquen solamente para la muestra; es decir, de que se vean afectados por un sesgo en relación con la población objetivo, ante la sub o sobrerrepresentación de una o más subpoblaciones importantes.

La incertidumbre que dio lugar al nombre del proyecto *Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos* motivó la ampliación de la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) del INEGI, cuyo fin era el de indagar el uso de redes sociales en estos. De esta manera, Bustos *et al*. (2021) pudieron identificar discrepancias sociodemográficas entre los usuarios de redes sociales y la población mexicana en general, con lo cual, además de que se justifica la elección del nombre del proyecto, quedan establecidas sus limitaciones y alcances. Destacan que, para el caso mexicano, las poblaciones de usuarios de redes sociales son, en promedio, más jóvenes, con un estatus socioeconómico más alto y un mayor nivel educativo que el resto de la población. Cabe destacar que, aun cuando subrepresentadas en términos relativos, hay en estas subpoblaciones usuarios de casi todos los grupos de edad, así como de todos los niveles de escolaridad y socioeconómicos.

Sobre el denominado sesgo de selección en redes sociales, es poco lo que se ha escrito (ver Golub, 2010 o Stark, 2015) y menos todavía acerca de sus implicaciones en la estadística oficial. Iacus *et al*. (2020) elaboran una propuesta metodológica para corregirlo a través de modelos comúnmente usados para estimación en dominios pequeños. En este caso, se ajustan los modelos mencionados a los agregados al nivel de dominio, afectados por el sesgo de selección, usando como variables explicativas agregados similares provenientes de la estadística oficial; en otras palabras, se busca corregir agregados y no a la base de datos misma. Por su parte, Kim y Tam (2020) recurren al mismo tipo de modelos, pero su enfoque, basado en lo que ha venido denominándose integración de datos, permite el tratamiento de errores de medición en las variables tanto para *Big Data* como para la muestra probabilística. Para la corrección del sesgo de selección en la estimación de proporciones, Tam y Kim (2019) introducen dos enfoques, para el caso en el que los registros de la encuesta pueden ser vinculados con los de la fuente *Big Data* y para lo contrario, recurriendo a modelos logísticos.

A causa de lo anterior, por el momento no ha sido posible instrumentar una estrategia que nos permita acercarnos al estado de ánimo de los mexicanos, según *Twitter*. Para ello, requeriríamos etiquetar adicionalmente los tuits capturados de acuerdo con características sociodemográficas (SD) de sus autores, como se muestra en la figura 2. Sin embargo, ya que la ENDUTIH tiene como unidad de análisis a las personas, en términos de las cuales hemos caracterizado el sesgo de selección y su posible corrección, es deseable dejar de trabajar con mensajes individuales. Si las características SD de los tuiteros fueran conocidas, en el momento de calcular totales para la construcción de índices se podrían reponderar sus etiquetas temáticas según la relación guardada entre los tamaños relativos de la clase a la que pertenecen en la población abierta, por un lado, y en la de tuiteros que publicaron durante el periodo en cuestión, por el otro. Para la clase identificada con las etiquetas $i, j,\ldots, k$, se representa su tamaño relativo en relación con la población de referencia (R) por $\%P^{R}_{ij\ldots k}$ y el de la de la muestra de redes sociales (SN) para el periodo por $\%P^{SN}_{ij\ldots k}$. Entonces, el factor de corrección para

representantes de esa subpoblación se presenta en la expresión (1):

$$\omega_{ij\ldots k} = \frac{\%P^{R}_{ij\ldots k}}{\%P^{SN}_{ij\ldots k}} \qquad (1)$$

Un valor menor a 1 del cociente (1) indicará que la subpoblación está sobrerrepresentada en esa red social. En caso contrario, se estará frente a una subrepresentación. A partir de ese momento, ese usuario será ponderado de acuerdo con el anterior cociente al ser incluido en agregaciones espaciales y/o temporales, tomando en cuenta sus publicaciones recientes. Sin embargo, tal información no se encuentra a nuestra disposición en general.

Si la información requerida estuviera disponible para cada tuitero *t* que publicó al menos un tuit en la región y en el periodo considerados, se estaría en condiciones de recalcular el Índice de Positividad del estado de ánimo de los tuiteros como se muestra en (2). En dicha expresión, los índices $I_{+}(t)$ e $I_{-}(t)$ toman valor 1 según si el ánimo del tuitero en el periodo es considerado positivo o negativo, respectivamente:

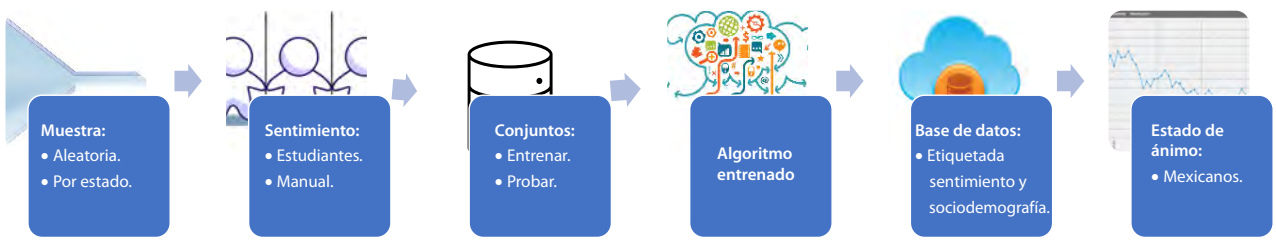
$$\textit{Índice de Positividad} = \frac{\sum_t \omega_{ij\ldots k}(t) I_{+}(t)}{\sum_t \omega_{ij\ldots k}(t) I_{-}(t)} \qquad (2)$$

Consulta directa a los usuarios

Otro asunto no resuelto se refiere la incertidumbre introducida cuando *expertos* etiquetan los mensajes. Ejemplo de ella es el hecho de que un mismo mensaje no reciba la misma etiqueta al ser evaluado por distintos individuos. En tanto se estudia la manera de incorporar dicha incertidumbre en el entrenamiento de algoritmos, parece deseable explorar alternativas. Por ejemplo, podría haberse considerado preguntar de manera directa al autor de la publicación sobre su estado de ánimo al momento de escribirla, como se muestra en la figura 3. Es claro que ello no elimina totalmente la incertidumbre, ya que el propio proceso de recordación carece de certezas.

Figura 2

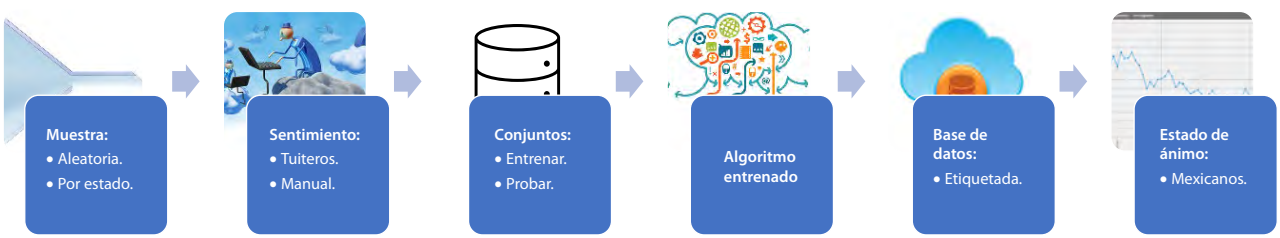
Proceso del estado de ánimo de los mexicanos reduciendo sesgo de selección



Fuente: elaboración propia.

Figura 3

Proceso del estado de ánimo de los mexicanos consultando autores



Fuente: elaboración propia.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Sin embargo, existen otras formas de acercarse a los usuarios de *Twitter* que, a nuestro juicio, reducen algunos de los mencionados riesgos. Las comentaremos más adelante como parte de nuestra propuesta.

Predicción de la información requerida

El hecho de que nuestra base de datos de tuits georreferenciados no cuente con información sobre la edad, el sexo, el nivel académico ni el estrato socioeconómico impide llevar a cabo la indicada ponderación de tuiteros. En *Twitter*, los usuarios pueden decidir mantener confidencial su perfil. En consecuencia, no podemos acudir a dicha información.

La segunda opción obvia es la de predecir las características de interés de los tuiteros cuyos tuits públicos almacenamos. Ejemplificaremos cómo proceder haciendo uso de la ENDUTIH, aunque cualquier otra encuesta que recopile la información necesaria podría servir. A lo largo del llenado del cuestionario de la Encuesta se han registrado datos sociodemográficos de cada integrante del hogar. Adicionalmente, los informantes que declaran ser usuarios recientes de alguna red han sido identificados.[10] Se requiere vincular aquella información con los mensajes públicos de cada usuario. Para ello, sugerimos que al concluir la entrevista después de darle una breve explicación del objetivo que perseguimos, así como de señalarle los artículos de la *Ley del Sistema Nacional de Información Estadística y Geográfica* que garantizan el uso confidencial de sus datos (ver anexo), se solicite al tuitero que nos proporcione su *username* o que publique un mensaje que contenga un código personalizado y conocido solo por ambas partes para evitar errores al registrar su nombre de usuario. El informante está en total libertad de atender nuestra solicitud. Si lo hiciera, nos aportará su nombre de usuario, lo que nos permitirá acudir a sus publicaciones. Estas podrán ahora ser etiquetadas con la información socioeconómica recogida en el cuestionario de la Encuesta. De esta manera, habremos creado un conjunto de entrenamiento, y uno menor de prueba, que serán insumo para la prueba de diversos algoritmos de aprendizaje automatizado. Los resultados de estos se evaluarán para determinar una combinación adecuada. En su primera mitad, la figura 4 ilustra el anterior procedimiento. En ella, el resultado intermedio en la forma de un algoritmo entrenado se destaca con un color de fondo diferente; ello nos permitirá señalar dónde y cuándo se usará posteriormente dicho algoritmo. A partir de este momento, dejaremos de hacer uso individual de los datos de los informantes de la ENDUTIH.

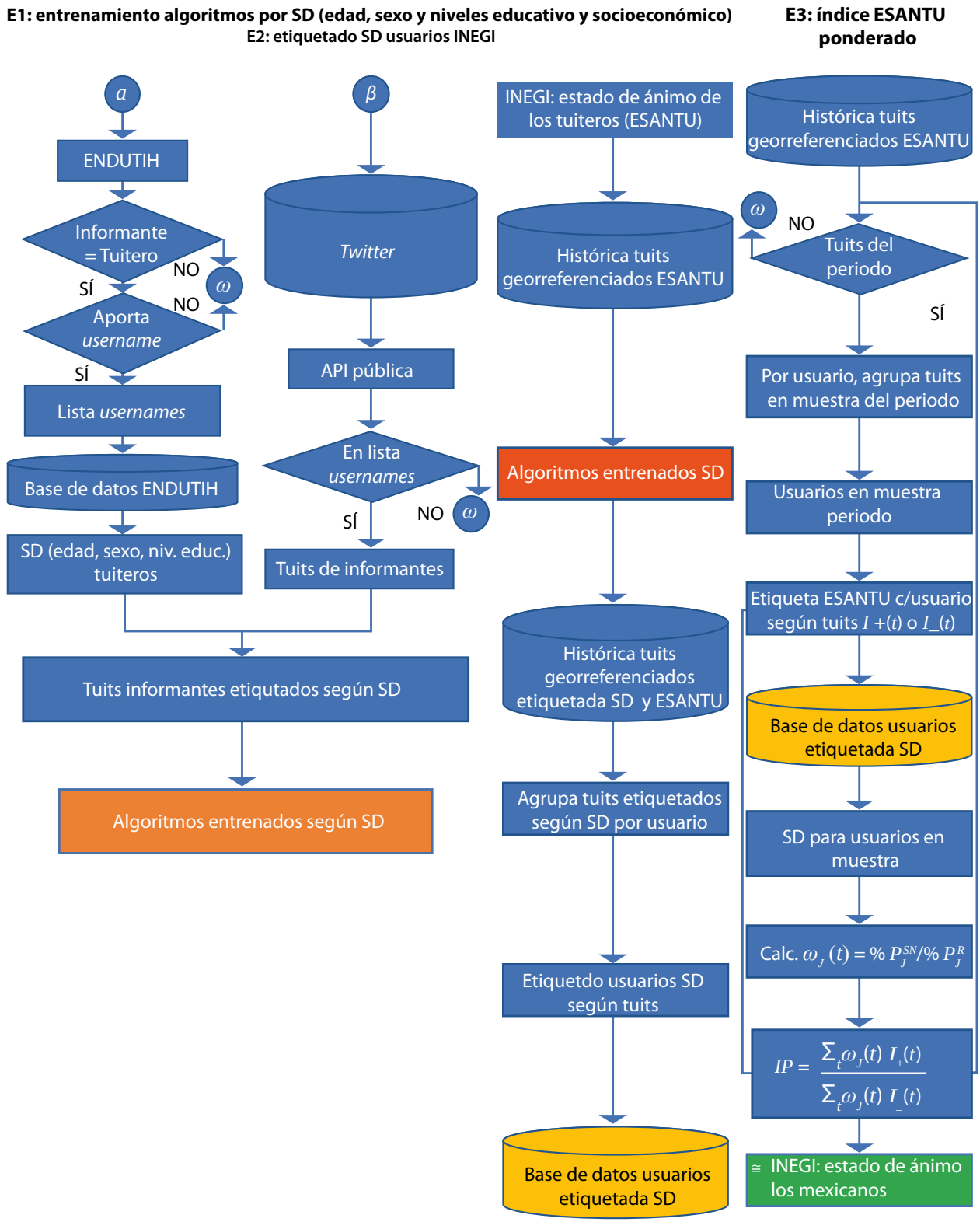
Si la calidad de los resultados lo permite, se estará en condiciones de proceder a (E2 en figura 4) etiquetar sociodemográficamente (SD) a los usuarios autores de tuits georreferenciados en la base de datos en poder del INEGI. Inicialmente, será necesario aplicar el algoritmo a los mensajes contenidos en la base de datos histórica para predecir la información SD requerida. Acto seguido, todos los mensajes etiquetados de un mismo usuario se reunirán para decidir las etiquetas que, a su vez, le correspondan de acuerdo con esa información. Ello se debe a que no hay garantía de que todos sus mensajes hayan recibido las mismas etiquetas, por lo cual habrá que determinar las que se le asignarán. A partir de este momento se tendrá un nuevo resultado intermedio en la forma de una base de datos de usuarios etiquetados según sus características sociodemográficas. Recurriremos a esta para asignar una ponderación a cada usuario incluido en las muestras por periodo y ubicación, que se usará en el cálculo de estadísticas o indicadores.

De acuerdo con el etiquetado del tuitero, será posible revisar los resultados de proyectos, como el *Estado de ánimo de los tuiteros*. Sin pérdida de generalidad, supongamos que se desea producir resultados en un día particular para una entidad federativa. Primero se determinará el estado de ánimo más probable de cada tuitero $(I_+(t) \circ I_-(t))$ con base en el etiquetado automático de sus publicaciones, de ese día y lugar, por el algoritmo en uso actualmente. Acto seguido, será determinado el tamaño relativo de cada una de las subpoblaciones $J = A,B,\ldots,Z$ representadas

10 En caso de recurrir a otra encuesta, sería necesario lograr dicha identificación.

Figura 4

Reducción del sesgo de selección en el estudio del estado de ánimo en México



Fuente: elaboración propia.

en la muestra de tuits publicados el mismo día y en el mismo lugar. A partir de la expresión (1), se determinará la ponderación que permitirá mitigar el sesgo de selección en los resultados. La relación de positividad se calculará como el cociente de la suma de ponderaciones de los tuiteros positivos entre la de las correspondientes a los negativos, según muestra la expresión (2).

Se procederá de manera similar para actualizar resultados mediante la captura diaria de tuits. Cuando se obtengan publicaciones de un usuario que no se encuentre en la base de datos, este será etiquetado en alguna subpoblación sociodemográfica y añadido a la información histórica.

Para el caso en el que el informante rehúse atender nuestra petición, estaremos ante la posibilidad de un nuevo sesgo de selección por no respuesta, semejante al estudiado en la teoría de muestreo. En esta ocasión, no obstante, la propia Encuesta aporta información suficiente que nos permite comparar entre lo que debería haberse recibido y lo que finalmente se recibió. En tanto se cuente con respuestas favorables de representantes de cada una de las clases definidas, será posible reponderar lo recibido si esto fuera necesario.[11] Cabe destacar, sin embargo, que en realidad no tenemos claro si, con el fin de entrenar un algoritmo, es o no útil reponderar los casos disponibles ni conocemos resultados en la literatura disponible que nos indiquen cómo hacerlo en la fase de entrenamiento, en su caso. Contra lo que ocurre al calcular indicadores con datos de la muestra, podemos equiparar la fase de entrenamiento de algunos algoritmos de ML (ej., redes neuronales) con el ajuste de modelos no lineales a esta. Para este último caso, no es un requisito que la muestra sea representativa, por lo que no es usual reponderar las unidades muestrales. En todo caso, se requiere que estén presentes dos o más casos para todas y cada una de las subpoblaciones.

11 Durante el proceso de revisión de este trabajo se recibieron indicaciones de que este tema ha sido estudiado en la literatura de ciencia de datos. Sin embargo, no fueron puestas a nuestra disposición las correspondientes referencias.

Seguimiento continuo a temas sociales

Además de permitirnos el etiquetado de la base de datos de usuarios de *Twitter*, el procedimiento descrito para el *Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos* nos aporta un camino a seguir para lograr la producción de información estadística con base en el uso combinado de encuestas y publicaciones en redes sociales. De esta manera, es posible sugerir que también, en el caso de otras encuestas, se solicite el permiso de acceso a las publicaciones de usuarios de *Twitter*, residentes en las viviendas seleccionadas, algunas de cuyas características en su carácter de informante han sido recogidas durante la entrevista. La identificación de aquellas con etiquetas para los textos publicados en fechas cercanas a la de la entrevista permitirá entrenar uno o más algoritmos cubriendo múltiples temas. El etiquetado automático mediante dicho algoritmo dará lugar a etiquetas adicionales para cada tuit. Bajo estas nuevas condiciones y la reponderación obtenida, será posible hacer un seguimiento estadístico frecuente a los temas estudiados por cada encuesta.

De esta manera, por ejemplo, es de esperar que cada levantamiento de la Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) aportará etiquetas acerca del delito y sus causas, o su repercusión sobre sus víctimas. Con base en sus publicaciones, dichas etiquetas serán automáticamente asignadas a los autores que publiquen en fechas cercanas y posteriores al levantamiento, lo que permitirá dar seguimiento estadístico frecuente a esos temas. Del mismo modo, la Encuesta Nacional de Ocupación y Empleo (ENOE) aportará elementos para seguir la evolución de la ocupación, además de la de los recogidos en la hoja de datos sociodemográficos (defunciones, nacimientos, así como migración y sus causas). De manera adicional, el uso de paneles rotatorios abre la posibilidad de estudiar la evolución de la forma en que tuitea una persona cuyo estatus ocupacional cambia durante el tiempo que permanece en muestra.

En términos de la prevención del delito, la Encuesta de Cohesión Social para la Prevención de la Violencia y la Delincuencia (ECOPRED) aportaría el etiquetado acerca de los factores de riesgo y exposición a situaciones de violencia y delincuencia que enfrentan los jóvenes de 12 a 29 años de edad. Por su parte, la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) lo haría sobre aquellas situaciones de violencia emocional, económica, patrimonial, física y sexual ejercida en contra de las mujeres de 15 años y más, ocurrida en distintos ámbitos (escolar, laboral, comunitario, familiar y de la pareja). Con ello se abre, además, la posibilidad de diseñar encuestas sobre temas emergentes, como los contemplados por algunos de los indicadores para los ODS de la ONU, o que no hemos podido estudiar con anterioridad, como la salud mental en adolescentes, para vincularlos con publicaciones en redes sociales y darles también seguimiento frecuente.

Figura 5

Proceso del seguimiento diversos tópicos de los mexicanos usando encuestas

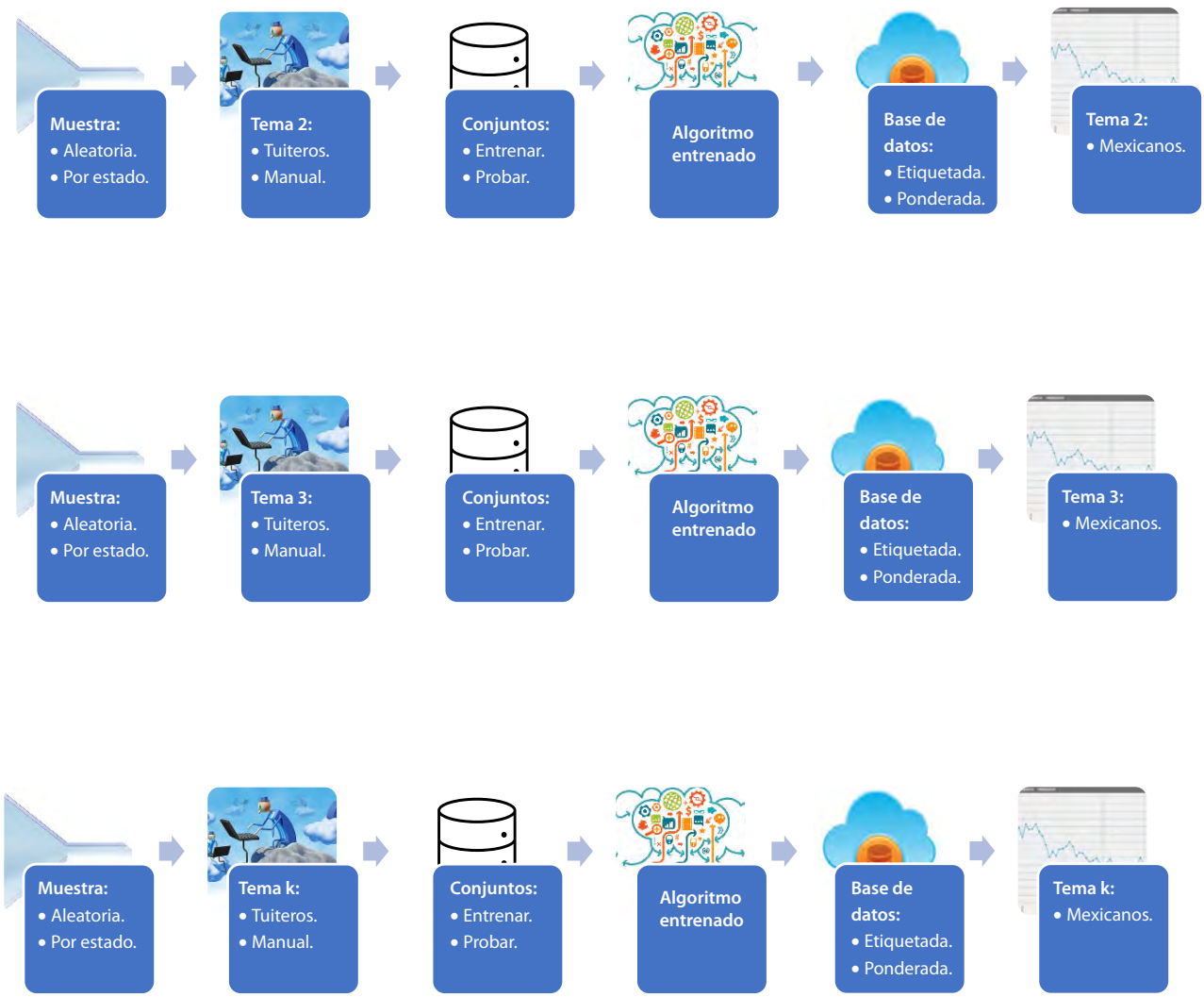
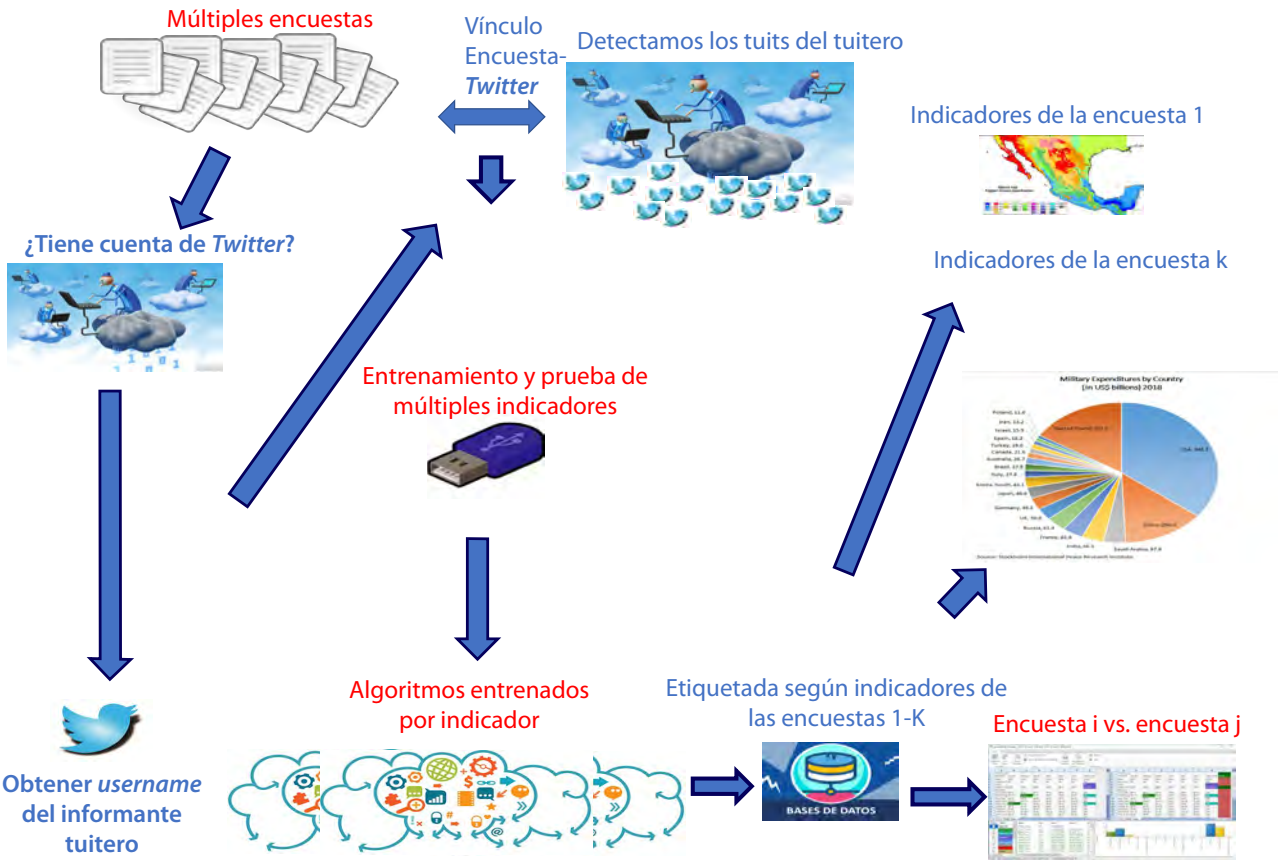


Figura 6

Proceso del seguimiento de diversos tópicos y de sus interrelaciones usando *Twitter*



Fuente: elaboración propia.

En consecuencia, cada usuario será clasificado desde muy diversos puntos de vista. De esta manera, se abre la posibilidad de relacionar las temáticas de distintas encuestas en formas hasta ahora impensables. Por ejemplo, es posible pensar en dar seguimiento a la salud mental de aquellas personas que han perdido el empleo o que han sufrido algún otro tipo de pérdida; otro podría ser el cambio en la confianza de los consumidores o víctimas recientes de algún delito. En fin, se extiende una gama interesante de posibilidades que no estaban a nuestro alcance ni en el caso del levantamiento de un censo de población.

Más aún, diversas encuestas permiten identificar cambios en el entorno inmediato del usuario/informante, como: el nacimiento de un nuevo integrante del hogar, el retorno de un migrante o el matrimonio de un pariente, que pueden influenciar la forma en la que el tuitero se expresa en la red. Nos gustaría pensar que el uso de toda o parte de esta información como etiquetas asociadas a sus mensajes permitirá entrenar con razonable precisión nuevos algoritmos. Esto nos llevaría a relacionar sus publicaciones con eventos no asociados directamente con el usuario/informante.

Comentarios finales

La experiencia acumulada en el uso de información de redes sociales ha sido invaluable para

su explotación en la producción de estadística oficial. El grupo de técnicos y profesionales que laboran en el INEGI y que se han capacitado en el uso de las técnicas relevantes alcanza ya un número importante. Asimismo, los planes para el fortalecimiento de la infraestructura muestran avances valiosos. Es preciso reconocer que a lo largo de los diferentes trabajos reportados en este documento se ha cumplido con el propósito didáctico que alentó los primeros esfuerzos. Esta propuesta es una más de las derivadas de dicha experiencia.

En ella hemos delineado un camino para el seguimiento continuo de temas sociales y demográficos de interés para las oficinas nacionales de estadística. Esta ruta se apoya en los datos recolectados dentro de los programas tradicionales de producción de información estadística oficial de cualquier ONE a través de encuestas por muestreo. De esta manera, se reduce el sesgo de selección en los indicadores con los que se dará seguimiento prácticamente continuo a diversos temas; asimismo, se reduce la incertidumbre detrás de esos indicadores, pues se evita depender de la opinión de expertos para el etiquetado de los conjuntos de entrenamiento y prueba.

En general, tanto si se trata del etiquetado en la base de datos de usuarios como si lo es del seguimiento de algún tema objeto de estudio de alguna encuesta, la hoja de ruta propuesta para optimizar la combinación de datos de encuestas con los de las redes sociales debe abarcar los siguientes pasos, como mínimo:

a) Aplicar el cuestionario de la encuesta.
b) Al finalizar el llenado, indagar si algún informante es usuario de redes sociales.
c) Solicitarle la publicación de un mensaje.
d) En su caso, vincular sus respuestas en el cuestionario con sus publicaciones.
e) Incorporarlo al conjunto de entrenamiento.
f) Entrenar un algoritmo.
g) Predecir las etiquetas para la base de datos, o para las nuevas publicaciones, a través del mismo algoritmo.
h) Elaborar agregados usando las reponderaciones correspondientes.

Parece útil sugerir una postura proactiva que informe a los usuarios de redes sociales acerca de los alcances del ejercicio que la oficina nacional de estadística se propone realizar. Se sugiere el uso de las propias redes sociales para difundir los alcances y las limitaciones en el uso de la información, así como los análisis y resultados que se derivarán de su uso. Es de esperarse que los mensajes transmitidos por esta vía faciliten la labor de los entrevistadores. Por supuesto, y para aportar una descripción más amplia de las acciones a realizar, las publicaciones de la ONE harán referencia a alguna sección en su página oficial. En ella, el informante encontrará, además, la base legal que garantice el uso con fines estadísticos de su información, y ninguno otro.

La adaptación de nuestra propuesta a la explotación de registros administrativos con fines similares no nos resulta inmediata. Para este caso, habrá que desarrollar formas de vincular información contenida en un registro administrativo con las publicaciones del ciudadano. Sin embargo, será necesario ser cautos para evitar hacer públicos datos cuyo uso inadecuado por terceros pueda implicar un daño a nuestro informante. Por ejemplo, sugerimos evitar el uso de la Clave Única del Registro de Población (CURP) utilizada en México para fines oficiales, pero cuya publicación en un tuit permitiría vincular la información de diversos registros públicos con la publicada en la red social.

Fuentes

Brakel, J.; E. van den Söhler, P. Daas y B. Buelens. "Social media as a data source for official statistics; the Dutch Consumer Confidence Index", en: *Survey Methodology*. Vol. 43, No. 2, December. Statistics Canada, 2017, pp. 183-210. 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.

Bustos y de la Tijera, V. A., A. A., Coronado Iruegas, S. L. Fraustro Velhagen, G. Leyva Parra, N. López Delgado, R. A. Olvera Navarro, A. M. Romo Anaya y

V. Silva Cuevas. "Caracterización del sesgo de selección en redes sociales en México a través de algunas características sociodemográficas de sus usuarios", en: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. En prensa 2022.

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: estado del arte y análisis comparativo de estrategias nacionales de Big Data, noviembre de 2017* (DE) https://datapopalliance.org/wp-content/uploads/2018/09/Documento1_VersionFinal_DNP.pdf, consultado el 11 de noviembre 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-organizacioneSDeben-implementar-una-estrategia-centrada-en-loSDatos, consultado el 21 de mayo de 2021.

Golub, B., and M. O. Jackson. "Using selection bias to explain the observed structure of Internet diffusions", en: *Proc Natl Acad Sci USA*. Vol. 107, No. 24, June 15, 2010, pp. 10833-10836.

Iacus, S., G. Porro, S. Salini y E. Siletti. "Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal", en: *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 21 de 2021.

INEGI. *Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos. Documento metodológico V 2.0.* 2017 (DE) https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825099718, consultado el 29 de octubre 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.

Kim, J. K., and S. M. Tam. "Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference", en: *International Statistical Review*. John Wiley & Sons Ltd on behalf of International Statistical Institute, 2020, doi:10.1111/insr.12434.

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.

Smith, H., *Big Data for Official Statistics*. Workshop on Big Data for Economic Statistics: Challenges and Opportunities, 11 September 2018, Rio de Janeiro, Brazil, https://unstats.un.org/Unsd/nationalaccount/workshops/2018/rio/UNSD.PDF, consultado el 11 de noviembre 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. y P. Daas. "Quality approaches to big data in official statistics, European Conference on Quality", en: *Official Statistics*. 2014 (DE) http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf, consultado el 21 de mayo de 2021.

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

Tam, S. M., and J. K. Kim. "Big Data ethics and selection-bias: An official statistician's perspective", en: *Statistical Journal of the IAOS*. 34, 2018, pp. 577-588, DOI 10.3233/SJI-170395.

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, 3-6 March, 2015 (DE) https://documentSDds-ny.un.org/doc/UNDOC/GEN/N14/692/71/PDF/N1469271.pdf?OpenElement, consultado el 21 de mayo de 2021.

Anexo

Módulo Experimental Aprendizaje Automatizado basado en Encuestas en Hogares (MAAEH) 2022

Texto que será leído por el entrevistador al informante seleccionado en la ENDUTIH 2022 después de finalizar el llenado del cuestionario de la encuesta:

"Con el fin de informar a los mexicanos y las mexicanas, el INEGI explora nuevas formas de producir

estadísticas. En este nuevo módulo experimental queremos entrenar algoritmos de inteligencia artificial con su información para **establecer si la forma en la que escriben los usuarios mexicanos de *Twitter* depende de su edad, sexo, y escolaridad**. Si gracias a este proyecto se establece que sí hay relación, más adelante, usando solo los tuits que publican en el país **millones de tuiteros que NO están en la muestra de la ENDUTIH**, podremos hacer un seguimiento frecuente de la evolución de estas y otras características en la población de México.

Solicitamos su apoyo para avanzar en este proyecto de la siguiente forma: le invitamos a que nos comparta ahora su nombre de usuario en *Twitter* o, si lo prefiere, a que envíe dentro de los siguientes siete días un tuit público, o uno directo, a @INEGI_INFORMA, que incluya un número personalizado que le daré, y que solo usted y el INEGI conocerán. Si acepta nuestra invitación, esta información será la vía: a) para leer todos sus tuits de dominio público hasta esta fecha, ____ de ____ de 2022; b) para vincularlos con sus respuestas a la ENDUTIH; y c) para establecer si hay o no una relación entre los primeros y estas. Los textos de sus tuits públicos serán tratados conforme a las disposiciones del **artículo 37, párrafo primero, de la *Ley del Sistema Nacional de Información Estadística y Geográfica*** en vigor, al igual que la información que nos brindó en la ENDUTIH 2022. Los resultados de esta investigación no serán divulgados por ningún medio, ya que serán utilizados estrictamente para análisis interno del Instituto.

El número que le solicitamos que incluya en su mensaje es el XXXXXXXXXX".

Propuesta de indicadores alternativos **para medir la confianza del consumidor**

Proposed Alternative Indicators **for Measuring Consumer Confidence**

Jesús López-Pérez,* Francisco de Jesús Corona Villavicencio* y José Manuel Lecuanda Ontiveros**

En este trabajo realizamos propuestas para mejorar la correlación del Indicador de Confianza del Consumidor (ICC) con variables relevantes de la situación económica actual, como la actividad económica y el consumo privado. A partir de las 15 preguntas que produce la Encuesta Nacional sobre Confianza del Consumidor (ENCO) y adicionalmente de indicadores no tradicionales, específicamente de *Google Trends*, estimamos ICC alternativos al actual. En ambos casos, la técnica estadística para la generación de estos se realiza usando mínimos cuadrados parciales (PLS, por sus siglas en inglés). Los resultados indican que los indicadores propuestos están más correlacionados con las variables de actividad económica que el ICC actual, sobre todo el basado en *Google Trends*, aunque la interpretabilidad de los resultados es más consistente cuando se usan las preguntas de la ENCO. En consecuencia, se concluye que el mejor resultado es estimar el ICC usando todas las preguntas de la ENCO y PLS como técnica de estimación.

Palabras clave: confianza del consumidor; consumo privado; mínimos cuadrados parciales; información no tradicional.

In this paper we make proposals to improve the correlation of the Consumer Confidence Indicator (CCI) with relevant variables of the current economic situation, such as economic activity and private consumption. Based on the 15 questions produced by the National Survey on Consumer Confidence (ENCO in Spanish) and additionally on non-traditional indicators, specifically Google Trends, we estimate alternative CCIs to the current one. In both cases, the statistical technique for the generation of these is performed using partial least squares (PLS). The results indicate that the proposed indicators are more correlated with the economic activity variables than the current CCI, especially the one based on Google Trends, although the interpretability of the results is more consistent when using the ENCO questions. Consequently, it is concluded that the best result is to estimate the CCI using all the ENCO questions and PLS as the estimation technique.

Key words: consumer sentiment; private consumption; partial least squares; nontraditional information.

Recibido: 2 de agosto de 2021.
Aceptado: 9 de noviembre de 2021.

* Instituto Nacional de Estadística y Geografía (INEGI), jesus.lopezp@inegi.org.mx y franciscoj.corona@inegi.org.mx, respectivamente.
** CETYS Universidad, manuel.lecuanda@cetys.mx

Online shopping. Internet Shopping Concept. Shopping From Home/ Nanci Santos/ iStock

Introducción

El consumo privado representa una gran parte de la economía de una nación, usualmente las dos terceras partes del Producto Interno Bruto (PIB) en los países desarrollados y en los emergentes. En México, durante el 2019, este representó 67.9 % del PIB.[1] Por ello, es razonable pensar que los tomadores de decisiones promuevan políticas económicas para incentivar el consumo de un país, y es deseable conocer las expectativas del consumidor, lo cual, presumiblemente, nos de elementos para tener una mayor capacidad de anticipación sobre el consumo futuro. En el contexto nacional, el Instituto Nacional de Estadística y Geografía (INEGI) y el Banco de México, cada mes y de manera conjunta, llevan a cabo la Encuesta Nacional de Confianza del Consumidor (ENCO), cuyo objetivo es medir las percepciones de la situación económica actual y futura para las familias y el país, la capacidad de comprar productos duraderos y de anticipar el comportamiento a largo plazo de la demanda agregada y el crecimiento económico.[2]

El Indicador de Confianza del Consumidor (ICC) es el resultado principal de esta encuesta y se compone a partir de cinco de las 15 preguntas que forman parte de la ENCO. Desde el inicio de su publicación en abril del 2001, ha atraído la atención de economistas y presentadores de noticias.[3] En Leyva *et al*. (2016) se investiga la naturaleza de su construcción y se sugiere publicar los datos como indicadores de balance en lugar de en términos de índice, recomendación que el INEGI cumplió

1 INEGI. *Oferta y demanda global*. Series desestacionalizadas. Precios del 2013.

2 INEGI. *Encuesta Nacional sobre Confianza del Consumidor 2015. Documento metodológico.*

3 En la fecha de elaboración de esta investigación, la búsqueda en *Google News* "Índice de Confianza del Consumidor México" arroja 208 mil resultados.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

en el 2019 ya que, a partir de esa fecha, los datos se publican primordialmente en balance.

Sin embargo, los mecanismos mediante los cuales el comportamiento y las expectativas de los hogares se reflejan en la economía son menos comprendidos (Ludvigson, 2004). De hecho, una preocupación común en la prensa financiera es que los niveles del ICC no se traducen en mayor o menor consumo, más bien es mucho más habitual pensar que están más ligados al contexto político, como la aprobación presidencial (ver, por ejemplo, Aldrete y Flores, 2019 y BANAMEX, 2019). Por lo tanto, no es claro si el ICC en realidad está correlacionado con el consumo o el estado actual de la economía.

El objetivo del presente trabajo es proponer ICC alternativos que estén correlacionados con el Indicador Mensual del Consumo Privado del Mercado Interno (IMCPMI, en adelante Consumo) y con el Indicador Global de la Actividad Económica (IGAE), este último considerado una variable *proxy* del PIB mensual. Para lo anterior, proponemos usar las 15 preguntas de la ENCO y seleccionar la contribución de cada variable según la correlación bivariada que tienen estas con el Consumo y el IGAE. En consecuencia, la técnica de estimación seleccionada es por mínimos cuadrados parciales (PLS, por sus siglas en inglés), la cual permite reducir dimensionalidad similar a componentes principales, pero considerando, también, variables dependientes.

Además, se propone un ICC alternativo con características similares que se correlacione con el consumo privado. Dicho indicador se estimará utilizando fuentes no tradicionales de información, de manera específica a través de una serie de tópicos de *Google Trends* relacionados con el consumo privado. Lo anterior se realiza con el objetivo de comparar entre un ICC tradicional y otro estimado mediante fuentes no convencionales, parecido a como lo proponen Aprigliano *et al*. (2019).

Los resultados de este trabajo permitirán a los analistas de información macroeconómica contar con ICC potencializados que estén estructuralmente más relacionados con variables importantes de la economía mexicana. De esta forma, y dada la oportunidad del ICC respecto al Consumo y el PIB, por ejemplo, se pueden tener estimaciones más certeras del comportamiento de estas variables en el corto plazo.

El resto del artículo está organizado de la siguiente manera: en el primer apartado se describe cómo se ha comportado el consumo privado en los últimos años; enseguida, se destaca la literatura relevante; la tercera sección presenta los datos a utilizar y su tratamiento; en el cuarto apartado se muestra la metodología estadística empleada para obtener un indicador del consumidor altamente correlacionado con el consumo y la actividad económica; después, en el quinto, se expone la construcción del Indicador de Expectativas del Consumidor; en la sexta sección están las conclusiones; y, por último, los apéndices.

Evolución del consumo y la actividad económica y su relación con el ICC

El consumo privado en México ha estado en una senda de deterioro desde septiembre del 2019 y se ha agravado por el confinamiento derivado de la pandemia de la COVID-19. En mayo del 2020, el Consumo registró su mayor caída de -22.5 % en términos anuales y, a pesar de una leve recuperación en los meses siguientes, hasta septiembre del 2020, los niveles de dicho indicador se encuentran aún en el terreno negativo con -5.5 por ciento. De manera similar, durante la contingencia sanitaria, la actividad económica registró un declive considerable, a saber, el IGAE alcanzó una tasa anual de -25.3 % en mayo del 2020, para comenzar una lenta recuperación, aún por debajo de los niveles del año anterior, donde a septiembre del 2020 todavía se encontraba en -5.5 % a tasa anual.

Para ilustrar la relación del ICC con el Consumo y el IGAE, la gráfica 1 presenta la evolución de estas tres series desde abril del 2002 hasta septiembre del 2020. Podemos caracterizar ampliamente cua-

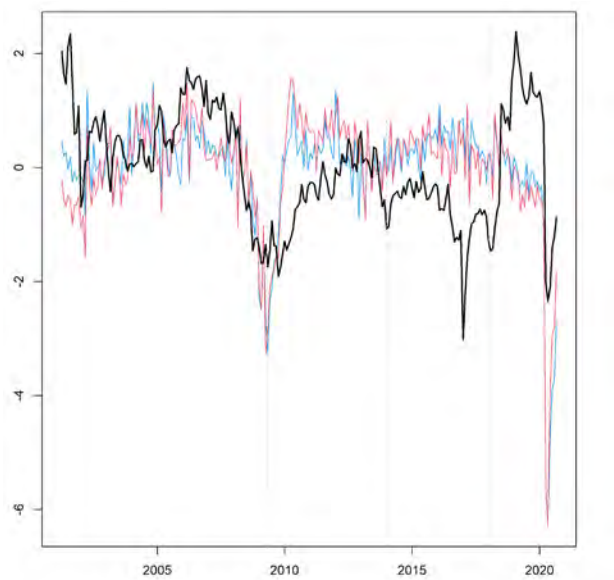
tro periodos basados en la dinámica del ICC con respecto a esos indicadores. En el primero, desde el 2002 hasta mediados del 2009, el ICC sigue una tendencia y volatilidad similar con el par de series; luego, en el segundo lapso, desde mediados del 2009 hasta principios del 2014, este parece estar rezagado, pero con la misma tendencia; después, del 2014 hasta agosto del 2019, muestra una mayor volatilidad y un comportamiento contemporáneo con las otras dos variables; y al final, desde julio del 2018, donde el ICC llegó a niveles máximos tras el cambio de gobierno, aun cuando la economía empezó a mostrar signos de ralentización desde septiembre del 2018, al término de este subperiodo, el ICC registró también un ciclo de declive con la llegada de la COVID-19 a inicios del 2020.

Revisión de la literatura

Una interpretación general de la confianza del consumidor es que se correlaciona con el consumo privado porque captura las expectativas de los hogares sobre los ingresos futuros, mientras que la hipótesis del ciclo de vida del consumo (Modigliani, 1949) considera que el aumento del valor presente del ingreso es mayor cuando el cambio es transitorio; por otro lado, la hipótesis del ingreso permanente (Friedman, 1957) postuló que el ingreso, y, está integrado por dos componentes: uno permanente (y^p) y otro transitorio (y^t), y que el consumo está determinado solo por el permanente y, dado que los permanentes son mucho más suaves en relación con los actuales o transitorios, las innovaciones en los ingresos generan variaciones relativamente pequeñas en el ingreso permanente y, eventualmente, en el consumo.

Gráfica 1

ICC, Consumo, IGAE, abril del 2002 a septiembre del 2020



Notas: la línea azul representa el Consumo, la roja es IGAE y la negra es ICC; tasas de crecimiento anual para IMCPMI e IGAE, niveles para ICC; series estandarizadas para facilitar la comparabilidad.

Los modelos de consumo recién descritos arrojan luz para los hacedores de política pública (Fernández-Corugedo, 2009). Por ejemplo, con la hipótesis del ciclo de vida se puede analizar el papel que juegan la visión a futuro de los consumidores, o bien, se puede describir cómo reacciona el consumidor ante cambios en la tasa de interés. Asimismo, es posible identificar con estas teorías otros aspectos de la economía que también son relevantes para determinar el consumo, como la distribución de edades, la edad de retiro, la esperanza de vida, entre otras.

Ahora bien, los artículos clásicos que analizan si la confianza del consumidor contiene información sobre modificaciones en el gasto de este son los trabajos de Carroll *et al*. (1994), quienes encuentran que la variable rezagada tiene algún poder explicativo para los cambios actuales en el gasto de los hogares. Además, Ludvigson (2004) da cuenta de que la confianza del consumidor puede pronosticar el consumo. Otro trabajo que se adentra en la anticipación de este es Slacalek (2004), que investiga el desempeño de pronósticos fuera de muestra usando modelos basados en la confianza del consumidor, los cuales se comparan favorablemente con aquellos que no lo consideran.

Artículos más recientes utilizan información no tradicional para hacer *nowcast* del consumo, como Vosen y Schmidt (2012) y Gil *et al*. (2018). En fecha reciente, el uso de información de alta frecuencia

se relaciona con el comportamiento del consumidor, en particular el empleo de índices de consultas en internet de *Google Trends*, que es analizado por Aprigliano *et al*. (2019), utilizando datos de sistemas de pago para pronosticar la actividad económica, incluida la tasa de crecimiento del consumo. En este trabajo también se construye un indicador sintético de sentimiento del consumidor que utiliza datos de búsqueda en la web para explorar alternativas al ICC.

De esta forma, dado nuestro objetivo de maximizar la correlación entre la confianza del consumidor, el consumo privado y la actividad económica en general, proponemos examinar diferentes medidas del sentimiento del consumidor basadas en la ENCO y en tópicos de búsqueda en *Google Trends*.

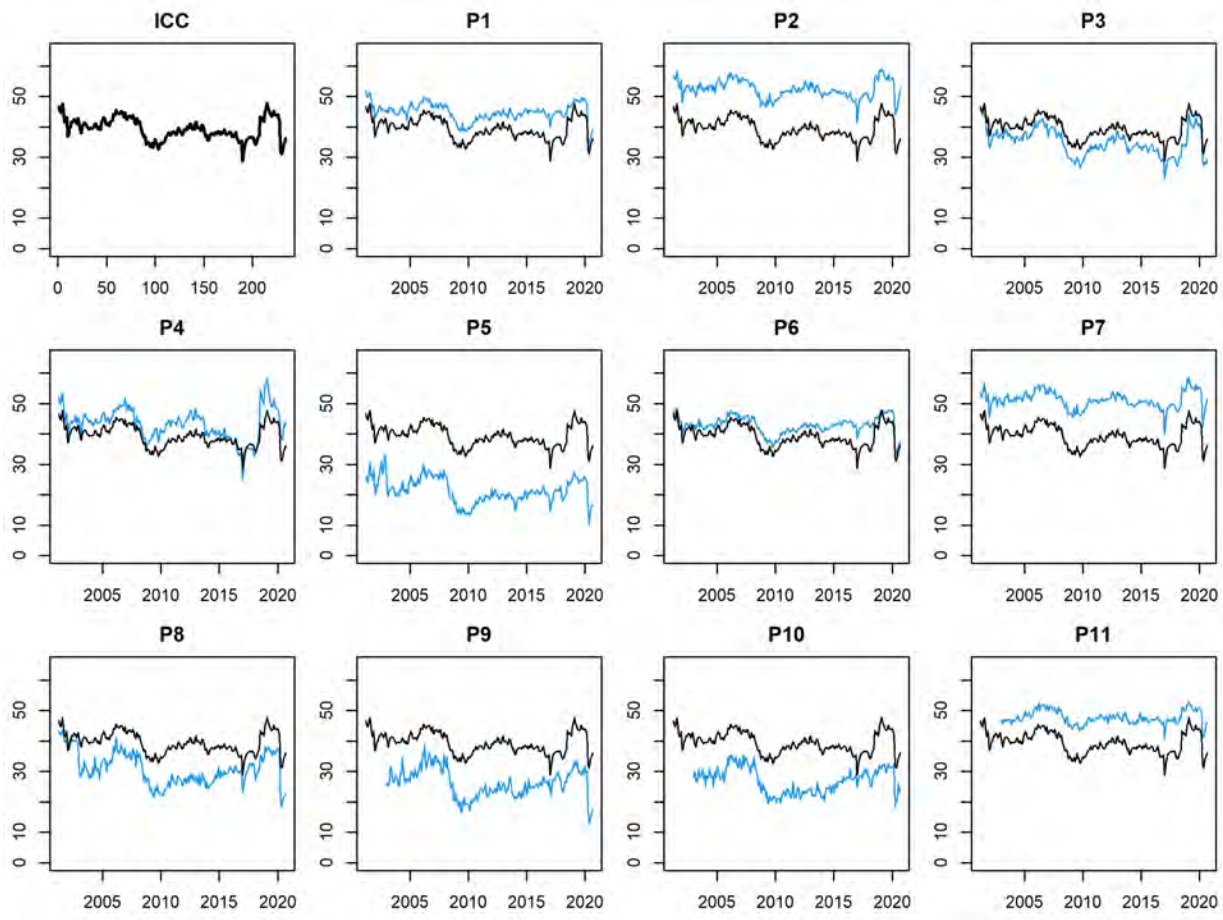
Datos: ENCO y *Google Trends*

Como ya se mencionó, el ICC se construye a partir de cinco preguntas de la ENCO, a las cuales se les aplica un peso igual para cada uno como $ICC = (P1 + P2 + P3 + P4 + P5) / 5$;[4] es decir, para su cálculo se siguen las prácticas internacionales y se le asignan pesos homogéneos a las cinco preguntas que lo componen (INEGI, 2016), no obstante, es posible construir más índices con valor analítico propio, incluso considerando las otras 10 de la Encuesta (Heath, 2012). Las gráficas 2 muestran la evolución de los 15 reactivos de abril del 2001 a septiembre del 2020.

4 En el *Apéndice A* se muestra la redacción de las 15 preguntas.

Gráficas 2 Continúa

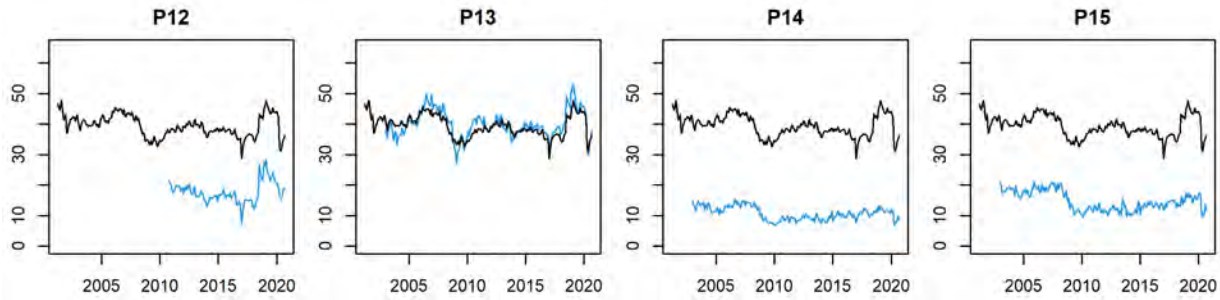
Variables de la ENCO, periodo abril del 2001 a septiembre del 2020



REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Gráficas 2 Concluye

Variables de la ENCO, periodo abril del 2001 a septiembre del 2020



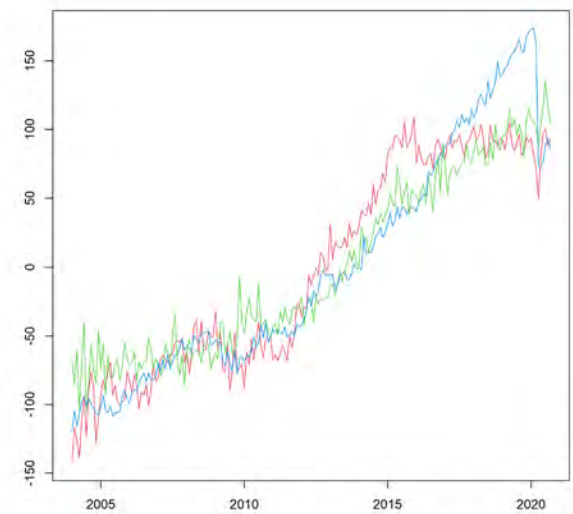
Nota: la línea negra muestra el ICC y la azul es la pregunta correspondiente de la ENCO.

Podemos apreciar que todas las series están linealmente relacionadas, en especial la P3, relativa a la situación económica actual del país comparada con la de hace un año; la P4, que enfatiza en cómo se espera que sean las condiciones respecto a los últimos 12 meses; la P6, relacionada con la situación económica actual del encuestado respecto a hace un año; y, finalmente, la P13, sobre la percepción del encuestado acerca del empleo en el país en los próximos 12 meses.

Por otra parte, la información de *Google Trends* proporciona un índice del volumen relativo de búsquedas realizadas a través de *Google*. El procedimiento para extraer información relacionada con el consumo se basa en definir un conjunto inicial de palabras para las tres categorías de consumo basadas en Aprigliano *et al*. (2019): bienes no duraderos (19 palabras clave), bienes duraderos (21) y servicios (25).[5] Para ampliar el conjunto inicial de estas para cada grupo, se buscan también las consultas relacionadas sugeridas por el motor de búsqueda de *Google*, para finalizar con 62, 73 y 86 palabras, respectivamente; en total, se obtiene información para un total de 221 de enero del 2004 a septiembre del 2020. Estos índices de búsqueda individuales se agregan dentro de cada categoría utilizando la técnica de componentes principales y se desestacionalizan con la metodología *X-13ARIMA-SEATS* (se utiliza la librería *seasonal* de Sax y Eddelbuettel, 2018). Los indicadores de sentimiento del consumidor resultantes basados en *Google Trends* se presentan en la gráfica 3.

Gráfica 3

Indicadores de búsquedas en internet de productos de consumo, periodo enero del 2004 a septiembre del 2020



Nota: la línea roja representa los bienes de consumo duraderos; la verde, los no duraderos; y la azul, los servicios.

Se puede apreciar que las tres series están altamente correlacionadas, aunque la de servicios es la que más resiente la caída al final del periodo de muestra, es decir, a partir de la pandemia de la COVID-19. Las otras dos muestran más resiliencia

5 En el *Apéndice B* se muestran estas 65 palabras.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

a este fenómeno, lo cual puede resultar intuitivo, puesto que es presumible que los servicios es la actividad económica más afectada por esta contingencia sanitaria dadas las restricciones de movilidad impuestas por el gobierno, no así el consumo no duradero. Estas tres series son la base para formular un indicador alternativo de confianza del consumidor.

Metodología

Para construir un indicador de confianza del consumidor altamente correlacionado con el Consumo y el IGAE, utilizamos mínimos cuadrados parciales o PLS (ver, por ejemplo, Wold, 1982 y Wold *et al*., 1987). A través de este tipo de análisis estadístico multivariado podemos modelar la relación entre dos matrices X e Y, es decir, analizar datos de numerosas variables dependientes (X) usualmente correlacionadas entre sí y, al mismo tiempo, diversas de respuesta (Y). El término *parcial* proviene del proceso de estimación, en el cual cada parámetro del modelo es estimado de manera iterativa como la pendiente de una regresión multivariada (por mínimos cuadrados). El resultado son combinaciones lineales de las X, las cuales denotamos como $\hat{f}$, y que pueden ser interpretadas como variables latentes que maximizan la correlación con las Y. De hecho, esta técnica se ha conocido también como proyección a estructuras latentes.

Tal es el caso que nos ocupa, donde X es una matriz T x K que representa el conjunto de preguntas de la ENCO (K=15), o bien, el de indicadores sintéticos obtenidos de *Google Trends* ($K = 3$) con T igual al número de observaciones mensuales, y Y es una matriz de dimensión T x M, que representa al conjunto de variables que miden el estado de la economía, específicamente $M = 2$, que son el Consumo y el IGAE.

De esta forma, seguimos la siguiente estrategia empírica para la construcción de un indicador basado en la ENCO que resulte altamente correlacionado con la situación económica actual:

1. Para la matriz $X_{l,t}$ con $l = 1, 2$, donde la primera matriz es el conjunto de las series de la ENCO y la segunda, la de *Google Trends*, fijamos $l = 1$.
2. Se elige si la transformación de las series económicas, Y, es lineal, o bien, se considera su transformación en diferencias logarítmicas.
3. Para evitar el riesgo de correlaciones espurias, cuando utilizamos series en niveles, verificamos que la regresión entre las variables económicas y $X_{l,t}$ generen residuales estacionarios. Lo anterior a través de la prueba aumentada de Dickey-Fuller.
4. Para $j = 1, 2$ que hace referencia al Consumo y el IGAE, respectivamente, fijamos $j = 1$. Encontramos la t^* en la cual se maximiza el promedio entre $\rho(y_{1,}, X_1)$ donde y_1 es, en este caso, el Consumo. Para determinar t^*, se realiza una búsqueda iterativa para $t = h, h + 1,\ldots,T$, donde $h = 1,\ldots,T_h$, y $T_h = (T|H < T)$, es decir, se realiza una búsqueda iterativa tipo *rolling window* tal que T tenga un número mínimo de muestra. En este caso, definimos $H = 50$.
5. Estimar $\hat{f}$ usando PLS y la matriz (Y, X_{t^*}). Se verifica el porcentaje de varianza explicada por las componentes obtenidas y extraemos la primera componente estimada $\hat{f}_1$.
6. Se calculan las correlaciones entre la matriz de variables económicas y el nuevo indicador estimado, $\rho(Y, \hat{f}_1)$.
7. Se repiten los pasos 4 al 6 para $j = 2$, es decir, se estima un ICC para el t^* tomando como referencia al IGAE.
8. Para el indicador basado en *Google Trends*, se repiten los pasos 1 al 7 para $l = 2$.

Propuesta del Indicador de Expectativas del Consumidor

Con el fin de evaluar los indicadores extraídos con la estrategia empírica, calculamos las correlaciones entre el factor estimado, $\widehat{f_1}$, con las variables económicas de interés, Consumo e IGAE, o sea, $\rho(Y, \widehat{f_1})$. Si la correlación es mayor que la calculada para

$\rho(Y, ICC)$, podemos decir que el nuevo indicador cumple con el objetivo de maximizar la correlación de manera simultánea para estas variables.

En el cuadro 1 se muestran estos cálculos; en la parte superior se presentan los resultados considerando la información de la ENCO y en la inferior, los obtenidos con la información de *Google Trends*. En la primera columna se aprecia el indicador de consumo a considerar, ya sea el obtenido de la estimación , ($\widehat{f_1^{ENCO}}$o $\widehat{f_1^{GT}}$) o bien, el ICC actual. En las columnas 2 y 3 tenemos la correlación cuando el factor se calcula con el tamaño de muestra que maximiza la correlación entre las variables del ICC y el Consumo, y en las 4 y 5 es la que se obtiene cuando el periodo de muestra maximiza la correlación entre las variables del ICC y el IGAE. De la misma manera, las 6 a 9 muestran las correlaciones similares a las columnas anteriormente descritas, pero cuando se especifican las matrices en primeras diferencias logarítmicas.

Para el caso de la información proveniente de la ENCO, la primera componente del PLS es $\widehat{f_1}$, la cual denominaremos ICC'; podemos ver en la parte superior del cuadro 1 que las correlaciones de este indicador con el Consumo y el IGAE rondan sobre el 0.72, aunque las series son más cortas que el ICC actual, el cual comienza desde abril del 2001. Sin embargo, para el periodo de muestra común, podemos ver que las correlaciones se comparan de manera favorable con respecto a las que obtendríamos con el ICC oficialmente publicado por el INEGI, pues, en este caso, se obtienen correlaciones de alrededor de 0.50. Nótese que, en todos los casos, las regresiones entre el factor y las variables económicas generan residuales estacionarios, por lo tanto, las correlaciones no son espurias. Si consideramos las primeras diferencias logarítmicas, las correlaciones son radicalmente más grandes para el ICC' que para el ICC actual, lo cual genera más evidencia en favor del ICC'.

Cuando utilizamos información de *Google Trends*, los resultados en niveles son incluso mejores que el ICC', incluido el hecho de que tenemos series más largas, pero en primeras diferencias, los resultados son algo inciertos, pues no encontramos concordancia en las correlaciones obtenidas cuando seleccionamos el periodo de muestra que maximi-

Cuadro 1

Resultados de la estrategia empírica

Con información de la ENCO								
	Variables en niveles				Variables en diferencias logarítmicas			
	$t^*_{y_1}$ = 2016/07		$t^*_{y_2}$ = 2013/11		$t^*_{y_1}$ = 2016/08		$t^*_{y_2}$ = 2016/08	
Variables	Consumo	IGAE	Consumo	IGAE	Consumo	IGAE	Consumo	IGAE
$ICC' = \widehat{f_1^{ENCO}}$	0.75	0.70	0.73	0.72	0.61	0.57	0.61	0.57
	(0.01)	(0.02)	(0.03)	(0.07)				
ICC	0.56	0.50	0.51	0.48	0.26	0.24	0.19	0.17
Con información de *Google Trends*								
	Variables en niveles				Variables en diferencias logarítmicas			
	$t^*_{y_1}$ = 2004/01		$t^*_{y_2}$ = 2004/01		$t^*_{y_1}$ = 2008/08		$t^*_{y_2}$ = 2016/08	
Variables	Consumo	IGAE	Consumo	IGAE	Consumo	IGAE	Consumo	IGAE
$GT = \widehat{f_1^{GT}}$	0.89	0.92	0.89	0.92	-0.03	-0.14	0.82	0.80
	(0.01)	(0.02)	(0.01)	(0.02)				
ICC	-0.57	-0.60	-0.57	-0.60	-0.14	-0.13	0.13	0.18

Nota: los números entre paréntesis son los valores p de la prueba de raíz unitaria de Dickey Fuller aplicada a los residuos de y_i con el factor estimado $\widehat{f_1}$, donde $\widehat{f_1}$ es la primera componente estimada por PLS para las matrices$X_{l,t}$, mientras que $t^*_{y_1}$ indica que la serie se considera desde el periodo t^*, donde se maximiza la correlación entre $\rho\,(y_j, X_{l,t})$.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

za la correlación cuando la guía es el Consumo o el IGAE. En el primer caso, no se puede garantizar que exista una correlación positiva.

Adicionalmente, se encuentra que la varianza explicada por el primer factor para todos los casos es de, al menos, 79 % de la variabilidad de la matriz Y.

Los resultados del cuadro 1 pueden corroborarse a partir de las gráficas C1 a C4 del *Apéndice C*; ahí se puede observar que el indicador ICC′, generado a partir de la ENCO, tiene un comportamiento más correlacionado cuando consideramos las variables Y en niveles que en primeras diferencias. De la misma manera, para el indicador GT, tiene más correlación cuando tomamos también las variables en niveles.

Además, en las gráficas D1 a D4 del *Apéndice D* se observan las contribuciones de cada una de las variables de la ENCO y *Google Trends* a los indicadores ICC′ y GT construidos. Se aprecia que, para las variables de la ENCO, todas contribuyen de manera positiva. Por último, en las mismas gráficas podemos ver que para el indicador GT, la componente asociada a los servicios tiene mayor peso, seguida de los bienes de consumo duradero y, finalmente, de los no durables; también, para la especificación en primeras diferencias, el *subfactor* atribuible a los no duraderos tiene carga negativa, lo que resulta complicado interpretar de forma económica.

En consecuencia, podemos concluir que el ICC′ muestra resultados más consistentes en todos los casos, pues las correlaciones son mayores al ICC y positivas tanto en niveles como en primeras diferencias; asimismo, la conformación de los indicadores tiene una explicación estructural.

Conclusiones

Este trabajo contribuye al mejor aprovechamiento de la información contenida en la ENCO y utiliza, además, fuentes no tradicionales para construir indicadores alternativos del ICC altamente correlacionados con la situación económica. Podemos ver que utilizando la metodología PLS es posible aprovechar de mejor manera toda la información disponible de la ENCO y de otros insumos para correlacionar al mismo tiempo un conjunto de variables dependientes con otro de variables independientes.

En específico, utilizamos toda la información de la Encuesta para formar un nuevo indicador al que denominamos ICC′ en lugar de solo las cinco preguntas que en la actualidad se consideran. Se encontró que es posible construir uno alternativo para el ICC que presente una correlación de, al menos, 0.73 con el Consumo y 0.70 con el IGAE, la cual es superior a la que muestra el ICC actual de, al menos, 0.45 y 0.42, respectivamente.

También, se explotó información alternativa a la ENCO, la cual proviene de índices de búsquedas en internet de productos de consumo y servicios, *Google Trends*, a partir de la cual se puede elaborar otro Indicador de Confianza del Consumidor muy correlacionado con el par de variables económicas antes mencionadas. En este caso, el denominado GT alcanza correlaciones de 0.89 y 0.92 con el Consumo y el IGAE, respectivamente. No obstante, al analizar las cargas que forman a cada uno de los indicadores, los resultados son más consistentes si usamos la ENCO dado que la especificación en primeras diferencias genera cargas no interpretables; para el GT, en particular, se obtienen cargas negativas para los bienes no duraderos. De esta forma, es más recomendable usar el ICC′.

Un aspecto relevante para futuras investigaciones es la elaboración de indicadores de confianza del consumidor, pero que maximicen el poder predictivo utilizando para ello modelos de pronóstico de corto plazo, como los de *nowcasting*. También, se puede explorar las posibilidades de combinar los resultados de la ENCO y las búsquedas de *Google Trends*, ya que se encontró en este trabajo que ambas fuentes presentan altas correlaciones con las variables de interés.

Fuentes

Aldrete, J. C. y F. Flores. *Confianza del consumidor: retroceso por segundo mes consecutivo en abril. 2019* (DE) https://www.banorte.com/cms/casadebolsabanorteixe/analisisyestrategia/analisiseconomico/mexico/20190506 ICC Abr 19.pdf, consultado el 14/10/2020.

Aprigliano, V., G. Ardizzi y L. Monteforte. "Using the payment system data to forecast the economic activity", en: *International Journal of Central Banking*. WP, 2019, p. 1098.

BANAMEX. *Examen de la situación económica de México, segundo trimestre 2019*. Número 1073, volumen XCIV, 2019 (DE) https://www.notimx.mx/2019/06/citibanamex-examen-de-la-situacion.html, consultado el 14/10/2020.

Carroll, C. D., J. C. Fuhrer, and D. W. Wilcox. "Does consumer sentiment forecast household spending? if so, why?", en: *The American Economic Review*. 84(5), 1994, pp. 1397-1408.

Friedman, M. *A theory of the consumption function*. Princeton University Press, 1957.

Fernández-Corugedo, E. *Teoría del consumo*. Centro de Estudios Monetarios Latinoamericanos, 2009.

Gil, M., J. J. Pérez, A. J. Sánchez Fuentes, and A. Urtasun. *Nowcasting private consumption: traditional indicators, uncertainty measures, credit cards and some internet data*. 2018.

INEGI. *Encuesta Nacional sobre Confianza del Consumidor 2015*. *Documento metodológico*. Aguascalientes, México, INEGI, 2016.

Leyva, G., O. Páez, and S. M. Esperanza. "Un umbral empírico y otras recomendaciones para el reporte de la confianza del consumidor en México", en: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. 7(1), 2016, pp. 112-121.

Ludvigson, S. C. "Consumer confidence and consumer spending", en: *Journal of Economic perspectives*. 18(2), 2004, pp. 29-50.

Modigliani, F. "Fluctuations in the Saving-Income Ratio: A Problem in Economic Forecasting", en: *Studies in Income and Wealth*. Vol. 11. National Bureau of Economic Research, 1949, pp. 371-443.

Sax, C., y D. Eddelbuettel. "Seasonal adjustment by x-13arima-seats", en: *Journal of Statistical Software*. 87(1), 2018, pp. 1-17.

Slacalek, J. *Forecasting consumption*. German Institute for Economic Research, DIW, Department of Macro Analysis and Forecasting, 2004.

Vosen, S. and T. Schmidt. "A monthly consumption indicator for Germany based on internet search query data", en: *Applied Economics Letters*. 19(7), 2012, pp. 683-687.

Wold, H. "Soft modeling: the basic design and some extensions", en: Joreskog, K.-G. y H. Wold, Eds. *Systems Under Indirect Observation*. Vol. 2. North-Holland, Amsterdam, 1982, pp. 1-53.

Wold, S., S. T. L. Hellberg, M. Sjostrom y H. Wold. *PLS Model Building: Theory and applications. PLS modeling with latent variables in two or more dimensions*. Frankfurt am Main, 1987.

Apéndice A. ENCO

- P1. Comparada con la situación económica que los miembros de este hogar tenían hace 12 meses, ¿cómo cree que es su situación en este momento?
- P2. ¿Cómo considera usted que será la situación económica de los miembros de este hogar dentro de 12 meses, respecto a la actual?
- P3. ¿Cómo considera usted la situación económica del país hoy en día comparada con la de hace 12 meses?
- P4. ¿Cómo considera usted que será la condición económica del país dentro de 12 meses, respecto de la situación actual?
- P5. Comparando la situación económica actual con la de hace un año, ¿cómo considera en el momento actual las posibilidades de que usted o alguno de los integrantes de este hogar realice compras tales como muebles, televisor, lavadora, otros aparatos electrodomésticos, etcétera?
- P6. ¿Cómo describiría usted su situación económica comparada con la de hace 12 meses?
- P7. ¿Y cómo cree usted que será su situación económica dentro de 12 meses respecto de la actual?
- P8. ¿En este momento tiene usted mayores posibilidades de comprar ropa, zapatos, alimentos, etc., que hace un año?
- P9. ¿Considera usted que durante los próximos 12 meses usted o alguno de los integrantes de este hogar tendrán posibilidades económicas para salir de vacaciones?
- P10. ¿Actualmente usted tiene posibilidades de ahorrar alguna parte de sus ingresos?
- P11. ¿Cómo considera usted que serán sus condiciones económicas para ahorrar dentro de 12 meses comparadas con las actuales?
- P12. Comparando con los 12 meses anteriores, ¿cómo cree usted que se comporten los precios en el país en los siguientes 12 meses?
- P13. ¿Cree usted que el empleo en el país en los próximos 12 meses (aumentará, permanecerá igual, disminuirá...)?

- P14. ¿Algún miembro de este hogar o usted están planeando comprar un automóvil nuevo o usado en los próximos 2 años?
- P15. ¿Algún miembro de este hogar o usted están planeando comprar, construir o remodelar una casa en los próximos 2 años?

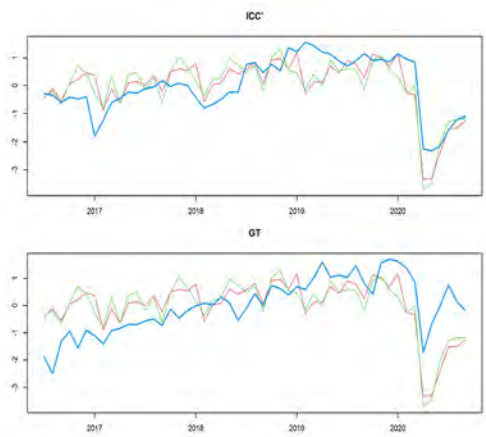
Apéndice B. Palabras clave *Google Trends*

Categorías de consumo	Palabras clave *Google Trends*
Durables	
Vehículos	Compra auto, auto usado, Mazda, Volkswagen, Ford, Chevrolet, Nissan, Tsuru.
Muebles	Electrodomésticos, muebles, decoración interior, diseño interior.
Recreación	Videojuego.
Otros	Celular, libro electrónico.
No durables	
Alimentos y bebidas	Alimentos y bebidas, comida, bebidas.
Ropa y calzado	Ropa, ropa segunda mano, zapatos, zapatos segunda mano, lencería, ropa interior, playera.
Energéticos	Electricidad, pago luz, gasolina, pago agua.
Otros	Medicinas, cuidado del cuerpo, cuidado facial, productos de belleza, cigarros.
Servicios	
Vivienda	Seguro de casa, decoración interior, agencia bienes raíces.
Salud	Salud, seguro de gastos médicos, doctores.
Transportación	Metro, avión, renta carro.
Recreación	Descanso, videojuegos, película en línea, comprar película, mirar película.
Alojamiento y alimentación	Venta de boletos, hoteles, moteles, restaurante, terraza, bienestar.
Servicios financieros y de seguros	Comisiones bancarias, créditos.
Otros	Telecomunicaciones, seguro de vida, servicios de asistencia.

Apéndice C

Gráficas C1

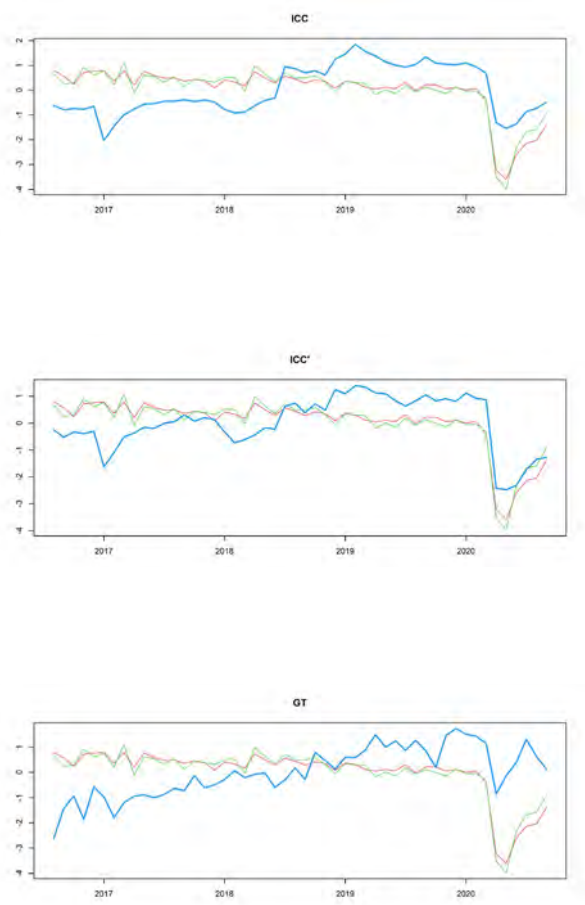
Estimación usando $t^*_{y_1}$, series en niveles



Nota: parte superior ICC original, media ICC alternativo y baja indicador construido con GT denotados con la línea azul; las líneas roja y verde representan el Consumo y el IGAE, respectivamente.

Gráficas C2

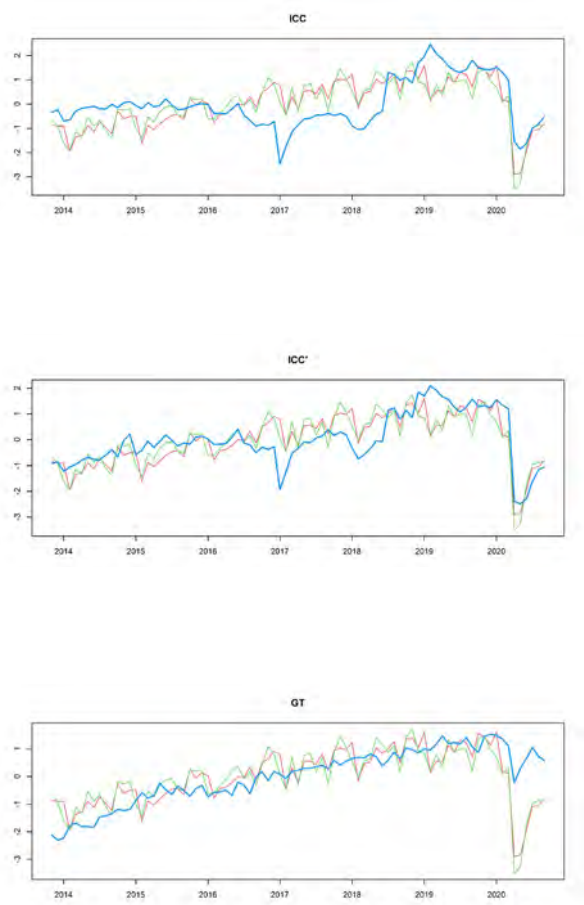
Estimación usando t^*_{y1}, primeras diferencias logarítmicas



Nota: parte superior ICC original, media ICC alternativo y baja indicador construido con GT denotados con la línea azul; las líneas roja y verde representan el Consumo y el IGAE, respectivamente.

Gráficas C3

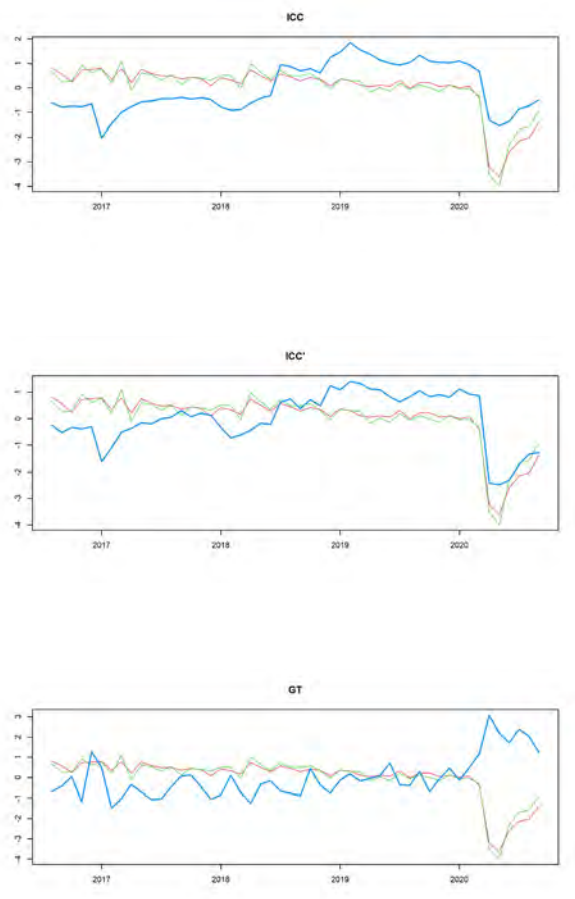
Estimación usando t^*_{y2}, series en niveles



Nota: parte superior ICC original, media ICC alternativo y baja indicador construido con GT denotados con la línea azul; las líneas roja y verde representan el Consumo y el IGAE, respectivamente.

Gráficas C4

Estimación usando $t^*_{y_2}$, primeras diferencias logarítmicas

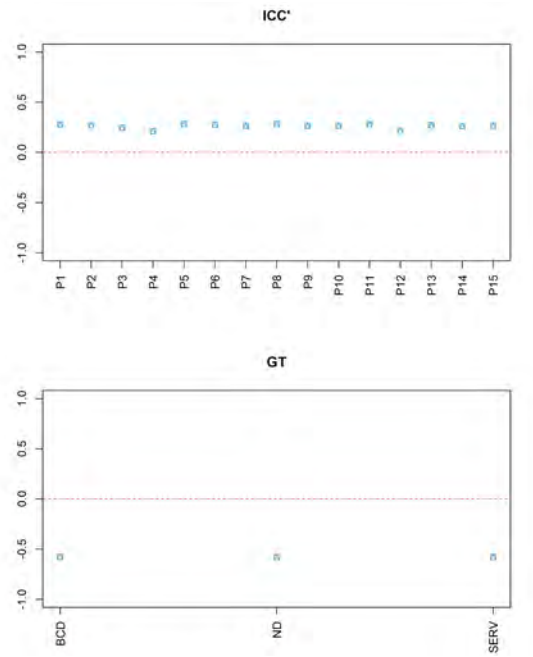


Nota: parte superior ICC original, media ICC alternativo y baja indicador construido con GT denotados con la línea azul; las líneas roja y verde representan el Consumo y el IGAE, respectivamente.

Apéndice D

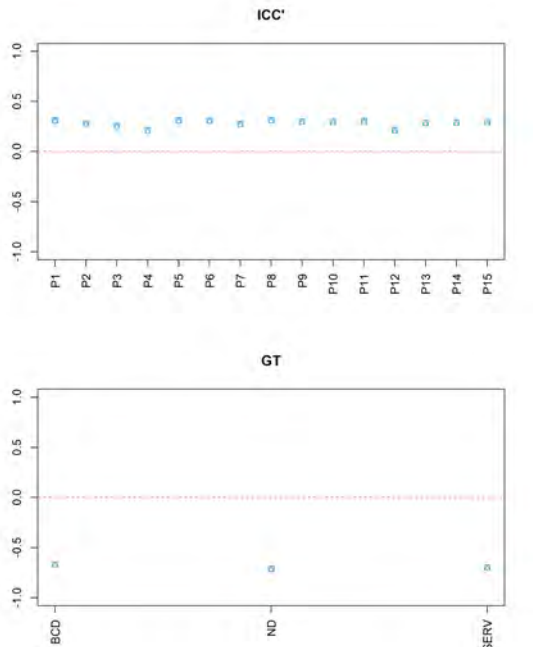
Gráficas D1

Cargas PLS usando $t^*_{y_1}$, series en niveles



Gráficas D2

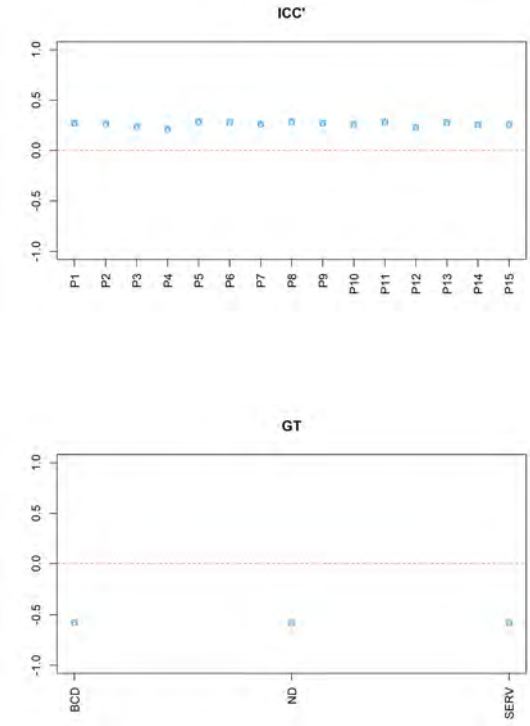
Cargas PLS usando $t^*_{y_1}$, primeras diferencias logarítmicas



REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

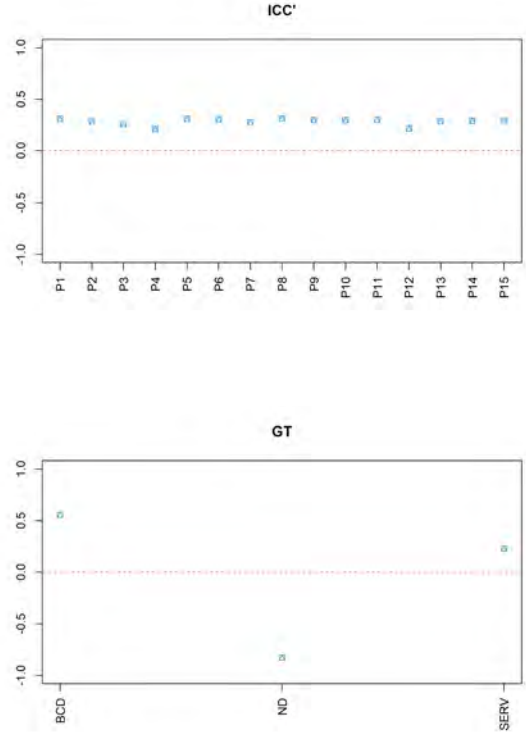
Gráficas D3

Cargas PLS usando $t^*_{y_2}$, series en niveles



Gráficas D4

Cargas PLS usando $t^*_{y_2}$, series en primeras diferencias logarítmicas



Esperanza de vida ***sin limitaciones físicas ni mentales*** en México

Life Expectancy ***without Physical or Mental Limitations*** in Mexico

Olinca Páez*

* Instituto Nacional de Estadística y Geografía (INEGI), olinca.paez@inegi.org.mx
Nota del autor: agradezco la asistencia de Claudia Cerón, así como los comentarios de Gerardo Leyva, Mauricio Rodríguez, Abigail Rojas y Adriana Pérez.

Health Insurance Concept arranging wood block stacking with icon healthcare medical./ stockphoto/ iStock

En este artículo se presentan estimaciones de esperanza de vida sin limitaciones físicas ni mentales en México producto de la aplicación del método de Sullivan en tablas de vida periodo construidas directamente con información censal y estadísticas de mortalidad. Se trata de una propuesta original para el uso inmediato de la información que produce y concentra el Instituto Nacional de Estadística y Geografía, que considera la diversidad nacional según sexo y entidad de residencia en el diseño de este indicador sintético de salud, que refleja la calidad de vida y no solo la expectativa de su duración.

El indicador está basado en la información sobre prevalencia de limitaciones físicas y mentales junto con estadísticas de mortalidad, y puede considerarse un *proxy* de la esperanza de vida saludable. Los resultados obtenidos son consistentes tanto con las estimaciones nacionales del Consejo Nacional de Población como con las internacionales de la Organización Mundial de la Salud y el Instituto para la Medición y Evaluación de la Salud, con las ventajas de oportunidad y desagregación simultáneas en nuestro caso. Además, en los periodos intercensales, es factible estimar este indicador gracias a que la información necesaria se genera hoy en día en varias encuestas nacionales, como la ENADID y la ENASEM.

Palabras clave: esperanza de vida saludable; EVISA; esperanza de vida libre de discapacidad; Sullivan; indicadores de salud.

Recibido: 4 de septiembre de 2020.
Aceptado: 10 de noviembre de 2021.

This article presents estimates of Life Expectancy Without Physical or Mental Limitations (EVSL for its Spanish acronym) in Mexico, resulting from the application of Sullivan's method to period-life tables constructed directly with census information and mortality statistics. This is an original proposal for the immediate use of the information produced and concentrated by INEGI, which considers the national diversity by sex and place of residence in the design of a synthetic health-indicator that reflects the quality of life and not only life expectancy.

This indicator is based on information on the prevalence of physical and mental limitations together with mortality statistics and can be considered a proxy for healthy life expectancy (HALE). The results obtained are consistent with the national CONAPO estimates, and those of the WHO and the Institute for Health Metrics and Evaluation (IHME), with the advantages of simultaneous timeliness and disaggregation in our case. In addition, in intercensal periods, it is feasible to estimate this indicator because the necessary information is now generated in several national surveys, such as ENADID and ENASEM.

Key words: healthy life expectancy; HALE; disability-free life expectancy; Sullivan; health indicators.

Esperanza de vida saludable en México

Los trabajos para desarrollar indicadores que integran la prevalencia de padecimientos no fatales a las estadísticas de mortalidad datan de la década de los 60 del siglo pasado (Chiang, 1965; Moriyama, 1968; Sullivan, 1966 y 1971), y han resultado a la fecha en la consolidación de dos mediciones estándar en el ámbito internacional: la esperanza de vida saludable (EVISA o HALE, por sus siglas en inglés) y la esperanza de vida ajustada por discapacidad (EVAD o DALE). La idea detrás de la EVISA es la de descontar del promedio de vida esperado el tiempo vivido en malas condiciones de salud, por enfermedad o lesiones. Por su parte, la EVAD equipara esas malas condiciones con la discapacidad. Es decir, aunque ambas derivan de la esperanza de vida, la EVAD considera solo una fuente de pérdida de años saludables, aunque quizá la más evidente, la falta o limitación de alguna facultad física o mental.

La Organización Mundial de la Salud (OMS) estima y difunde cada cinco años los indicadores de EVISA al nacer y a los 60 años de edad, para los países miembros, las regiones y el mundo. Por su parte, el Instituto para la Medición y Evaluación de la

Salud (IHME, por sus siglas en inglés), un centro de investigación sobre la salud mundial en la Universidad de Washington, produce estimaciones anuales de EVISA a diferentes edades, desagregadas por entidad y sexo.

El objetivo de este trabajo es mostrar la capacidad nacional de producir una estadística semejante, pero más oportuna, directamente con los datos generados y concentrados por el Instituto Nacional de Estadística y Geografía (INEGI). Algunas investigaciones han hecho ejercicios semejantes para México, aunque para un subconjunto de la población, como la de derechohabientes del Instituto Mexicano del Seguro Social (Rodríguez-Abrego *et al*., 2006) o enfocándose en enfermedades prevalentes, como la diabetes mellitus (Andrade, 2009). En la esfera internacional, la estimación de este tipo de indicadores ha servido para estudiar la desigualdad entre grupos de la población y países (Santosa *et al*., 2016; Seuc & Domínguez, 2003; Sullivan, 1971).

El artículo explica el procedimiento para calcular de manera periódica un indicador del número de años que pueden esperar vivir las personas sin limitación física ni mental, por entidad y sexo; señala los aspectos a considerar para producir estimaciones robustas que al mismo tiempo puedan replicarse fácilmente; expone resultados de la aplicación de esta metodología, así como un breve análisis de los cambios reflejados por el indicador; comenta acerca de la validez de estas estimaciones, las limitaciones del ejercicio y algunas posibilidades de aplicación; y, por último, presenta conclusiones y algunas recomendaciones.

Tablas de vida periodo ajustadas por el método de Sullivan

Aquí se explica de forma concisa una técnica para estimar la esperanza de vida, el ajuste propuesto por Sullivan y su aplicación con datos para México.

La esperanza de vida puede calcularse con tablas de vida periodo bajo el supuesto de que la experiencia de la mortalidad es relativamente estable en el corto plazo. Consiste en calcular las tasas de mortandad por grupo de edad, en el momento de referencia, para inferir el patrón específico en esa población en particular y estimar cuántos años viviría una persona si a cada edad se enfrentara a las pautas que han experimentado las generaciones previas.

Para un año específico, se parte del volumen estimado de la población a mitad de año (${}_nP_x$), distribuido por edades (x, $x+n$) y del número de defunciones (${}_nD_x$) ocurridas en el periodo a cada edad, para obtener las tasas de mortalidad (${}_nm_x$):

$$m_x^n = \frac{D_x^n}{P_x^n}$$

Estas tasas son transformadas en cocientes de mortalidad (${}_nq_x$), que expresan el riesgo de muerte entre las edades x y $x+n$ en una cohorte sintética de 100 mil personas (l_0):

$$q_x^n = \frac{m_x^n}{l_x}$$

donde l_x es el número de sobrevivientes en la cohorte sintética a cada edad x.

Con el número de sobrevivientes a cada edad (l_x) y el tiempo que sobrevivieron (n) se obtienen los años-persona vividos en cada edad (${}_nL_x$) y, a partir de ahí, su acumulado (T_x), que es la suma de años vividos por el conjunto de sobrevivientes de la cohorte sintética:

$${}_nL_x = l_{x+n} \bullet n$$

$$T_x = \sum_{x=0}^{\omega} {}_nL_x$$

La esperanza de vida al nacer (E_0) se obtiene dividiendo la suma de años vividos por el conjunto de la población inicial (T_0) entre 100 mil, mientras que la esperanza de vida a cualquier edad (E_x) se puede calcular dividiendo el número de años vividos por el total de la población que alcanzó esa edad (T_x) entre el número de sobrevivientes (l_x):

$$E_0 = \frac{T_0}{100\,000}$$

$$E_x = \frac{T_x}{l_x}$$

El cuadro 1 muestra la aplicación de la técnica al caso de las mujeres en México en el 2020.

Sullivan (1966 y 1971) propone el uso de estadísticas de discapacidad para el desarrollo de un indicador de salud que integre los patrones de mortalidad y morbilidad de forma conjunta. Su ajuste consiste en multiplicar la esperanza de vida a cada edad (E_x) por la proporción de años-persona que ese grupo de individuos vive sin discapacidad (I_x):

$$E_s = E_x \cdot I_x$$

$$I_x = 1 - \frac{w_x}{365}$$

donde w_x es el número de días con discapacidad por persona por año a cada edad.

Cuadro 1

Tabla de vida periodo para el cálculo de la esperanza de vida de las mujeres en México, 2020

Edad, x	n	${}_nP_x$	${}_nD_x$	${}_nm_x$	${}_nq_x$	${}_np_x$	$l(x)$	${}_nd_x$	${}_nL_x$	$T(x)$	$E(x)$
< 1	1	902 250	7 074	0.0078	0.01	0.99	100 000	778	99 280	7 848 314	78.5
1 a 4	4	4 097 628	1 886	0.0005	0.00	1.00	99 222	182	396 432	7 749 034	78.1
5 a 9	5	5 343 344	1 091	0.0002	0.00	1.00	99 039	101	494 929	7 352 602	74.2
10 a 14	5	5 421 806	1 254	0.0002	0.00	1.00	98 938	114	494 431	6 857 673	69.3
15 a 19	5	5 376 796	2 433	0.0005	0.00	1.00	98 824	223	493 604	6 363 241	64.4
20 a 24	5	5 287 934	3 376	0.0006	0.00	1.00	98 600	314	492 257	5 869 637	59.5
25 a 29	5	5 162 568	4 162	0.0008	0.00	1.00	98 286	395	490 482	5 377 380	54.7
30 a 34	5	4 922 633	4 730	0.0010	0.00	1.00	97 891	469	488 338	4 886 899	49.9
35 a 39	5	4 717 044	6 203	0.0013	0.01	0.99	97 421	639	485 620	4 398 560	45.1
40 a 44	5	4 468 087	8 747	0.0020	0.01	0.99	96 783	943	481 746	3 912 940	40.4
45 a 49	5	4 154 996	12 966	0.0031	0.02	0.98	95 840	1 485	475 748	3 431 194	35.8
50 a 54	5	3 727 732	17 291	0.0046	0.02	0.98	94 355	2 165	466 780	2 955 446	31.3
55 a 59	5	3 021 106	23 084	0.0076	0.04	0.96	92 190	3 460	452 841	2 488 667	27.0
60 a 64	5	2 578 670	29 251	0.0113	0.06	0.94	88 730	4 901	432 019	2 035 825	22.9
65 a 69	5	1 949 925	33 300	0.0171	0.08	0.92	83 829	6 877	402 680	1 603 807	19.1
70 a 74	5	1 422 381	36 372	0.0256	0.12	0.88	76 953	9 268	362 429	1 201 127	15.6
75 a 79	5	972 518	39 426	0.0405	0.18	0.82	67 685	12 486	307 995	838 698	12.4
80 a 84	5	655 484	42 538	0.0649	0.28	0.72	55 199	15 422	237 650	530 703	9.6
85 a 89	5	378 163	39 231	0.1037	0.41	0.59	39 776	16 310	157 216	293 052	7.4
90 a 94	5	160 410	27 383	0.1707	0.59	0.41	23 466	13 813	80 916	135 837	5.8
95 +		70 665	12 421	0.1758	1.00	0.00	9 654	9 654	54 921	54 921	5.7

Fuente: elaboración propia con datos del Censo de Población y Vivienda 2020 y la Base de Datos Nacional del Registro Civil.

Con este cálculo se obtiene la esperanza de vida libre de discapacidad (E_s) y por diferencia con la esperanza de vida, la esperanza de años vividos con discapacidad (Z_x):

$$Z_x = E_x - E_s$$

La información disponible para el caso mexicano obliga a deducir la proporción de años-persona vividos sin discapacidad (I_x) a partir de la prevalencia de esta, suponiendo que tales limitaciones se mantuvieron constantes durante todo el año. El cuadro 2 muestra el ajuste de Sullivan para obtener la esperanza de vida sin discapacidad para las mujeres en México en el 2020.

Estrategia para un diseño robusto y replicable

El ajuste de Sullivan puede aplicarse considerando diferentes definiciones de discapacidad, y de ahí que pueda haber distintas estimaciones resultantes. En algunos países, el instrumento a partir del cual se le define es el que considera las actividades básicas de la vida diaria (ADL, por sus siglas en inglés). Por

Cuadro 2

Tabla de vida periodo ajustada por el método de Sullivan para el cálculo de la esperanza de vida sin discapacidad de las mujeres en México, 2020

Edad, x	n	${}_nP_x$	…	${}_nL_x$	$T(x)$	$E(x)$	Individuos con discapacidad	Prevalencia discapacidad I_x	Años-persona con discapacidad ${}_nL_x * I_x$	Años-persona sin discapacidad ${}_nL_x*(1-I_x)$	$T(x\|I')$	$E(s)$
< 1	1	902 250	…	99 280	7 848 314	78.5	23 298	0.03	2 564	96 716	7 092 559	70.9
1 a 4	4	4 097 628	…	396 432	7 749 034	78.1	103 018	0.03	9 967	386 465	6 995 842	70.5
5 a 9	5	5 343 344	…	494 929	7 352 602	74.2	123 580	0.02	11 447	483 482	6 609 377	66.7
10 a 14	5	5 421 806	…	494 431	6 857 673	69.3	136 010	0.03	12 403	482 028	6 125 894	61.9
15 a 19	5	5 376 796	…	493 604	6 363 241	64.4	145 385	0.03	13 347	480 257	5 643 866	57.1
20 a 24	5	5 287 934	…	492 257	5 869 637	59.5	137 493	0.03	12 799	479 458	5 163 609	52.4
25 a 29	5	5 162 568	…	490 482	5 377 380	54.7	128 116	0.02	12 172	478 310	4 684 151	47.7
30 a 34	5	4 922 633	…	488 338	4 886 899	49.9	125 878	0.03	12 487	475 851	4 205 841	43.0
35 a 39	5	4 717 044	…	485 620	4 398 560	45.1	129 091	0.03	13 290	472 330	3 729 990	38.3
40 a 44	5	4 468 087	…	481 746	3 912 940	40.4	159 543	0.04	17 202	464 544	3 257 660	33.7
45 a 49	5	4 154 996	…	475 748	3 431 194	35.8	202 765	0.05	23 217	452,531	2 793 116	29.1
50 a 54	5	3 727 732	…	466 780	2 955 446	31.3	253 340	0.07	31 723	435 057	2 340 585	24.8
55 a 59	5	3 021 106	…	452 841	2 488 667	27.0	271 117	0.09	40 638	412 203	1 905 528	20.7
60 a 64	5	2 578 670	…	432 019	2 035 825	22.9	309 321	0.12	51 822	380 197	1 493 325	16.8
65 a 69	5	1 949 925	…	402 680	1 603 807	19.1	311 531	0.16	64 334	338 345	1 113 129	13.3
70 a 74	5	1 422 381	…	362 429	1 201 127	15.6	299 887	0.21	76 412	286 016	774 783	10.1
75 a 79	5	972 518	…	307 995	838 698	12.4	277 816	0.29	87 984	220 011	488 767	7.2
80 a 84	5	655 484	…	237 650	530 703	9.6	255 243	0.39	92 540	145 110	268 755	4.9
85 a 89	5	378 163	…	157 216	293 052	7.4	191 825	0.51	79 748	77 467	123 645	3.1
90 a 94	5	160 410	…	80 916	135 837	5.8	99 863	0.62	50 374	30 542	46 178	2.0
95 +		70 665	…	54 921	54 921	5.7	50 547	0.72	39 285	15 636	15 636	1.6

Fuente: elaboración propia con datos del Censo de Población y Vivienda 2020 y la Base de Datos Nacional del Registro Civil.

ejemplo, las personas con discapacidad pueden ser quienes reportan dificultad moderada, severa o extrema, en los últimos 30 días, para caminar una larga distancia, bañarse, vestirse, moverse de una habitación a otra, comer e ir al baño (Santosa *et al*., 2016).

En los programas estadísticos del INEGI (2010, 2014 y 2020) en los que se ha incluido el tema de discapacidad, el diseño conceptual sigue las recomendaciones del Grupo de Washington (Washington Group on Disability Statistics, 2017), de manera que la pregunta abarca, además de las actividades básicas de la vida diaria, otras actividades y funciones.

El cuadro 3 resume las semejanzas y diferencias en los cuestionarios, así como en la operacionalización de las variables; además, contiene anotaciones acerca de los detalles que debieron tomarse en cuenta para equiparar la medición de las prevalencias en los tres programas estadísticos.

Por ejemplo, mientras que la información proveniente de los censos de población puede consultarse desagregada por edad, sexo y entidad directamente a través de la herramienta de tabulados interactivos, para el caso de la Encuesta Nacional de la Dinámica Demográfica (ENADID) fue necesario acceder a los microdatos, calcular las prevalencias de discapacidad, y aplicarlas al volumen y estructura estimados de la población en el 2014.

En el 2014 y 2020, la opción de declarar poca dificultad para llevar a cabo las actividades enuncia-

Cuadro 3 **Continúa**

Preguntas sobre discapacidad en diferentes programas estadísticos

Censo de Población y Vivienda 2020	Respuestas	Notas
En su vida diaria, ¿(NOMBRE) cuánta dificultad tiene para: ver, aun usando lentes? oír, aun usando aparato auditivo? caminar, subir o bajar? recordar o concentrarse? bañarse, vestirse o comer? hablar o comunicarse (por ejemplo: entender o ser entendido por otros)? ¿Tiene algún problema o condición mental? (autismo, síndrome de Down, esquizofrenia, etcétera)	1. No tiene dificultad 2. Lo hace con poca dificultad 3. Lo hace con mucha dificultad 4. No puede hacerlo ***** 5. Sí 6. No	La consulta en tabulados dinámicos permite seleccionar a las personas que tienen alguna limitación o discapacidad y además algún problema o condición mental. Para calcular la prevalencia, de ese conjunto de personas se excluyó a quienes solo presentaban limitación (sin discapacidad ni problema o condición mental).
Censo de Población y Vivienda 2010		
En su vida diaria, ¿(NOMBRE) tiene dificultad al realizar las siguientes actividades: caminar, moverse, subir o bajar? ver, aun usando lentes? hablar, comunicarse o conversar? oír, aun usando aparato auditivo? vestirse, bañarse o comer? poner atención o aprender cosas sencillas? ¿Tiene alguna limitación mental? Entonces, ¿no tiene dificultad física o mental?	0. No 1. Sí	Para calcular la prevalencia, se consideraron todas las personas con limitación en la actividad.

Cuadro 3 Concluye

Preguntas sobre discapacidad en diferentes programas estadísticos

Encuesta Nacional de la Dinámica Demográfica 2014		
Por algún problema de nacimiento o de salud, ¿cuánta dificultad tiene (NOMBRE), para:	1. No puede hacerlo 2. Lo hace con mucha dificultad 3. Lo hace con poca dificultad 4. No tiene dificultad	Para calcular la prevalencia, se consideraron todas las personas que no pueden hacer o hacen con mucha dificultad cualquiera de las actividades mencionadas.
caminar, subir o bajar usando sus piernas? ver (aunque use lentes)? mover o usar sus brazos o manos? aprender, recordar o concentrarse? escuchar (aunque use aparato auditivo)? bañarse, vestirse o comer? hablar o comunicarse (por ejemplo, entender o ser entendido por otros)?	***** 5. Sí 6. No	
Por problemas emocionales o mentales, ¿cuánta dificultad tiene (NOMBRE), para realizar sus actividades diarias (con autonomía e independencia)? Problemas como: autismo, depresión, bipolaridad, esquizofrenia, etcétera.		

Fuente: elaboración propia con información de los censos de población y vivienda 2010 y 2020, así como de la Encuesta Nacional de la Dinámica Demográfica 2014.

das resultó en la identificación de un volumen de personas con algún grado de limitación que no necesariamente significaba la presencia de una discapacidad. Para asegurar correspondencia entre los indicadores provenientes de las distintas fuentes, la esperanza de vida sin discapacidad fue estimada considerando solo a quienes manifestaron mucha dificultad o no poder realizar la actividad.

Pese a lo anterior, el indicador que aquí se propone es llamado esperanza de vida sin limitaciones físicas ni mentales (EVSL) con la finalidad de aproximarlo conceptual y estadísticamente a la EVISA, pues algunas limitaciones consideradas en los instrumentos de recolección de información podrían derivarse de ciertas comorbilidades.

En síntesis, para estimar la EVSL en un periodo determinado se necesitan tres tipos de datos desagregados según sexo, grupo de edad y entidad federativa: volumen de la población, defunciones, y prevalencia de limitaciones físicas y mentales (ver cuadro 4). Pero es hasta que se han equiparado conceptual y estadísticamente las mediciones de prevalencia de limitaciones físicas y mentales provenientes de distintas fuentes que se puede integrar la información para el ajuste de las tablas de vida periodo que pretendan ser analizadas en conjunto.

Cambios en la esperanza de vida sin limitaciones físicas y mentales entre el 2010 y 2020

La particularidad del 2020 en cuanto al exceso de mortalidad asociado con la pandemia por COVID-19 significó pérdidas en la esperanza de vida que otras publicaciones han detallado (Gallardo, 2020; García-Guerrero & Beltrán-Sánchez, 2021). Como la intención de este apartado es mostrar los cambios en la EVSL debidos a la evolución de la prevalencia de limitaciones físicas y mentales, intencionalmente excluiremos el efecto producido por el cambio en la esperanza de vida, suponiendo que las defunciones asociadas a COVID-19 no hubieran ocurrido. Para ello, empleamos las bases de datos para el análisis del exceso de mortalidad de la Secretaría de Salud (Dirección General de Información en Salud, 2021).

Cuadro 4

Tipos de datos y fuentes de información

Datos	Fuentes	Notas
Población clasificada según sexo y edad	Censo de Población y Vivienda	Cifras con estimación y prorrateo de quienes no especificaron su edad. Se proyecta a mitad de año considerando la tasa de crecimiento poblacional intercensal. La calidad de los datos fue evaluada con técnicas para identificar la preferencia por declarar edades terminadas en cero y cinco, y la estructura por edad y sexo, en general. Los datos probaron ser suficientemente buenos, por lo que su estructura no fue suavizada.
Defunciones clasificadas según sexo y edad	Estadísticas de mortalidad	En el año y lugar donde se registraron. En algunas entidades federativas y para los primeros grupos de edad, el número de defunciones no llega a 20. En esos casos, la tasa específica de mortalidad puede ser altamente variable, así que la interpretación basada en esas estimaciones debe hacerse con cautela.
Limitaciones físicas y mentales clasificadas según sexo y edad	Censo de Población y Vivienda	No se prorratea la población que no especificó su edad, pues la prevalencia suele aumentar con la edad y podría resultar en una sobreestimación en los primeros grupos.

Fuente: elaboración propia.

Los cambios reflejados por el indicador son los siguientes: entre el 2010 y 2020, disminuyó la esperanza de vida sin limitaciones físicas ni mentales en casi todas las entidades federativas, tanto para los hombres como para las mujeres; solo en Sinaloa, Chihuahua y Durango, la situación mejoró para ellos, quienes tenían los niveles más bajos de esperanza de vida al nacer en el 2010 en el país (ver gráfica 1).

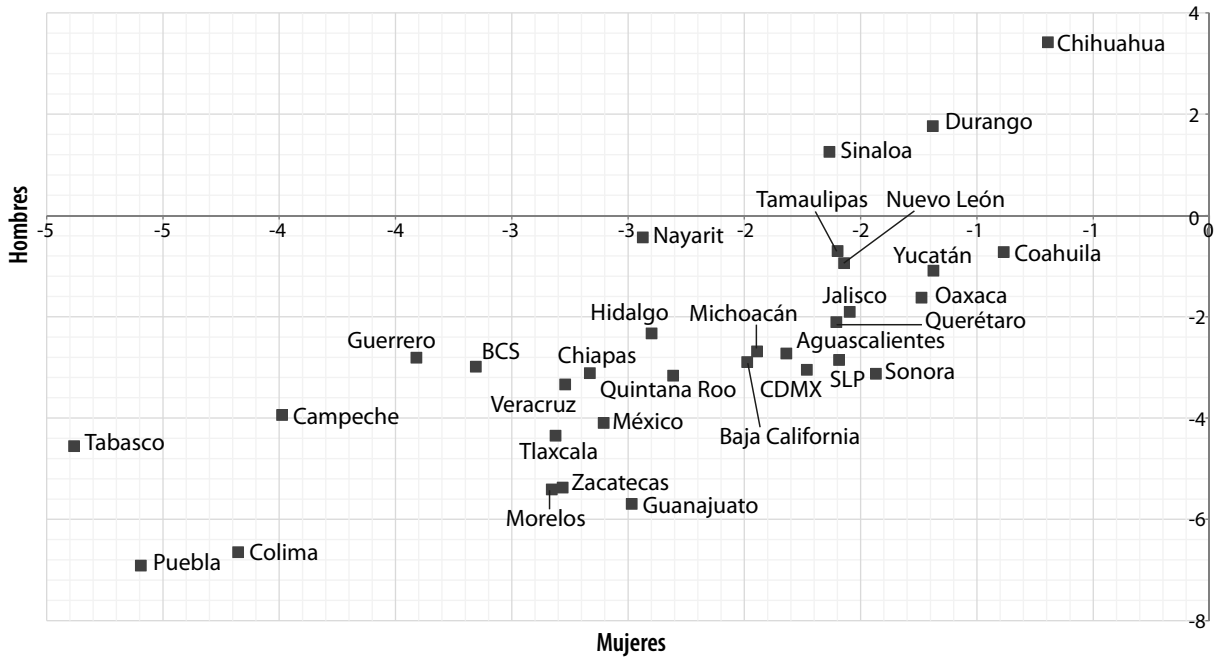
El indicador a los 60 años de edad (ver gráficas 2-A y 2-B) muestra pérdidas en salud de diferentes dimensiones en las entidades. Para los y las residentes en Campeche, Chiapas, Guerrero, Puebla y Tabasco, la esperanza de una vida sin limitaciones al llegar a esa edad disminuyó entre 2.4 y 3.2 años del 2010 al 2020. En el otro extremo, en Chihuahua, Coahuila de Zaragoza, Tamaulipas y Yucatán, la EVSL(60) de las mujeres y de los hombres disminuyó menos de un año.

Las gráficas 3-A y 3-B muestran que las brechas de género en el promedio de años vividos con alguna limitación física o mental en cada estado se agudizaron en el 2020. Las razones de la expansión de esa brecha pueden encontrarse en la ganancia de esperanza de vida para las mujeres, aunque en malas condiciones de salud. Un análisis cuidadoso requiere considerar el punto de partida de cada entidad, primero en el promedio de años que se espera vivir y, luego, en las circunstancias que pueden ser determinantes del distinto deterioro en las condiciones de salud y, en especial, en la aparición de limitaciones en el curso de vida.

Como es posible apreciar, un indicador como este, desagregado por entidad y sexo, y con posibilidad de generarse a partir de encuestas levantadas entre los censos de población que recogen información sobre discapacidad, puede informar la toma de decisiones en materia de salud, toda vez que permite el monitoreo regional y con perspectiva de género de los avances y retrocesos en esta dimensión. Estudiar las causas de las limitaciones, así como sus intensidades, y las razones de las des-

Gráfica 1

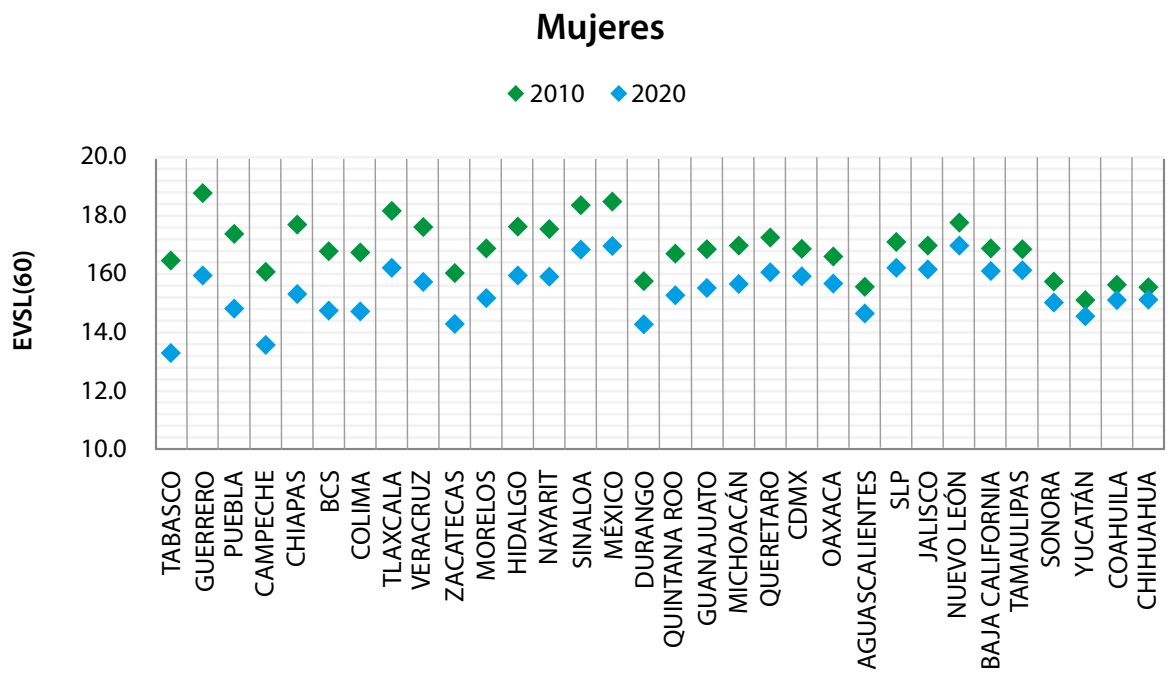
Cambios en las entidades federativas entre el 2010 y 2020 en la esperanza de vida sin limitaciones físicas ni mentales según sexo



Fuente: elaboración propia con datos de los censos de población y vivienda y estadísticas de mortalidad.

Gráfica 2-A

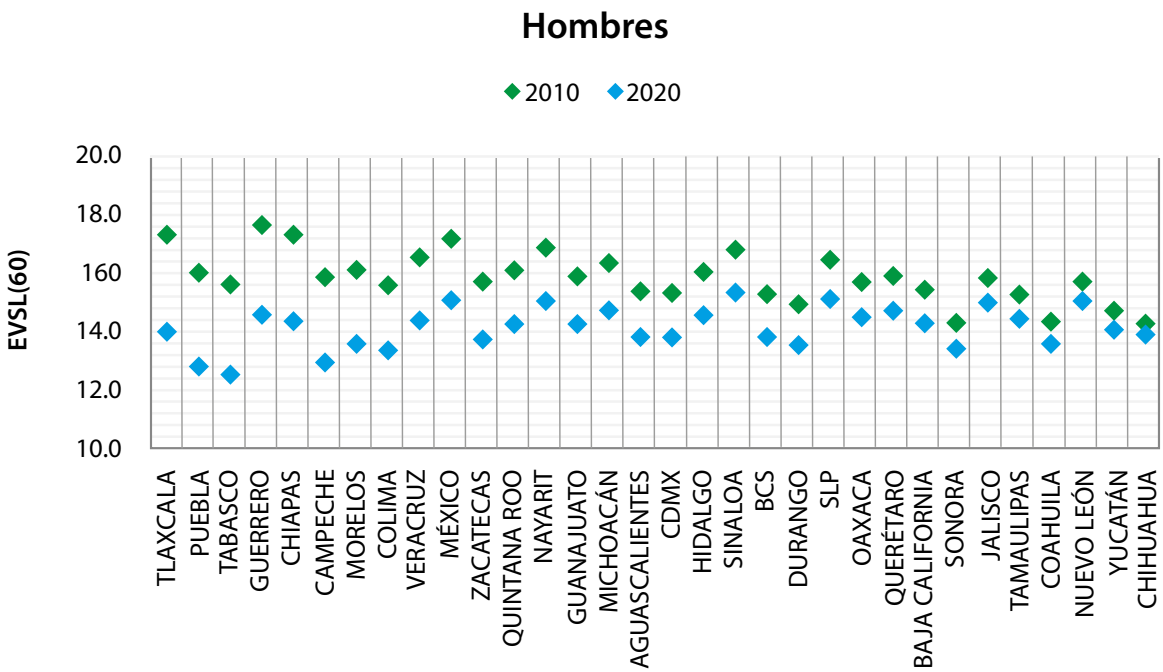
Esperanza de vida sin limitaciones físicas ni mentales a los 60 años, según sexo y entidad, en el 2010 y 2020



Fuente: elaboración propia con datos de los censos de población y vivienda y estadísticas de mortalidad.

Gráfica 2-B

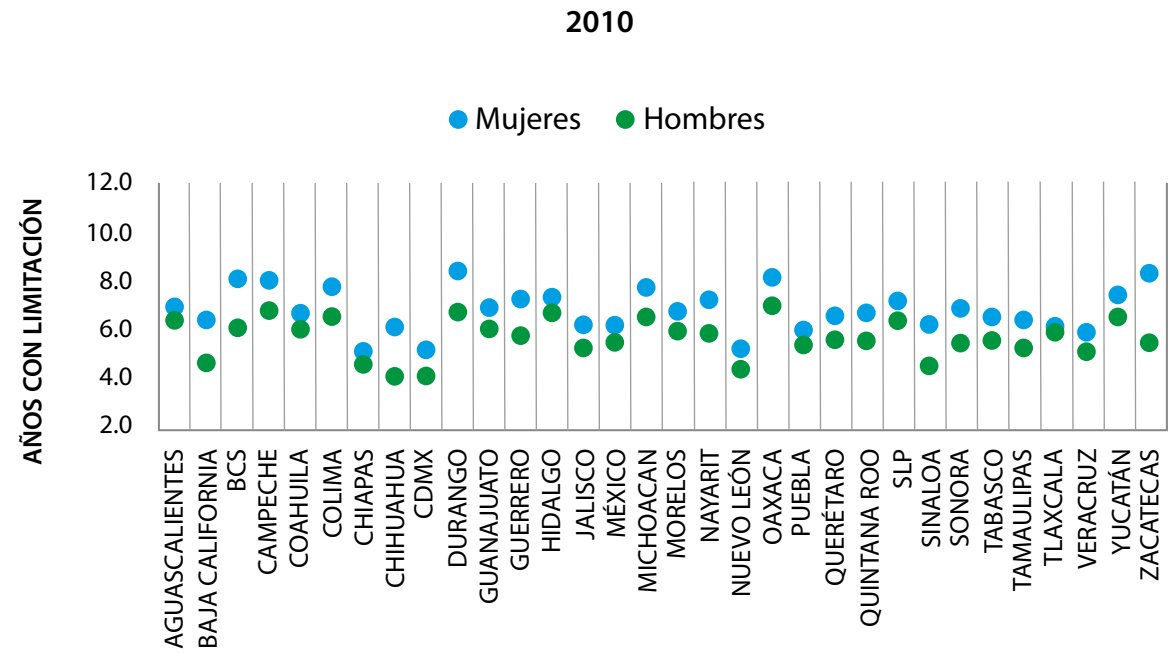
Esperanza de vida sin limitaciones físicas ni mentales a los 60 años, según sexo y entidad, en el 2010 y 2020



Fuente: elaboración propia con datos de los censos de población y vivienda y estadísticas de mortalidad.

Gráfica 3-A

Promedio de años vividos con alguna limitación física o mental, por entidad y sexo, en el 2010 y 2020

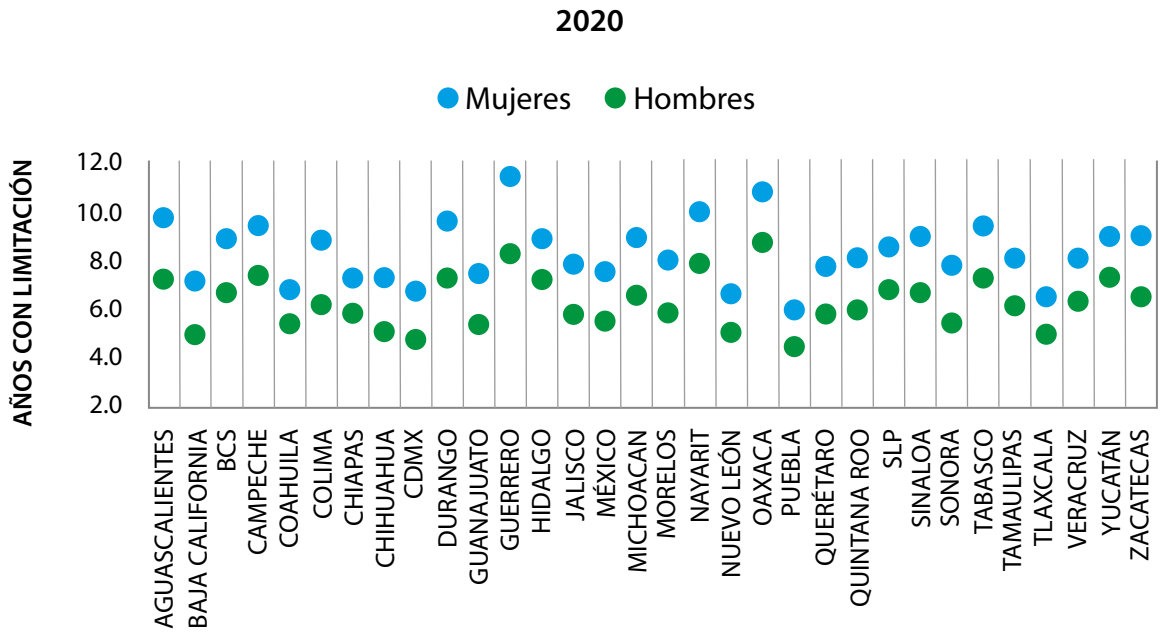


Fuente: elaboración propia con datos de los censos de población y vivienda y estadísticas de mortalidad.

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Gráfica 3-B

Promedio de años vividos con alguna limitación física o mental, por entidad y sexo, en el 2010 y 2020



Fuente: elaboración propia con datos de los censos de población y vivienda y estadísticas de mortalidad.

igualdades se vuelve primordial para diseñar intervenciones públicas que promuevan que, a lo largo del territorio nacional y para ambos sexos, la mayor extensión de la vida venga acompañada de la capacidad de funcionar de forma autónoma para vivirla con mejor calidad.

Validez de las estimaciones y limitaciones

La EVSL puede entenderse como una aproximación a EVISA y, de hecho, aunque hay diferencias conceptuales, las estimaciones resultado del diseño que aquí se propone son muy próximas a las de EVISA de la OMS y del IHNM.

De acuerdo con la estadística oficial, la esperanza de vida al nacer en México en el 2010 fue de 71.6 años para los hombres y de 77.9 años para las mujeres (Consejo Nacional de Población, CONAPO, 2018). Las estimaciones de la OMS y las del método empírico seguido en esta investigación difieren por menos de un año, como se puede observar en las columnas (a) y (c) del cuadro 5. Desafortunadamente, el CONAPO no publica la esperanza de vida a los 60 años de edad, pero la estimación aquí propuesta, columna (f), es muy próxima a la de la OMS, columna (d).

Las pequeñas diferencias en la esperanza de vida al nacer muy probablemente se relacionan con la forma en la que se producen las estimaciones. El CONAPO emplea proyecciones de los distintos componentes demográficos que son resultado del uso de modelos que, además, consideran el ejercicio de conciliación demográfica que se lleva a cabo en México de manera periódica. En cuanto a las es-

timaciones de la OMS, aunque también se emplean proyecciones basadas en las tendencias demográficas, se prefiere la información proveniente de registros civiles cuando esta es completa; ello implica que los procedimientos para la estimación de tablas de vida varían de país en país, en función de la disponibilidad y calidad de los datos sobre mortalidad infantil y adulta (World Health Organization, 2020). En el caso de esta investigación, se constató, con las técnicas de análisis demográfico tradicionales, que la distribución por edad y sexo de la población censal era congruente y se asumió que el registro de las defunciones es aproximadamente completo, por lo que las tablas de vida se construyeron empleando este dato de forma directa para derivar así las probabilidades de fallecer en cada grupo de edad.

En cuanto a la medición de la EVSL y EVISA, se advierten diferencias de 2.8 en el 2010 y de 2.4 años en el 2019/2020 entre el indicador al nacer derivado de esta propuesta y el de la OMS. Sin embargo, la diferencia va reduciéndose con la edad hasta que, a los 60 años, la estimación es igual en términos estadísticos, como se revela en las columnas (i) y (j) del cuadro 5.

El indicador de la EVSL es de construcción más simple que la EVISA. El primero, además de omitir la experiencia de comorbilidades, pondera igual toda clase de limitación física o mental e ignora que algunas limitaciones, o la coexistencia de varias, implican peor calidad de vida. Por eso es importante resaltar que un diseño de esta naturaleza sirve al propósito de obtener un buen indicador agregado, frecuente y oportuno de la ausencia de limitaciones importantes, pero que, para un estudio más fino de la discapacidad en México en particular, y de la calidad de vida en general, se requieren otro tipo de información y otros métodos.

El hecho de que las esperanzas de vida al nacer y a los 60 años, así como la EVISA/EVSL a los 60 años, sean consistentes con las respectivas estimaciones de la OMS es prueba del poder de la técnica y de su utilidad para producir un indicador sintético de la calidad de la salud, desagregado por entidad y sexo, con la información estadística oficial.

En cuanto a la esperanza de vida saludable, el CONAPO no la estima, y aunque el IHME sí lo hace para las distintas entidades de México anualmente, el dato disponible más reciente es para el 2019. En este sentido, cabe destacar que varias encuestas nacionales recogen información sobre la incidencia y prevalencia de limitaciones físicas y mentales, y también acerca de otras mediciones de la disca-

Cuadro 5

Esperanza de vida, EVISA y EVSL para México en el 2010 y 2019/2020: estimaciones de la OMS, el CONAPO y Páez

2010	Esperanza de vida						EVISA / EVSL			
	Al nacer			A los 60 años			Al nacer		A los 60 años	
	OMS (a)	CONAPO (b)	Páez (c)	OMS (d)	CONAPO (e)	Páez (f)	OMS (g)	Páez (h)	OMS (i)	Páez (j)
Total	**75.2**	**74.8**	**75.3**	**21.5**	**-**	**21.6**	**66.5**	**69.3**	**16.5**	**16.6**
Hombres	72.3	71.6	72.5	20.2	-	20.6	64.4	67.0	15.6	15.9
Mujeres	78.1	77.9	78.0	22.8	-	22.6	68.6	71.6	17.4	17.1
2019 / 2020 sin muertes por COVID-19										
Total	**76.0**	**75.2**	**75.0**	**21.8**	**-**	**21.4**	**66.0**	**68.4**	**16.1**	**16.1**
Hombres	73.1	72.4	71.6	20.5	-	19.7	64.0	65.9	15.3	15.2
Mujeres	78.9	78.1	78.5	23.1	-	22.9	67.0	70.9	16.8	16.8

Nota: las estimaciones en celdas sombreadas corresponden al 2019.
Fuente: elaboración propia con datos de la OMS, el CONAPO, el INEGI y el Registro Civil.

pacidad, durante los periodos intercensales, lo que ofrece una indiscutible oportunidad para un seguimiento más puntual de la evolución de la EVSL en cada entidad y para cada sexo. La gráfica 4 ilustra la similitud de las estimaciones del IHME y de esta autora con información de la ENADID.

La generación de indicadores a partir de encuestas para distintos años amerita, sin embargo, un cuidado adicional, el de verificar que el diseño conceptual y estadístico de cada uno de los programas sea compatible o pueda homologarse, para que cualquier discrepancia o cambio revelado en las estimaciones no se deba a diferencias en los cuestionarios o en las muestras. Aun en ese caso, es posible que las estimaciones no sean comparables entre sí y que, por lo tanto, no sean fiables para hacer análisis en el tiempo. Siempre que sea posible, la revisión de fuentes longitudinales puede ser de gran utilidad para constatar si los cambios en el tiempo mostrados por fuentes de corte transversal son consistentes.

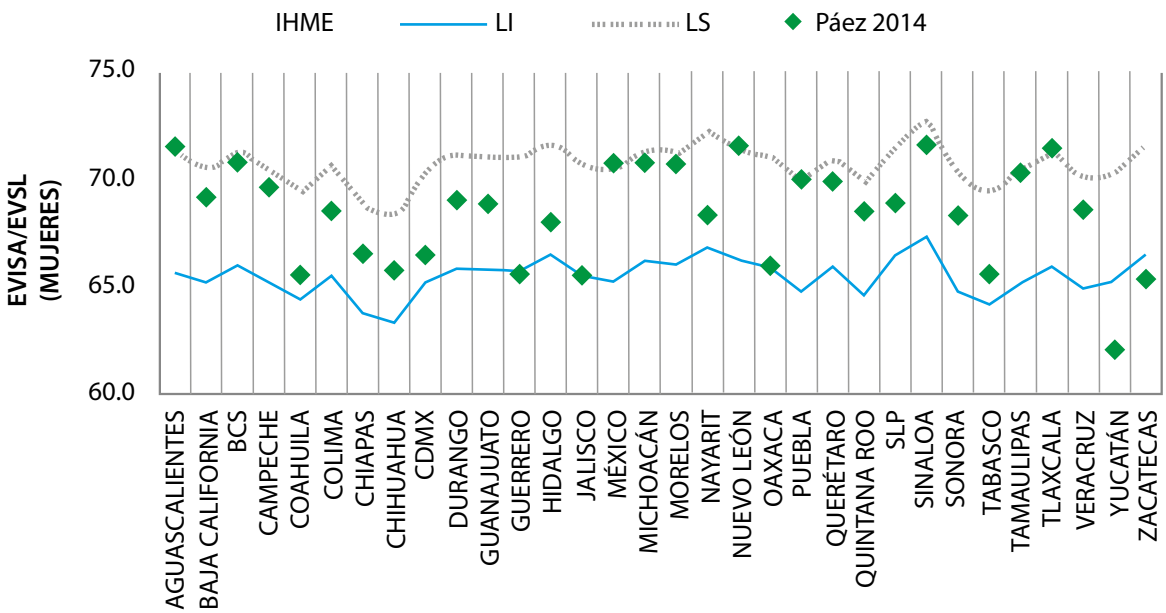
Probada la consistencia de dos o tres observaciones consecutivas por unidad geográfica, se pueden examinar los cambios ocurridos en los últimos años en la calidad de la salud de las mujeres y los hombres de cada entidad.

Conclusiones y recomendaciones

Este trabajo explota la información que el INEGI pone a disposición de los usuarios para generar indicadores de salud, desagregados por entidad federativa y sexo, derivados de la esperanza de vida al nacer. La metodología propuesta se utiliza para estimar la esperanza de vida sin limitaciones físicas ni mentales (al nacer y a los 60 años), así como el promedio de años vividos con tales limitaciones.

Gráfica 4

Esperanza de vida sin limitaciones físicas ni mentales para las mujeres de cada entidad: comparación de las estimaciones del IHME y Páez para el 2014



Fuente: elaboración propia con datos del IHME y el INEGI.

Las estimaciones son consistentes con las producidas por organismos internacionales e instituciones especializadas, como la OMS y el IHME. Además, la investigación en fuentes alternativas revela que existe el potencial para producir estimaciones con mejor oportunidad, toda vez que programas estadísticos varios ahora incluyen al menos la pregunta esencial para captar información sobre la prevalencia de discapacidad en México a nivel de las entidades federativas.

Asimismo, se mostró la utilidad de contar de manera periódica con los indicadores propuestos, para contribuir con el análisis de las brechas nacionales en salud mediante la intersección de las dimensiones temporal, geográfica y de género.

Por lo tanto, la primera recomendación es la de continuar con el levantamiento de información acerca de limitaciones físicas y mentales en programas estadísticos representativos a nivel entidad, en especial en aquellos que recogen información longitudinal, como la Encuesta Nacional sobre Salud y Envejecimiento en México (ENASEM) y la Encuesta Demográfica Retrospectiva (EDER).

En segundo lugar, se propone la inclusión de las estimaciones de esperanza de vida sin limitaciones físicas ni mentales (al nacer y a los 60 años de edad) y el promedio de años vividos con tales limitaciones en el *Visor dinámico de bienestar* (INEGI, 2015), una herramienta en la que el usuario elige el número y la ponderación de indicadores relacionados con el bienestar para obtener un ordenamiento jerárquico de las entidades del país, y que actualmente solo incluye la esperanza de vida al nacer y el subíndice de salud del Índice de Desarrollo Humano como variables de la dimensión salud.

En tercero, convendría explorar el potencial de la autopercepción del estado de salud como indicador de vida saludable. En el Reino Unido, por ejemplo, la esperanza de vida saludable se define como el número de años vividos en autoevaluada buena salud (Public Health England, 2017), mientras que el reporte subjetivo del estado de salud, levantado en algunos programas del INEGI, ha sido subutilizado hasta ahora.

Finalmente, es necesario continuar aprovechando la información longitudinal que produce el INEGI para refinar los indicadores de salud. Los datos transversales ofrecen una buena aproximación del patrón de morbilidad en las poblaciones y, en algunos casos, el costo en precisión de no emplear datos longitudinales es cubierto por la ganancia de contar con mayor desagregación regional y sociodemográfica; no obstante, el uso de información longitudinal permitiría profundizar en el análisis del tiempo vivido con algún padecimiento, identificando momentos de inicio, término, cambio en la intensidad de la limitación y duración de las fases.

Fuentes

Andrade, F. "Estimating diabetes and diabetes-free life expectancy in Mexico and seven major cities in Latin America and the Caribbean", en: *Revista Panamericana de Salud Publica (Pan American Journal of Public Health)*. 26(1), 2009, pp. 9-16.

Chiang, C. L. "AN INDEX OF HEALTH: MATHEMATICAL MODELS", en: *Vital and Health Statistics. Ser. 1, Programs and Collection Procedures*. 3, 1965, pp. 1-19.

CONAPO. *Indicadores demográficos de México de 1950 a 2050* (DE) https://tinyurl.com/tyj2ydl, consultado el 7 de agosto de 2020.

Dirección General de Información en Salud. *Exceso de mortalidad en México.* 2021 (DE) https://tinyurl.com/yhp9vzqc, consultado el 13 de julio de 2021.

Gallardo, A. "Efecto del COVID-19 en la expectativa de vida al nacer en México", en: *Nexos. Taller de datos*. 2020 (DE) https://tinyurl.com/yg3tcwrj, consultado el 29 de julio de 2020.

García-Guerrero, V. M., & H. Beltrán-Sánchez. "Heterogeneity in Excess Mortality and Its Impact on Loss of Life Expectancy due to COVID-19: Evidence from Mexico", en: *Canadian Studies in Population*. 2021, pp. 1-36.

INEGI. *Censo de Población y Vivienda 2010* (DE) https://tinyurl.com/yjnrvqxc, consultado el 7 de agosto de 2020.

_____ *Censo de Población y Vivienda 2020* (DE) https://tinyurl.com/yelut9ab, consultado el 20 de julio de 2021.

_____ *Encuesta Nacional de la Dinámica Demográfica* (*ENADID*) *2014* (DE) https://tinyurl.com/yel596c7, consultado el 7 de agosto de 2020.

_____ *Investigación—Visor dinámico de bienestar*. 2015 (DE) https://tinyurl.com/yhmhekoo, consultado el 8 de noviembre de 2021.

Moriyama, I. M. "Problems in the measurement of health status", en: *Indicators of social change: Concepts and measurements*. Russel Sage Foundation, 1968, pp. 573-600.

Public Health England. *Chapter 1: Life expectancy and healthy life expectancy*. 2017 (DE) https://tinyurl.com/yk5hbsl4, consultado el 13 de julio de 2017.

Rodríguez-Abrego, G. *et al*. "Esperanza de vida saludable en la población mexicana con seguridad social", en: *Perinatol Reprod. Hum*. 20(1), 2006, pp. 4-18.

Santosa, A., J. Schröders, M. Vaezghasemi, & N. Ng. "Inequality in disability-free life expectancies among older men and women in six countries with developing economies", en: *Journal of Epidemiology and Community Health*. 70(9), 2016, pp. 855-861.

Seuc, A. H., & E. Domínguez. "Introducción a la esperanza de vida ajustada por discapacidad", en: *Revista Cubana de Higiene y Epidemiología*. 41, 2003.

Sullivan, D. F. "Conceptual problems in developing an index of health", en: *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*. 17, 1966, pp. 1-18.

_____ "A single index of mortality and morbidity", en: *HSMHA Health Reports*. 86(4), 1971, pp. 347-354.

Washington Group on Disability Statistics. *The Washington Group Short Set on Functioning (WG-SS) (*DE*)* https://tinyurl.com/ygn2elka, consultado el 23 de octubre de 2017.

World Health Organization. *WHO methods and data sources for life tables 1990-2019*. Geneva, 2020.

Anexos

Cuadro A1 **Continúa**

Esperanza de vida al nacer y EVSL por entidad federativa según sexo en el 2010

Entidad federativa	Hombres		Mujeres	
	Esperanza de vida al nacer	EVSL	Esperanza de vida al nacer	EVSL
Aguascalientes	74.1	67.7	77.5	70.6
Baja California	69.9	65.2	77.5	71.1
Baja California Sur	74.2	68.1	80.0	71.9
Campeche	74.9	68.1	79.3	71.3
Coahuila de Z.	72.4	66.3	76.9	70.2
Colima	73.9	67.4	79.7	72.0
Chiapas	73.6	68.9	77.1	71.9
Chihuahua	62.5	58.3	75.0	68.9
Ciudad de México	69.3	65.2	74.9	69.7
Durango	70.8	64.1	78.3	69.9
Guanajuato	73.8	67.7	78.5	71.6
Guerrero	73.8	68.0	80.9	73.6
Hidalgo	74.7	68.0	79.9	72.5
Jalisco	72.5	67.2	77.9	71.7

Cuadro A1 Concluye

Esperanza de vida al nacer y EVSL por entidad federativa según sexo en el 2010

Entidad federativa	Hombres		Mujeres	
	Esperanza de vida al nacer	EVSL	Esperanza de vida al nacer	EVSL
México	75.1	69.6	79.7	73.5
Michoacán de O.	73.9	67.3	79.5	71.8
Morelos	73.4	67.4	78.4	71.6
Nayarit	73.4	67.5	79.9	72.6
Nuevo León	72.4	67.9	78.3	73.0
Oaxaca	73.4	66.4	78.9	70.8
Puebla	72.5	67.1	77.2	71.2
Querétaro	73.4	67.8	78.5	71.9
Quintana Roo	73.8	68.3	78.6	71.9
San Luis Potosí	74.7	68.3	79.1	71.9
Sinaloa	70.6	66.0	80.0	73.7
Sonora	70.9	65.4	76.9	70.0
Tabasco	71.3	65.7	76.8	70.3
Tamaulipas	71.4	66.1	78.1	71.7
Tlaxcala	75.1	69.2	78.9	72.7
Veracruz de I. de la Ll.	72.8	67.6	77.9	71.9
Yucatán	73.3	66.7	77.1	69.6
Zacatecas	72.5	67.0	78.8	70.5

Fuente: elaboración propia con datos del Censo de Población y Vivienda 2010 y estadísticas de mortalidad.

Cuadro A2 Continúa

Esperanza de vida al nacer y EVSL por entidad federativa según sexo en el 2014

Entidad federativa	Hombres		Mujeres	
	Esperanza de vida al nacer	EVSL	Esperanza de vida al nacer	EVSL
Aguascalientes	74.1	68.9	79.4	71.5
Baja California	71.6	65.8	78.4	69.1
Baja California Sur	74.4	68.6	80.9	70.7
Campeche	75.0	65.5	79.4	69.6
Coahuila de Z.	72.0	61.5	76.9	65.5
Colima	72.7	64.4	79.1	68.5
Chiapas	73.6	63.7	76.8	66.5
Chihuahua	70.0	61.9	76.3	65.7
Ciudad de México	70.3	63.5	76.1	66.4
Durango	73.6	65.4	80.0	69.0
Guanajuato	73.4	64.8	78.9	68.8
Guerrero	74.2	62.8	81.1	65.5
Hidalgo	75.3	64.3	80.3	67.9
Jalisco	72.8	64.4	78.4	65.5
México	75.3	69.1	80.2	70.7
Michoacán de O.	74.4	67.1	80.1	70.7
Morelos	73.8	65.7	79.4	70.7
Nayarit	74.9	64.5	80.4	68.3
Nuevo León	73.9	67.4	78.8	71.5
Oaxaca	74.2	62.0	79.3	65.9
Puebla	72.7	65.4	78.3	69.9
Querétaro	73.7	67.5	78.2	69.8
Quintana Roo	75.0	65.1	79.7	68.5
San Luis Potosí	74.8	64.5	79.6	68.8
Sinaloa	73.7	68.2	81.0	71.6
Sonora	71.3	63.0	78.1	68.3

Cuadro A2 Concluye

Esperanza de vida al nacer y EVSL por entidad federativa según sexo en el 2014

Entidad federativa	Hombres		Mujeres	
	Esperanza de vida al nacer	EVSL	Esperanza de vida al nacer	EVSL
Tabasco	71.9	63.3	76.9	65.5
Tamaulipas	72.8	65.4	78.5	70.3
Tlaxcala	75.8	66.9	79.6	71.4
Veracruz de I. de la Ll.	73.0	65.5	78.6	68.5
Yucatán	73.5	60.7	74.6	62.0
Zacatecas	74.7	63.1	75.8	65.3

Fuente: elaboración propia con datos de la Encuesta Nacional de la Dinámica Demográfica 2014 y estadísticas de mortalidad.

Cuadro A3 Continúa

Esperanza de vida al nacer y EVSL por entidad federativa según sexo en el 2020

Entidad federativa	Hombres		Mujeres	
	Esperanza de vida al nacer	EVSL	Esperanza de vida al nacer	EVSL
Aguascalientes	72.2	65.0	78.4	68.8
Baja California	67.3	62.3	76.2	69.1
Baja California Sur	71.8	65.1	77.6	68.8
Campeche	71.6	64.2	76.7	67.3
Coahuila de Z.	71.0	65.6	76.1	69.3
Colima	66.9	60.7	76.6	67.8
Chiapas	71.7	65.8	76.5	69.3
Chihuahua	66.9	61.8	75.5	68.2
Ciudad de México	66.9	62.1	74.7	68.0
Durango	73.1	65.8	78.3	68.8
Guanajuato	67.4	62.0	76.5	69.1
Guerrero	73.4	65.2	81.6	70.2
Hidalgo	72.9	65.7	79.0	70.1

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

Cuadro A3 Concluye

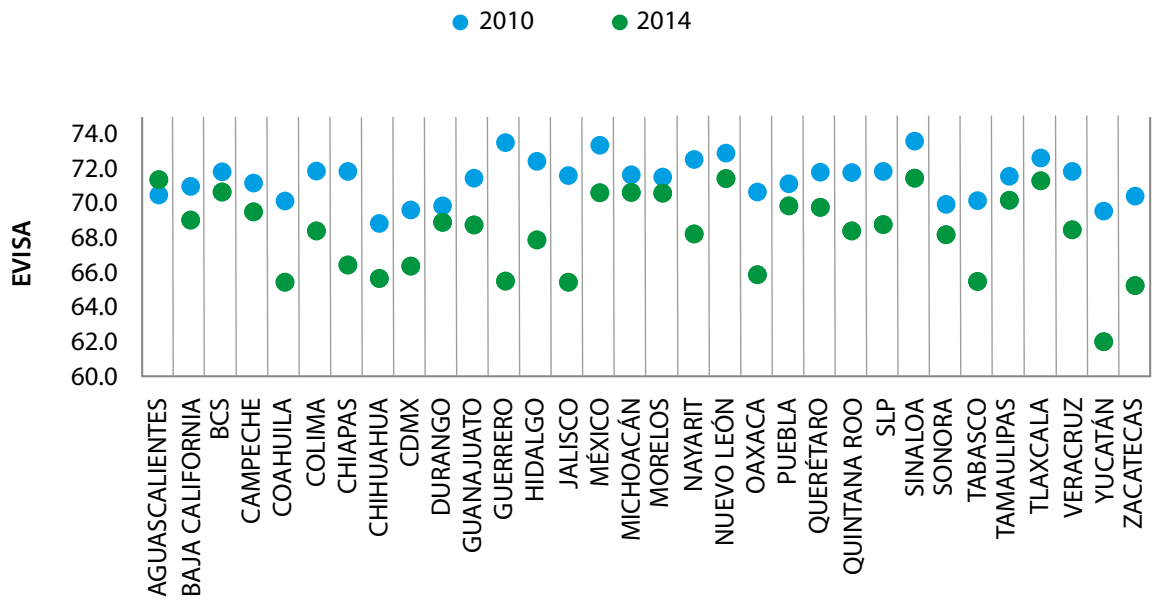
Esperanza de vida al nacer y EVSL por entidad federativa según sexo en el 2020

Entidad federativa	Hombres		Mujeres	
	Esperanza de vida al nacer	EVSL	Esperanza de vida al nacer	EVSL
Jalisco	71.1	65.3	78.0	70.2
México	71.0	65.5	78.4	70.9
Michoacán de O.	71.2	64.6	78.7	69.8
Morelos	67.9	62.0	76.8	68.8
Nayarit	74.9	67.1	80.1	70.2
Nuevo León	72.1	67.0	78.1	71.5
Oaxaca	73.5	64.8	80.3	69.5
Puebla	64.7	60.2	72.6	66.6
Querétaro	71.5	65.7	78.0	70.3
Quintana Roo	71.1	65.1	77.7	69.6
San Luis Potosí	72.3	65.5	78.9	70.4
Sinaloa	73.9	67.3	81.0	72.1
Sonora	67.8	62.3	76.4	68.6
Tabasco	68.4	61.1	74.7	65.4
Tamaulipas	71.6	65.4	78.1	70.1
Tlaxcala	69.9	64.9	76.4	69.9
Veracruz de I. de la Ll.	70.6	64.3	77.2	69.2
Yucatán	72.9	65.6	77.4	68.4
Zacatecas	68.1	61.6	76.7	67.7

Fuente: elaboración propia con datos del Censo de Población y Vivienda 2020 y estadísticas de mortalidad.

Gráfica A1

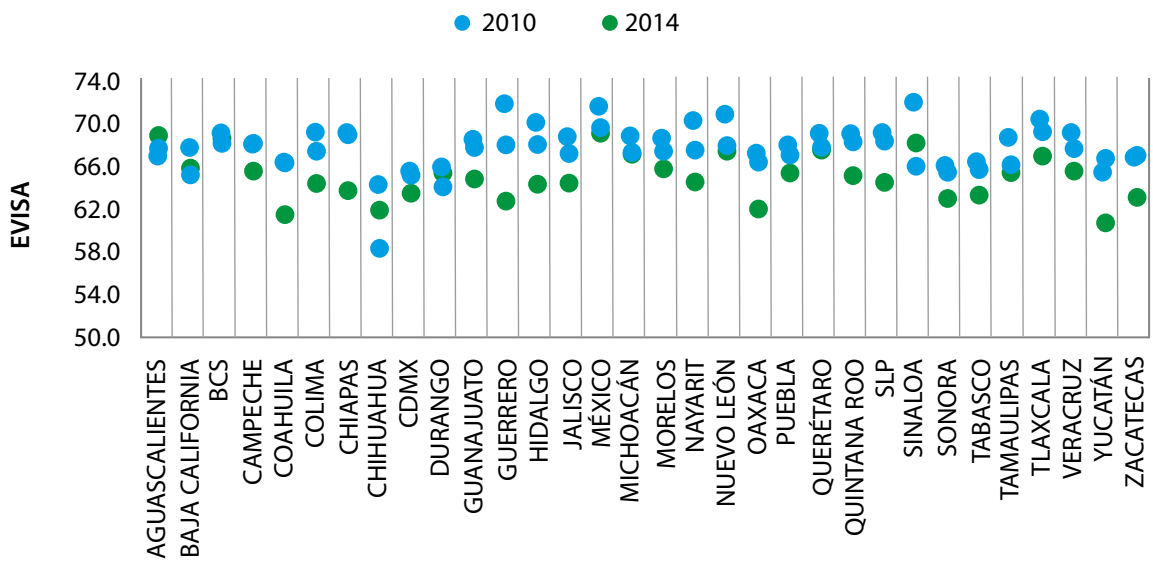
Esperanza de vida sin limitaciones físicas ni mentales para las mujeres de cada entidad en el 2010 y 2014



Fuente: elaboración propia con datos del Censo de Población y Vivienda 2010 y la Encuesta Nacional de la Dinámica Demográfica 2014.

Gráfica A2

Esperanza de vida sin limitaciones físicas ni mentales para los hombres de cada entidad en el 2010 y 2014



Fuente: elaboración propia con datos del Censo de Población y Vivienda 2010 y la Encuesta Nacional de la Dinámica Demográfica 2014.

Lecturas en lo que indican los indicadores

Readings in What Economic Indicators Indicate

Reseña

Jonathan Heath*

Este año se cumple una década de la publicación del libro *Lo que indican los indicadores. Cómo utilizar la información estadística para entender la realidad económica de México*, editada por el Instituto Nacional de Estadística y Geografía (INEGI).

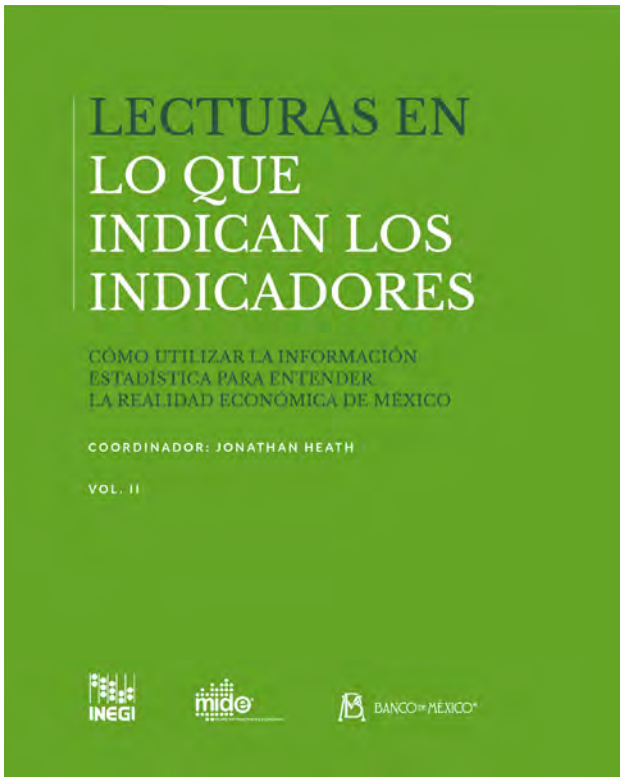
Tras muchos años impartiendo clases sobre el uso adecuado de indicadores económicos para analizar y entender la coyuntura económica, consideré útil y necesario escribir un libro que explayara este arte específicamente para México. Sin embargo, jamás imaginé el amplio interés que generaría entre el público general, académico, profesionista e institucional.

Así, durante mi estancia como investigador invitado en el INEGI, tuve la oportunidad de acercarme aún más a los que hacen los indicadores, diseñan las encuestas, calculan las cifras y divulgan los resultados al público para platicar sobre las metodologías, discutir las alternativas y entender más a fondo las limitaciones y alcances de cada indicador. Asombrosamente, el libro ha llegado a ser de utilidad no solo como referencia obligada del tema, sino también como texto principal o complementario en casi todas las universidades del país. El INEGI lo estableció como lectura prerrequisito para el examen de admisión para trabajar en el Instituto. Al poco tiempo de su publicación, el Foro Educativo del Museo Interactivo de Economía (MIDE) me invitó a diseñar el diplomado Indicadores Macroeconómicos de Coyuntura en México, que giraría alrededor del libro.

La primera edición del diplomado en el MIDE, en 2013, consistió en 15 módulos, los cuales básicamente seguían el índice del libro. Decidí que el diplomado fuera complementándose con la participación de expertos con el fin de que cada uno compartiera su dominio para analizar los indicadores económicos y que los alumnos no se aburrieran con un solo profesor. Este formato fue un éxito, por lo que, con los años, fui añadiendo temas e invitando a más especialistas, hasta llegar a 24 módulos en su séptima edición en 2019,

* jonathanheath54@gmail.com

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022



todos con temas relevantes para analizar y entender la coyuntura económica de nuestro país.

En estos años, los indicadores han evolucionado y han aparecido muchos otros que, pese a estar considerados en el contenido de los módulos, hacían necesaria una segunda edición del libro original. Sin embargo, me quedaba claro que, para realizar tal esfuerzo, necesitaría de nuevo de un año sabático, algo difícil al considerar mis obligaciones actuales. Aun así, no dejé de pensar en el proyecto cuando llegó la pandemia del COVID-19, que llevó a la inevitable cancelación de la octava edición del diplomado. Dado que todos estábamos confinados en casa, pensé que, en vez de que cada experto diera su módulo presencial, esta era la oportunidad para que cada quien lo escribiera en forma de una lectura. El objetivo fue que estas lecturas pudieran ser integradas en una nueva obra, que no solo complementaría y actualizaría al libro original, sino que sirviera de referencia tanto para cuando se volviera a impartir el diplomado como para aquellos que me han insistido en publicar ya la segunda edición.

En esta ocasión, nuevamente el INEGI, con el apoyo de Julio Santaella, entonces presidente del INEGI, y el MIDE, con el apoyo de Silvia Singer, lanzamos *Lecturas en lo que indican los indicadores*. Esta edición abarca tres volúmenes con 12 lecturas, cada una escritas por verdaderos expertos en la materia. Para esta edición, he logrado reunir a algunos de los mejores economistas del país, incluyendo al que fue el gobernador del Banco de México (Banxico), hasta finales del año pasado; dos de sus subgobernadores; al anterior presidente del INEGI; afamados directores en los sectores público y privado; economistas en jefe de instituciones prestigiadas; así como a importantes personalidades de la academia en México. Todos ellos no solo nos presentan sus visiones sobre varios temas relacionados con el análisis de la coyuntura económica de nuestro país, sino también sus perspectivas en torno a diversos indicadores. Además, estos tres volúmenes contienen una gran variedad de temas en los que también encontrarán una diversidad sobre cómo explicar y entender los indicadores.

El primer volumen inicia con una lectura sobre la estadística oficial en México, escrita por Julio Santaella, presidente del INEGI hasta finales de 2021. En ella, se destaca el marco institucional que hace que las estadísticas sean confiables. Posteriormente, diversos autores exponen una diversidad de temas conceptuales, como la lectura de estacionalidad de las cifras económicas (de Edwin Tapia), la medición de los ciclos económicos en nuestro país (de Pablo Mejía) y la programación financiera (de Luis Foncerrada), que emplea el Fondo Monetario Internacional para entender la problemática de cada país. Este volumen también incluye lecturas que ponen énfasis en cómo entender la actividad económica actual, por ejemplo, la lectura sobre el mercado laboral mexicano, de mi autoría, y la lectura de Gabriel Casillas, economista en jefe para América Latina de Barclays, sobre cómo usar los indicadores macroeconómicos para identificar en qué parte del ciclo nos encontramos.

Asimismo, en este volumen participan expertos en temas muy específicos. Tal es el caso de Federico Rubli, quien nos comparte dos lecturas: una sobre

los indicadores de ahorro e inversión y otra sobre la problemática del Sistema de Ahorro para el Retiro (SAR). Salvador Bonilla, experto en temas de Balanza de Pagos quien en dos lecturas hace un análisis de la cuenta corriente y de la cuenta financiera, respectivamente. Jesús Cervantes, coordinador del Foro de Remesas en el Centro de Estudios Monetarios Latinoamericanos (CEMLA), quien nos enseña sobre el ingreso por remesas y la migración mexicana. Este volumen cierra con mi lectura sobre el Indicador de Confianza del Consumidor, uno de los pocos indicadores que el INEGI ha cambiado su forma de presentar para que sea más transparente y fácil de utilizar.

En el segundo volumen comenzamos con una diversidad de temas, algunos de los cuales no estaban incluidos en el libro original, pero sí se impartían como módulos del diplomado. El texto abre con una lectura de mi autoría sobre el Rey de los Indicadores, el Producto Interno Bruto. Posteriormente, se abarcan textos de diversos especialistas, como Mariana Campos, de México Evalúa, quien nos da una guía sobre las finanzas públicas a través de diversos indicadores para entender la política fiscal; así como un texto de Sergio Martin, en su momento asesor de la Junta de Gobierno del Banco de México, sobre la percepción empresarial y los indicadores de producción manufacturera. De ahí, este volumen continúa con dos lecturas sobre la inflación, primero de Alejandrina Salcedo, directora de Precios en el Banco de México, y Ernesto Rattia, investigador en esa dirección, y después una lectura de Javier Salas, q. e. p. d., y que era un experto incuestionable en este tema, la cual denomina una enfermedad social. En lo personal considero un honor que el último texto que escribió Javier en vida esté incluido en este volumen.

Las siguientes cuatro lecturas abordan temas relacionados con los mercados financieros y la política monetaria. Ociel Hernández, estratega en jefe para México de BBVA (antes conocido mundialmente como Bancomer), ilustra con las crisis de 2007-2009 y 2020 la relevancia de conocer los indicadores sobre estos mercados financieros. Posteriormente, hay dos lecturas de los asesores de la Junta de Gobierno del Banco de México: la de Ernesto Sepúlveda con un texto introductorio al tema de estabilidad financiera y la de Jaime Acosta, quien explica el concepto de riesgo-país y los indicadores de riesgo soberano. Este bloque de lectura incluye una sobre la política monetaria escrita por el mismísimo Alejandro Díaz de León, gobernador de Banco de México de 2018 a 2021.

El último bloque de lecturas en este volumen se destina a temas más relacionados con la política social. Rodolfo de la Torre comparte su dominio de conocimientos sobre la medición oficial de la pobreza en México. Gerardo Leyva, director general adjunto de Investigación del INEGI, aborda el relacionado con el bienestar desde una concepción más amplia, que incluye aspectos ambientales y sociales del progreso. Este bloque concluye con el el tema de la desigualdad y distribución del ingreso con un texto de Gerardo Esquivel, subgobernador del Banco de México, quien domina como pocos la materia en el país.

Al momento de elaborar estos dos primeros volúmenes de lecturas, me percaté de que había muchos temas aún pendientes que fácilmente podían ser considerados en un tercer volumen. Por ello, me tomé la libertad de invitar a otros especialistas a escribir lecturas sobre las temáticas que dominan. Este tercer volumen inicia con un texto de Gerardo Leyva, en conjunto con Francisco Corona y Jesús López —investigadores del INEGI—, que nos ofrece una explicación del Indicador Oportuno de la Actividad Económica —elaborado por el INEGI—, el cual, sin duda, es una herramienta muy útil para aquellos interesados en seguir la coyuntura económica. Enseguida, los responsables de la publicación del *Reporte sobre las Economías Regionales*, Daniel Chiquiar, que fue director de Investigación Económica del Banco de México hasta el año pasado, y Juan Carlos Chávez, investigador en la misma dirección, nos exponen los indicadores regionales comúnmente analizados en dicho reporte.

En este volumen no podía dejar de lado los temas relacionados con el mercado laboral. Por ello,

David Kaplan, un reconocido experto en economía laboral, expone una lectura sobre cómo interpretar los indicadores laborales y, de igual manera, Edgar Vielma, director general de Estadísticas Sociodemográficas del INEGI, analiza el papel de la mujer en el mercado laboral en México. Finalmente, en un tema muy relacionado, Mario Correa abarca el tema de la productividad.

De igual manera, varios temas relacionados con los mercados financieros fueron considerados. Los indicadores del mercado accionario son analizados por Nur Cristiani, la estratega principal para banca privada de América Latina de JP Morgan, mientras que Gabriela Siller, economista en jefe del Grupo Financiero Base, destaca "… todo lo que se tiene que saber sobre el peso mexicano". Estas lecturas son complementadas por tres autores del Banco de México con una amplia experiencia en las operaciones monetarias: en la primera de ellas, Jaime Acosta comparte su experiencia en el desarrollo de la nueva tasa TIIE de Fondeo a un día como un indicador clave en el mercado interbancario. Por su parte, Juan García, director de Operaciones Nacionales, nos explica el proceso de instrumentación de la política monetaria que él dirige día a día. De la misma manera, Rodrigo Cano, director de Apoyo a las Operaciones, explica el análisis de las reservas internacionales.

La penúltima lectura de este volumen es de Sofía Ramírez, directora de México ¿Cómo vamos?, quien presenta el Índice de Progreso Social explicando tanto su base conceptual como su interpretación. Finalmente, en la última lectura escribo sobre los indicadores en los tiempos de pandemia. En esta detallo las distorsiones que se tienen en los indicadores económicos actualmente y el reto que tenemos en su interpretación.

Como se puede apreciar, este ha sido un esfuerzo descomunal plasmado en tres volúmenes que incluyen 36 lecturas escritas por 32 autores distintos y un servidor bajo una absoluta libertad en el estilo y énfasis de cada uno. Por ello, los invito a leer y acompañarnos en esta narrativa heterogénea sobre la evolución de los indicadores que complementa a la primera edición con interpretaciones, comentarios y nuevos enfoques que no existían cuando escribí el libro original. Con estos tres volúmenes de lecturas tenemos una referencia única y actualizada para seguir conociendo exactamente lo que indican los indicadores.

Colaboran en este número

José Román Herrera Morales

De nacionalidad mexicana. Es doctor en Tecnologías de la Información por la Universidad de Guadalajara. Fue director del Centro Nacional Editor de Discos Compactos (CENEDIC) y encargado del desarrollo del Sistema Integral de Automatización de Bibliotecas (SIABUC) de la Universidad de Colima. En la actualidad, es profesor-investigador de tiempo completo en la Facultad de Telemática de esta misma casa de estudios. Sus intereses de investigación son búsqueda y recuperación de información, tecnología web, aprendizaje artificial, minería de datos, IoT y bases de datos.
Contacto: rherrera@ucol.mx

Luis Francisco Barbosa Santillán

De nacionalidad mexicana. Es doctor en Tecnologías de la Información por la Universidad de Guadalajara. Trabajó en la docencia desde el 2008 en el Tecnológico de Estudios Superiores de Coacalco (TESCo) del estado de México y la Universidad Politécnica de Tapachula en Chiapas (UPTap); ha realizado estancias de investigación dentro de las universidades Complutense de Madrid, la Politécnica Salesiana sede Cuenca (Ecuador) y de Colima (México). En la actualidad, es profesor-investigador en la Universidad Tecnológica de Puebla. Tiene un total de cinco publicaciones académicas, de las cuales dos fueron en revistas indexadas en el *Journal Citation Reports (JCR)*.
Contacto: luis.barbosa@utpuebla.edu.mx

Jorge Rafael Gutiérrez Pulido

De nacionalidad mexicana. Es doctor en Computación por la Universidad de Nottingham, Reino Unido. Ha participado en comités editoriales internacionales para revistas de Elsevier, IEEE, IGI Global e Inderscience. Fue miembro del Sistema Nacional de Investigadores (SNI) con nivel 1, del Consejo Nacional de Ciencia y Tecnología (CONACYT). En la actualidad, es profesor-investigador en la Facultad de Telemática de la Universidad de Colima (México) y, desde el 2008, colabora en la columna "Estado del Arte" de la revista *Komputer Sapiens* de la Sociedad Mexicana de Inteligencia Artificial (SMIA).
Contacto: jrgp@ucol.mx

María Andrade Aréchiga

De nacionalidad mexicana. Tiene el Doctorado en Ciencias de la Computación por el Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Baja California. Sus intereses de

investigación se enfocan en la tecnología educativa, ambientes virtuales de aprendizaje, minería de datos y educación en matemáticas. Desde 2011, es miembro del SNI con nivel 1. En la actualidad, es profesora-investigadora de tiempo completo en la Facultad de Telemática de la Universidad de Colima (México).
Contacto: mandrad@ucol.mx

Sara Sandoval Carrillo

De nacionalidad mexicana. Es maestra en Telemática por la Universidad de Colima (México). En la actualidad, es profesora-investigadora de tiempo completo en la Facultad de Telemática de esta misma casa de estudios. Sus intereses de investigación se basan en el desarrollo de *software* dirigido por modelos y las tecnologías de la información.
Contacto: sary@ucol.mx

Víctor Alfredo Bustos y de la Tijera

De nacionalidad mexicana. Es licenciado en Actuaría por la Facultad de Ciencias de la Universidad Nacional Autónoma de México (UNAM), maestro en Estadística e Investigación de Operaciones por el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la UNAM y doctor en Estadística por la London School of Economics. Ingresó al Instituto Nacional de Estadística y Geografía (INEGI) en 1991, donde actualmente es investigador en la Dirección General Adjunta de Investigación. Sus áreas de interés son, entre otras: el ajuste de modelos a partir de encuestas, donde destaca la estimación de la distribución del ingreso con datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), compatible con las cuentas nacionales; la estimación para dominios pequeños en temas como la desocupación y la migración; la clasificación multivariada con aplicación a la medición del bienestar y la marginación; y la construcción de índices temporales multivariados para seguir la evolución de una economía mediante criterios explícitos.
Contacto: alfredo.bustos@inegi.org.mx

Abel Alejandro Coronado Iruegas

Nació en México. Es ingeniero en Sistemas Computacionales por la Universidad Autónoma de Aguascalientes (UAA) y maestro en Ciencias en Estadística Oficial por el Centro de Investigaciones en Matemáticas (CIMAT), A. C.; en la actualidad es candidato a doctor en Ciencia de Datos en el Centro de Investigación e Innovación en

	Tecnologías de la Información y Comunicación (INFOTEC). Tiene una trayectoria de 20 años en el INEGI, donde ha participado como líder técnico y arquitecto de múltiples sistemas informáticos, incluyendo proyectos de ciencia de datos y *Big Data*. Actualmente, es subdirector de Investigación en Ciencia de Datos en el Laboratorio de Ciencia de Datos de la Dirección General Adjunta de Investigación del INEGI, donde realiza estudios enfocados en el desarrollo de técnicas de *Machine Learning* para el aprovechamiento de grandes volúmenes de datos no estructurados como imágenes de satélite y texto. **Contacto:** abel.coronado@inegi.org.mx
Silvia Laura Fraustro Velhagen	De nacionalidad mexicana. Es ingeniera en Cibernética y Ciencias de la Computación por la Universidad La Salle. Bajo su supervisión como directora general de Tecnologías de la Información y la Comunicación (TIC) en el Poder Judicial mexicano, se estableció la primera Red Nacional de Telecomunicaciones; desde 1995 trabaja para el INEGI, donde ha participado en una serie de proyectos innovadores de TIC, como: la primera página web del Instituto, reemplazo de comunicaciones satelitales por fibra óptica, el prototipo de un DWH institucional y conceptualización de la Red Nacional de Información; en fechas más recientes, ha trabajado en proyectos internacionales y su implementación en el INEGI, por ejemplo: la adopción del estándar para el intercambio de datos y metadatos (SDMX); es parte del *Social Media Data Task Team*; es responsable del soporte tecnológico del Modelo de Producción Estadística y Geográfica (MPEG) del Instituto; y es parte del *Task Team* para alinear los procesos transversales de GSBPM con GAMSO, del equipo de ciencia de datos del INEGI que busca formas diferentes de producir estadísticas (actualmente experimentales), utilizando fuentes alternativas de *Big Data*. **Contacto:** silvia.fraustro@inegi.org.mx
Gerardo Leyva Parra	De nacionalidad mexicana. Es licenciado en Economía por la UAA, tiene una maestría en la misma disciplina por el Instituto Tecnológico Autónomo de México (ITAM) y otra en Ciencia Regional por la Cornell University, donde también obtuvo el Doctorado con Especialización en Crecimiento y Desarrollo Económicos; además,

cuenta con el Diplomado en Psicología Positiva por la Universidad Iberoamericana. En el ámbito laboral, ha impartido cursos de Teoría Económica en varias universidades y tiene 22 años de experiencia profesional en el INEGI, donde ha sido analista, asesor de tres presidentes, director de Censos Económicos, director general adjunto de Estadísticas Económicas y, a partir del 2009, director general adjunto de Investigación. Fue integrante del Grupo de Expertos en Medición de la Pobreza de la ONU (conocido como Grupo de Río) y del Comité Técnico para la Medición de la Pobreza en México. Es miembro de los comités de Estudios Económicos del Instituto Mexicano de Ejecutivos de Finanzas (IMEF), del indicador IMEF del Entorno Económico Empresarial y del de Coyuntura de la Asociación Nacional de Tiendas de Autoservicio y Departamentales (ANTAD), así como del Consejo Asesor Técnico del Centro de Estudios Económicos del Sector Privado (CEESP). Participa en los consejos editoriales de las revistas *Políticas Públicas* de la Escuela de Graduados en Administración Pública y Política Pública (EGAP), *Coyuntura Demográfica* de la Sociedad Mexicana de Demografía (SOMEDE) e *Investigación Económica* de la UNAM, y es editor técnico de *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía* del INEGI.

Contacto: gerardo.leyva@inegi.org.mx

Noemí López Delgado

De nacionalidad mexicana. Es licenciada en Matemáticas Aplicadas por la UAA y concluyó la Maestría de Análisis Estadístico en el CIMAT. Labora en el INEGI desde el 2009, donde actualmente es investigadora en la Dirección General Adjunta de Investigación, entre otras áreas, en la construcción de una metodología alternativa para el análisis de series temporales. Fue catedrática en la UAA impartiendo diversos cursos de matemáticas

Contacto: nohemi.delgado@inegi.org.mx

Ricardo Antonio Olvera Navarro

De nacionalidad mexicana. Es ingeniero en Tecnologías de la Información y la Comunicación con especialidad en Sistemas por la Universidad Tecnológica de Aguascalientes. Desde el 2010, ha trabajado en diversos proyectos para el INEGI, entre los que destacan el desarrollo de: una plataforma transversal para el resguardo de

evidencias de cualquier modelo de procesos de generación de información, herramientas de clasificación manual de textos, *software* orientado a servicios web y a clientes que exploten dichos servicios con *frameworks* responsivos y tolerantes a fallas, así como de paquetes de R para consulta de servicios de tuits y clasificación de contenido para visualizaciones de información y explotación de web API. Tiene experiencia en análisis de información masiva, en diseño y desarrollo de arquitecturas de explotación de información y en desarrollo de visualizaciones; además, es experto en el estándar de intercambio de información estadística SDMX entre instituciones nacionales e internacionales tanto en el análisis de estructuras de datos y su modelado como en herramientas informáticas que soportan el estándar.

Contacto: ricardo.olvera@inegi.org.mx

Ana Miriam Romo Anaya

De nacionalidad mexicana. Es licenciada en Matemáticas Aplicadas por la UAA y maestra en Ciencias en Estadística Oficial por el CIMAT. Desde el 2009, ha participado en la Dirección General Adjunta de Investigación del INEGI en el desarrollo y análisis de técnicas estadísticas, como: muestreo, modelos lineales y no lineales, técnicas de agrupamiento y clasificación, optimización de funciones paramétricas y el desarrollo computacional estadístico; asimismo, ha sido catedrática en la UAA y en la Universidad Politécnica de Aguascalientes impartiendo cursos de matemáticas y estadística.

Contacto: miriam.romo@inegi.org.mx

Víctor Silva Cuevas

De nacionalidad mexicana. Es ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Zacatecas y estudiante de la Maestría en Ciencia de Datos por el INFOTEC. Desde el 2015, ha participado en la Dirección General Adjunta de Investigación del INEGI para la prueba e implementación de las primeras bases en la adopción de nuevas tecnologías relacionadas con *Big Data* y ciencia de datos, así como en proyectos relacionados con *Big Data*, bases de datos NoSQL, análisis de sentimientos y de movilidad.

Contacto: victor.silvac@inegi.org.mx

Mauricio Montiel

De nacionalidad mexicana. Se tituló en Ingeniería Financiera por la Universidad Politécnica de Pachuca (UPP) y tiene la Maestría en

Economía Aplicada por el Colegio de la Frontera Norte (COLEF). Actualmente, se desempeña como jefe corporativo de Finanzas en la compañía Lala Administración y Control, S. A. de C. V.
Contacto: mmontiel.mea2018@colef.mx

Francisco de Jesús Corona Villavicencio

De nacionalidad mexicana. Es licenciado en Economía por la Universidad Autónoma de Baja California (UABC), maestro en Estadística Aplicada por el Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) y doctor en Economía y Métodos Cuantitativos por la Universidad Carlos III de Madrid (UC3M); fue mérito académico en sus estudios de licenciatura, recibió la beca excelencia para realizar su maestría y la calificación Sobresaliente Cum Laude en sus estudios de doctorado. En la actualidad, es investigador en el INEGI y sus líneas de estudio están relacionadas con el análisis econométrico y pronóstico de series de tiempo. Ha publicado trabajos en revistas arbitradas de circulación internacional y nacional. Pertenece al SNI, nivel I.
Contacto: franciscoj.corona@inegi.org.mx

Jesús López-Pérez

De nacionalidad mexicana. Es licenciado en Economía por el ITESM y maestro en Estadística Aplicada por la misma institución. En la actualidad, es subdirector de investigación en el INEGI en temas relacionados con el análisis econométrico de series de tiempo; anteriormente, ocupó diversos cargos en áreas de administración y análisis de riesgo crediticio en instituciones financieras de los sectores público y privado.
Contacto: jesus.lopezp@inegi.org.mx

José Alejandro Ruiz Sánchez

De nacionalidad mexicana. Cuenta con estudios de Maestría en Economía por el Centro de Investigación y Docencia Económicas (CIDE), A. C., y en Teoría Económica por el ITAM. En el ámbito profesional, su trayectoria se ha desarrollado en la administración pública federal: desde enero del 2016 se incorporó al INEGI como investigador dentro de la Dirección General Adjunta de Investigación. Sus áreas de especialización incluyen métodos de aprendizaje de máquina aplicados a las ciencias sociales, nuevas fuentes de información y econometría aplicada.
Contacto: jose.ruizs@inegi.org.mx

Jael Pérez Sánchez

Nació en México. Es licenciado en Economía por la UAA y concluyó la Maestría de Estadística Oficial en el CIMAT; además, cuenta con un diplomado en Demografía por El Colegio de México (COLMEX). Actualmente, ocupa el puesto de subdirector de Estandarización de Clasificaciones y Estrategias de Codificación y es vicepresidente regional de la Federación Nacional de Colegios de Economistas.
Contacto: jael.perez@inegi.org.mx

Adrián Pastor López Monroy

Mexicano. Obtuvo el Doctorado en Ciencias de la Computación por el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), recibiendo el premio a la mejor Tesis Doctoral otorgado por la Sociedad Mexicana de Inteligencia Artificial (SMIA). Es profesor-investigador de tiempo completo en el CIMAT, Guanajuato, y su área de interés es el *Deep Learning* aplicado al procesamiento del lenguaje natural (PLN); en el 2017 y 2018, se desempeñó como posdoctorado y catedrático en la Universidad de Houston, en Texas, Estados Unidos de América. Actualmente, es miembro de la Asociación Mexicana del PLN y del Sistema Nacional de Investigadores.
Contacto: pastor.lopez@cimat.mx

Matthew Miller

Nació en Estados Unidos de América (EE. UU.). Obtuvo el título de maestro en Estadística por la Universidad Texas A&M en EE. UU. En el ámbito laboral, es un bioestadístico en el Centro Sealy de Envejecimiento en la Universidad de Texas, Campus Médico; antes de esa adscripción, fue asistente de investigación en el Departamento de Química y Bioingeniería de la Universidad Rice.
Contacto: mrmiller@utmb.edu

Alejandra Michaels-Obregón

De nacionalidad colombiana. Es economista por la Pontificia Universidad Javeriana en Bogotá, Colombia, con Maestría en Economía de la misma institución, y cuenta con un *MBA* por la Universidad de St. Thomas en Houston, Texas. En la actualidad, es coordinadora de proyectos de investigación en el Centro Sealy de Envejecimiento de la Universidad de Texas Campus Médico.
Contacto: almichae@utmb.edu

Karina Orozco Rocha	De nacionalidad mexicana. Es economista por la Universidad de Colima, doctora en Estudios de Población por COLMEX y con posdoctorado por esa misma universidad. Actualmente, es profesora en esta última institución académica. Es miembro del SNI del CONACYT con nivel 1, y sus líneas de investigación examinan las desigualdades de género en el trabajo remunerado y no remunerado, así como el bienestar económico y redes de protección en las edades avanzadas en México. **Contacto:** korozco9@ucol.mx
Rebeca Wong	Nació en México. Es actuaria por la UNAM y cursó el Doctorado en Economía en la Universidad de Michigan. Cuenta con amplia experiencia en la demografía económica del país. Es profesora en la Universidad de Texas, Campus Médico, y realiza investigación sobre salud y envejecimiento en México, así como en otras poblaciones latinas en EE. UU. y América Latina. **Contacto:** rewong@utmb.edu
José Manuel Lecuanda Ontiveros	De nacionalidad mexicana. Es licenciado en Matemáticas Aplicadas por la UABC y estudió la Maestría en Teoría Económica y el Doctorado en Economía en el ITAM. En el sector público se desempeñó como titular de las direcciones de Análisis Económico en la Secretaría de Finanzas del Gobierno del Distrito Federal y en la Presidencia de la República del Gobierno Federal; actualmente, es profesor en la Escuela de Graduados en Administración y Negocios en CETYS Universidad. **Contacto:** manuel.lecuanda@cetys.mx
Olinca Páez	Demógrafa y economista mexicana (El Colegio de México/Universidad Veracruzana), doctoranda en Ciencias de la Salud Pública (Universidad de Guadalajara), con diplomados en *Género, Sexualidad y Derecho* y *Gobierno, Gestión y Políticas Públicas* (CIDE Región Centro). Obtuvo el segundo lugar del Premio Nacional de Investigación Social y de Opinión Pública 2012 y en el 2018 fue beneficiaria del *ISI-World Bank Trust Fund for Statistical Capacity Building*. Participó en el *SLLS Summer School 2019 on Longitudinal and Life Course Research* (Université de Genève) con el objetivo de desarrollar una línea de investigación sobre la diferenciación de las trayectorias vitales por gé-

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFIA. Vol. 13, Núm. 2, mayo-agosto, 2022. 2022

nero y su impacto en el bienestar de las poblaciones. Es investigadora desde el 2003 y desde septiembre del 2014 es subdirectora de Investigación de Información Econométrica en el INEGI.
Contacto: olinca.paez@inegi.org.mx

Jonathan Heath

Es economista, egresado de la Universidad Anáhuac y tiene estudios de posgrado en Economía por la Universidad de Pensilvania. Posee 40 años de experiencia en el análisis de la economía mexicana y sus perspectivas. En este tiempo, ha sido economista principal de México en varias instituciones financieras globales y consultorías internacionales, además de profesor en las universidades Panamericana, Anáhuac, de las Américas e Iberoamericana, así como en el ITESM, donde ha impartido cursos relacionados con la economía mexicana, su historia y sus perspectivas. De igual modo, ha impartido materias de macroeconomía, política monetaria, inflación y empleo en la Universidad Autónoma Metropolitana Azcapotzalco como profesor invitado de tiempo completo. Ha sido conferencista en más de 50 universidades en México y Estados Unidos de América. Es autor de *Lo que indican los indicadores*, *Lecturas en lo que indican los indicadores, Para entender al Banco de México, La maldición de las crisis sexenales* y *El dinero*. Es el creador de los indicadores IMEF Manufacturero y No Manufacturero, al igual que de la Encuesta Mensual de Expectativas IMEF.
Contacto: jonathanheath54@gmail.com

Política y lineamientos editoriales

Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía es una publicación cuatrimestral que sirve de enlace entre la generación de la información estadística y geográfica oficial y la investigación académica para compartir el conocimiento entre especialistas e instituciones con propósitos similares.

Se publicarán sólo artículos inéditos y originales relacionados con la situación actual del uso y aplicación de la información estadística y geográfica a nivel nacional e internacional.

Es una revista técnico-científica, bilingüe, cuyos trabajos son arbitrados por pares (especialistas), bajo la metodología doble ciego, con los siguientes criterios de evaluación: trabajos inéditos, originalidad, actualidad y oportunidad de la información, claridad en la definición de propósitos e ideas planteadas, cobertura de los objetivos definidos, estructura metodológica adecuada y congruencia entre la información contenida en el trabajo y las conclusiones.

El resultado del proceso de dictaminación se comunica por correo electrónico y contempla tres variantes: recomendado ampliamente (con modificaciones menores), recomendado (pero condicionado a modificaciones sugeridas) y no recomendado (rechazado). Dos dictámenes aprobados, se notifica al autor que se publica y se envía a corrección de estilo; un aprobado y uno rechazado, se le solicita realizar cambios; y dos rechazados, se notifica la no publicación.

***Indizaciones* y registros**

- LATINDEX Catálogo (Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal).
- CLASE (Citas Latinoamericanas en Ciencias Sociales y Humanidades).
- REDIB (Red Iberoamericana de Innovación y Conocimiento Científico).

Lineamientos para publicar

Se publicarán trabajos en español e inglés: artículos de investigación, revisión y divulgación; ensayos; metodologías; informes técnicos; comunicaciones cortas; reseñas de libros; revisiones bibliográficas y estadísticas, entre otros.

1. El artículo —o cualquier otro tipo de escrito de los mencionados— deberá entregarse con una carta dirigida al editor responsable de REALIDAD, DATOS Y ESPACIO. REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA en la que se proponga el texto para su publicación, que se declare que es inédito y que no ha sido postulado de manera paralela en otro medio. Asimismo, deben incluirse los datos completos del(os) autor(es), nacionalidad(es), institución(es) de adscripción y cargo(s) que ocupa(n), domicilio(s) completo(s), correo(s) electrónico(s) y teléfono(s). Esto debe dirigirse a la atención de la M. en C. Virginia Abrín Batule, virginia.abrin@inegi.org.mx (tel. 5278 10 00, ext. 1161).
2. El trabajo se debe presentar en versión electrónica (formato *Word* o compatible) con: a) extensión no mayor de 20 cuartillas; b) letra Helvética, Arial o Times de 12 puntos y c) interlineado de 1.5 líneas. El material adicional al texto se requiere por separado: a) las imágenes, con resolución de 300 ppp y un tamaño no menor a 17 centímetros de base (ancho) en formato JPG o TIF —no remuestrear (ampliar) imágenes de menor resolución—; si son líneas o mapas, deben entregarse en formato vectorial (EPS o Ai), en caso de incluirse imágenes en mapa de bits, incrustarlas o enviarlas con el nombre con el cual se creó el vínculo (conservando los requerimientos de resolución y tamaño estipulados); y para fotografías, éstas no deben ser menores a 5 megapixeles; b) las fórmulas o expresiones matemáticas tienen que elaborarse con el editor de ecuaciones propio de *Microsoft*™, pero en caso de usar *software* de terceros, incluir en la entrega PDF testigo en el cual figuren exactamente cómo deben representarse; c) las gráficas, que incluyan el archivo en *Excel* con el cual se desarrollaron o, en su defecto, la imagen JPG legible, de origen, en alta resolución; y d) los cuadros, que sean editables, no se deben insertar como imagen.
3. La colaboración debe incluir: título del trabajo (en español e inglés o viceversa); resúmenes del trabajo en español e inglés (que no excedan de un párrafo de 10 renglones); palabras clave en español e inglés (mínimo tres, máximo cinco); bibliografía u otras fuentes; así como breve(s) semblanza(s) del(os) autor(es) que no exceda(n) de un párrafo de cinco renglones y que incluya(n) nacionalidad(es), grado(s) académico(s), principal(es) experiencia(s) profesional(es), adscripción(es) laboral(es) actual(es) y dirección(es) electrónica(s) de contacto.
4. Las referencias bibliográficas u otras fuentes deberán presentarse al final del artículo de la siguiente manera: nombre(s) del(os) autor(es) comenzando por el(los) apellido(s); título de la publicación con cursivas (si se trata de un artículo, debe estar entrecomillado, seguido de coma y la preposición en con dos puntos y, enseguida, el título de la revista o libro donde apareció publicado, con cursivas); país de origen; editorial; lugar y año de edición; página(s) consultada(s). En el caso de las fuentes electrónicas (páginas *web*) se debe seguir el mismo orden que en las bibliográficas, pero al final se pondrá entre paréntesis DE (dirección electrónica), la fecha de consulta y la liga completa. Se tienen que omitir aquellas que se mencionen como notas a pie de página. Si se aplica la opción de incluir en cuerpo de texto la referencia de nombre de autor y año de la fuente consultada entre paréntesis, sí deben aparecer todas las referencias mencionadas.

Página electrónica: **http://rde.inegi.org.mx**

Editorial Guidelines and Policy

Reality, Data and Space International Journal of Statistics and Geography is a four-monthly publication that connects statistics and geographic official information with academic research in order to share knowledge among specialists and institutions with similar aims.

We will publish only original and unpublished articles related to the current use and appliance of statistical and geographical information at both national and international levels.

It is a technical-scientific and bilingual magazine, with articles previously peer-reviewed by specialists under a double-blind methodology with the following evaluation criteria: unpublished works, originality, information related to opportunity and current affairs, we expect clarity in the definition of aims and ideas stated, defined objectives coverage, accurate methodological structure and coherence between the information of the paper as well as its conclusions.

The result of the paper-assessment process is delivered by email, and it involves three possibilities: fully recommended (with slight modifications), recommended (on condition of suggested modifications) and not recommended (i.e. rejected). When there are two reports of approval, the author gets notified that his/her paper will be published and it is sent to a style editing process. When one report approves the paper for publication and another one rejects it, the author is requested to make some changes for the text to be published. If the text submitted receives two non-favourable reports, the author is notified that the text will not be published.

Index and Registers

- LATINDEX Catalogue (Online Regional Information System for Scientific Journals from Latin America, the Caribbean, Spain and Portugal).
- CLASE (Latin American Quotations in Humanities and Social Sciences)
- REDIB (Latin American Net of Innovation and Scientific Knowledge)

Publishing Guidelines

Articles will be published in Spanish or English: research, revision and scientific-spreading articles; methodologies; technical reports; short texts; book reviews; and bibliographical and statistical revisions, among others.

1. The article —or any other kind of text from those aforementioned— must be delivered with an attached letter addressed to the chief editor of Reality, Data and Space. International Statistics and Geography Magazine in which the text intended for publication will be submitted. There it must be stated that the text has not been published, and that it has not been submitted for publication in any other media. The names in full of the authors must be included, as well as their nationalities, adscription institutions, position in those institutions, postal address, e-mail address, and telephone numbers. This must be addressed to MSc Virginia Abrín Batule, Virginia.abrin@inegi.org.mx (tel (+52) (55) 52.78.10.00, extension 1161.
2. The article must be submitted in an electronic version (a Microsoft Word file or a compatible one) with the following format: a) the text should not exceed the 20 pages of length; b) typography must be Helvetic, Arial or Times (12 points); and c) there should be a 1.5 line spacing in each paragraph. Additional material to the text will be delivered separately: a) images with a resolution of 300 ppp and no smaller than 17 cm width will be delivered in format JPG or TIF —please do not amplify images with lower resolution—. If the added materials are lines or maps, these must be delivered in vectorial format (EPS or Ai). If there are images in bits map, these must be embedded or attached with the name of the original file with which the link was created (keeping the resolution and size requirements above stated). As regards to photographs, these should not be inferior as 5 megapixels; b) mathematical expressions or formulae have to be created with the equations editor by Microsoft™, but in case of using third-parties software, please attach a witness PDF in which the exact representation of mathematical formulae or expressions is contained; c) graphics must include the Excel file in which they were created or a legible image in the original JPG format in high resolution; and d) charts must be editable, and must not be inserted as images.
3. The text must include the following: the article's title (both in English and Spanish); the abstract of the article—both in English and Spanish (not longer than a 10-line paragraph); key words—both in English and Spanish (three as minimum and five as maximum); bibliography and other sources; as well as brief biographical sketches of the authors not exceeding a five-line paragraph each including nationalities, academic titles, main professional experiences, current work-related affiliations, and electronic addresses for the authors to be contacted.
4. Bibliographical references and other sources must be included at the end of the article in the following way: author's name (Surname first), and publication's title (in italics). If it is an article, the tile must be in quotation marks followed by a coma and the preposition "en" with semicolon (in Spanish), then it should appear the title of the book or magazine in which the article was published (in italics); country of origin; publishing house, edition year, and consulted pages. As regards to electronic sources (web pages) the same order of the bibliographical references must be followed, but at the end the word "EA" (as for Electronic Address) ("DE" in Spanish) must be added within parenthesis followed by consultation date and the complete reference link. Those web links referred previously as footnotes, must be omitted in this section. However, if the name of the author and the year of the consulted source were included in the main body of the text within parenthesis, all these must be included as part of the bibliographical references.

Webpage: **http://rde.inegi.org.mx**

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

MÉXICO