Vol. 10, Núm. 3, septiembre-diciembre, 2019.

ISSN 2007-2961

REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

INTERNATIONAL JOURNAL OF STATISTICS AND GEOGRAPHY

REALITY, DATA AND SPACE

INEGI

EL CENSO SIRVE PARA SABER CÓMO ES TU ENTORNO

¡PREGÚÚÚNTAME!

YA VIENE EL CENSO
MARZO 2020

INEGI

2020 censo de Población y Vivienda

Vol. 10, Núm. 3, septiembre-diciembre, 2019.

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

Contenido

INEGI

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Presidente del Instituto
Julio Alfonso Santaella Castell

Vicepresidentes
Enrique de Alba Guerra
Paloma Merodio Gómez
Enrique Jesús Ordaz López
Adrián Franco Barrios

Dirección General de Estadísticas Sociodemográficas
Edgar Vielma Orozco

Dirección General de Estadísticas de Gobierno, Seguridad Pública y Justicia
Óscar Jaimes Bello

Dirección General de Estadísticas Económicas
José Arturo Blancas Espejo

Dirección General de Geografía y Medio Ambiente
María del Carmen Reyes Guerrero

Dirección General de Integración, Análisis e Investigación
Sergio Carrera Riva Palacio

Dirección General de Coordinación del Sistema Nacional de Información Estadística y Geográfica
María Isabel Monterrubio Gómez

Dirección General de Comunicación, Servicio Público de Información y Relaciones Institucionales
Eduardo Javier Gracida Campos

Dirección General de Administración
Marcos Benerice González Tejeda

Contraloría Interna
Manuel Rodríguez Murillo

CONSEJO EDITORIAL

Enrique de Alba Guerra
Presidente del Consejo

Fernando Cortés Cáceres
Profesor Emérito de FLACSO
PUED de la UNAM
México

Gerardo Bocco Verdinelli
Universidad Nacional Autónoma de México
México

Juan Carlos Chávez Martín del Campo
Banco de México
México

Lidia Bratanova
UNECE Statistical Division
Switzerland

Tonatiuh Guillén López
El Colegio de la Frontera Norte
México

Víctor Manuel Guerrero Guzmán
Instituto Tecnológico Autónomo de México
México

Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía

Editor responsable
Sergio Carrera Riva Palacio

Editor técnico
Gerardo Leyva Parra

Coordinación editorial
Virginia Abrín Batule y Mercedes Pedrosa Islas

Corrección de estilo
José Pablo Covarrubias Ordiales y Laura Elena López Ortiz

Corrección de textos en inglés
Gerardo Piña

Diseño y formación edición impresa
Juan Carlos Martínez Méndez y Eduardo Javier Ramírez Espino

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. 9, Núm. 3, septiembre-diciembre, 2019, 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, imprenta y distribución: 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.

El contenido de los artículos, así como sus títulos y, en su caso, fotografías y gráficos utilizados son responsabilidad del autor, lo cual no refleja necesariamente el criterio editorial institucional. Asimismo, la Revista se reserva el derecho de modificar los títulos de los artículos, previo acuerdo con los autores. La mención de empresas o productos específicos en las páginas de la Revista no implica el respaldo por el Instituto Nacional de Estadística y Geografía.

Se permite la reproducción total o parcial del material incluido en la Revista, sujeto a citar la fuente. Esta publicación consta de 400 ejemplares y se terminó de imprimir en noviembre del 2019.

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

Editorial

El número final del 2019 presenta, en primer lugar, *Comparación de metodologías de imputación aplicadas a ingresos laborales de la ENOE* (*Comparison of Imputation Methodologies Applied to Labor Income of the ENOE*), estudio que el autor plantea porque la no declaración de ingresos y la declaración en intervalos de múltiplos de salarios mínimos en su conjunto en esa encuesta del INEGI es una situación que está creciendo y pude ser insostenible a futuro, pues serán más las observaciones con datos imputados que con observados; en este sentido, sugiere poner a consideración de los usuarios de la ENOE la adopción de esta propuesta de procedimiento como parte de la preparación de datos para la generación de indicadores, aunque lo deseable es que el propio INEGI lo haga y que ponga a disposición de estos tanto los datos observados como los imputados y toda la información relacionada con la imputación; sin embargo, para revertir esa tendencia, sería necesario que el Instituto tome las medidas necesarias en el diseño de instrumentos de captación y en el operativo de campo.

Enseguida, *Análisis jerárquico de las emisiones de gases efecto invernadero en México* (*Hierarchical Analysis of Greenhouse Gas Emissions in Mexico*) es una herramienta para entender la mecánica de interacción entre la información económica y ambiental de los sectores industriales desde 1999 al 2012, la cual permite plantear nuevas hipótesis sobre su comportamiento al analizar la evidencia estadística existente. Las técnicas de modelado jerárquico bayesiano que se emplearon no se han encontrado en la literatura del ámbito nacional o internacional, por lo que constituye una aplicación novedosa.

En *Movilidad laboral internacional en el caso mexicano* (*International Labor Mobility. The Case of Mexico*) se plantea que la medición de este fenómeno es relevante para las políticas (en materia de integración, empleo y provisión de servicios públicos) de corto, mediano y largo plazo que los países receptores deben diseñar e instrumentar. Se expone de forma breve el acuerdo internacional en su conceptualización y las recomendaciones para su adecuada medición. Con ello como referencia, se analiza el caso para México, principalmente como nación receptora, desde la localización de las fuentes disponibles, la descripción de la información que compilan y los límites de la misma, hasta la caracterización de volúmenes y flujos de trabajadores migrantes internacionales y extranjeros no residentes.

A continuación, en *Actualización de la matriz total de insumo-producto de México del 2003. Aplicación de los métodos de doble deflación y RAS* (*An Update of the Mexican Input-Output Table of 2003. An Application of the RAS and the Double Deflation Methods*) se menciona que la comparación de diversas matrices permite conocer la evolución de la estructura productiva de una economía durante un periodo determinado, sin embargo, requiere del uso de aquellas valuadas a los mismos precios. Se encontró que, de las disponibles para México, la del 2003 podía homologarse para ser comparada con las del 2008 y 2012 (ya que estas se encuentran valuadas con igual nivel de precios y consideran el mismo sistema de clasificación industrial), pero requería la estimación a precios del 2008. Aquí se detallan los pasos para su obtención y se presenta la información utilizada en ambos métodos.

El artículo *Registro de los nacimientos en México. Una mirada crítica de su evolución en las últimas tres décadas* (*The Birth Registry in Mexico. A Critical View of its Evolution During the Last Three Decades*) tiene el objetivo de evaluar el progreso en el tiempo de la captación de datos de estos hechos en México, cuya justificación radica en que las estadísticas vitales, derivadas del Registro Civil, constituyen información estratégica para la planeación del país. El análisis se hizo utilizando la temporalidad del registro y el contraste con otras fuentes, de las cuales resalta la pertinencia, como referencia para su validación, de los censos de población y del Subsistema de Información sobre Nacimientos, recabado este último por la Secretaría de Salud.

Para cerrar la edición, *Funcionamiento en muestras finitas de técnicas de imputación y retropolación: caso de las series de encuestas económicas nacionales del INEGI* (*Finite Sample Performance of Imputation and Retropolation Techniques: the INEGI's National Economic Surveys' case*) es un estudio orientado a evaluar el desempeño de dichas técnicas en la construcción de indicadores adecuados en el contexto de series de tiempo y bajo diferentes estructuras estocásticas cuyos resultados puedan ser útiles para los generadores de información oficial. Aun cuando existen investigaciones previas que lo han hecho, no hay una que sea integral y que valore a fondo; por ello, este artículo pretende cubrir esa laguna en la literatura.

http://rde.inegi.org.mx

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Comparación de metodologías de imputación aplicadas a ingresos laborales de la ENOE

Comparison of Imputation Methodologies Applied to Labor Income of the ENOE

Benito Durán Romo*

* Instituto Nacional de Estadística y Geografía, benito.duran@inegi.org.mx

Nota: el autor agradece los valiosos comentarios de Gerardo Leyva Parra y la colaboración de Lilia Guadalupe Luna Ramírez para la elaboración de este documento.

Snow geese, Anatidae/De Agostini/Getty Images

En las encuestas por muestreo aparece un problema muy común: la no respuesta; esta puede ser completa cuando no se consigue la entrevista o parcial cuando falta información de alguna sección o de tan solo una pregunta. La solución del primer caso no crea mayor conflicto, pues se resuelve ajustando los factores de expansión por no respuesta; sin embargo, el segundo presenta ciertas complicaciones.

La Encuesta Nacional de Ocupación y Empleo del INEGI no está exenta de estas dificultades, ya que un porcentaje importante y creciente de personas ocupadas no responde a cierto número de preguntas pero, sobre todo, omite la declaración de sus ingresos por trabajo, presentando para este tema 6.7% de no respuesta en el primer trimestre del 2005 con un incremento permanente que alcanzó 16.8% en el cuarto del 2017.

Las prácticas más frecuentes para tratar los casos con no respuesta parcial son la eliminación por lista o eliminación por pares, aunque ambas tienen sus inconvenientes. Otra forma de lidiar con datos faltantes en el análisis es mediante la imputación de estos utilizando metodologías de imputación simple o múltiple.

Por lo anterior, el objetivo de este trabajo es mostrar los resultados de un ejercicio comparativo de algoritmos y metodologías de imputación de ingresos laborales de la Encuesta para ponerlas a consideración de los usuarios de esta y valorar su posible adopción como solución a los ingresos laborales faltantes, ejercicio que está basado en algunas medidas de desempeño y de los efectos que la imputación puede tener en el Índice de Tendencia Laboral de la Pobreza emitido por el Consejo Nacional de Evaluación de la Política de Desarrollo Social cada trimestre. Con esta investigación se encontraron resultados muy relevantes, como que los ingresos per cápita obtenidos por el Consejo están subestimados entre 16.7 y 23.5% en promedio, dependiendo de la metodología que se utilice, provocando así una diminución promedio entre 1 y 4% en el Índice.

Palabras clave: no respuesta; no respuesta completa; no respuesta parcial; datos faltantes; imputación; imputación simple; imputación múltiple; Reglas de Rubin; eficiencia relativa; ENOE; ingresos laborales; ITLP.

Recibido: 24 de octubre de 2018.
Aceptado: 15 de febrero de 2019.

A very common problem appears in sampling surveys: the "non-response". This can be absolute when the interview is not taken or partial when information is missing from any section or from just one question. The resolution of the first case does not create greater conflict, as it is resolved by adjusting the expansion factors for non-response. However, the second one presents certain complications.

The National Survey of Occupation and Employment of the INEGI is not exempt from these difficulties, since a significant and growing percentage of employed people do not answer a certain number of questions but, above all, omit the declaration of their income. On this subject, there was a 6.7% of non-response in the first quarter of 2005 with a permanent increase that reached 16.8% in the fourth quarter of 2017.

The most frequent practices to treat cases with partial non-response are elimination by list or by pairs, although both have their drawbacks. Another way to deal with missing data in the analysis is by imputing them using single or multiple imputation methodologies.

Therefore, the objective of this work is to show the results of a comparative exercise of algorithms and methodologies of imputation of labor income of the Survey to be put to the consideration of its users and to assess its possible adoption as a solution to the missing labor income, an exercise that is based on some performance measures and the effects that the imputation can have on the Labor Trend Index of Poverty issued by the National Council for the Evaluation of Social Development Policy every quarter. With this research, very relevant results were found, such as that the per capita income obtained by the Council is underestimated between 16.7 and 23.5% on average, depending on the methodology used, thus causing an average decrease between 1 and 4% in the Index. This study found very relevant results, such as the per capita income obtained by the CONEVAL are underestimated between 16.7 and 23.5% on average, depending on the methodology used, thus causing an average decrease between 1 and 4% in the ITLP.

Key words: non-response; complete non-response; partial non-response; missing data; imputation; simple imputation; multiple imputation; Rubin's Rules; relative efficiency; ENOE; labor income; ITLP.

Introducción

En las encuestas por muestreo es común que se presenten problemas de no respuesta en dos sentidos: completa y parcial. La primera ocurre cuando no se logra la entrevista con la unidad de observación debido a que, por ejemplo, no pudo ser localizada por un problema de actualización del marco (una vivienda que en este aparece habitada y al momento de visitarla ya no lo está) o porque, aunque fue ubicada no fue posible contactarla, o bien, fue contactada pero sus ocupantes se negaron a proporcionar información. Por su parte, la parcial se da cuando aun lograda la entrevista no se dan datos para alguna sección o algunas preguntas de la entrevista, bien porque el informante no los tiene o, simplemente, no la(s) quiso contestar.

Al concentrarnos en el segundo caso, y en específico en las preguntas relacionadas con el ingreso de los individuos, este fenómeno se ha vuelto muy recurrente y se ha incrementado de manera persistente en México. Esta no respuesta en el reporte de ingresos puede deberse, principalmente, a que la información no es proporcionada por el perceptor directo, sino por un tercero que la conoce de forma parcial o que la desconoce en su totalidad; pero también puede ser causada por el miedo que provoca la creciente percepción de inseguridad en el país.

La Encuesta Nacional de Ocupación y Empleo (ENOE) no está exenta de este fenómeno, ya que reporta un porcentaje importante y al alza de personas ocupadas (como trabajadores subordinados remunerados, empleadores y trabajadores por cuenta propia) que no declaran sus ingresos por trabajo.

La variable de los ingresos laborales en la ENOE presentó 6.7% de no respuesta en el primer trimestre del 2005 y mantuvo un incremento permanente hasta alcanzar 16.8% durante el cuarto del 2017.

Pero el problema realmente no termina ahí, pues todo tipo de análisis realizados con información de esta encuesta y que involucra al ingreso por trabajo se llevan a cabo, la mayoría de las veces, eliminando las observaciones de los individuos con no respuesta en esta variable, o bien, considerándolos como cero ingreso, provocando que los resultados de esos análisis presenten sesgo.

La forma más común de resolver el problema de no respuesta completa de la unidad de observación es distribuyendo su factor de expansión en las unidades con respuesta del mismo conglomerado —el cual puede ser la Unidad Primaria de Muestreo (UPM)—, pero este procedimiento no es aplicable cuando falta respuesta de alguna pregunta durante la entrevista o, por lo menos, no es práctico: aquí lo recomendable sería imputar la respuesta.

La forma de proceder para preguntas con no respuesta parcial o datos faltantes rellenarlas usando metodologías de imputación simple o múltiple.

Aunque los procedimientos de imputación de datos faltantes han sido adoptados por un gran número de oficinas nacionales de estadística (ONE) desde hace varios años, el Instituto Nacional de Estadística y Geografía (INEGI), de México, no los ha implementado en ninguno de sus proyectos estadísticos en hogares.

Ante esto, el objetivo de este trabajo es mostrar los resultados de un ejercicio comparativo de algoritmos y metodologías de imputación de ingresos laborales de la ENOE para ponerlas a consideración de los usuarios de esta y valorar su posible adopción como solución a los ingresos laborales faltantes. Esta comparación se basa en algunas medidas de desempeño, pero también en la evaluación de los efectos que la imputación puede tener en el Índice de Tendencia Laboral de la Pobreza (ITLP) que emite el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) cada trimestre y que es el uso más visible que se la ha dado a los ingresos laborales de la ENOE.

La presente investigación está organizada como sigue: en la primera sección se aborda una descripción general de la fuente de datos y sus problemas de no respuesta en los ingresos; algunas generalidades sobre imputación de datos faltantes y algunos casos de usos se tratan en la segunda; en la

tercera se describen las metodologías de imputación a comparar; los resultados se muestran en la cuarta; y las conclusiones, en la quinta.

1. Los datos

La fuente que se usó para este trabajo fue la ENOE, como ya se había mencionado; esta proporciona amplia información sobre la fuerza laboral en México permitiendo su identificación y clasificación.

El objetivo general de la Encuesta es garantizar que se cuente con una base de información estadística acerca de las características ocupacionales de la población en el ámbito nacional, así como con la infraestructura sociodemográfica que permita profundizar en el análisis de los aspectos laborales.

La ENOE es un levantamiento continuo en hogares con una muestra de más de 120 mil viviendas organizadas en cinco paneles rotatorios que permanecen activos durante cinco trimestres consecutivos: se hace una visita a la vivienda por trimestre y se reemplaza cada panel de manera escalonada. Con ese tamaño de muestra se logra hacer inferencia estadística con desglose nacional, urbano y rural, por entidad federativa y 40 ciudades autorrepresentadas.

La Encuesta contempla un número importante de preguntas que permiten tanto determinar la condición de ocupación (ocupado, buscador de trabajo o no económicamente activo) de la población como conocer el contexto laboral de los ocupados en su empleo principal, así como las características de la unidad económica donde laboran y, también, algunos aspectos del empleo secundario, cuestiones relacionadas con antecedentes laborales, apoyos económicos, etcétera.

Dentro del ámbito laboral de los ocupados se incluyen dos preguntas para registrar el ingreso por trabajo que declare el informante: en la primera se indaga el monto y la frecuencia de pago; en caso de que no se proporcione la información, se hace la segunda, la cual presenta algunos intervalos de múltiplos de salarios mínimos, dando opción a que se seleccione cualquiera de estos.

Aun y con la presencia de las dos formas de proporcionar datos de los ingresos laborales por parte de los informantes, la ENOE ha registrado un incremento permanente en la no respuesta de estos. Como se puede observar en la gráfica 1a, esta comenzó con 6.7% en el primer trimestre del 2005 y alcanzó 17.2% en el tercero del 2017 (nivel más alto de la serie), volviendo a disminuir un poco (16.8%) en el cuarto de ese año.

En la misma gráfica se puede notar que la declaración en intervalos también se incrementó, aunque no en la misma magnitud que la no respuesta, al comenzar con 3.9% en el primer trimestre del 2005 y llegar a 10.7% en el cuarto del 2017.

Estos dos fenómenos han provocado que la declaración de ingreso en monto haya disminuido de forma constante al pasar de 89.4 a 72.5% entre el primer trimestre del 2005 y el cuarto del 2017, por lo que, si se va más allá y se considera a la declaración de ingresos en intervalos como no respuesta, entonces esta se incrementaría todavía más, teniendo 10.6% al inicio de la serie y alcanzando 27.5% en el cuarto trimestre del 2017, como se aprecia en la gráfica 1b.

Y con esto último nos quedaremos para este trabajo, es decir, se considerará como no respuesta tanto a la falta de declaración de ingresos como a la declaración en intervalos de múltiplos de salarios mínimos. Este último caso es debido a que el tratamiento más común que se da es imputarle el punto medio del intervalo, lo cual es impreciso y arbitrario.

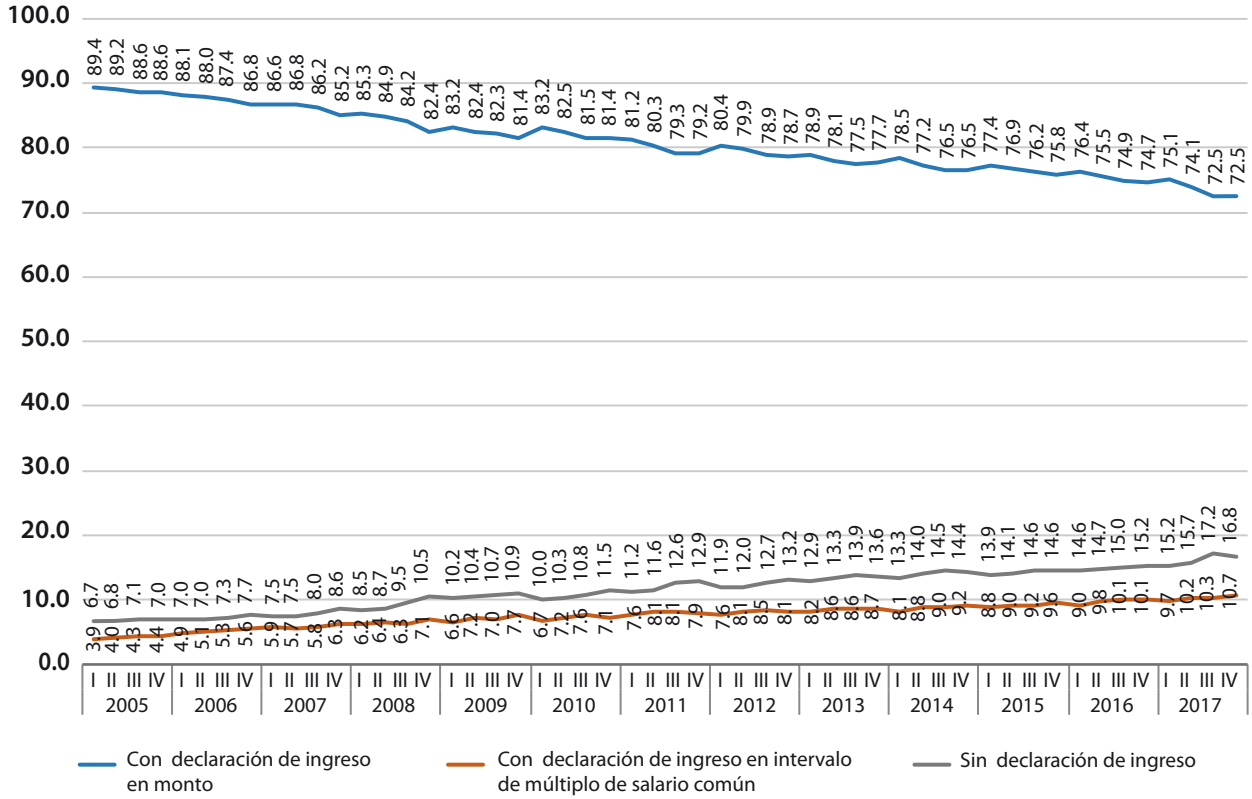
Cabe hacer mención de que estos dos fenómenos en su conjunto han estado creciendo 0.3% en promedio por trimestre, por lo que, de seguir esta tendencia, en 10 años estará rondando en 40 por ciento.

La falta de declaración de ingresos por trabajo puede deberse a múltiples factores, siendo uno de

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Gráfica 1a

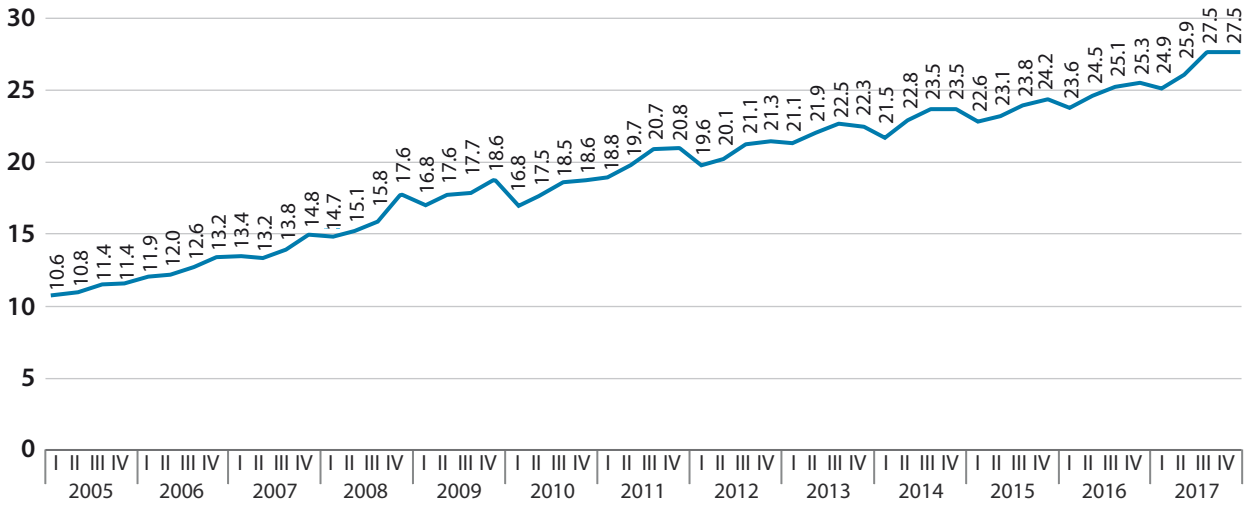
Ocupados por declaración de ingresos por trabajo (distribución porcentual)



Fuente: INEGI. *Encuesta Nacional de Ocupación y Empleo.*

Gráfica 1b

Ocupados con no respuesta de ingresos por trabajo (distribución porcentual)



Nota: incluye a los que no declararon ingreso más los que declararon ingreso en intervalos de múltiplos de salarios mínimos.
Fuente: INEGI. *Encuesta Nacional de Ocupación y Empleo.*

los más comunes cuando el informante adecuado no siempre es el informante directo, es decir, este es el que percibe en realidad los ingresos y es quien debería declararlos, pero por alguna razón no pudo ser entrevistado, y el adecuado es un integrante del hogar que, supuestamente, conoce la información del resto, sin embargo, esto no siempre ocurre así.

Otro factor que ha cobrado importancia en los últimos años es el constante incremento en la percepción de inseguridad en México, lo cual ha ocasionado que los informantes subdeclaren el monto del ingreso percibido, o bien, se nieguen a proporcionar el dato por miedo a ser víctimas de algún delito.

Aun y con este fuerte problema, se han hecho trabajos de análisis/generación de indicadores derivados de los ingresos laborales, y la manera de tratar los faltantes ha sido eliminando las observaciones, o bien, considerándolos como si no hubieran percibido ingresos (como cero ingreso), a pesar de saber que estas prácticas arrojan estimadores sesgados.

El Índice de Tendencia Laboral de la Pobreza es un caso donde se eliminaron las observaciones con ingresos faltantes. El Índice tiene como objetivo conocer la tendencia del poder adquisitivo del ingreso laboral, usando la ENOE como fuente de información.

Con lo expuesto anteriormente, es evidente que la ENOE tiene un problema importante de ingresos faltantes; con esto en mente, el presente ejercicio propone la alternativa de rellenarlos con datos imputados, por lo que se probarán diversos algoritmos/metodologías para llevarlo a cabo.

Una vez imputados los ingresos faltantes se realizará un ejercicio comparativo de esos algoritmos/metodologías en dos etapas, la primera a través de un conjunto de medidas de desempeño (error estándar, R^2, raíz del error cuadrático medio y el error medio absoluto) y la segunda usando el propio ITLP y algunos indicadores derivados de este.

La imputación de ingresos faltantes en la ENOE es una opción aceptable para solucionar este problema, pero, de seguir al alza la falta de declaración de ingresos en esta encuesta, en el futuro serán más las observaciones con datos imputados que con observados, por lo que es imperativo tomar las medidas necesarias en el operativo de campo para revertir esa tendencia.

2. Imputación de ingresos: conceptos y casos de uso

Las prácticas más comunes para llevar a cabo análisis con datos faltantes son la eliminación por lista (*listwise deletion*) y por pares (*pairwise deletion*). La primera, también llamada análisis de casos completos, elimina todas las observaciones que tengan por lo menos una variable con datos faltantes. En la segunda, denominada asimismo selección por variable o análisis de casos disponibles, solo descarta aquellas que presentan datos faltantes en la variable involucrada en el análisis.

Ambas tienen sus inconvenientes, por ejemplo, la eliminación por lista requiere que el mecanismo de datos faltantes sea *MCAR*,[1] como lo mencionan Peugh y Enders (2004), ya que de ser *MAR* se generan estimadores sesgados, además de que la muestra puede reducirse lo suficiente como para producir estimadores poco confiables. En la que se hace por pares, también indican que la comparabilidad dentro de un estudio es problemática debido a las diferencias que pueden resultar en el tamaño de los subconjuntos de datos, además de requerir el supuesto de *MCAR* para producir estimadores insesgados.

Una alternativa a la eliminación de observaciones con datos faltantes en análisis es la imputación de estos.

1 Peugh y Enders (2004) describen los mecanismos de datos faltantes como:
- *Missing Completely at Random (MCAR)* cuando los valores perdidos de la variable X no están relacionados con los valores de las demás variables ni tampoco con los valores subyacentes de la misma variable X.
- *Missing at Random (MAR)* cuando los valores perdidos de la variable X sí están relacionados con los valores de las demás variables, pero no con los valores subyacentes de la misma variable X.
- *Missing not at Random (MNAR)* cuando los valores perdidos de la variable X sí están relacionados con los valores subyacentes de la misma variable X.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

En Durán (2018) menciono que con la imputación se asigna un valor a una variable con no respuesta para que el cuestionario pase el proceso de validación en la entrada de datos y que la imputación puede ser una buena solución al trabajar con datos incompletos o faltantes (por no respuesta), pero que el procedimiento debe hacerse con el mayor de los cuidados porque, de no hacerlo, los datos completos pueden acabar muy sesgados y no mostrar la realidad que se pretende descubrir.

Para llevar a cabo este procedimiento, se pueden usar métodos de imputación simple, o bien, múltiple.

Imputación simple

En esta se asigna un solo valor a la variable con datos faltantes en cada una de las observaciones, dando como resultado un solo conjunto de datos completo.

El método tiene dos importantes características según Rubin (1987): la primera es que se pueden usar los métodos estándar de análisis de datos completos en el conjunto de datos imputado y la segunda es que cuando el conjunto de datos es de uso público, las imputaciones deben ser llevadas a cabo por el productor de los datos para que de esta forma sea incorporado su conocimiento sobre los mismos; pero Rubin también enumera dos desventajas: el valor imputado no refleja la variabilidad del muestreo sobre el valor real ni la incertidumbre que adicionan los datos faltantes.

La imputación basada en modelos (asignando una media, una mediana o el resultado de una regresión) y las *técnicas Deck* (*Hot Deck* y *Cold Deck*) son metodologías usadas en la imputación simple.

Imputación múltiple

Siguiendo el pensamiento de este mismo autor, esta conserva las virtudes de la simple, pero también corrige sus fallas, es decir, pueden usarse los métodos estándar de análisis de datos completos e incorporar el conocimiento del productor de los datos, el cual se ve reflejado en la incertidumbre sobre cuál dato imputar.

Con imputación múltiple se asignan m valores a la variable con datos perdidos para cada observación, dando como resultado m conjuntos de datos completos, es decir, se producen varios conjuntos de datos imputados (digamos m) donde cada uno de ellos contiene un valor imputado diferente para cada dato faltante. Se realiza el análisis por separado para los m conjuntos de datos y el valor del estimador será el promedio de los resultados de esos análisis, por ejemplo, el estimador del total será el promedio de los totales de los m conjuntos de datos, la media estará dada por el promedio de las medias de los m conjuntos de datos y así con otros estimadores. El cálculo de los errores estándar estará sujeto a las Reglas de Rubin (1987).[2]

Según Alisson (2012), hay dos razones por las que se requiere más de un conjunto de datos imputados: la primera se refiere a que, con un solo conjunto de datos, los estimadores serán altamente ineficientes debido a que tendrá más variabilidad de la necesaria, al promediar los resultados sobre varios conjuntos de datos esta se reduce; la segunda es que la variabilidad de las estimaciones en diversos conjuntos de datos proporciona la información necesaria para que los errores estándar reflejen con precisión la incertidumbre sobre los valores faltantes. También, menciona que estas dos razones tienen implicaciones en el número de imputaciones a efectuar.

Dado esto, Rubin (1987) introduce el término Eficiencia Relativa (ER) de los estimadores, la cual está dada por:

2 Reglas de Rubin (1987), ver también Rubin y Schenker (1986).

Sea $\hat{Q}_l$ el estimador puntual del i-ésimo conjunto de datos imputado con $i=1,2,\ldots,m$. Entonces, el estimador para Q sobre las múltiples imputaciones estará dado por el promedio de los m conjuntos de datos completos: $\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_l$.

La varianza estimada del estimador puntual estará dada por la combinación de las varianzas intra-imputación y entre-imputación de la siguiente forma: $T = \bar{W} + \left(\frac{m+1}{m}\right)B$, donde $\bar{W} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_l$ representa las varianzas intra-imputación y $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_l - \bar{Q})^2$, la entre-imputación.

El intervalo de confianza estará dado por $\bar{Q} \pm t_{v,1-\alpha/2}\, T^{1/2}$, con grados de libertad $v = (m-1)\left(1+\frac{m}{m+1}\frac{\bar{W}}{B}\right)^2$.

$$ER = (1 + \lambda/m)^{-1/2}$$

medida en unidades de desviaciones estándar y donde λ es la fracción de información faltante (algunos autores la refieren como la proporción de observaciones con datos faltantes) y m es el número de imputaciones.

Para entender de qué se está hablando, revisemos el cuadro 1, donde se observa la ER como porcentaje, que resulta de aplicar la fórmula anterior usando un número finito de imputaciones y cierta proporción de fracción de información faltante. Al usar como ejemplo la ENOE del cuarto trimestre del 2017 que alcanzó 27.5% de ingresos faltantes (casi 30%), se requieren 10 datos imputados para lograr una ER de 98.5%, es decir, con 30% de ingresos faltantes son suficientes 10 imputaciones para obtener estimadores 98.5% tan eficientes como los que se obtendrían con un número infinito de imputaciones. Es por esto que la mayoría de los expertos proponen que entre cinco y 10 imputaciones son suficientes para obtener estimadores eficientes.

Casos de uso

Algunas ONE usan, con frecuencia, los métodos de imputación simple para corregir los problemas de datos faltantes; por ejemplo, *Statistics Canada* usa la metodología *Hot Deck* para imputar en la Encuesta de la Fuerza Laboral (LFS, por sus siglas en inglés) y el Buró de Censos de Estados Unidos de América la utiliza en la Encuesta de Población Actual (CPS, por sus siglas en inglés) y en la Encuesta de Ingresos y Participación en el Programa (SIPP, por sus siglas en inglés).

En la Encuesta de Panel de Hogares Británica (BHPS, por sus siglas en inglés), *Hot Deck* se emplea en variables con poca presencia de valores como ingresos provenientes de inversiones o ahorros.

Cuadro 1

ER, en porcentaje, usando un número finito de imputaciones de acuerdo con cierta fracción de información faltante

m	Fracción de información faltante (λ)								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	95.3	91.3	87.7	84.5	81.6	79.1	76.7	74.5	72.5
2	97.6	95.3	93.3	91.3	89.4	87.7	86.1	84.5	83.0
3	98.4	96.8	95.3	93.9	92.6	91.3	90.0	88.9	87.7
4	98.8	97.6	96.4	95.3	94.3	93.3	92.3	91.3	90.4
5	99.0	98.1	97.1	96.2	95.3	94.5	93.7	92.8	92.1
10	99.5	99.0	98.5	98.1	97.6	97.1	96.7	96.2	95.8
15	99.7	99.3	99.0	98.7	98.4	98.1	97.7	97.4	97.1
20	99.8	99.5	99.3	99.0	98.8	98.5	98.3	98.1	97.8
25	99.8	99.6	99.4	99.2	99.0	98.8	98.6	98.4	98.2
30	99.8	99.7	99.5	99.3	99.2	99.0	98.9	98.7	98.5
35	99.9	99.7	99.6	99.4	99.3	99.2	99.0	98.9	98.7
40	99.9	99.8	99.6	99.5	99.4	99.3	99.1	99.0	98.9
45	99.9	99.8	99.7	99.6	99.4	99.3	99.2	99.1	99.0
50	99.9	99.8	99.7	99.6	99.5	99.4	99.3	99.2	99.1

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

En los casos donde la frecuencia de valores es alta se utiliza el método de *pareamiento por medias predictivas*. Al aplicar estos dos métodos pueden introducirse sesgos en las tasas de cambio de los valores entre rondas de levantamiento. Para evitarlo, se hace una imputación de rondas cruzadas, de tal forma que al seleccionar a la observación del donador no solo se escoge aquel con características similares al receptor, sino con valores similares obtenidos de las otras rondas de levantamiento (University of Essex, 2018).

Un caso interesante de imputación múltiple de ingreso es el que se realiza en la Encuesta de Gastos del Consumidor (CE, por sus siglas en inglés) recolectada por el Buró de Censos para el Buró de Estadísticas del Trabajo del Departamento del Trabajo de Estados Unidos de América, operativo que provee datos sobre gastos, ingresos y características demográficas de los consumidores estadounidenses.

De acuerdo con Paulin *et al*. (2006), el proceso de imputación múltiple en la CE comenzó en el 2004 con el fin de rellenar los huecos causados por no respuesta, logrando preservar la media de todas las fuentes de ingreso e incluso incorporar al cálculo de la varianza la incertidumbre causada por la imputación. Este proceso se lleva a cabo por medio de una regresión, donde los coeficientes obtenidos son distorsionados cada vez agregando ruido, los cuales son usados para estimar los datos faltantes (cinco estimaciones diferentes). Una vez estimados los datos, estos también son distorsionados agregándoles ruido con el fin de dar variabilidad a los resultados.

También, mencionan cómo obtener estimadores a partir de los datos imputados, pero enfatizando que las medidas de precisión de esos estimadores deben ser generadas tomando en cuenta las Reglas de Rubin (1987) usando los datos de las cinco estimaciones.

Cabe hacer mención que el Buró de Estadísticas del Trabajo pone a disposición del público en general los microdatos de la CE, los cuales incluyen los resultados de la imputación múltiple de las diferentes fuentes de ingreso y del ingreso total.

Otro caso para tomarse en cuenta es el presentado por Starick (2005), donde hace un análisis comparativo de metodologías de imputación aplicadas a los ingresos de la Encuesta sobre la Dinámica de los Hogares, los Ingresos y el Trabajo en Australia (HILDA, por sus siglas en inglés), la cual es una encuesta longitudinal que pone especial atención en la información de las viviendas y los hogares, además del trabajo e ingreso de los individuos.

Starick menciona que, aunque las metodologías de imputación para estudios transversales pueden ser una muy buena opción, tienen una desventaja: pueden introducir distorsiones en la tendencia de las estimaciones entre rondas de levantamiento. Ante esto, presenta la comparación de tres métodos de imputación para encuestas longitudinales: de regresión del vecino más cercano, *Little and Su* y *Little and Su* extendido.

En el primero se usa un paquete estadístico para construir los modelos de regresión que dependen de las rondas en que se observaron datos utilizando los valores predichos para buscar el vecino que tenga el valor más cercano al dato observado.

El segundo es un método de imputación simple para estudios longitudinales en variables continuas, el cual involucra información sobre la tendencia de los datos a través del tiempo y el dato de cada observación, de tal manera que el valor imputado está formado por los efectos de columna (que representa los cambios en la media a través del tiempo) y renglón (que es el nivel de la observación corregida por el de columna), además de un componente aleatorio como residual (calculado a partir de un valor observado cercano al de renglón), es decir:

$$imputación = efecto\ columna * efecto\ renglón * residual$$

Los efectos de este método (para mayores detalles sobre este, ver Eurostat, 2014) se calculan

con base en todos los datos observados en cada ronda de levantamiento.

En *Little and Su* extendido se usa el mismo procedimiento anterior, salvo que los efectos se calculan con base en los datos observados de receptores y donadores con características similares.

Starick basa la comparación en algunos criterios que son evaluados con sus respectivas métricas, diferentes a las que aquí se presentan. De acuerdo con los resultados de esos criterios, ella recomienda el uso del *Little and Su*, pero haciéndole mejoras cada vez.

En México se han hecho esfuerzos para subsanar el problema de ingresos laborales faltantes en la ENOE llevando a cabo algunos ejercicios de imputación, pero todos ellos han tenido fines de investigación, por ejemplo, Campos-Vázquez (2013) imputó ingresos a la ENOE (del 2005 al 2012) usando cuatro técnicas: *pareamiento por puntajes de propensión, Hot Deck, imputación en la mediana de un grupo más ruido* y *pareamiento por promedios predictivos*, esto con el objetivo de medir la pobreza laboral y compararla con las mediciones del CONEVAL; en ese estudio se encontró que la pobreza laboral es menor entre 6 y 7% a la reportada por el Consejo.

También, Rodríguez y López (2015) imputan ingresos faltantes a la ENOE para después analizar las diferencias en pobreza laboral y el posible sesgo en estimaciones de capital humano cuando se ignoran dichas observaciones. Las técnicas que usaron son *Hot Deck* aleatorio y *Hot Deck* con función de distancia utilizando características observables demográficas de cada individuo y de su hogar para encontrar el donante. Ellos encontraron que, al no considerar las observaciones con ingresos faltantes, la pobreza laboral está sobrestimada y que los retornos a la educación serían alrededor de medio punto porcentual más bajos si se contempla el conjunto de datos completo (con imputados) y el sesgo no sería significativo.

Los anteriores son ejemplos del tratamiento alternativo que se puede hacer a los datos faltantes en las encuestas en hogares con el fin de disminuir al mínimo posible el sesgo al generar resultados a partir de ellas. Por ello, el presente trabajo es un ejercicio comparativo de metodologías de imputación de ingresos laborales de la ENOE para que usuarios de esta lo consideren, pero lo deseable es que el INEGI es quien deba proporcionar la solución a esta problemática a través de sus áreas sustantivas, adoptando la imputación como parte de sus actividades de procesamiento y, al hacerlo, el Instituto incorporaría el conocimiento de sus propios datos a ese procedimiento, que es una de las características que enumera Rubin (1987).

Cabe destacar que, aunque la ENOE es una encuesta de panel rotatorio (con cinco rondas de levantamiento cada panel), para esta investigación se le da el tratamiento de encuesta transversal, por lo que solo son comparadas metodologías de imputación para eventos transversales, dejando la comparativa para datos longitudinales para estudios futuros.

3. Metodologías de imputación a comparar

En este ejercicio se aplicaron la simple y la múltiple cuidando que el(los) valor(es) a imputar fuera(n) estimado(s) a partir de un conjunto de variables que pudieran explicar los ingresos laborales de los individuos.

Aunque la ENOE incluye un conjunto importante de variables, solo algunas aplican a las personas ocupadas y de estas se buscaron aquellas que tuvieran hasta 10 categorías y que todos los códigos presentaran un número importante de ocurrencias debido a las restricciones que imponían los algoritmos. De esta forma, se buscó que todas las metodologías se compararan bajo las mismas condiciones y restricciones. Además, se descartaron todas las variables continuas debido a que dificultaban el pareamiento en *Hot Deck*.

De esta manera, las variables seleccionadas fueron:

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

- Estrato sociodemográfico (cuatro categorías).
- Sexo.
- Edad (seis categorías) [*eda7c*].
- Nivel de instrucción (cuatro categorías) [*niv_ins*].
- Posición en la ocupación (tres categorías) [*pos_ocu*].
- Ocupación (10 categorías) [*c_ocu11c*].
- Rama de actividad (cinco categorías) [*rama*].
- Tipo de unidad económica (tres categorías) [*tue1*].
- Duración de la jornada laboral (cinco categorías) [*dur_est*].

Dado que todas las variables explicativas son categóricas, se descartaron aquellas metodologías que, para su funcionamiento, requieren solo variables continuas y las que imputan a partir de un donante y usan una medida de distancia para encontrarlo, o bien, aquellas otras basadas en Análisis de Componentes Principales (ACP).

Cabe hacer mención que la mayoría de estas variables incluyen códigos para *No especificado*, *No sabe* o *No contestó*, por lo que debieron considerarse como datos perdidos y codificados como NA (*Not Available*, que es como se identifica un dato perdido en *R*). También, la variable de ingresos laborales (*ingocup*) con valor 0 fue codificada con NA.

Además, para este trabajo solo se tomaron en cuenta a los individuos de 15 y más años de edad que reportaron estar ocupados y con códigos en la variable *pos_ocu*: 1) trabajadores subordinados y remunerados, 2) empleadores y 3) trabajadores por cuenta propia, excluyendo a 4) trabajadores sin pago.

Por otro lado, en la actualidad, experimentar con metodologías de imputación no es tan complicado. Algunos paquetes estadísticos (como *Stata*, *SPSS* y *SAS*) ya traen incorporadas estas funcionalidades, la desventaja es que son piezas de *software* que representan costos monetarios altos y no incluyen gran variedad de opciones metodológicas; sin embargo, existe *R*, que es totalmente gratuito y sí tiene un número importante de algoritmos con variedad de metodologías de imputación.

Entonces, tomando como *software* estadístico base a *R*, se experimentó con un número importante de algoritmos/metodologías, además de incluir dos algoritmos propios, que implementan la metodología *Hot Deck* aleatorio y la de *bosques aleatorios* en dos etapas, y desarrollados también en *R*, para al final trabajar con los siguientes algoritmos:

- *Hot Deck* aleatorio.
- *Multivariate Imputation by Chained Equations (MICE).*
- *Amelia II.*
- *missForest.*
- *Hmisc.*
- *Mi.*
- *Rf2e.*

Otros fueron descartados, ya sea porque no convergieron, o bien, por el exceso de recursos que requieren para funcionar.

Todos los algoritmos, con excepción de *Hot Deck* aleatorio, imputan todas las variables especificadas en el modelo, pero en este ejercicio solo se reportan los resultados en el ingreso laboral de los individuos.

Las metodologías de imputación aplicadas se basan en el supuesto de que el mecanismo de datos faltantes de ingresos laborales es *MAR*,[3] pudiéndose comprobar mediante un modelo de regresión logística donde la variable dependiente es binomial con valor 0 cuando no hay dato faltante y 1 cuando sí lo hay, y como covariables las enlistadas con anterioridad.

A continuación, se describen cada una de las metodologías usadas por algoritmo.

Hot Deck aleatorio

Para su implementación se desarrolló un algoritmo en *R*, de imputación simple, el cual tuvo sus inicios en

3 Ya Rodríguez y López (2015) demostraron que la probabilidad de los ingresos faltantes de la ENOE no se da de forma completamente aleatoria (*MCAR*, ver p. 9 primera columna).

uno desarrollado con *MS Visual FoxPro 9.0* usado para hacer comparaciones de ingresos laborales entre la ENOE y el Módulo de Condiciones Socioeconómicas (MCS) 2015 como parte de los trabajos de investigación que derivaron en el Modelo Estadístico 2015 para la continuidad del MCS-Encuesta Nacional de Ingresos y Gastos de los Hogares.[4]

Al inicio, este algoritmo adiciona un número aleatorio entre 0 y 1 (de una distribución uniforme) a cada observación del conjunto de datos; después, crea dos subconjuntos, uno de donadores con datos completos y otro de receptores con las observaciones con datos faltantes.

A los receptores que declararon menos de 1 salario mínimo se les imputó el salario mínimo multiplicado por el número aleatorio ya asignado. A los que declararon 1 salario mínimo se les imputó este.

Para cada individuo contenido en el subconjunto de datos de receptores que aún no se le ha imputado ingreso se busca uno o varios donadores posibles en el subconjunto correspondiente. Para la búsqueda, se usan las variables antes enlistadas y deben coincidir los valores del receptor con los valores de los posibles donadores. Si se encuentra más de un donador, entonces es seleccionado aquel que tenga el número aleatorio mayor (que le fue asignado al principio) como donador; entonces, el ingreso de este donador es imputado a quien tiene el dato faltante.

Dado que es casi imposible que todos los receptores encuentren un donador incluyendo todas las variables de empate (más de 95% sí encuentra), fue necesario repetir el proceso seis veces más, pero eliminando ciertas variables cada vez.

MICE

Creada por Stef Van Buuren y Groothuis-Oudshoorn (2011), es un algoritmo que implementa la imputación múltiple a múltiples variables a través del método de ecuaciones encadenadas. También, incluye funciones que permiten analizar los datos imputados y apilar esos resultados de acuerdo con las Reglas de Rubin (1987).

Van Buuren y Groothuis-Oudshoorn (1999) explican que *MICE* asume que para cada variable con datos incompletos se especifica una distribución condicional para los datos faltantes, por ejemplo, para una variable binaria incompleta podría usarse una regresión logística, una politómica para datos categóricos y una lineal para datos numéricos.

Para entender esta metodología, consideremos a X como un vector de variables ($X_1, X_2, \ldots, X_j$). Con ecuaciones encadenadas se imputa una primera variable X_1 usando X_2 a X_j como explicativas; después, se imputa la variable X_2 usando la X_1 imputada y las X_3 a X_j también como explicativas; luego, se imputa la X_3 usando X_1 y X_2 imputadas y de X_4 a X_j como explicativas y así sucesivamente. Esto lo hace en k iteraciones hasta lograr un conjunto de datos imputado, repitiendo el proceso para las m imputaciones.

MICE implementa 23 métodos de imputación, de los cuales 11 se pueden observar en el cuadro 2. El método a usar dependerá del tipo de la variable con datos faltantes.

Para lo que aquí se presenta, se usaron dos metodologías no paramétricas: *pareamiento por medias predictivas* (*pmm*) y *bosques aleatorios* (*rf*). Se recomienda el uso de estas, por ejemplo, cuando la variable en cuestión no presenta una distribución normal, que es el caso de los ingresos laborales de la ENOE.

Pmm, en realidad, es un método *Hot Deck* con una forma distinta de seleccionar el donador. Para llevarlo a cabo, *MICE* primero ajusta una variante de un modelo de mínimos cuadrados ponderados; después, agrega ruido a los coeficientes de la regresión; luego, obtiene una predicción de los ingresos en el conjunto de datos imputado (tanto en los casos con ingreso faltante como en los del

4 Material publicado en el sitio del INEGI en la sección de *Investigación*: http://www.beta.inegi.org.mx/proyectos/investigacion/eash/2015/

Cuadro 2

MICE, métodos de imputación

Método	Descripción	Tipo de dato	*Default*
pmm	Pareamiento por medias predictivas (*predictive mean matching*).	numérico	Sí
norm	Regresión lineal bayesiana.	numérico	
norm.nob	Regresión lineal no bayesiana.	numérico	
mean	Imputación de la media no condicionada.	numérico	
2L.norm	Modelo lineal de dos niveles.	numérico	
logreg	Regresión logística.	categórica, dos niveles	Sí
polyreg	Modelo logístico multinomial.	categórica, > dos niveles	Sí
polr	Modelo logístico ordinal.	ordinal, > dos niveles	Sí
lda	Análisis discriminante lineal.	categórica	
sample	Muestra aleatoria a partir de los datos observados.	cualquiera	
rf	Bosques aleatorios.	cualquiera	

Fuente: Van Buuren y Groothuis-Oudshoorn (2011).

observado); enseguida, calcula una matriz de distancias entre el ingreso predicho de los casos con ingreso observado y el predicho de los del faltante; a continuación, elige el donador dentro de los d casos con ingreso observado que tengan la distancia más pequeña; el caso seleccionado dona el ingreso observado. Para mayor detalle sobre *pmm*, ver Van Buuren (2018), Little (1988) y Allison (2015).

El de *bosques aleatorios* está basado en la técnica de clasificación y regresión implementada en *R* como *randomForest*, la cual hace uso de esta. Esta ajusta un bosque aleatorio para cada variable a imputar y el valor asignado al dato faltante será aquel que resulte de un árbol seleccionado aleatoriamente.

Amelia II

Es una herramienta en *R* para imputación múltiple que puede ser aplicada a datos transversales, series de tiempo y series de tiempo transversales. El algoritmo de imputación múltiple en *Amelia* asume que los datos tienen una distribución normal multivariada que emplea el algoritmo *Expectation Maximization* (EM) basado en *bootstrap* (*EMB*), descrito por Honaker y King (2010), para imputar los valores. Además, cuenta con funciones de diagnóstico que permiten validar el modelo de imputación.

missForest

Este método para imputación, al igual que *MICE rf*, hace uso de la técnica de bosques aleatorios implementada en *randomForest*. Es de imputación simple y funciona muy bien con tipos de variables mezclados, con relaciones no lineales, interacciones complejas y con alta dimensionalidad (cuando hay más variables que observaciones).

Para cada variable, el algoritmo ajusta un bosque aleatorio sobre los datos observados y después predice los faltantes. El proceso se lleva a cabo de forma iterativa actualizando la matriz imputada y midiendo las diferencias entre el valor previo y el nuevo. El proceso iterativo se detiene cuando la diferencia comienza a crecer o cuando se cumple el número de iteraciones indicadas por el usuario. El dato a imputar a la observación con dato faltante será aquella categoría que más se repita en el bosque aleatorio para el caso de variables cualitativas, o bien, el promedio cuando se trate de variables cuantitativas.

Hmisc

Contiene un conjunto de herramientas para la reducción de datos, la imputación, el cálculo de potencia y tamaño de muestra, la creación avanzada de tablas, variables de recodificación, importación e inspección de datos y gráficos generales.

El algoritmo de imputación múltiple de *Hmisc* toma en cuenta la incertidumbre de las imputaciones mediante *bootstraping* para aproximarse a la predicción de los valores a partir de la distribución predictiva bayesiana completa. Está basado en modelos semiparamétricos que utilizan los métodos *regresión aditiva, bootstraping* y *pareamiento por medias predictivas*.

El algoritmo funciona de la siguiente forma:

1. Para cada variable con NA, inicializa esta con valores de una muestra aleatoria de los valores observados.
2. Realiza lo siguiente "burnin"+"n.impute" veces:
   a. Extrae una muestra con reemplazo del conjunto de datos completo y ajusta un modelo aditivo flexible para predecir los valores de todos los casos.
   b. Imputa cada valor faltante con el observado de la observación, cuyo valor predicho es más cercano al predicho del valor faltante (*pareamiento por medias predictivas*).

Mi

Es un conjunto de herramientas que permite manipular datos, imputar valores faltantes y analizar conjuntos de datos múltiples imputados, entre otras funciones.

Para que *Mi* pueda llevar a cabo la imputación múltiple, requiere de una matriz de información construida al configurar el conjunto de datos para la imputación, donde se guarda el tipo de la variable y se propone una función de regresión por aplicar en la imputación, las cuales se muestran en el cuadro 3. Esta matriz puede ser modificada de acuerdo con las necesidades del usuario.

Las funciones *mi.continuous*(), *mi.binary*(), *mi.count*() y *mi.polr*() ajustan modelos lineales generalizados bayesianos adicionando una distribución *t* de *Student a priori* a los coeficientes de regresión. De esta forma *mi.polr*() usa la función *bayespolr*() de la librería *arm*, mientras que *mi.continuous*(), *mi.binary*() y *mi.count*() utilizan *bayesglm*() de la misma librería, con *family=gaussian*, *family=binomial* y *family=quasipoisson*, respectivamente, para predecir los valores a imputar. La función *mi.fixed*() solo copia el valor de los observados y *mi.categorical* emplea *multinom*(), de la librería *nnet*, para imputar variables categóricas escalares.

El método de *pareamiento de medias predictivas* (*mi.pmm()*), que es el que se eligió en este algoritmo para desarrollar el presente ejercicio, usa *bayesgml()* para predecir los valores del conjunto de

Cuadro 3 Continúa

_Mi_ , tipos de variables y funciones de regresión correspondientes

Tipo de variable	Descripción	Función de regresión
binary	Variable que contiene dos valores únicos.	*mi.binary*
continuous	Variable numérica continua sin transformación.	*mi.continuous*
count	Variable especificada por el usuario.	*mi.count*
fixed	Variable que contiene un valor único.	*mi.fixed*
log-continuous	Variable continua *log*-escalada.	*mi.continuous*
nonnegative	Variable numérica no negativa con más de cinco valores únicos.	*mi.continuous*
ordered-categorical	Variables que tienen atributo de ordenación.	*mi.polr*

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Cuadro 3 Concluye

***Mi*, tipos de variables y funciones de regresión correspondientes**

Tipo de variable	Descripción	Función de regresión
unordered-categorical	Variable factor o carácter.	*mi.categorical*
positive-continuous	Variable positiva con más de cinco valores.	*mi.continuous*
proportion	Variable numérica cuyos valores están entre 0 y 1, sin incluirlos.	*mi.continuous*
predictive-mean-matching	No es un tipo, solo se usa para invocar la función.	*mi.pmm*

Fuente: Su *et al.* (2011).

datos completo e imputa cada valor faltante con el observado de la observación cuyo valor predicho es más cercano al predicho del valor faltante.

Rf2e

Este algoritmo de imputación múltiple fue creado específicamente para esta investigación y emplea la técnica de bosques aleatorios de *randomForest* en dos etapas. En la primera se usa el algoritmo *missForest* para una primera imputación de todas las variables involucradas, esto con el fin de tomar ventaja del trabajo de imputación de las variables categóricas (desechando la imputación del ingreso).

En la segunda etapa se utiliza como insumo un conjunto de datos creado a partir de las variables categóricas imputadas, en la primera etapa, y la variable de ingreso con datos faltantes. Después, para cada i-ésima (con $i = 1, 2, 3, \ldots, m$) imputación se lleva a cabo lo siguiente:

1. Se crean dos subconjuntos: uno con datos completos y otro con faltantes.
2. Se extrae una submuestra aleatoria con reemplazo con tamaño igual a 20% del total de observaciones con datos completos (dado que en la mayoría de los trimestres hay más de 100 mil observaciones, la submuestra resultante incluye más de 20 mil).
3. Se entrena el algoritmo *randomForest* con la submuestra que fue extraída en el paso anterior teniendo a los ingresos como variable a predecir y las otras como predictoras.
4. Usando los resultados del entrenamiento que se hizo en el paso anterior, se predicen los ingresos del subconjunto de datos con ingresos faltantes.
5. El ingreso predicho para cada observación con el faltante es imputado al conjunto de datos resultante de la primera etapa (en su respectiva observación).

4. Resultados

En este ejercicio se contrastaron siete algoritmos diferentes con seis metodologías de imputación, de las cuales dos son para imputación simple (*Hot Deck* aleatorio y missForest) y cuatro para la múltiple (*pmm* en *MICE*, *Hmisc* y *Mi*, *MICE rf* , *Amelia II* y *Rf2e*).

Los resultados que aquí se presentan fueron analizados en dos vertientes: en la primera se revisan algunas medidas de desempeño de las metodologías, las cuales pueden funcionar como criterio para elegir la que mejor se adapte a las características de la ENOE; en la segunda se muestran los efectos que puede tener la imputación de ingresos laborales en la ENOE en el ITLP del CONEVAL.

Medidas de desempeño

Para valorar el desempeño de cada metodología, se usaron tres medidas que involucran solo a los

Cuadro 4

Medidas de desempeño promedio de la serie por método de imputación

Método	CV	ES	R^2	RECM	EMA
Hot Deck	1.124	15.285	0.206	4 385	1 244
missForest	0.998	13.617	0.342	3 465	1 274
Amelia II	0.982	13.876	0.285	3 531	1 458
MICE pmm	1.120	15.051	0.258	3 197	1 174
MICE rf	1.108	15.045	0.264	3 359	1 237
Hmisc	1.123	15.224	0.265	3 332	1 224
Mi	1.126	15.490	0.276	4 401	1 306
Rf2e	0.997	13.585	0.338	3 347	1 265
Observados	1.074	15.602	0.205		

conjuntos de datos completos: el coeficiente de variación (CV), el error estándar (ES) y el coeficiente de determinación (R^2), así como dos que involucran tanto a los conjuntos de datos con valores faltantes como los completos: la raíz del error cuadrático medio (RECM) y el error medio absoluto (EMA), de las cuales puede observarse un resumen en el cuadro 4 expresado en promedios de toda la serie de la ENOE aquí estudiada. Cabe hacer mención que para obtener estas medidas se aplicaron las Reglas de Rubin (1987) para el caso de metodologías de imputación múltiple.

Si se comienza revisando el coeficiente de variación que se observa en la gráfica 2, lo primero que podemos notar es que los ingresos por trabajo de la ENOE (en los datos observados) han tenido una alta variabilidad con una clara tendencia a la baja a través del tiempo; esto contrasta con el incremento que ha reportado la no respuesta de ingresos (faltantes) observada en la gráfica 1b, pudiéndose decir que, conforme incrementa la no respuesta, la variabilidad de los ingresos se compacta (tan es así que la correlación entre ambos es de -0.81).

En la gráfica 2 también podemos notar que los métodos que presentan menor variabilidad entre los ingresos completos son *Amelia*, *Rf2e* y *missForest* (con un CV de 0.98, 1.00 y 1.00, respectivamente), incluso presentando menor variabilidad que los datos observados (con CV de 1.07). Con mayor variabilidad que estos últimos, y muy cercanos entre ellos, están *MICE rf, MICE pmm*, *Hmisc*, *Hot Deck* y *Mi* (con un CV promedio de 1.11, 1.12, 1.12, 1.12 y 1.13, en ese orden).

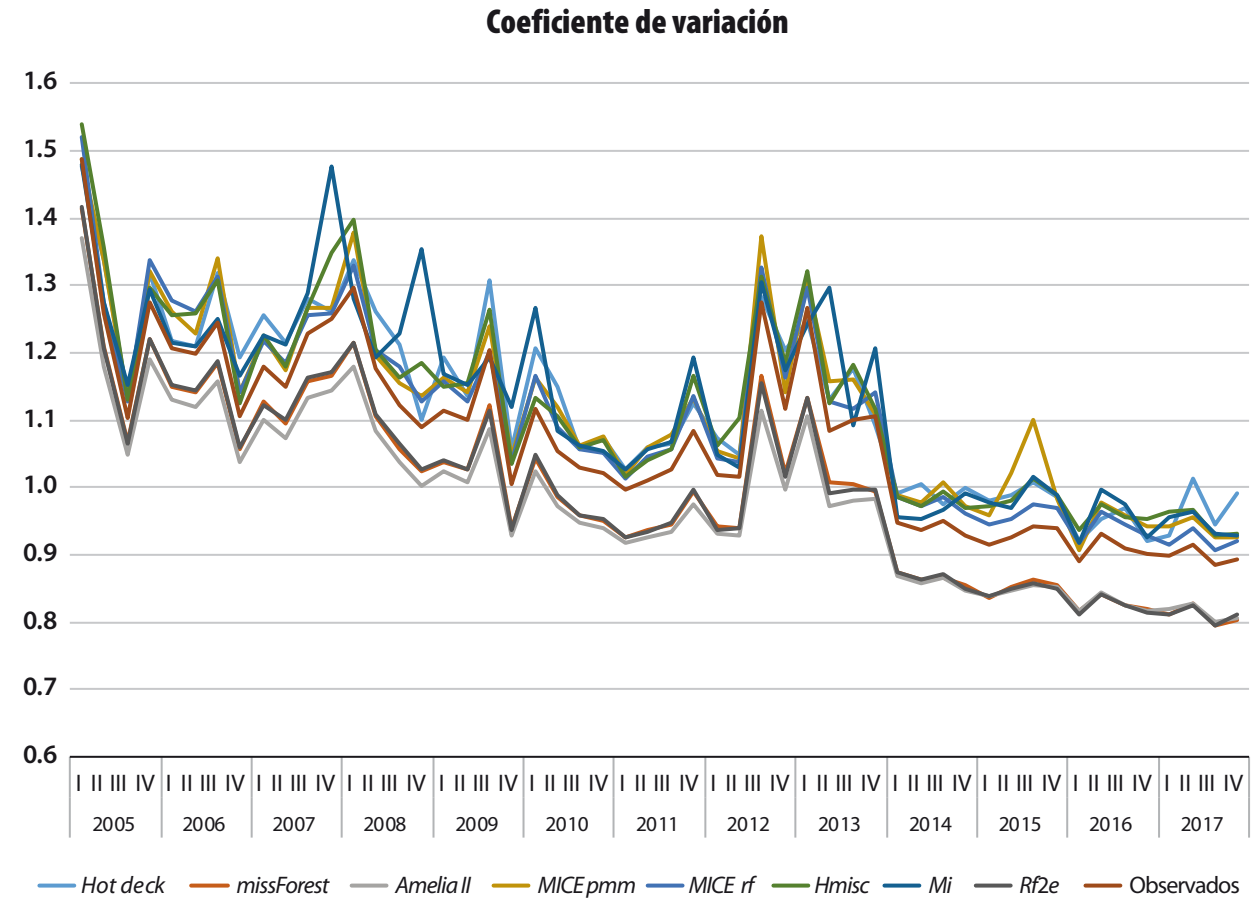
Dado lo anterior y si el criterio de selección del método fuera aquel que arroje la variabilidad más cercana a los datos observados, el elegido sería *MICE rf*, aunque los otros con mayor variabilidad también son una buena opción.

Ahora bien, si lo que se quiere es medir la precisión de la media muestral, entonces es necesario revisar los errores estándar, los cuales se presentan en la gráfica 3. Lo primero que podemos notar es que los ES de los ingresos por trabajo de la ENOE (en los datos observados) han tenido una tendencia estable a la baja a través del tiempo, aunque con una leve tendencia al alza en los últimos años.

También, al observar dicha gráfica, se puede notar que quien arroja mejor precisión es *Rf2e* (con un ES promedio de 13.59) seguido de *missForest* (13.62), *Amelia* (13.88) y, con menos precisión, *MICE rf* (15.05), *Hmisc* (15.22), *Hot Deck* (15.29) e, incluso con la precisión menor, *Mi* (15.49); por lo tanto, si el criterio de selección es el método que proporcione la mayor precisión de la media, se elegiría *Rf2e*.

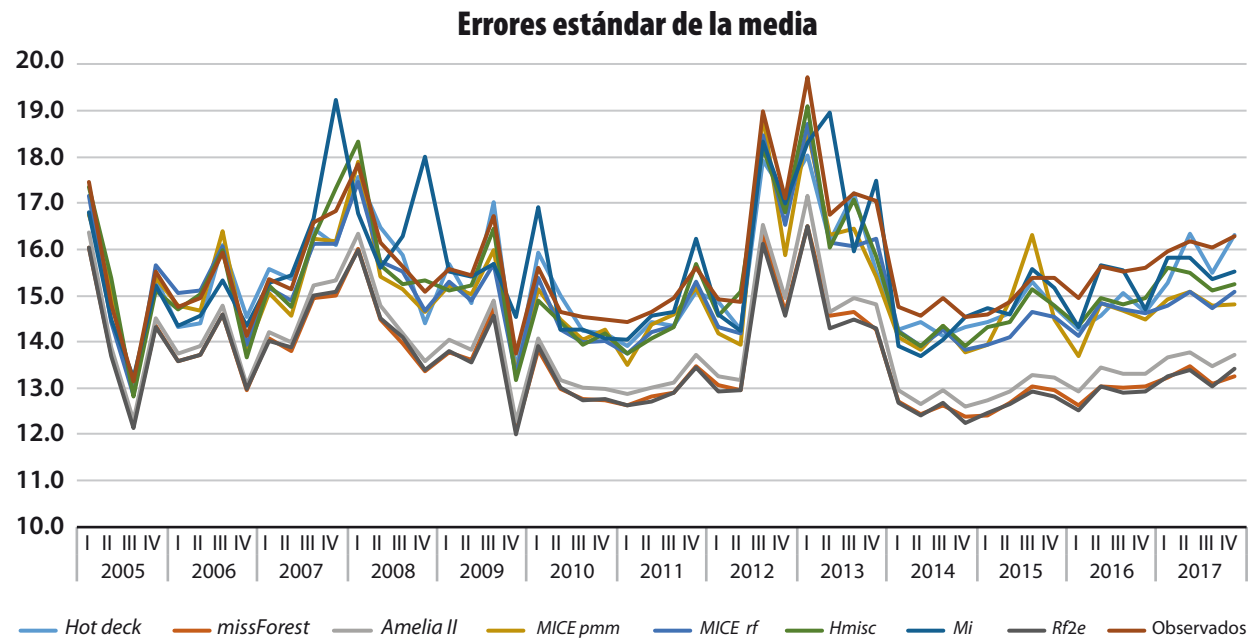
INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Gráfica 2



Fuente: cálculos propios a partir de la ENOE antes y después de la imputación.

Gráfica 3



Fuente: cálculos propios a partir de la ENOE antes y después de la imputación.

Por otro lado, sabiendo que el coeficiente de determinación se define como la proporción de variación de la variable Y (ingreso) que es explicada por las variables X (las enlistadas anteriormente) en un modelo de regresión, este puede ser usado como una medida de desempaño de la imputación.

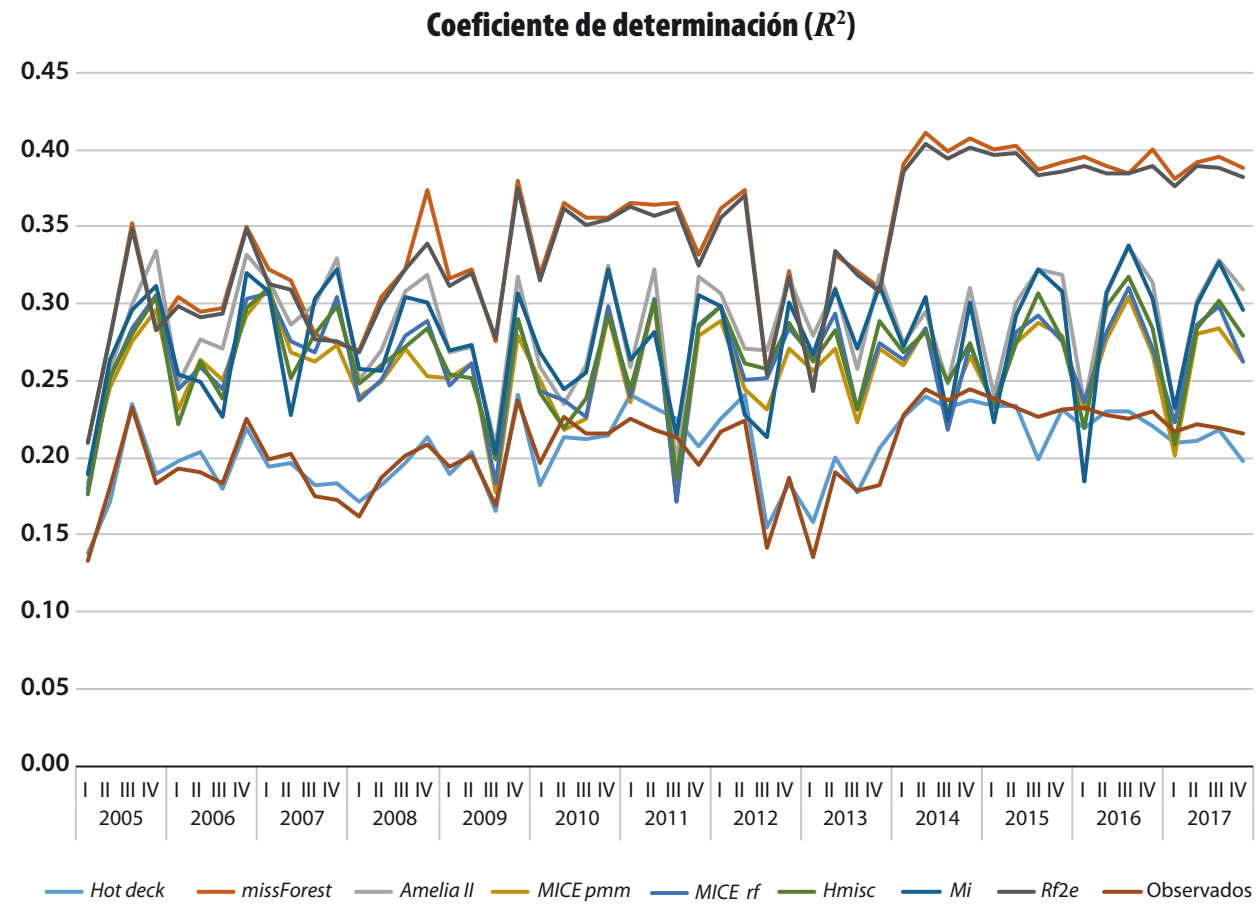
Como se puede observar en la gráfica 4, los R^2 son relativamente bajos, sugiriendo que las estimaciones del modelo no ajustan tan bien a la variable real; esto se debe a la gran variabilidad que se observa en los ingresos laborales mostrada en la gráfica 2, tan es así que, conforme la variabilidad disminuye a través del tiempo, los R^2 van incrementando; incluso, la correlación entre ambos es de -0.88.

La gráfica 4 también muestra que las metodologías que mejor ajustan son *Rf2e* y *missForest* con un promedio en toda la serie de 0.34 (presentando un comportamiento muy similar a lo largo del tiempo), con un repunte entre el 2014 y 2017 (0.39 de promedio); además, son las que siguen el patrón de los datos observados. Después, es seguido por un grupo compacto formado por *Amelia*, *Mi*, *Hmisc*, *MICE rf* y *MICE pmm* con promedios de R^2 de 0.29, 0.28, 0.26, 0.26 y 0.26, respectivamente. Algo separado queda *Hot Deck* con 0.21 que, incluso, es el más cercano al promedio de R^2 de los conjuntos de datos observados.

Sabiendo que esta medida no da muy buenos resultados y si, aun así, se usara como criterio de selección, *Rf2e* o *missForest* serían dos buenas opciones a elegir.

RECM y EMA son dos medidas que se basan en la distancia entre el valor observado y el imputado de una misma observación, por lo que para

Gráfica 4



Fuente: cálculos propios a partir de la ENOE antes y después de la imputación.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

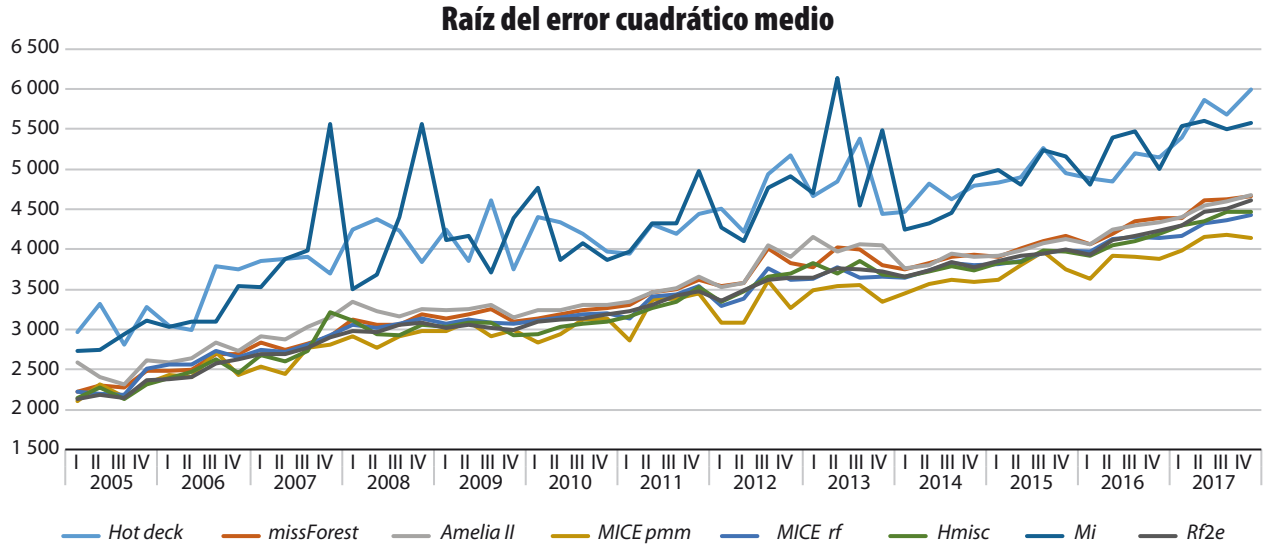
poder calcularlos debió ser necesario considerar como 0 los valores faltantes.

Al examinar la gráfica 5, que presenta los RECM, podemos observar un primer grupo muy compacto de métodos con los peores RECM que son *Mi* y *Hot Deck* con promedios en la serie de 4 401 y 4 385, respectivamente. En un segundo grupo, también muy compacto, se encuentran *Amelia*, *missForest*, *Rf2e*, *MICE rf*, *Hmisc* y *MICE pmm* con los *RECM* más bajos en promedio (3 531, 3 465, 3 347, 3 359, 3 332 y 3 197, en ese orden).

Si se decide seleccionar el método con el menor RECM, se elegiría *MICE pmm*, aunque cualquiera de los otros cercanos a este representaría una buena decisión.

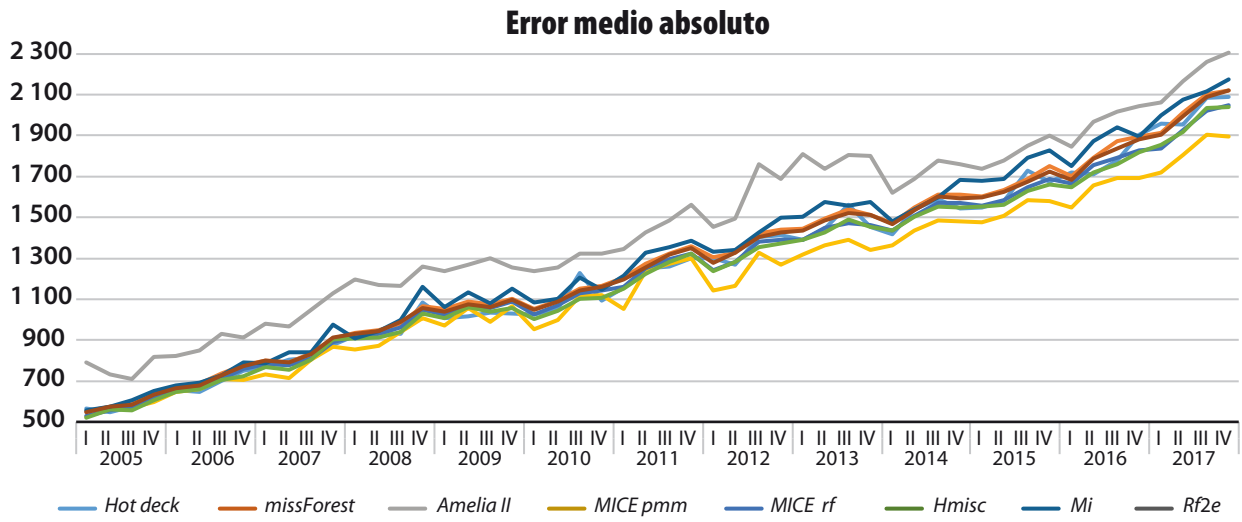
Por otro lado, al observar los EMA en la gráfica 6, podemos notar que estos presentan, al igual que el RECM, una tendencia creciente en el tiempo (normal si pensamos en los incrementos en el ingreso de los individuos) y que la mayoría de los métodos arrojan EMA muy cercanos los dos primeros años, comenzando a dispersarse a partir del tercero.

Gráfica 5



Fuente: cálculos propios a partir de la ENOE antes y después de la imputación.

Gráfica 6



Fuente: cálculos propios a partir de la ENOE antes y después de la imputación.

También, podemos percibir que el método con los EMA más bajos es *MICE pmm* (con un promedio en el tiempo de 1 174); después, separándose un poco, están *Hmisc*, *MICE rf*, *Hot Deck*, *Rf2e* y *missForest* (con 1 214, 1 237, 1 244, 1 265 y 1 274, respectivamente); un poco más arriba y con los peores EMA quedan *Mi* y *Amelia* (con 1 306 y 1 458, en ese orden).

Al igual que RECM, *MICE pmm* puede ser la mejor elección si el criterio de selección del método a usar de forma definitiva es aquel que tenga el EMA menor.

Dado que los expertos se inclinan más por metodologías de imputación múltiple (descartando las de la simple) y observando el cuadro 5, se puede decir que la metodología que mejores resultados presenta es *MICE pmm*, por su mayor frecuencia. Aunque cualquiera de las que ahí aparecen, excepto *missForest*, puede ser elegida debido a que los resultados de las métricas son muy cercanos entre ellos.

Efectos de la imputación en el ITLP

Anteriormente se mencionó que el CONEVAL elimina las observaciones con ingresos faltantes en la ENOE a la hora de hacer el cálculo del ITLP; además, imputa el punto medio de los ingresos que son reportados como rango de múltiplos de salarios mínimos. También, se dijo que hacerlo de esa forma agrega un fuerte sesgo a cualquier estimación que se lleve a cabo a partir de esos ingresos, estimaciones que, al ser sesgadas, no reflejan la verdadera realidad que se quiere medir.

De acuerdo con lo anterior, para poner en evidencia los efectos que pudiera registrar la imputación de ingresos en la ENOE, se analizaron el ITLP y dos indicadores derivados de este: el ingreso per cápita promedio del hogar y el porcentaje de población con ingreso laboral inferior al costo de la canasta alimentaria.

Al revisar primero el ingreso per cápita por hogar, que se aprecia en la gráfica 7, se puede observar que los promedios han tenido una tendencia a la baja a partir del 2007; esto puede deberse, en gran medida, a que los ocupados están subdeclarando sus ingresos. También, se nota que *Amelia* es la metodología que reporta los incrementos más altos, respecto a los reportados por el CONEVAL en este indicador, con un aumento promedio en el tiempo de 23.5%; después, un poco por debajo, le sigue *Mi* con 19.9% en promedio; asimismo, formando un grupo compacto, siguen *missForest*, *Rf2e*, *MICE rf*, *Hot Deck*, *Hmisc* y *MICE pmm* con incrementos promedio de 18.7, 18.5, 18.2, 17.9, 17.8 y 16.7%, respectivamente. Dado esto, se puede decir que los ingresos per cápita obtenidos por el CONEVAL están subestimados entre 16.7 y 23.5% en promedio, dependiendo de la metodología que se aborde.

El incremento en el ingreso per cápita debido a la imputación trae como consecuencia una disminución en el porcentaje de población con ingreso laboral inferior a la canasta alimentaria, como se percibe en la gráfica 8, donde se puede notar que *Amelia* es la que reporta la disminución más alta con 9.4% en promedio con respecto a lo publicado por el CONEVAL; después, formando un grupo compacto, están *missForest*, *Rf2e*, *MICE rf*, *Hmisc*, *MICE pmm* y *Mi* con disminuciones que van de 7.5 a 6.5% en promedio; *Hot Deck* aparece con la

Cuadro 5

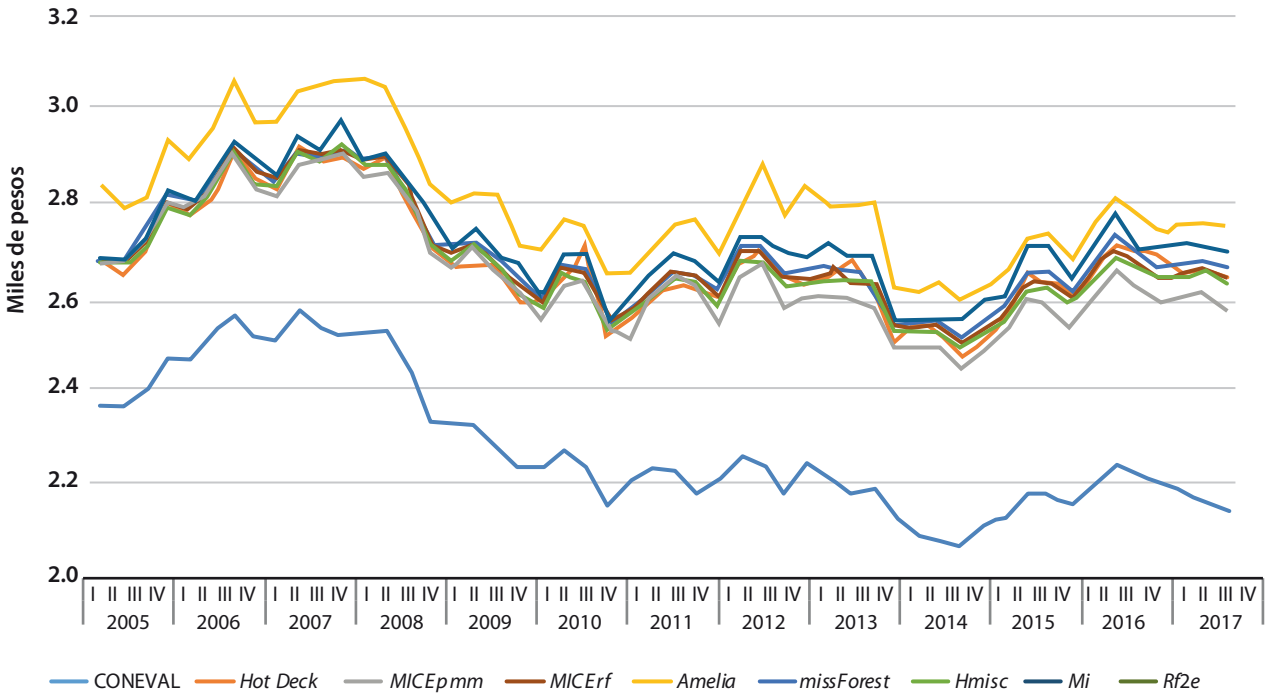
Mejores metodologías según el criterio y la métrica

Métrica/criterio	Mejor metodología	Segundo lugar
Variabilidad más cercana a los datos observados (CV más cercano)	*MICE rf*	*MICE pmm*
Mayor precisión (SE menor)	*Rf2e*	*missForest*
Mayor R^2	*missForest*	*Rf2e*
Menor RECM	*MICE pmm*	*Hmisc*
Menor EMA	*MICE pmm*	*Hmisc*

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Gráfica 7

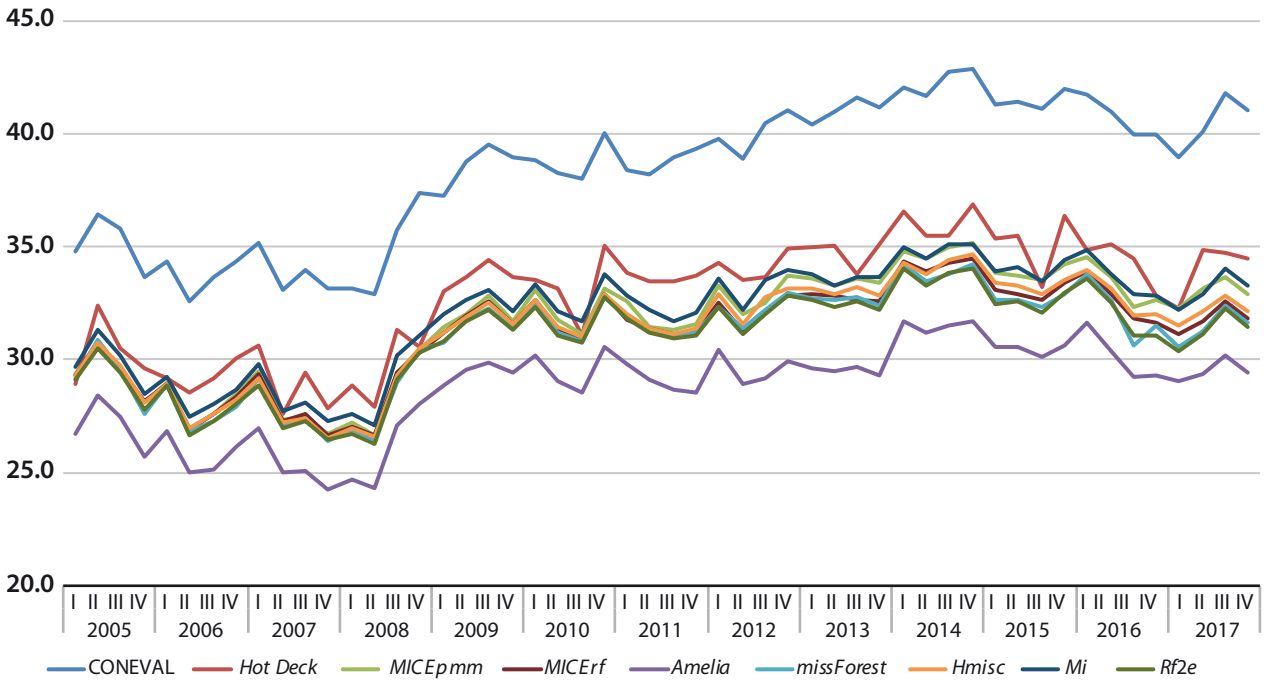
Ingreso per cápita promedio por hogar (pesos constantes, base *t*417)



Fuente: cálculos propios aplicando el algoritmo del CONEVAL a los datos de la ENOE antes y después de imputar.

Gráfica 8

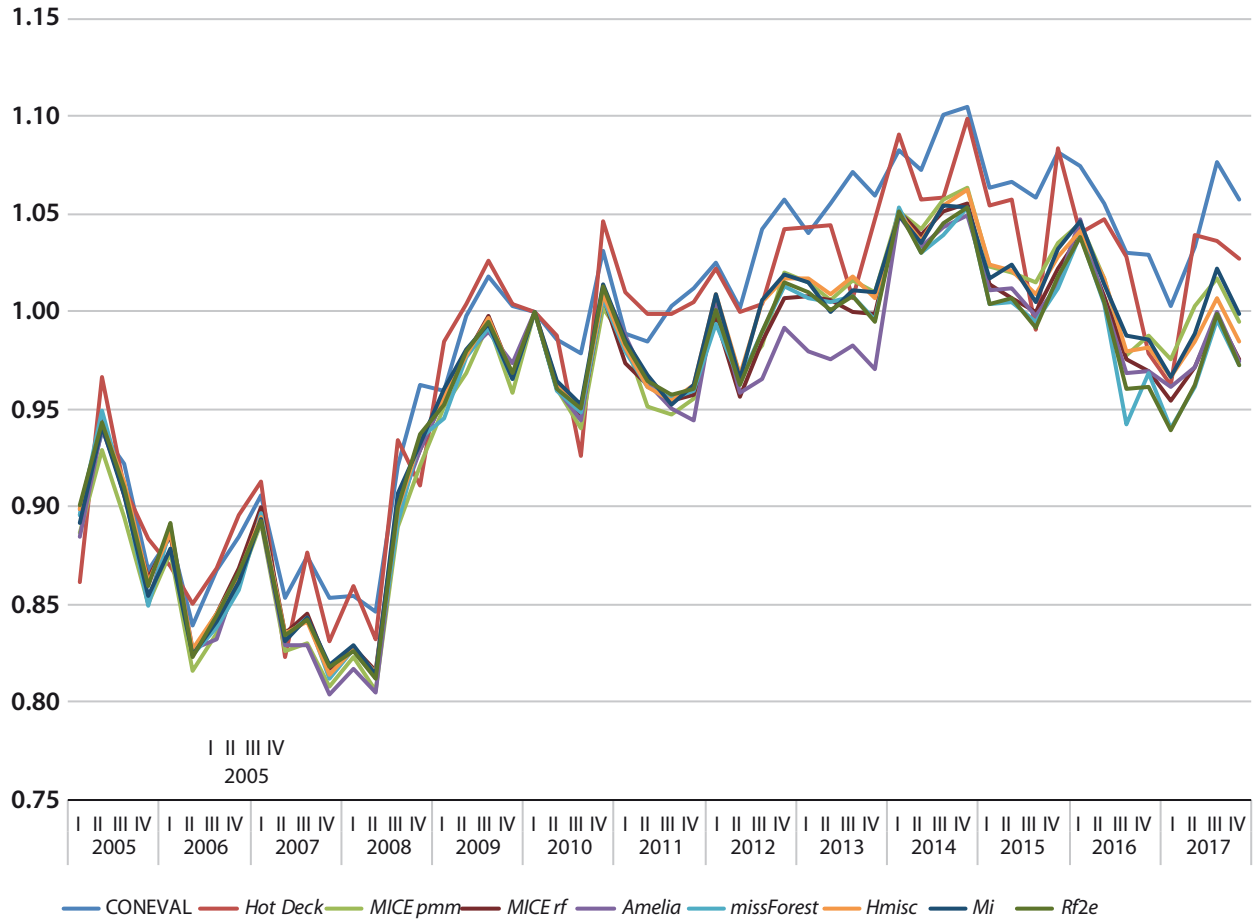
Porcentaje de población con ingreso laboral inferior al costo de la canasta alimentaria



Fuente: cálculos propios aplicando el algoritmo del CONEVAL a los datos de la ENOE antes y después de imputar.

Gráfica 9

Índice de Tendencia Laboral de la Pobreza (base *t*110)



Fuente: cálculos propios aplicando el algoritmo del CONEVAL a los datos de la ENOE antes y después de imputar.

menor disminución (5.6% en promedio) considerando el indicador del CONEVAL.

En la gráfica 9 podemos apreciar que la imputación de ingresos en la ENOE provoca una disminución en el ITLP. Incluso, observamos que el método que menos disminuye, con respecto a lo publicado por el CONEVAL, es *Hot Deck* (alrededor de 1% en promedio), aunque muestra una gran variabilidad en el tiempo. El resto disminuye entre 3 y 4% en promedio, siendo *Amelia* la que aparece con la mayor disminución.

La imputación de ingresos laborales en la ENOE, entonces, puede contribuir a explicar de mejor manera lo que se quiere medir a través del ITLP.

5. Conclusiones

A lo largo del tiempo, la ENOE ha presentado un incremento permanente en la no respuesta de los ingresos laborales, incluyendo los casos que se reportan en rangos de múltiplos de salarios mínimos, al iniciar en el primer trimestre del 2005 con 10.6% y alcanzando 27.5% en el cuarto del 2017.

La no declaración de ingresos y la declaración en intervalos de múltiplos de salarios mínimos en su conjunto en la ENOE han crecido 0.3% en promedio por trimestre, haciendo que estos dos fenómenos en el futuro sean insostenibles, y aunque la imputación de ingresos faltantes es una opción aceptable para solucionarlo, de seguir al alza estos

dos fenómenos, serán más las observaciones con datos imputados que con observados. Ante esto, es necesario que el INEGI tome las medidas necesarias en el diseño de instrumentos de captación y en el operativo de campo para revertir esa tendencia.

Dado que el INEGI publica los microdatos de la Encuesta con esos ingresos faltantes, se planteó hacer un ejercicio de comparación de metodologías de imputación tanto simples como múltiples para poner a consideración de los usuarios de la ENOE la adopción de este procedimiento como parte de la preparación del conjunto de datos para la generación de indicadores, aunque lo deseable es que sea el Instituto quien lo adopte como parte del procesamiento de las encuestas en hogares y ponga a disposición de los usuarios tanto los datos observados como los imputados y toda la información relacionada con el procedimiento de imputación.

Para medir el desempeño de cada metodología, se reportaron un conjunto de métricas (coeficiente de variación, error estándar, R^2, raíz del error cuadrático medio y el error medio absoluto) que permiten dar una idea de cuál pueda ser la más adecuada por adoptar. En este sentido, ya que los expertos en el tema se inclinan más por las de imputación múltiple, elegir cualquiera de entre *MICE pmm*, *Rf2e*, *Hmisc* y *MICE fr* puede ser una muy buena opción debido a que los resultados de las métricas son muy cercanos entre ellos, aunque *MICE pmm* es la que arroja los mejores.

Para cada metodología, también se analizó el efecto que provoca la imputación de ingresos en la ENOE al ITLP y otros indicadores que calcula y difunde el CONEVAL, encontrando lo siguiente:

1. Que los ingresos per cápita obtenidos por el CONEVAL están subestimados entre 16.7 y 23.5% en promedio, dependiendo de la metodología que se revise.
2. Que el incremento en el ingreso per cápita debido a la imputación trae como consecuencia una disminución en el porcentaje de población con ingreso laboral inferior a la canasta alimentaria. Este decremento representa entre 5.6 y 9.4% en promedio, dependiendo de la metodología utilizada.
3. Que el aumento en el ingreso per cápita también provoca una disminución entre 1 y 4% en promedio del ITLP.

Algo que es necesario recalcar es que el ingreso per cápita por hogar ha tenido una tendencia a la baja a partir del 2007 y que esto puede deberse, en gran medida, a una creciente subdeclaración de ingresos por parte de los ocupados. Además, el proceso de imputación, por sí mismo, no corrige esa subdeclaración, ya que se imputan valores restringidos a los que se observaron en cada trimestre.

La imputación de ingresos laborales en la ENOE, por lo tanto, tiende a disminuir el sesgo y permite explicar de mejor manera lo se quiere medir con el ITLP.

Fuentes

Allison, P. *Imputation by Predictive Mean Matching: Promise & Peril*. 2015, marzo 5 (DE), recuperado en junio 12 del 2018 de https://statisticalhorizons.com/predictive-mean-matching

_____ *Why You Probably Need More Imputations Than You Think*. 2012, noviembre 9 (DE), recuperado en junio 12 del 2018 de https://statisticalhorizons.com/more-imputations

Campos-Vazquez, Raymundo. *Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en México*. Serie documentos de trabajo del Centro de Estudios Económicos, El Colegio de México, Centro de Estudios Económicos, 2013 (DE), recuperado el 11 de junio de 2018 de https://EconPapers.repec.org/RePEc:emx:ceedoc:2013-04

Durán Romo, B. "Ajuste demográfico por imputación", en: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. Número especial. México, INEGI, 2018, pp. 28-57 (DE), recuperado el 14 de septiembre de 2018 de http://www.inegi.org.mx/rde/2018/08/27/ajuste-demografico-imputacion/

Eurostat. "Imputation-Little and Su Method", en: *Memobust Handbook on Methodology of Modern Business Statistics*. Eurostat, 2014 (DE), recuperado de https://ec.europa.eu/eurostat/cros/content/little-and-su-method-method_en

James Honaker, J. & G. King. "What to do About Missing Values in Time Series Cross-Section Data", en: *American Journal of Political Science*. 54, 3, 2010, pp. 561-581 (DE), recuperado de https://gking.harvard.edu/files/abs/pr-Abs.shtml

Little, R. J. A. "Missing-Data Adjustments in Large Surveys", en: *Journal of Business & Economic Statistics*. 6(3), 1988, pp. 287-296 (DE), recuperado de www.jstor.org/stable/1391878

Paulin, G., J. Fisher, & S. Reyes-Morales. *User's Guide to Income Imputation in the CE [Ebook]*. Washington D.C.: US Department Of Labor. Bureau of Labor Statistics, 2006 (DE), recuperado de https://www.bls.gov/cex/csxguide.pdf

Peugh, J. & C. Enders. "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement", en: *Review of Educational Research*, 74(4), 2004, pp. 525-556 (DE), recuperado de http://www.jstor.org/stable/3515980

Reiter, J. & T. Raghunathan. "The Multiple Adaptations of Multiple Imputation", en: *Journal of the American Statistical Association*. 102(480), 2007, pp.1462-1471 (DE), recuperado de http://www.jstor.org/stable/27639995

Rodríguez-Oreggia, Eduardo & Bruno López-Videla. "Imputación de ingresos laborales. Una aplicación con encuestas de empleo en México", en: *El Trimestre Económico*. 82(325), 2015, pp. 117-146 (DE), recuperado el 8 de junio de 2018 de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2448-718X2015000100117&lng=es&tlng=es

Rubin, Donald B. *Multiple imputation for nonresponse in surveys*. New York, Wiley, 1987.

Rubin, Donald B., & Nathaniel Schenker. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse", en: *Journal of the American Statistical Association*. 81:394, 1986, pp. 366-374, DOI:10.1080/01621459.1986.10478280.

Starick, R. *Imputation in Longitudinal Surveys*: *The Case of HILDA*. Research Paper of the Australian Bureau of Statistics. 2005 (DE), recuperado de http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.075

Su, Y., A. Gelman, J. Hill, & M. Yajima. "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box", en: *Journal of Statistical Software*. 45(2), 2011, pp. 1-31 (DE), recuperado de doi:http://dx.doi.org/10.18637/jss.v045.i02

University of Essex. Institute for Social and Economic Research. British Household Panel Survey: Waves 1-18, 1991-2009. [data collection]. 8th Edition. UK Data Service, 2018, SN: 5151 (DE), recuperado de http://doi.org/10.5255/UKDA-SN-5151-2

Van Buuren, S. *Mice*. 2018 (DE), recuperado el 19 de junio de 2018 de https://www.rdocumentation.org/packages/mice/versions/3.0.0/topics/mice.impute.pmm

Van Buuren S. & Groothuis-Oudshoorn K. "Mice: Multivariate Imputation by Chained Equations in R.", en: *Journal of Statistical Software*. 45(3), 2011, pp. 1-67 (DE), recuperado de http://www.jstatsoft.org/v45/i03/

_____ *Flexible multivariate imputation by MICE* (Tech. rep. TNO/VGZ/PG 99.054). Leiden: TNO Preventie en Gezondheid, 1999 (DE), recuperado de http://publications.tno.nl/publication/34618574/FW469e/buuren-1999-flexible.pdf

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Análisis jerárquico de las emisiones de gases efecto invernadero en México

Hierarchical Analysis of Greenhouse Gas Emissions in Mexico

Carlos Samuel Pérez Pérez* y Luis Enrique Nieto Barajas**

En este trabajo se lleva a cabo un análisis mediante un modelado jerárquico de las emisiones de gases de efecto invernadero (GEI) de los principales sectores industriales de México desde 1999 hasta el 2012, el cual se utiliza para obtener estimadores de eficiencia económico-ambiental para cada sector. El objetivo es cuantificar la relación entre indicadores económicos clave y las series GEI al analizar su comportamiento dinámico. La información es obtenida de los Censos Económicos del Instituto Nacional de Estadística y Geografía y de los inventarios nacionales de emisiones del Instituto Nacional de Ecología y Cambio Climático.

Palabras clave: eficiencia ambiental; emisiones GEI; inferencia bayesiana; modelos jerárquicos.

This work presents a hierarchical analysis of Greenhouse Gas Emissions (GHG) of the main industrial sectors in Mexico from 1999 to 2012. Hierarchical dynamic models are used to obtain estimates of economic-environmental efficiency indicators for each sector. The objective is to quantify the relationship between key economic indicators and the GHG series by analyzing their dynamic behavior. Information was obtained from the economic censuses of the National Institute of Statistics and Geography (INEGI) and the national emissions inventories of the National Institute of Ecology and Climate Change (INECC).

Key words: Environmental efficiency; GHG emissions; Bayesian inference; hierarchical models.

Recibido: 18 de septiembre de 2018.
Aceptado: 26 de febrero de 2019.

* Instituto Tecnológico Autónomo de México (ITAM), carlos.perez@itam.mx
** ITAM, lnieto@itam.mx

Gráfico de Vector de cambio climático el calentamiento Global de México/bubaone/Getty Images

1. Introducción

De acuerdo con el Panel Intergubernamental del Cambio Climático[1] (IPCC, por sus siglas en inglés, 2013), la temperatura global de la Tierra ha aumentado de manera alarmante durante los últimos 200 años, en gran parte debido a un volumen desproporcionado de emisiones de gases de efecto invernadero (GEI), por lo que es vital para la humanidad frenar el cambio climático en los próximos 50 años. La búsqueda de mayor bienestar para la población ha inducido aumentos en la actividad económica; no obstante, más producción industrial también conduce a niveles de contaminación más altos y, en particular, a un nivel superior de emisiones de GEI. Los sectores industriales han tenido un impacto severo sobre la calidad del suelo, aire y agua, por lo que es relevante realizar un análisis que tenga como principal objetivo cuantificar la relación entre dicho deterioro ambiental y la actividad económica.

En este sentido, es de interés definir una cantidad que represente la noción de eficiencia ambiental y determinar cuáles son los sectores industriales que muestran mayores o menores valores de esta; el presente estudio establece un fundamento estadístico —desde un enfoque bayesiano— que permite mejorar la toma de decisiones en materia regulatoria al identificar aquellos que más contri-

1 Se creó en 1988 gracias al impulso de la Organización Meteorológica Mundial (WMO) y del Programa de Naciones Unidas para el Medio Ambiente (UNEP). Su función principal consiste en analizar la información científica relevante para entender el cambio climático.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

buyen a la emisión de GEI en relación con su nivel de actividad económica y, así, contribuir a la creación de políticas ambientales focalizadas en los menos eficientes.

En la literatura se han encontrado trabajos que analizan, sobre todo, el impacto económico que conlleva el cambio climático y, en específico, los efectos de una mayor concentración de emisiones de GEI, entre ellos destacan Nordhaus (1991) y Stern (2007), para el caso de Estados Unidos de América, y Galindo (2009) y SEGOB (2014), para el de México. También, hay estudios que utilizan matrices de insumo-producto para estimar los efectos de políticas de mitigación en relación con la actividad en el sector de la construcción para países como Australia, China, Irlanda y Noruega (Yu, 2017). Las investigaciones que más se asemejan a la que aquí se presenta analizan la relación causal que existe entre el crecimiento económico, el consumo de energía y las emisiones de GEI, en las cuales se da evidencia consistente con la hipotética curva ambiental de Kuznets y muestran, también, resultados de causalidad de Granger del crecimiento económico sobre la emisión de GEI para Canadá (Hamit-Haggar, 2012) y algunos países de la Unión Europea (Kasman & Duman, 2015).

Este trabajo ofrece contribuciones importantes para el ámbito de la Estadística Aplicada en México: primero, la realización de una correspondencia entre fuentes oficiales de información sin precedentes en México y, posiblemente, en toda América del Norte, en la cual se vinculan datos económicos y ambientales con el fin de cuantificar la eficiencia ambiental; segundo, una aplicación poco utilizada de los algoritmos de inferencia aproximada, que hacen uso intensivo de *software* estadístico para obtener estimaciones de los parámetros de interés; y, por último, se constituye como un novedoso referente en la literatura ambiental actual, ya que plantea el uso de modelado jerárquico bayesiano para cuantificar la relación entre los indicadores económicos clave y las emisiones de GEI en el pasado reciente. Más allá de la contribución técnica, se presenta como un punto de partida para la toma de decisiones sistemática en materia ambiental, las cuales son pertinentes para la conservación del mundo tal como lo conocemos.

2. Fuentes de información

2.1 Datos ambientales

Forman parte del Inventario Nacional de Emisiones de Gases Efecto Invernadero (INEGEI). De acuerdo con la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT, 2012), las mediciones a nivel subsector del IPCC se presentan cada año y son responsabilidad del Instituto Nacional de Ecología y Cambio Climático (INECC) de México.

La composición del inventario nacional consta de las emisiones por compuesto químico para los principales rubros ambientales determinados por el IPCC, los cuales se dividen en categoría, subcategoría, sector y subsector. El conjunto inicial de información consta de los resultados de emisiones para el periodo 1990-2012 a nivel nacional; esto quiere decir que se cuenta con datos anuales de cada compuesto químico de las emisiones medidas en gigatoneladas (Gt) para cada uno de los rubros, según la clasificación del IPCC a nivel nacional. El catálogo completo se puede consultar en SEMARNAT (2013).

Las emisiones clasificadas dentro de estas categorías se reportan en los inventarios con base en los seis gases de efecto invernadero incluidos en el Protocolo de Kyoto (UNFCCC, 1998):

- Bióxido de carbono (CO_2).
- Metano (CH_4).
- Óxido nitroso (N_2O).
- Hidrofluorocarbonos (HFC).
- Perfluorocarbonos (PFC).
- Hexafluoruro de azufre (SF_6).

Para cada uno de estos compuestos, se obtuvieron equivalencias de potencial de calentamiento global a partir de los documentos generados por el IPCC. A través de ellas, se homologaron las unidades para que todas las emisiones fueran medidas

en términos de gigatoneladas de dióxido de carbono equivalente (Gt CO_2eq), de modo que las emisiones entre sectores fueran comparables. Estos potenciales se encuentran en el cuarto reporte de la Convención Marco de las Naciones Unidas sobre el Cambio Climático (IPCC, 2007).

2.2 Datos económicos

Los referentes para cada sector o actividad industrial son un subconjunto de indicadores de los Censos Económicos (CE) de 1999, 2004, 2009 y 2014, que son producidos por el Instituto Nacional de Estadística y Geografía (INEGI). Los sectores que se toman en los CE son los que se encuentran en el Sistema de Clasificación Industrial de América del Norte (SCIAN), el cual tiene génesis entre 1994 y 1997 —a raíz del Tratado de Libre Comercio— con el fin de asegurar la comparabilidad y unificación de la información económica entre los tres países (SEGOB, 2009).

A pesar de que la clasificación del SCIAN incluye el sector agropecuario, la información económica de este sector no está presente en los CE, sino que se encuentra en los correspondientes censos y encuestas agropecuarios. Los datos no están sistematizados de la misma forma y, desafortunadamente, no se tiene, incluso, información pública digitalizada, lo cual representa un área de oportunidad para la gestión de información agrícola y pecuaria en el país. Existen sistemas externos al INEGI, por ejemplo, el de Información Agroalimentaria de Consulta (SIAP, 2018), que contiene lo referente a estos sectores primarios; no obstante, no cuenta con los mismos indicadores económicos, por lo cual no se toman en cuenta en el presente análisis. Por otro lado, el sector agropecuario no es una industria de jurisdicción federal, en términos de materia ambiental, respecto a las emisiones de GEI (SEGOB, 2012), por lo cual el nivel de política regulatoria no sería el mismo que para los otros sectores; sin embargo, eso no debe minusvalorar que la actividad agropecuaria es un componente importante en la contaminación atmosférica y deberá ser incluida en un trabajo posterior para obtener un panorama completo de los generadores de GEI.

Los indicadores obtenidos para cada rama de agregación sectorial son:

- Formación bruta de capital fijo (miles de pesos).
- Personal ocupado (número de personas).
- Remuneraciones (miles de pesos).
- Unidades económicas (número de empresas).
- Valor agregado censal bruto (miles de pesos).

Básicamente, el conjunto inicial de datos económicos comprende la información de los indicadores mencionados para las 303 ramas de la economía mexicana, según año censal, el cual fue filtrado para obtener las ramas necesarias (INEGI, 2013).

Debido a las diferencias en las frecuencias de medición entre los datos ambientales y los económicos, se obtuvieron datos para los años intermedios mediante una interpolación temporal para poseer información anual para cada sector. Se realizó una interpolación cúbica por pedazos para cada uno de los indicadores económicos de modo que se obtuvieran valores en los periodos intercensales. Si bien esta forma de estimar observaciones intermedias no es la más apropiada en términos económicos, este es un primer acercamiento para subsanar la falta de información y poder cumplir con el objetivo de análisis.

2.3 Correspondencia entre sistemas

A pesar de que el SCIAN e IPCC tienen más de dos décadas estableciendo la pauta de clasificación en las disciplinas económicas y ambientales, no existen trabajos de investigación que hayan intentado establecer una relación formal entre ambos, más allá de algunos documentos descriptivos de organismos públicos responsables de la eficiencia económico-ambiental (Environnement et Changement Climatique Canada, 2017). En México no hay, como tal, un trabajo que realice un mapeo entre los dos sistemas, por lo que una de las aportaciones del presente estudio es proponer dicho vínculo a partir de analizarlos de forma meticulosa, lo cual se logró a través de las definiciones establecidas en cada rubro. Es importante mencionar que

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

se tomaron en cuenta las ramas SCIAN y subsectores IPCC denominados *fuentes puntuales fijas de jurisdicción federal* (SEGOB, 1988).

El mapeo que se realizó entre estos dos conjuntos —que, a grandes rasgos, origina un mapeo entre subconjuntos de los sistemas de clasificación— no fue de forma directa ni uno a uno, pues cada sistema tiene niveles de desagregación distintos. A menudo, los rubros del IPCC son más generales por lo que, en su mayoría, cada subsector comprende una o varias ramas del SCIAN que deben ser agregadas. No obstante, existen casos en los cuales varios subsectores del IPCC corresponden a varias ramas del SCIAN por lo cual, en términos de datos, se consideraron de forma agregada en ambos sistemas para poder establecer una relación lo más completa posible. En este sentido, se hace notar que se agregaron en un solo rubro las industrias de las bebidas y alimentos con la de tabaco, ya que presentan un comportamiento temporal similar. Por facilidad, se utilizó el nombre y código que proviene del SCIAN a nivel subsector, sin que esto signifique que se están incluyendo todas las ramas del subsector.

Una vez realizada la agregación correspondiente, se tiene que el conjunto de datos resultante consta de datos anuales en los años en los que coinciden ambas fuentes de información, es decir, de 1999 al 2012, para un total de 14 subsectores industriales descritos en el cuadro 1.

Cuadro1

Relación de clave con descripción para cada uno de los sectores

CVE	Descripción
211	Extracción de petróleo y gas
212	Minería de minerales metálicos y no metálicos excepto petróleo y gas
221	Generación, transmisión y distribución de energía eléctrica
222	Suministro de agua y gas por ductos al consumidor final
312	Industria de las bebidas, alimentaria y del tabaco
322	Industria del papel
324	Fabricación de productos derivados del petróleo y del carbón
325	Industria química
326	Industria del plástico y del hule
327	Fabricación de productos a base de minerales no metálicos
331	Industrias metálicas básicas
335	Fabricación de aparatos eléctricos y generadores de energía eléctrica
336	Fabricación de equipo de transporte
562	Manejo de desechos y servicios de remediación

3. Metodología de análisis

3.1 Inferencia bayesiana

La Estadística permite estudiar los fenómenos a través de propiedades que presentan las observaciones de estos. Los datos se utilizan para generar un juicio de relativa certidumbre acerca de algunas características claves inherentes al fenómeno, las cuales están presentes como parámetros desconocidos en el modelo que representa el fenómeno y, por lo tanto, es necesario realizar inferencia estadística sobre los mismos. En el proceso inferencial bayesiano, la incertidumbre sobre los parámetros desconocidos se cuantifica mediante distribuciones de probabilidad, las cuales pueden incluir conocimiento o desconocimiento previo. Esto tiene como resultado que tanto las variables aleatorias observables (que forman la base de nuestro modelo) como los parámetros que se buscan conocer son descritos mediante distribuciones de probabilidad (Nieto-Barajas, L. E. & De Alba, E., 2014).

El mecanismo de actualización de información se conoce como Teorema de Bayes y se presenta en términos matemáticos como sigue: sea $Y' = (Y_1, \ldots, Y_n)$ una muestra aleatoria del modelo $f(\boldsymbol{y} \mid \theta)$; entonces, la distribución final de θ se obtiene como $f(\theta|\boldsymbol{y}) = f(\boldsymbol{y}|\theta)\, f(\theta)/ f(\boldsymbol{y})$. Si bien esta expresión es simple, su obtención analítica se puede complicar debido al cálculo de la constante de normalización $f(\boldsymbol{y})$. No obstante, gracias a los recientes avances computacionales, así como al desarrollo de nuevos métodos y algoritmos de aproximación, es posible obtener características

de cualquier distribución final sin necesidad de calcular dicha constante (Chen *et al*., 2012). Este es el caso de los algoritmos de simulación Monte Carlo vía cadenas de Markov (MCMC) que se encuentran implementados en la mayoría de las herramientas computacionales, como *OpenBUGS* y *JAGS*, los cuales permiten obtener simulaciones de la distribución final mediante la implementación de una cadena de Markov, cuya distribución de equilibrio corresponde a la final de interés. Por conveniencia práctica, es mucho más sencillo acceder a dichos algoritmos de simulación MCMC desde el *software* estadístico *R* (R Core Team, 2018).

3.2 Modelado jerárquico

La Estadística se ha enfocado en construir modelos de probabilidad para los datos de forma directa, lo que permite definir la verosimilitud a través de la cual se puede realizar inferencia sobre los parámetros desconocidos. Sin embargo, la verosimilitud no captura de forma directa que los datos son ruidosos, incompletos o que tienen alguna característica no observable que es de importancia (Banerjee *et al*., 2014).

Lo anterior se ha resuelto construyendo lo que se conoce como modelo jerárquico, el cual representa una forma de expresar incertidumbre a través de niveles de probabilidades condicionales: de estos se usan modelos para los datos, dado el fenómeno que los genera y, a su vez, otro de probabilidad para el fenómeno en sí, es decir, el modelo de interés está constituido a partir de otros submodelos, cuya incertidumbre también está expresada en subniveles. Este enfoque es, en cierto sentido, una especie de descomposición del análisis de varianza, que es más general que la descomposición aditiva habitual (Scheffe, 1959).

Estas distribuciones condicionales dependen, como es usual, de parámetros desconocidos y, en algunos casos —si se plantea un nivel inferior— por debajo del modelo de datos y el del fenómeno, en el cual se especifica una distribución conjunta de todos los parámetros desconocidos; entonces, el modelo jerárquico se considera como bayesiano. Según Berliner (1996), el modelo de regresión jerárquico bayesiano se puede representar de forma escalonada como: datos $f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$, fenómeno $f(\boldsymbol{\theta}|\boldsymbol{\varphi})$ y parámetro $f(\boldsymbol{\varphi})$.

El supuesto importante a partir del cual se deriva el poder del enfoque jerárquico es la intercambiabilidad o simetría de los parámetros del mismo tipo dentro del modelo (Cappé *et al*., 2010). Esto simplifica la estimación simultánea de varios parámetros con la ventaja de combinar los datos para mejorar la precisión de las estimaciones y, al mismo tiempo, permitir incorporar incertidumbre. En términos prácticos, se vuelve computacionalmente intensiva la realización de inferencias y, debido a la complejidad de los modelos, se requiere utilizar métodos MCMC.

3.3 Especificación del modelo para las emisiones de GEI

A continuación, se define la notación de las variables que será utilizada para presentar el desarrollo del análisis. Se denotan por X_{jkt} las que representan el indicador económico (en escala logarítmica) para $j = 1, \ldots, 5$ y por Y_{kt}, el ambiental, correspondiente al rubro industrial k, en el año t, donde $k = 1, \ldots, 14$ y $t = 1999, \ldots, 2012$. La transformación logarítmica sobre los indicadores económicos fue requerida, ya que las unidades rondaban desde miles hasta centenares de millón. Los cinco que se midieron en cada rubro industrial son los descritos en la sección 2.2. El indicador ambiental está medido en Gt CO_2eq, también en escala logarítmica, y es la variable cuyo comportamiento se desea modelar.

El modelo que se propone para el indicador ambiental Y_{kt} es uno jerárquico normal, cuya especificación en el primer nivel es:

$$Y_{kt} | \mu_{kt} \sim N(\mu_{kt}, \tau_k), \qquad (1)$$

donde τ_k es la precisión (recíproco de la varianza) que se considera constante para todos los tiempos t del mismo sector k, y μ_{kt} es la media del indicador

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

ambiental en el sector k al tiempo t, la cual depende de los indicadores económicos de cinco formas, mismas que se compararon para elegir aquella que mejor describe el comportamiento de los datos. Tales especificaciones son:

(i) Indicadores económicos con efectos comunes para todos los sectores y tiempos:

$$\mu_{kt} = \alpha + \sum_{j=1}^{5} \beta_j x_{jkt.}$$

(ii) Indicadores económicos con efectos comunes para todos los tiempos y efectos diferenciados por sector:

$$\mu_{kt} = \alpha + \alpha_k + \sum_{j=1}^{5} \beta_{jk} x_{jkt.}$$

(iii) Igual que el (ii), más efecto temporal:

$$\mu_{kt} = \alpha + \alpha_k + \gamma_t + \sum_{j=1}^{5} \beta_{jk} x_{jkt.}$$

(iv) Igual que el (ii), más efecto de interacción por sector y efecto temporal:

$$\mu_{kt} = \alpha + \alpha_k + \gamma_{kt} + \sum_{j=1}^{5} \beta_{jk} x_{jkt.}$$

(v) Indicadores económicos con efectos diferenciados por variable, sector y tiempo, más efectos diferenciados por sector:

$$\mu_{kt} = \alpha + \alpha_k + \sum_{j=1}^{5} \beta_{jkt} x_{jkt.}$$

En un segundo nivel de la jerarquía se encuentran las distribuciones de los parámetros de la misma naturaleza, los cuales se consideran intercambiables para los distintos sectores, o dinámicos a través del tiempo. En específico, las distribuciones iniciales de los parámetros son: $\beta_j | b_\beta \sim N(b_\beta, v_\beta)$ para el modelo (i); $\beta_{jk} | b_{j\beta} \sim N(b_{j\beta}, v_{j\beta})$ y $\alpha_k | b_\alpha \sim N(b_\alpha, v_\alpha)$ para el (ii); $\gamma_t | \gamma_{t\text{-}1}, v_\gamma \sim N(\gamma_{t\text{-}1}, v_\gamma)$ con $\gamma_1 | v_\gamma \sim N(0, v_\gamma)$ de forma adicional para el (iii); $\gamma_{kt} | \gamma_{k,t\text{-}1}, v_{k\gamma} \sim N(\gamma_{k,t\text{-}1}, v_{k\gamma})$ con $\gamma_{k1} | v_{k\gamma} \sim N(0, v_{k\gamma})$ adicionalmente para el (iv); y $\beta_{jkt} | \beta_{j,k,t\text{-}1}, v_{jk\beta} \sim N(|\beta_{j,k,t\text{-}1}, v_{jk\beta})$ con $\beta_{jk1} | v_{jk\beta} \sim N(0, v_{jk\beta})$ en complemento para el (v).

Por último, el tercer nivel de la jerarquía se completa con las distribuciones iniciales de los hiperparámetros: $\alpha \sim N(0,0.001)$, $b_\alpha \sim N(0,0.001)$, $b_\beta \sim N(0,0.001)$, $b_{j\beta} \sim N(0,0.001)$, $v_\gamma \sim Ga(0.01,0.01)$, $v_{k\gamma} \sim Ga(0.01,0.01)$, $v_{jk\beta} \sim Ga(0.01,0.01)$ y $v_\alpha = v_\beta = v_{j\beta} = 0.01$

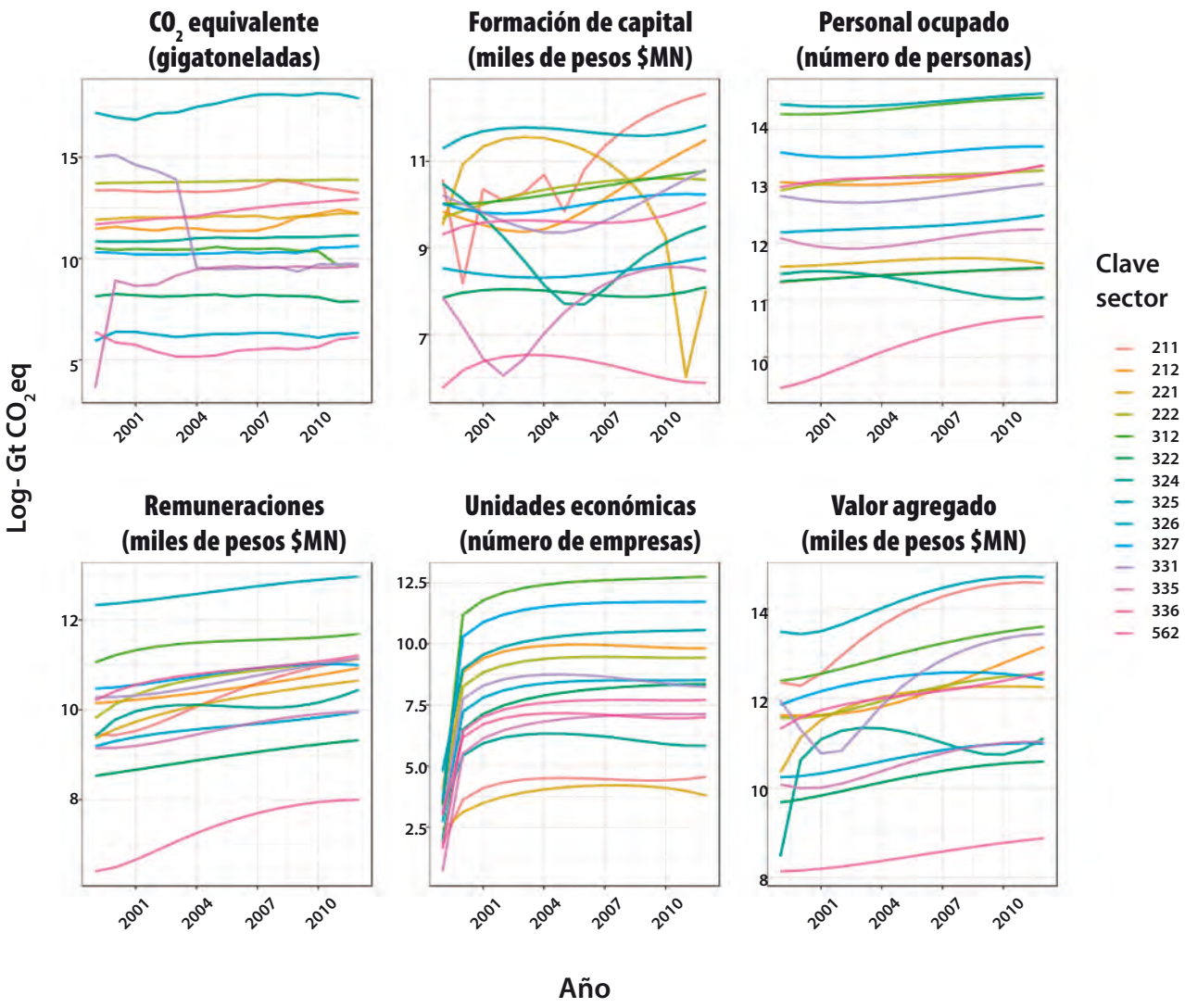
4. Análisis de datos

En la gráfica 1 se presentan, en forma de panel, todas las series de los sectores para cada variable coloreada por sector, siendo Y_{kt} el primer panel superior desde la izquierda, seguido de X_{1kt} en el segundo superior, hasta X_{5kt} en el último inferior. Se observa que algunos de los indicadores económicos se comportan de manera similar en cuanto a tendencia en los distintos sectores; este es el caso de *Personal ocupado*, *Remuneraciones* y *Unidades económicas*. Por otro lado, *Industria química* (CVE 325) y *Manejo de desechos y servicios de remediación* (CVE 562) presentan los valores más altos y más bajos, respectivamente, en todas las variables. Relativamente, existe mayor variabilidad en el indicador de emisiones que en los indicadores económicos.

De forma adicional, se calculó el coeficiente de correlación de Pearson entre la serie de emisiones Y_{kt} y cada indicador económico X_{jkt}, $j = 1, \ldots, 5$ para cada uno de los sectores $k = 1, \ldots, 14$. Estas correlaciones se muestran en el cuadro 2, donde se encuentran cinco bloques verticales, uno para cada variable. En medio de cada uno hay una línea punteada que representa el valor de correlación cero. El color azul denota una positiva, mientras que el rojo, una negativa. La línea gris continua representa el valor de la correlación muestral entre -1 y 1. Por ejemplo, la correlación entre *Personal ocupado* y el indicador de emisiones del sector *Fabricación de productos derivados del petróleo y el carbón* (CVE 324) es la más negativa de todas las calculadas con un valor muy cercano a -1; es decir, entre menor es el número de personas ocupadas en la industria de productos derivados del carbón y del petróleo, mayor es la cantidad de emisiones de GEI. Por su parte, las emisiones de GEI de *Suministro de agua y gas por ductos al consumidor final* (CVE 222) están muy correlacionadas positivamente con todos los indicadores económicos.

Gráfica 1

Series de todos los indicadores coloreados por sector para el periodo 1999-2012



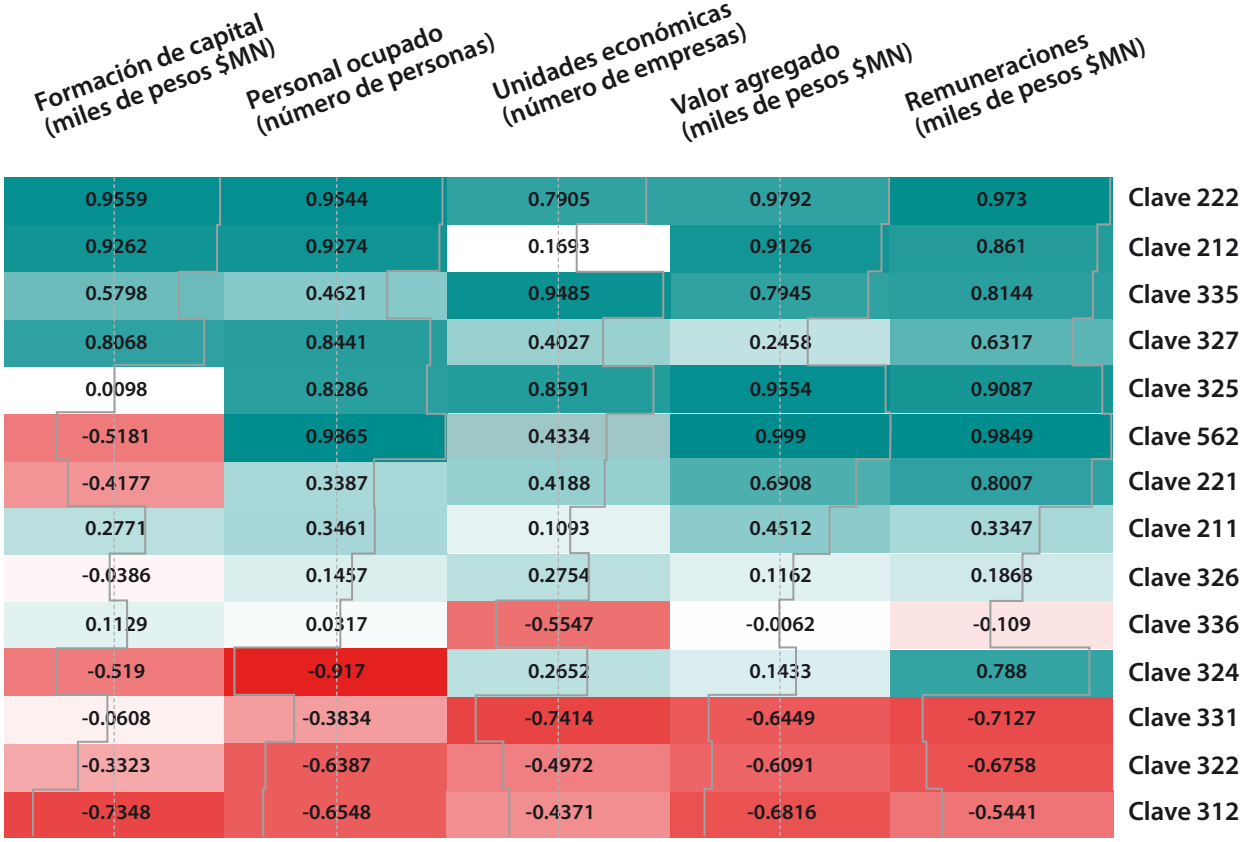
Nota: las claves de sector se encuentran en el cuadro 1.

Las distribuciones posteriores se calcularon a través de simulaciones MCMC, que se obtuvieron mediante *OpenBUGS* con el *software* estadístico *R*. El número de iteraciones de las cadenas fue de 10 mil con un periodo de calentamiento de 2 mil iteraciones y ocupando una simulación cada dos iteraciones para disminuir la autocorrelación de la cadena.

El desempeño de cada especificación se determinó con el criterio de información devianza (DIC, por sus siglas en inglés), el cual es una medida de bondad de ajuste basada en el valor esperado de menos dos veces la verosimilitud, más un factor de penalización por el número efectivo de parámetros incluidos (Spiegelhalter *et al*., 2002). Los valores DIC para cada especificación se presentan en el cuadro 3. Se observa que aquella que muestra el menor DIC es la (ii), la cual presenta efectos diferenciados por sector y coeficientes de regresión diferenciados por variable y sector. Es importante mencionar que los parámetros adicionales en (iii), (iv) y (v) no inducen mejoras en el DIC y esto indica

Cuadro 2

Mapa de color que muestra por sector la correlación de Pearson entre cada indicador económico y el indicador ambiental



Nota: las claves de sector se encuentran en el cuadro 1.

que el efecto temporal es captado sobre todo por la dinámica de las variables económicas explicativas. La explicación reside en que el DIC penaliza especificaciones más complejas —es decir, con más parámetros— y, en el caso de las especificaciones (iii) a (v), la inclusión de estos no contrarresta tal penalización en términos del ajuste.

La especificación completa del modelo ganador se forma mediante la verosimilitud (1), la media definida con la ecuación (ii) y sus correspondientes distribuciones iniciales descritas al final de la sección 3. A continuación, se presentan las estimaciones posteriores de los parámetros de este modelo.

La gráfica 2 presenta las estimaciones, puntuales y por intervalo a 95%, de los efectos de los sectores industriales α_k y de los coeficientes de regresión β_{jk} para $j = 1, \ldots, 5$. El valor agregado es el indicador

Cuadro 3

DIC para cada especificación de la media

Especificación	DIC
(i)	588
(ii)	-221
(iii)	652
(iv)	375
(v)	696

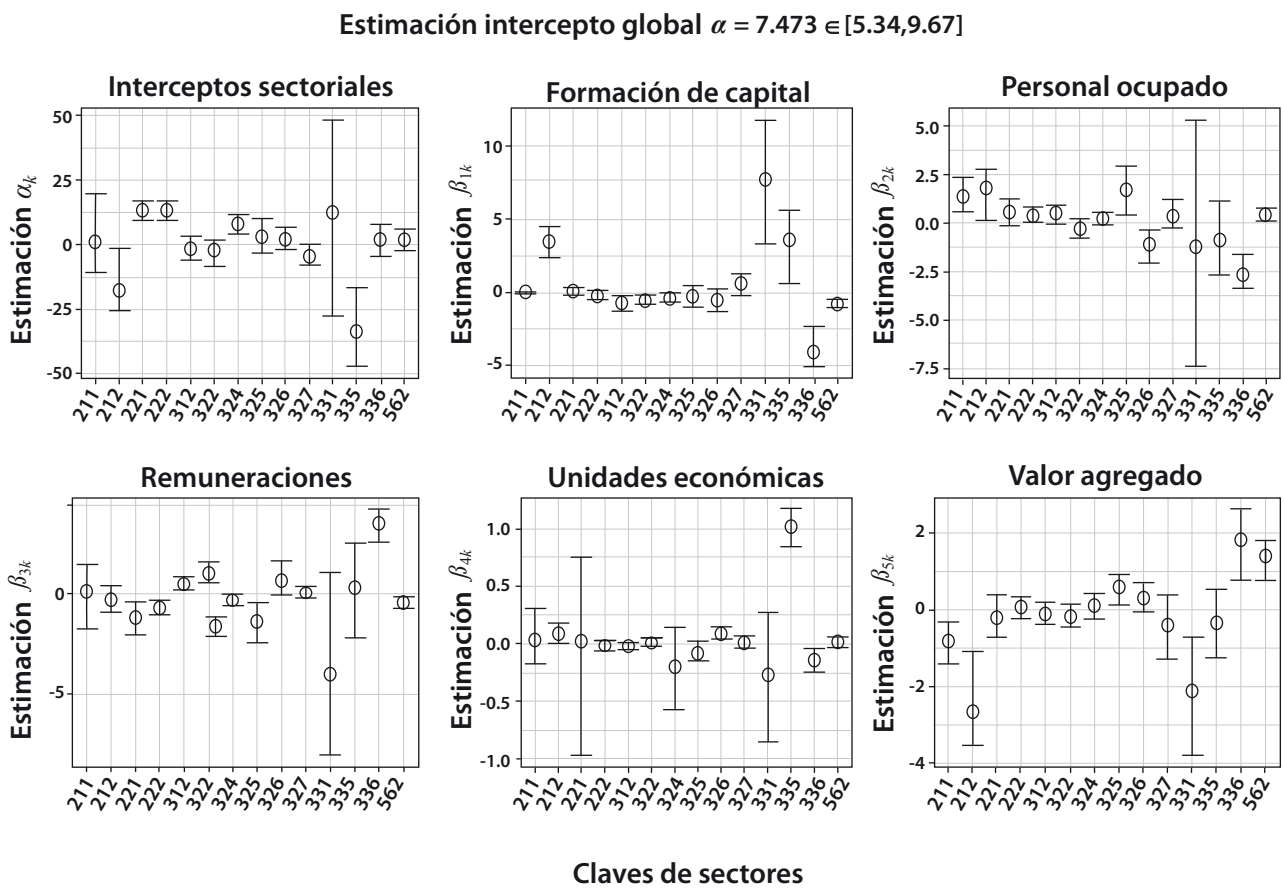
clave con la mayoría de los coeficientes (β_{5k}) con valores significativamente alejados del cero. Por otro lado, la variable cuyos coeficientes presentan intervalos posteriores pequeños es la formación bruta de capital fijo (β_{1k}). Merece mención que *Industrias metálicas básicas* (CVE 331) es, de manera consistente, el sector cuyos coeficientes presentan los mayores intervalos de incertidumbre, seguidos por los del 335, *Fabricación de aparatos eléctricos y generadores de energía eléctrica*.

En este caso, el número de empresas (β_{4k}) se presenta irrelevante (no significativo) para explicar el comportamiento de las emisiones en la mayoría de los sectores elegidos, siendo la excepción el sector 335, en el cual un incremento en el número de empresas tiende a aumentar de forma significativa las emisiones de gases de efecto invernadero.

La formación de capital fijo (β_{1k}) se presenta como *eficiente* en términos ecológicos para casi todos los sectores, con excepción de *Fabricación de aparatos eléctricos y generadores de energía eléctrica* (CVE 335), *Minería de minerales metálicos y no metálicos excepto petróleo y gas* (CVE 212) e *Industrias metálicas básicas* (CVE 331), en los cuales aumentos en el capital fijo producen incrementos significativos en las emisiones de GEI. Por el contrario, *Fabricación de equipo de transporte* (CVE 336) muestra un comportamiento muy eficiente en términos ambientales al reducir las emisiones de GEI por aumentos en la formación de capital fijo.

Gráfica 2

Estimaciones de los efectos de los sectores industriales α_k y de los coeficientes de regresión β_{jk}



Nota: las claves de sector se encuentran en el cuadro 1.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

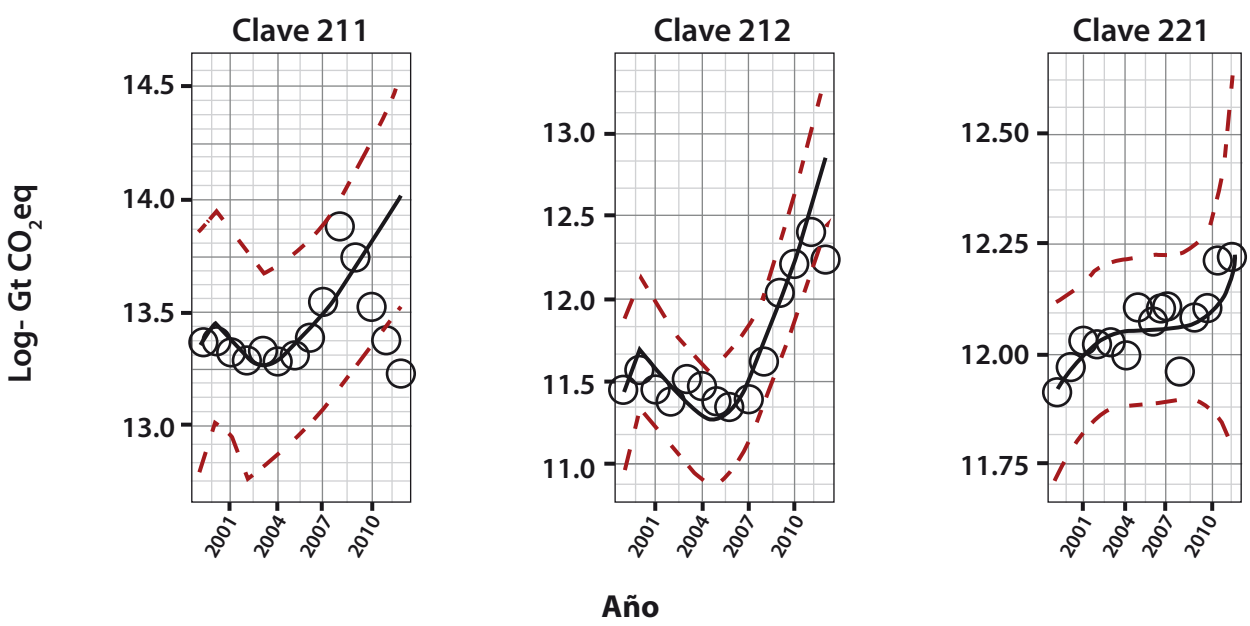
Los otros indicadores (β_{2k}, β_{3k}, β_{5k}) muestran una eficiencia mixta por sectores, mientras que *Fabricación de productos a base de minerales no metálicos* (CVE 327) no presenta relevancia estadística para ninguno de los indicadores.

El resultado inmediato es que los sectores 335, 212 y 331 son los primeros focos rojos en la lista de industrias más susceptibles a contaminar por aumentos en la formación de capital, en especial el último. En particular, un aumento de 1% en la formación bruta de capital en los dos primeros tendría como consecuencia un incremento promedio de alrededor de 0.035 Gt CO_2eq, mientras que para el 331 un aumento similar significaría 0.077 Gt CO_2eq. Una de las posibles explicaciones es que no están invirtiendo en tecnologías más limpias o no se encuentran bien regulados.

Las predicciones de Y_{kt} obtenidas con el modelo ganador —para los mismos valores de las variables explicativas observadas— se presentan en la gráfica 3, las cuales sirven como un indicador del ajuste del modelo: la línea continua corresponde a la predicción puntual, mientras que las punteadas en rojo forman un intervalo de predicción a 95% de probabilidad y los datos se muestran con puntos huecos. *Extracción de petróleo y gas* (CVE 211), *Industria del papel* (CVE 322) y *Fabricación de equipo de transporte* (CVE 336) son los sectores cuyas predicciones se alejan más de los datos, esto se debe, quizá, a que los datos muestran patrones caóticos moviéndose de forma drástica hacia arriba o abajo, siendo el problema principal los últimos dos años analizados (2011 y 2012). Para los demás sectores (la mayoría), los ajustes muestran intervalos angostos, mostrando predicciones cercanas a los datos, en especial para *Suministro de agua y gas por ductos al consumidor final* (CVE 222), *Fabricación de productos derivados del petróleo y del carbón* (CVE 324) y *Manejo de desechos y servicios de remediación* (CVE 562). Las predicciones para *Industrias del plástico y del hule* (CVE 326) y *Fabricación de productos a base de minerales no metálicos* (CVE 327) presentan intervalos anchos, sin embargo, son aceptables ya que los datos son capturados dentro de los intervalos.

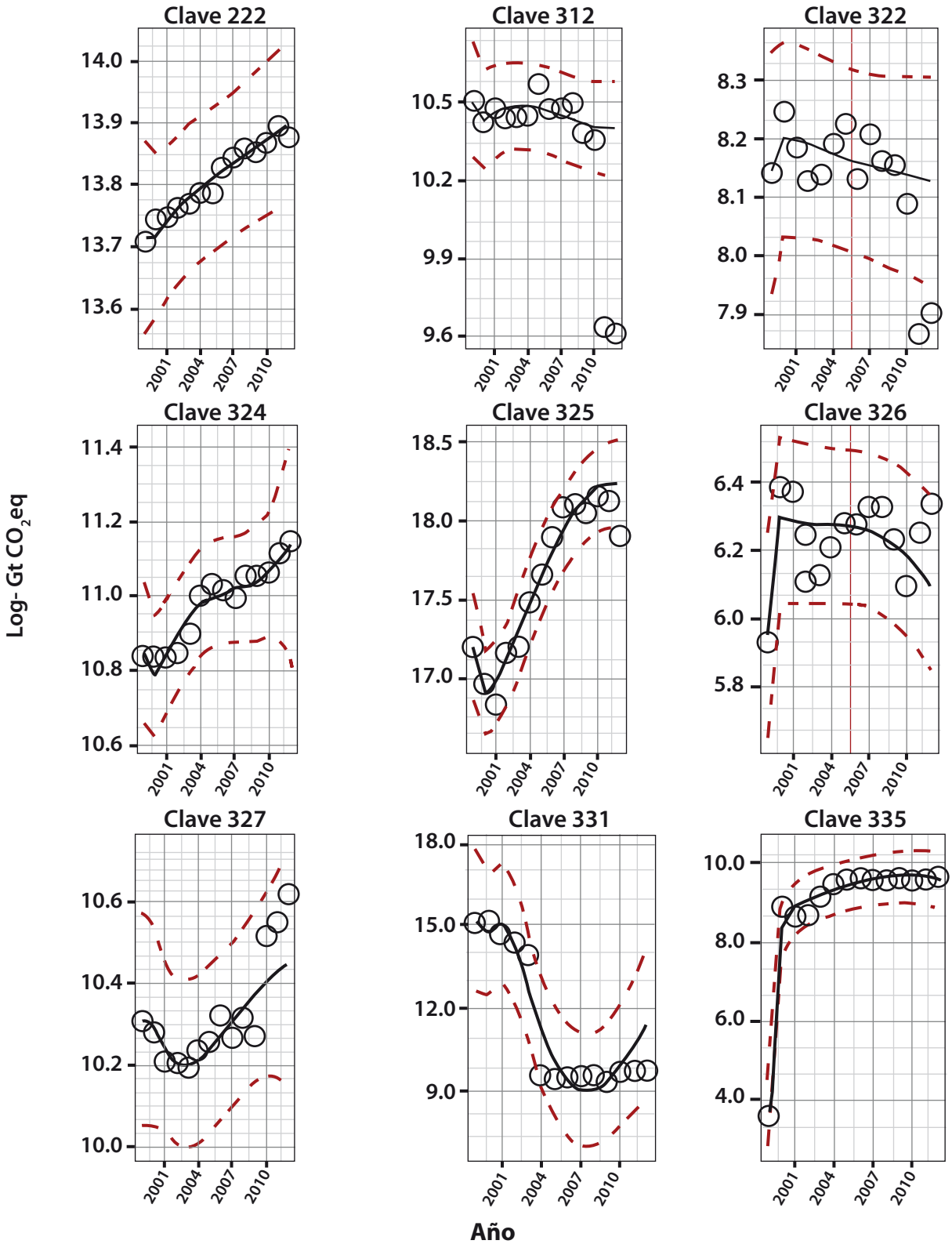
Gráfica 3 Continúa

Predicciones dentro de muestra de Y_{kt} obtenidas con el modelo definido por (1) y (ii) para cada sector



Predicciones dentro de muestra de Y_{kt} obtenidas con el modelo definido por (1) y (ii) para cada sector

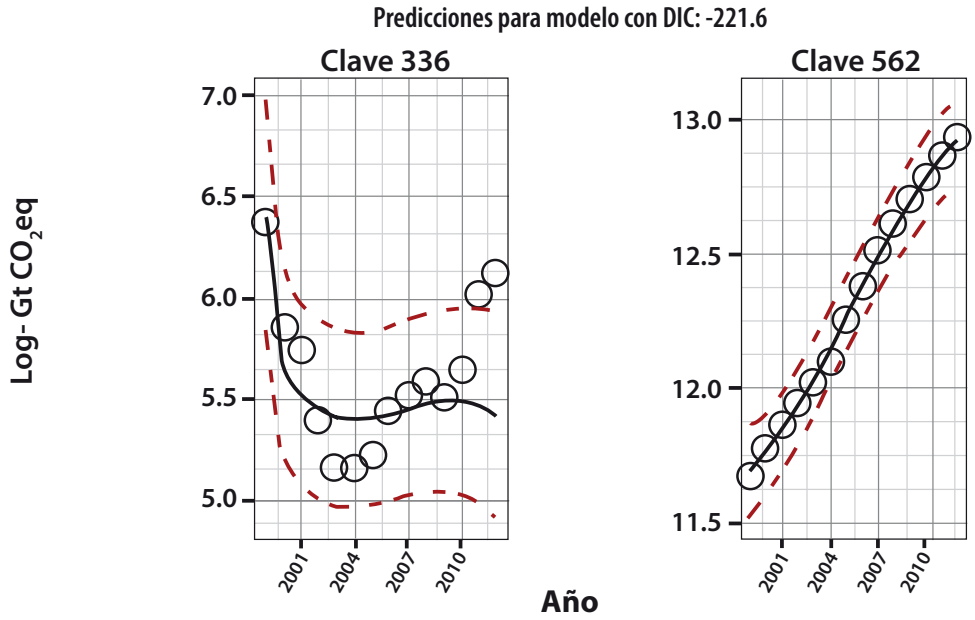
Predicciones para modelo con DIC: -221.6



INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Gráfica 3 Concluye

Predicciones dentro de muestra de Y_{kt} obtenidas con el modelo definido por (1) y (ii) para cada sector



Nota: las claves de sector se encuentran en el cuadro 1.

5. Conclusiones

En este proyecto se aprovecharon las fuentes oficiales de datos ambientales y económicos para realizar un análisis de eficiencia desde un punto de vista económico-ambiental. Para su realización, fue necesario un laborioso mapeo conceptual entre la clasificación ambiental del IPCC y el SCIAN. Esta correspondencia es la primera que se realiza en México con el fin de analizar de forma puntual la eficiencia que presentan los sectores industriales en materia ambiental.

La utilización de un enfoque estadístico bayesiano permite la estimación simultánea de varios coeficientes (alrededor de 80) cuando se tiene una cantidad relativamente reducida de datos; en este caso, se contaba con solo 12 datos temporales para cada combinación de sector e indicador económico. Esto se refleja en el buen desempeño de un modelo de regresión lineal jerárquico que describe las interacciones entre las variables macroeconómicas y las emisiones de GEI medidas en Gt CO_2eq. Las técnicas de modelado jerárquico bayesiano que se emplearon en este trabajo no se han encontrado en la literatura del ámbito nacional o internacional, por lo que constituye una aplicación novedosa con el fin de analizar la relación entre los indicadores económicos y las emisiones contaminantes de los principales sectores industriales a través de su comportamiento en el tiempo.

A partir de los resultados del modelo, se establecieron criterios simples, al final de la sección de análisis, que permitieron clasificar a tales sectores en términos de la eficacia positiva o negativa que presentan los indicadores clave de actividad económica en relación con las emisiones de GEI. Esto permitirá a quienes toman decisiones saber, por ejemplo, en qué sector habría evidencia de una propensión a ejecutar inversión limpia y en cuáles se debe poner mayor atención al uso de recursos.

Este análisis representa una herramienta para entender la mecánica de interacción entre la información económica y ambiental de los sectores, la cual permite plantear nuevas hipótesis sobre el comportamiento de las industrias al analizar la evidencia

estadística. Será relevante, en materia ambiental, investigar a qué se deben algunos comportamientos particulares, si son cambios de medición, efectos de una nueva legislación o bien, la inclusión de nuevas tecnologías más limpias. Más aún, será vital para la sociedad impulsar decisiones sustentadas en análisis interdisciplinarios con el fin de plantear mejores estrategias de desarrollo que, a su vez, permitirán alcanzar las metas de reducción de emisiones y mitigar así los catastróficos efectos del cambio climático.

Fuentes

Banerjee, S.; B. P. Carlin & A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall, Boca Raton, 2014.

Berliner L. M. "Hierarchical Bayesian Time Series Models", in: Hanson, K. M. & R. N. Silver (eds.). *Maximum Entropy and Bayesian Methods*. *Fundamental Theories of Physics*. Vol. 79. Dordrecht, Springer, 1996.

Cappé, O.; E. Moulines & T. Ryden. *Inference in Hidden Markov Models*. New York, Springer Verlag, 2010, 653 pp.

Chen, M.-H.; Q.-M. Shao & J. G. Ibrahim. *Monte Carlo methods in Bayesian computation*. New York, Springer, 2012, 386 pp.

Environnement et Changement Climatique Canada. *Programme de déclaration des émissions de gaz à effet de serre par les installations. Aperçu des émissions déclarées pour 2015*. Quebec, Canada, 2017 (DE) https://www.ec.gc.ca/ges-ghg/82BA1E22-9653-45F1-8EC2-BF8A2151555/ECCC_PDGES_Aper%E7uDes%C9missionsD%E9clar%E9esPour2015.pdf

Galindo, L. M. *La economía del cambio climático en México*. Síntesis. 2009.

Gelman, A.; J. B. Carlin; H. S. Stern & D. B. Rubin. *Bayesian data analysis*. Boca Raton, FL, USA; Chapman & Hall/CRC; 2014; 662 pp.

Hamit-Haggar, M. "Greenhouse gas emissions, energy consumption and economic growth: A panel cointegration analysis from Canadian industrial sector perspective", in: *Energy Economics*. 34, 2012, pp. 358-364.

Instituto Nacional de Estadística y Geografía (INEGI). *Sistema de Clasificación Industrial de América del Norte, México. SCIAN 2013*. México, INEGI, 2013.

IPCC. *Climate Change 2007: The Physical Science Basis*. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S.; D. Qin; M. Manning; Z. Chen; M. Marquis; K. B. Averyt; M. Tignor and H. L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007, 996 pp.

_____ *Climate Change 2013*: *The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T. F.; D. Qin; G.-K. Plattner; M. Tignor; S. K. Allen; J. Boschung; A. Nauels; Y. Xia; V. Bex and P. M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013, 1535 pp.

Kasman, A. & Y. S. Duman. "CO_2 emissions, economic growth, energy consumption, trade and urbanization in new EU member and candidate countries: a panel data analysis", in: *Economic Modelling*. 44, 2015, pp. 97-103.

Nieto-Barajas, L. E. & E. de Alba. "Bayesian regression models", in: *Predictive Modeling Applications in Actuarial Science*. Frees, E. W.; R. A. Derrig and G. Meyers (eds.). Cambridge University Press, 2014, pp. 334-366.

Nordhaus, W. D. "To slow or not to slow: the economics of the greenhouse effect", in: *The Economic Journal*. 101, 1991, pp. 920-937.

R Core Team. R: *A language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing, 2018 (DE) https://www.R-project.org/

Scheffe. H. *The Analysis of Variance*. New York, John Wiley & Sons, 1959, 470 pp.

Secretaría de Gobernación (SEGOB). "Acuerdo para el uso del Sistema de Clasificación Industrial de América del Norte (SCIAN) en la recopilación, análisis y presentación de estadísticas económicas", en: *Diario Oficial de la Federación*. México, 10 de julio de 2009 (DE) http://dof.gob.mx/nota_detalle.php?codigo=5098199&fecha=10/07/2009

_____ "Ley General de Cambio Climático (LGCC)", en: *Diario Oficial de la Federación*. México, 6 de junio de 2012 (DE) http://dof.gob.mx/nota_detalle.php?codigo=5249899&fecha=06/06/2012

_____ "Ley General de Equilibrio Ecológico y Protección al Ambiente (LGEEPA)", en: *Diario Oficial de la Federación*. México, 28 de enero de 1988 (DE) http://dof.gob.mx/nota_detalle.php?codigo=4718573&fecha=28/01/1988

_____ "Programa Especial de Cambio Climático (PECC) 2014-2018", en: *Diario Oficial de la Federación*. México, 28 de abril de 2014 (DE) http://dof.gob.mx/nota_detalle.php?codigo=5342492&fecha=28/04/2014

Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). *Inventario Nacional de Emisiones de Gases de Efecto Invernadero 1990-2010*. México, 2013.

_____ *Quinta comunicación nacional ante la Convención Marco de las Naciones Unidas sobre el Cambio Climático*. México, 2012.

Servicio de Información Agroalimentaria y Pesquera (SIAP). *Sistema de Información Agroalimentaria de Consulta (SIACON)*. 2018 (DE) https://www.gob.mx/siap/prensa/sistema-de-informacion-agroalimentaria-de- consulta-siacon?idiom=es

Spiegelhalter, D.; N. Best; B. Carlin & A. van der Linde. "Bayesian measures of model complexity and fit", in: *Journal of the Royal Statistical Society*. Series B. 64, 2002, pp. 583-639.

Stern, N. H. *The economics of climate change*: *The Stern review*. Cambridge University Press, 2007.

United Nations Framework Convention on Climate Change (UNFCCC). *Kyoto Protocol to the United Nations Framework Convention on Climate Change*. 1988 (DE) https://unfccc.int/sites/default/files/kpeng.pdf

Yu, M.; T. Wiedmann; R. Crawford & C. Tait. "The carbon footprint of Australia's construction sector", in: *Procedia engineering*. 180, 2017, pp. 211-220.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Movilidad laboral internacional en el caso mexicano

International *Labor Mobility* The Case of Mexico

Olinca Páez*

* Instituto Nacional de Estadística y Geografía, olinca.paez@inegi.org.mx

Illegal immigrants being escorted back across the border to Mexico /Loomis Dean/Getty Images

En el mundo, todos los días, las personas cruzan fronteras internacionales con el propósito de trabajar. La duración de su estancia en los lugares de destino y el eventual cambio de residencia permite distinguir, al menos de forma conceptual, a los trabajadores transfronterizos de los trabajadores migrantes temporales y a estos y aquellos, de los extranjeros residentes. En la práctica, sin embargo, la medición de la población global que se mueve por razones laborales no es sencilla.

La Comisión Económica de las Naciones Unidas para Europa propuso la integración de un grupo de trabajo orientado a la definición de la movilidad laboral internacional como objeto de estudio, así como a la identificación de fuentes de información en distintos países. Los acuerdos y recomendaciones surgidos del análisis conjunto se presentan aquí, junto con la descripción de las fuentes disponibles en el caso mexicano y la caracterización de volúmenes y flujos vinculados a este tipo de movilidad.

En México, la información para describirla es vasta, frecuente y variada, aunque está dispersa y difícilmente puede ser agregada o integrada por un usuario no experto. Para comprender el fenómeno a cabalidad, como coordinador del Sistema Nacional de Información Estadística y Geográfica, el Instituto Nacional de Estadística y Geografía debe orientar los esfuerzos para reunir de manera sistemática esta información proveniente de registros administrativos, censos y encuestas.

Palabras clave: trabajadores migrantes temporales; movilidad laboral internacional; trabajadores transfronterizos; trabajadores extranjeros no residentes.

Recibido: 30 de abril de 2018.
Aceptado: 4 de marzo de 2019.

In the world, every day, people cross international borders with the purpose of working. The duration of their stay in the places of destination and the eventual change of residence makes it possible to distinguish, at least in a conceptual way, cross-border workers from temporary migrant workers and these, in turn, from resident foreigners. In practice, however, measuring the global population that moves for labor reasons is not simple.

The United Nations Economic Commission for Europe proposed the integration of a working group aimed at the definition of international labor mobility as an object of study, as well as the identification sources of information in different countries. The agreements and recommendations arising from the joint analysis are presented here, together with the description of the sources available in the Mexican case and the characterization of volumes and flows linked to this type of mobility.

In Mexico, the information to describe it is vast, frequent and varied, although it is scattered and can hardly be added or integrated by a non-expert user. To fully understand the phenomenon, as coordinator of the National Statistical and Geographical Information System, the National Institute of Statistics and Geography must guide efforts to systematically gather this information from administrative records, censuses and surveys.

Key words: temporary migrant workers; international labour mobility; frontier workers; incoming commuters.

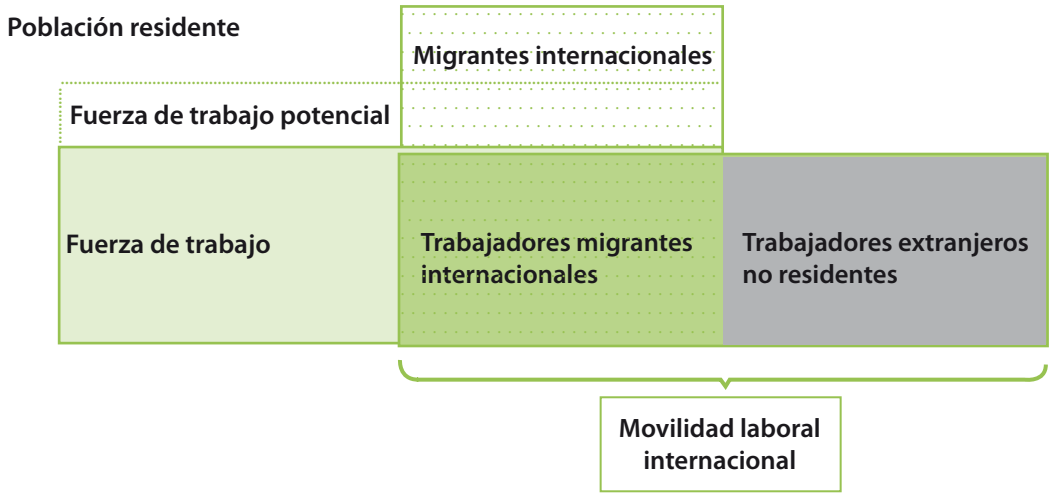
Introducción

La movilidad laboral internacional (MLI) implica desplazamiento, cruce de fronteras con otras naciones y motivaciones de trabajo; comprende todos los movimientos de personas de un país a otro con el propósito de emplearse u ofrecer servicios. Se nutre de la migración internacional, pero no son idénticas, porque esta puede estar motivada por otras razones, como: estudiar, reunirse con la familia o sobrevivir a las condiciones del propio país mediante asilo o refugio en la nación de llegada. Otra distinción importante es que la MLI puede ocurrir sin necesidad de cambiar de lugar habitual de residencia, lo que es factible donde la geografía lo permite y pertinente cuando existen fuertes disparidades económicas entre los países. La delimitación del concepto se ilustra en la figura 1.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Figura 1

Concepto de movilidad laboral internacional



Fuente: tomado del reporte *Measuring international labour mobility* preparado por el *Task Force on Measuring Labour Mobility.*

Aunque el fenómeno migratorio ha evolucionado de formas muy distintas en las regiones del mundo —acaparando la atención del público por su dimensión—, la MLI sin cambio de residencia habitual ha cobrado mucha relevancia en las últimas décadas como consecuencia de mayores libertades de tránsito para las personas en regiones específicas (como la Unión Europea) o en otras con cierta tradición de empleo transfronterizo (por ejemplo, los casos de Israel-Palestina o México-Guatemala).

La MLI se erige como un nuevo objeto de estudio debido a que los Estados ofrecen una diversidad de servicios sobre la base de la organización territorial y la población allí asentada, por lo que el cruce de fronteras internacionales de trabajadores que no cambian su residencia habitual implica desbalances en la oferta y demanda de servicios específicos. Por ello, la medición de la MLI es relevante para las políticas de corto, mediano y largo plazos que, en materia de integración, empleo y provisión de servicios públicos, los países receptores deben diseñar e instrumentar.[1]

La descripción de su volumen y flujos asociados está basada en la intersección de los conceptos y definiciones de migración internacional y fuerza laboral pues, como se menciona arriba, si bien la primera implica movilidad, no todos los migrantes se desplazan con fines laborales y, por otra parte, quienes trabajan en un país distinto del que residen no son, por definición, migrantes.

Así, el concepto de MLI distingue, a grandes rasgos, dos grupos poblacionales con presencia en el territorio de un país:

- Los *trabajadores migrantes internacionales*, que han cambiado de país de residencia habitual y forman parte de la fuerza de trabajo en el lugar de destino.
- Los *trabajadores extranjeros no residentes*, que cruzan las fronteras internacionales para emplearse u ofrecer servicios, pero no son residentes habituales en el país al que llegan.

Caracterizar la MLI requiere tanto de la identificación de los volúmenes de unos y otros en un momento en el tiempo como de la medición del

1 Un análisis exhaustivo de la situación actual y de la pertinencia de estudiar y definir de forma adecuada la MLI para solventar los vacíos estadísticos puede encontrarse en Statistics Austria (2017).

flujo de personas que, por razones de trabajo, se desplazan entre naciones conservando o no su residencia habitual. Como ya se mencionó, la dimensión de aquellos que no cambian su lugar de residencia habitual es creciente y muy significativa en ciertos países o regiones.

Además de insistir en que la migración relacionada con el trabajo no es la única forma de MLI, otro hecho a resaltar es que las naciones son, a la vez, emisoras y receptoras de población que se mueve a través de sus fronteras con fines laborales. En este sentido, la figura 1 define la MLI desde la perspectiva de un país receptor, es decir, refiere a una sola cara de la moneda. En el caso mexicano, la recolección de información, el análisis y las políticas públicas se han concentrado en la otra cara debido a que, históricamente, la nación ha sido emisora,[2] pero este perfil ha cambiado en las últimas décadas y cada vez más se reconoce la importancia de entenderlo como país de tránsito y destino.[3]

Entonces, las preguntas que guían este trabajo son: ¿disponemos en México de información sobre todos los tipos de movilidad laboral internacional?, ¿tenemos la capacidad de analizar el fenómeno como país emisor, de tránsito y de destino?, ¿estamos generando datos que permitan caracterizar el tema en toda su complejidad y, a la vez, desglosarlo con fines de política pública?

En este artículo se expone de forma sucinta el acuerdo internacional en la conceptualización de la MLI y las recomendaciones para su adecuada medición. Con ello como referencia, se analiza el caso para México, principalmente como país receptor, desde la localización de las fuentes disponibles, la descripción de la información que compilan y los límites de la misma, hasta la caracterización de volúmenes y flujos de trabajadores tanto migrantes internacionales como extranjeros no residentes. Hacia el final del documento, de manera muy elemental, se describen también los grupos de población que salen del territorio nacional hacia otros países por razones laborales. Por último, se emiten algunas recomendaciones para mejorar la captación de información sobre el tema en México.

Enfoque, metodología y fuentes

La construcción del marco estadístico para la medición de la MLI se basó en esencia en la integración de los marcos conceptuales y estadísticos ya existentes para la medición de la migración internacional[4] y el trabajo,[5] aunque considera también los desarrollados para el comercio internacional en servicios[6] y el turismo.[7] La labor del grupo internacional consistió en clarificar conceptos y definiciones comunes y complementarios que están vigentes en todos esos ámbitos para desarrollar un esquema armónico que permitiera el estudio de la MLI.

En paralelo, fueron identificadas las fuentes de información útil para la medición del fenómeno en Estados miembros de la Organización Internacional del Trabajo (OIT) y se revisaron con mayor detalle los casos de cuatro países (Israel, Italia, México y Noruega) bajo un esquema común de análisis a partir de tabulados diseñados en conjunto para tal fin.

El grupo reconoció las fuentes de información potenciales, a saber: encuestas en establecimientos, en hogares, de fuerza de trabajo, así como en fronteras o a pasajeros, además de registros administrativos y censos de población. Se destacó

2 Los estudios sobre la migración de mexicanos a Estados Unidos de América (EE. UU.) son vastos y de larga data, tal como ha sido el fenómeno en sí mismo. El abordaje es, además, diverso: se han estudiado las características de los hogares donde algún miembro es migrante; el impacto de la migración en las comunidades de origen y destino; la integración/asimilación de las primeras y segundas generaciones en los lugares de destino; el efecto de las remesas; el perfil sociodemográfico de los que migran y la migración circular; entre otros.

3 Vale la pena mencionar las investigaciones de Nájera (2011) enfocadas en los trabajadores transfronterizos en el sur del país, provenientes de Guatemala; Meza (2015), quien revisa la situación laboral de los centroamericanos en México; y Fernández y Rodríguez (2016), que estudian a la población de Honduras que transita por México con la intención de llegar a EE. UU. y que, eventualmente, se asienta en territorio mexicano.

4 *United Nations Recommendations on Statistics of International Migration (RSIM).*

5 13th International Conference of Labour Statisticians (ICLS). *Resolution I.*

6 *Manual on Statistics of International Trade in Services (MSITS) 2010.*

7 *International Recommendations for Tourism Statistics (IRTS) 2008.*

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

que, debido a que estas fuentes no suelen estar bien integradas o sistematizadas, el esfuerzo normalmente se concentra en reunir la información proveniente de registros de población, de migración y de empleo y seguridad social, además de los permisos de trabajo, de preferencia en las naciones de destino.

Los casos de estudio propuestos describen la disponibilidad de información, las características de las fuentes y un breve análisis del impacto de la MLI en los países,[8] y son la base para las recomendaciones sobre buenas prácticas en la medición de este tema.

Así, el estudio de la MLI en México estuvo guiado por los acuerdos conceptuales, técnicos y metodológicos del grupo internacional. El documento que describe el caso mexicano fue evaluado por los miembros del grupo y mejorado a partir de recomendaciones y apuntes específicos. Formó parte del reporte exhaustivo sobre este tema, avalado por la comunidad estadística internacional y presentado ante la 66.ª Sesión Plenaria de la Conferencia de Estadísticos Europeos en junio del 2018 en Ginebra, Suiza.

Este artículo sintetiza la propuesta internacional y expone lo hallado para el caso de México. El análisis parte de la compilación de información de distintas fuentes: estadísticas registradas y publicadas por el Instituto Nacional de Migración (INM); proyectos estadísticos oficiales a cargo del Instituto Nacional de Estadística y Geografía (INEGI) —Censo de Población y Vivienda 2010, Encuesta Intercensal (EI) 2015, además de las encuestas nacionales de la Dinámica Demográfica (ENADID), de Ocupación y Empleo (ENOE) y de Ingresos y Gastos de los Hogares (ENIGH)—; así como un proyecto dirigido por la comunidad académica: la Encuesta sobre Migración en la Frontera Sur de México (EMIF-Sur).

8 Cada uno de los casos de estudio es particular y relevante para entender la complejidad del fenómeno. Vale la pena revisarlos en https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2018/CES_3_LabourMobility_for_consultation_for_upload.pdf

Acuerdos conceptuales del grupo internacional

Respecto a los trabajadores migrantes internacionales

A pesar de que en la actualidad el concepto *trabajo* refleja una noción más amplia de actividades productivas que no necesariamente implican remuneración a cambio, para esta temática se consideró pertinente referirse a la participación en el mercado laboral, por lo que un trabajador migrante internacional se define como aquel que forma parte de la fuerza laboral en el país al cual llegó a residir.

Cabe destacar, también, que si bien otras categorías migratorias (como las de refugiados, dependientes o estudiantes) quedan, en principio, excluidas de esta perspectiva, es posible que de ellas se nutra el volumen de trabajadores migrantes futuro, por lo cual es fundamental tenerlas disponibles.

Acerca de los trabajadores extranjeros no residentes

Son las personas que llevan a cabo actividades económicas en un país del que no son residentes habituales, mientras que en la nación donde están asentadas son *residentes trabajando en el exterior*. Entre ellas podemos identificar trabajadores transfronterizos y temporales, oferentes de servicios no residentes, trabajadores de negocios que son remunerados por el país de origen y empleados en puestos de trabajo en el extranjero.

Los últimos tres tipos son considerados trabajadores extranjeros no residentes si su estancia en el país de destino es menor a 12 meses. En contraste, los primeros dos se consideran trabajadores extranjeros no residentes solo si su estancia es menor a tres meses, pues si su estadía se prolonga por más de ese tiempo, pero menos de 12 meses, se consideran migrantes de corto plazo con residencia habitual en el país de destino (United Nations, 1998).

A propósito del flujo internacional de personas

Está formado por el número de personas que cruzan las fronteras internacionales en un periodo determinado, típicamente es un año. En el tema que nos ocupa, el foco es la corriente de personas que ingresa al país con el objetivo de trabajar, su peso relativo en la fuerza de trabajo y la proporción que representan en el total de los migrantes internacionales.

Este flujo es difícil de medir por dos razones: 1) las personas pueden cruzan múltiples veces en un periodo y lo común es que en los controles fronterizos se registre el número de entradas y salidas, no el de personas; y 2) no todos los individuos cruzan fronteras por los puntos oficiales.

Asimismo, la medición del número de personas con el propósito de trabajar se dificulta porque ellas pueden preferir no declarar su intención de emplearse en el país de llegada, porque la razón principal de la migración puede ser otra, o debido a que no hay oportunidad de declarar el motivo de entrada (ya que esta no es oficial o los individuos no son sujetos de control migratorio). Además, se debe considerar que, aunque el acceso al país haya sido legal, tanto los trabajadores extranjeros no residentes como los migrantes internacionales pueden haberse incorporado al mercado laboral del país receptor con o sin autorización.

En cualquier caso, debe quedar claro también que quien declara intención de trabajar no siempre lo hace y que quien ingresa al país por otras razones primordiales, eventualmente puede incorporarse a la fuerza laboral.

Fuentes de información disponibles para describir la MLI en/desde México

Encuesta Intercensal

Ofrece datos para estimar el volumen de migrantes internacionales y el de trabajadores migrantes internacionales en el país, por duración de estancia y nacionalidad, así como el volumen de residentes trabajando en el exterior en la semana de referencia. Se pueden comparar las tasas de Población Económicamente Activa (PEA) y de desempleo entre nacidos en el país y residentes extranjeros o entre extranjeros con y sin nacionalidad mexicana. Al análisis de las diferencias se puede agregar, también, la duración de la estancia. Además, considerando el lugar de nacimiento y la nacionalidad de los residentes en México, es posible contrastar la distribución de personas empleadas en categorías de ocupación e industria.

La EI es el proyecto estadístico más grande de los últimos años en materia sociodemográfica: se levantó en el 2015 y recogió información en más de seis millones de viviendas particulares habitadas, registrando datos sobre sus residentes habituales (personas que normalmente la habitan y donde por lo general duermen, preparan su comida, comen y se protegen del ambiente, incluidas las que, al momento de la entrevista, estaban en la vivienda por no tener otro lugar donde vivir). Los microdatos están disponibles en la página del INEGI en internet.

Censo de Población y Vivienda

Con el levantamiento censal del 2010 se obtuvo la cifra de habitantes en México nacidos en EE. UU. o en otro país, aunque no fue posible distinguir a los mexicanos nacidos en el extranjero, pues no se añadió una pregunta sobre la nacionalidad, ni en el cuestionario básico ni en el ampliado.

La proporción de migrantes internacionales integrados a la fuerza de trabajo del país se puede calcular con la pregunta sobre la condición de actividad económica incluida en el cuestionario básico. También, es posible comparar tal proporción de aquellos que vivían en el extranjero cinco años atrás frente a los que ya residían en México desde entonces.

No obstante, solo en el cuestionario ampliado (aplicado en 2.9 millones de viviendas) es donde se

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

profundiza sobre las condiciones laborales, el sector de actividad económica, el trabajo en el exterior, el país de residencia actual de la población con experiencia migratoria y la fecha de retorno de los migrantes identificados.

Encuesta Nacional de la Dinámica Demográfica

La ENADID es una encuesta especial del INEGI que se ha levantado en 1992, 1997, 2006, 2009 y 2014. En la última ronda fueron observadas las personas que residían de forma habitual en una selección de viviendas particulares localizadas en el territorio nacional.

Permite calcular el volumen de migrantes internacionales y trabajadores migrantes internacionales en el país, así como analizar las diferencias en el volumen y tasas de PEA y desempleo respecto a las de los nacidos en México o por duración de la estancia.

Debido a que también ofrece información sobre las causas de la migración, de la ENADID se pueden obtener estimaciones de la cifra de refugiados y asilados,[9] así como una aproximación del flujo de inmigrantes (cuando se observan los efectivos con duraciones de estancia cortas), incluidos los migrantes con intenciones de trabajar.

Registros administrativos del Instituto Nacional de Migración

En el 2009 se llevó a cabo un inventario de registros administrativos del INM con el fin de sistematizar la información recolectada y estar en posibilidades de generar estadísticas sobre el volumen y perfiles sociodemográficos de los extranjeros en el país. El ejercicio consistió en el conteo de formas migratorias expedidas (o renovadas) a residentes temporales o permanentes en el país (Rodríguez-Chávez y Cobo, 2012). Los resultados de este inventario están disponibles en forma de tablas, pero desafortunadamente los microdatos no están abiertos.

Los reportes que cuantifican todos los tipos de permiso emitidos por el INM a visitantes o residentes extranjeros se difunden vía el *Boletín Mensual de Estadísticas Migratorias* y las series históricas son puestas a disposición. Es evidente que estos registros no incluyen a quienes están en el país sin haber entrado por los puntos oficiales. La *Ley de Migración* clasifica a la población extranjera en México como se muestra en el cuadro 1.

Encuesta sobre Migración en la Frontera Sur de México

Uno de los cuestionarios se aplica a las personas nacidas en Guatemala, El Salvador u Honduras que se desplazan desde México o EE. UU. después de haber permanecido ahí durante más de un mes. Están incluidos quienes residen en alguno de ellos y regresan de visita a su lugar de origen.

Varias instituciones públicas están involucradas en este proyecto anual coordinado por El Colegio de la Frontera Norte (COLEF). Se trata de la fuente de datos más rica sobre trabajadores extranjeros no residentes que cruzan la frontera sur, su dimensión y perfiles ocupacionales. Los microdatos están disponibles bajo registro.

Encuesta Nacional de Ocupación y Empleo

La ENOE se lleva a cabo de forma trimestral compilando información general de los residentes habituales de la vivienda e información reproductiva, conyugal y económica de las personas de 12 años y más de edad. Debido a modificaciones en la legislación mexicana, desde el cuarto trimestre del 2014 los indicadores laborales calculados y difundidos de manera oficial se refieren a la población de 15 años y más, al ser esta la edad legal para trabajar.

9 El refugio es un derecho humano que el gobierno mexicano reconoce a través de un procedimiento alineado a estándares internacionales y que se otorga a una persona en el territorio nacional con temores fundados de persecución por motivos de raza, religión, nacionalidad, género, opinión política o violencia generalizada, entre otros. El asilo es una facultad discrecional del Estado y un acto de política exterior para proteger a un individuo, situado en o fuera del país, que es perseguido por razones políticas y cuya vida, libertad o seguridad se encuentra en peligro (SRE, 2016).

Cuadro 1

Clasificación de la población extranjera en México

Categoría	Periodo permitido
Visitante sin permiso de trabajo	180 días
Visitante con permiso de trabajo	180 días
Visitante regional (sin permiso de trabajo)	Siete días
Trabajador fronterizo	Un año
Visitante por razones humanitarias	Durante el proceso legal
Visitante con fines de adopción	Durante el proceso legal
Residente temporal	Cuatro años
Residente temporal estudiante	Hasta obtener el grado
Residente permanente	Sin límite

Fuente: elaboración propia con base en el artículo 52 de la *Ley de Migración*.

El cuestionario ampliado de la ENOE (que se aplica solo el primer trimestre de cada año) incluye preguntas útiles para el estudio de la movilidad laboral: "Para conseguir o conservar este trabajo, ¿…se vio obligado a cambiar de ciudad o localidad?" y "Antes de este cambio, ¿en qué estado de la República o país vivía?". No obstante, ya que la muestra es relativamente pequeña (120 260 viviendas) y no está enfocada en la migración ni en la movilidad laboral, el número de personas que en el 2017 reportó haber cambiado de lugar de residencia desde algún otro país por razones laborales es muy pequeño ($n = 56$),[10] por lo que la información sobre condiciones de ocupación obtenida para esta subpoblación podría estar seriamente sesgada. Solo para ilustrar, las regiones de donde provienen los informantes de la ENOE con experiencia de movilidad laboral internacional hacia México se listan en el cuadro 2.

Con este número de personas entrevistadas no es posible producir estimaciones confiables sobre las características económicas y sociodemográficas del grupo, así que, para ser usada como fuente de información sobre MLI, el diseño de la ENOE tendría que ajustarse. Esta es la fuente oficial de los indicadores de ocupación y empleo en México por lo que, de considerarse prioritario el tema de la MLI, el diseño tendría que asegurar una sobre-

Cuadro 2

Número de entrevistados en la ENOE 2017, trimestre I, que llegaron a México desde el extranjero por motivos laborales según lugar de origen

Región/país	
EE. UU.	27
América, excepto Guatemala	13
Asia	10
Europa	3
Guatemala	3
Total	**56**

Fuente: elaboración propia con datos de: INEGI. *Encuesta Nacional de Ocupación y Empleo 2017.*

10 Cálculos de la autora basados en microdatos de la ENOE 2017 (trimestre I).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

muestra de las personas que cambiaron su lugar de residencia desde el exterior del país por motivos laborales.

Encuesta Nacional de Ingresos y Gastos de los Hogares

La ENIGH es un proyecto regular del INEGI que se levanta cada dos años. La población objetivo son los hogares de residentes nacionales y extranjeros que usualmente habitan viviendas particulares localizadas en territorio nacional. Los residentes en México que trabajan en el extranjero pueden identificarse, pero cualquier esfuerzo por caracterizarlos en términos de atributos ocupacionales debe hacerse con precaución, ya que representan una parte muy pequeña de la población y la movilidad laboral no es el foco del proyecto.

Integrando la información sobre MLI en México

Volumen de trabajadores migrantes internacionales

En la EI se identifica a los migrantes internacionales a partir de la pregunta: "¿En qué estado de la República Mexicana o en qué país nació (NOMBRE)?". Enseguida, el cuestionario distingue a los nacidos en EE. UU. de los que lo hicieron en otro país; esta estructura es consistente con la de la ENADID. De forma adicional, la EI recoge información para estimar el porcentaje de migrantes internacionales naturalizados o nacidos de padre o madre mexicana a partir de la pregunta: "¿(NOMBRE) tiene nacionalidad mexicana?".

Los migrantes internacionales de 12 años y más de edad que reportaron haber trabajado[11] la semana previa (preguntas en la EI y la ENADID) integran el volumen de trabajadores migrantes internacionales. La EI también permite saber si ellos trabajaron en México o en el exterior (en particular en EE. UU.).

En ambas encuestas se pregunta sobre el lugar de residencia cinco años antes. En la ENADID, además, se registra el del año anterior. A pesar de que no hay certeza acerca de los desplazamientos que pudieron haber ocurrido entre las fechas consideradas, los volúmenes de migrantes internacionales y trabajadores migrantes internacionales pueden ser clasificados según la [supuesta] duración de su estancia.

Otra cifra de trabajadores migrantes internacionales pudiera estimarse con una estrategia alternativa a partir de la pregunta de la ENADID sobre las causas de la migración de los individuos que vivían fuera del país uno y cinco años antes: aquellos cuya razón principal de desplazamiento fue conseguir trabajo o cambiar de empleo podrían ser considerados el *stock* básico de trabajadores migrantes internacionales. Evidentemente, la estimación vía preguntas específicas sobre la participación laboral es mejor, ya que toma en cuenta el volumen de personas que, en efecto, están incorporadas en actividades económicas, sin considerar sus razones originales para migrar. De hecho, es de esperar que la estimación vía preguntas específicas sobre la participación laboral sea mayor que la resultante de las preguntas sobre las causas de la migración.

Trabajadores extranjeros no residentes

Para estimar los indicadores sobre ellos, tenemos dos fuentes que se complementan: los registros del INM y la EMIF-Sur. De los primeros se puede obtener el número de tarjetas de trabajador fronterizo expedidas a personas provenientes de Belice y Guatemala y sus familiares, según sexo y grupos de edad del titular; y de la EMIF-Sur, una estimación que incluye y distingue la movilidad laboral no autorizada de personas, pues pregunta a quienes regresan a Guatemala si trabajaron durante su estancia en México y el tipo de documentación con

11 O haber llevado a cabo actividades de mercado, con o sin remuneración, o haber estado de vacaciones, incapacidad o licencia.

la que cuentan. La información de la EMIF-Sur enriquece la caracterización del fenómeno, pues se puede clasificar por duración de la estancia, rasgos básicos de la fuerza de trabajo y categorías de ocupación e industria.

Flujo de personas hacia el interior del país

La ENADID es útil para describir a los extranjeros que vivían en EE. UU. u otro país el año anterior al levantamiento. Debido a que estos individuos no habían alcanzado aún el umbral de 12 meses para ser considerados residentes habituales en México, dan cuenta del flujo de inmigrantes en un año particular, con posibilidad de clasificación según las causas de migración. Con los datos de la ENADID se estima, por ejemplo, que alrededor de 1 600 extranjeros tenían menos de un año en el país y se habrían desplazado debido a la violencia o inseguridad pública en sus lugares de origen; tres de cada cuatro serían menores de 12 años de edad.[12]

No obstante, la fuente adecuada para la medición anual de estos flujos son los registros del INM a partir de los cuales se publica el *Boletín Mensual de Estadísticas Migratorias*, que informa sobre los tipos de permiso emitidos a los extranjeros. En el 2016, por ejemplo, más de 100 mil personas acreditaron su estancia en México bajo la condición de visitantes (regionales, trabajadores fronterizos o por razones humanitarias).

La información de los boletines estadísticos permite el análisis de la distribución por países o regiones, así como por categoría de inmigración (familia, trabajo, rentista, estudiante u otra), además del seguimiento de la evolución de esos flujos, mes con mes y año con año.

En particular, la cifra aproximada de refugiados y asilados puede obtenerse a partir del número de tarjetas de residente permanente por reconocimiento de la Comisión Mexicana de Ayuda a Refugiados (COMAR) y del de tarjetas de visitante expedidas por razones humanitarias[13] que, en el 2016, ascendieron a 1 702 y 3 971, respectivamente. Las personas en estas categorías migratorias tienen permiso de trabajar a cambio de una remuneración.

El número de refugiados en México se ha incrementado en los últimos años: la información del INM muestra un aumento en el número de tarjetas emitidas por razones humanitarias de 2 096 en el 2015 a 5 392 en el 2016. La COMAR, por su parte, había reportado 939 personas reconocidas con ese estatus en el 2015 más 152 receptoras de protección complementaria.[14]

Sociodemografía de los grupos con experiencia de MLI en México

Trabajadores migrantes internacionales en México

El volumen de residentes nacidos en el exterior es de aproximadamente un millón de personas (INEGI, 2015). Constituyen una pequeña parte del total de la población del país y más de 40% de ellos tiene, de hecho, nacionalidad mexicana; la mayoría, por descender de padres mexicanos.

Más de siete de cada 10 personas definidas como migrantes internacionales (al haber nacido en un país diferente en el que de forma habitual residen) nacieron en EE. UU., aunque en buena parte son hijos de mexicanos residentes en la frontera norte o de migrantes mexicanos retornados,[15] de manera que cualquier esfuerzo por describir este *stock* debe tomar eso en consideración, pues el estricto uso de la definición, sin las ano-

12 La cifra es ilustrativa, pero es muy probable que esté subestimada como consecuencia del diseño de la muestra que, en principio, está integrada por residentes en viviendas particulares dentro del territorio nacional.

13 Además de personas cuya vida o integridad esté en peligro por violencia, incluye a las víctimas de catástrofe natural y a solicitantes de ingreso al país para prestar asistencia a familiares en estado grave de salud o llevar a cabo acciones de auxilio o rescate en situaciones de emergencia en el país (INM, 2016).

14 Por su parte, el Alto Comisionado de las Naciones Unidas para los Refugiados (UNHCR, por sus siglas en inglés) estimó una cifra de 2 923 para el mismo año y 1 350 casos pendientes. A finales del 2017, la cifra había ascendido a más de 19 mil personas y 10 368 casos pendientes.

15 De los residentes en México nacidos en EE. UU., 80% tiene ascendencia mexicana —estimaciones de la Unidad de Política Migratoria (UPM) de la Secretaría de Gobernación, basadas en la EI—.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

taciones pertinentes al caso, podría conducir a un análisis equivocado de sus características; por ejemplo, la proporción de personas que no está en la fuerza de trabajo es más alta en el caso de los nacidos en el exterior con nacionalidad mexicana, en comparación con la población nacida en México (0.61 *vs*. 0.49), lo que con certeza se relaciona con las diferencias en la estructura por edad de las dos subpoblaciones.[16]

Buena parte del resto de la población nacida en el exterior se asentó en México durante el siglo XX. Llegaron como refugiados de España, Guatemala, Argentina, Chile, Uruguay, Perú, Colombia, Brasil y El Salvador en tiempos de dificultades políticas en aquellos países (Somohano & Yankelevich, 2011). Hoy en día, la tasa de aceptación de las solicitudes de refugio ha disminuido de forma considerable,[17] a pesar de que el flujo de personas no se ha detenido, en especial el de los provenientes de Centroamérica. Aunque en principio los migrantes de esta región podrían estar transitando hacia EE. UU., algunas investigaciones recientes muestran un patrón de creciente asentamiento en México.[18] La Unidad de Política Migratoria de la Secretaría de Gobernación (UPM-SEGOB) estimó un total de 301 300 centroamericanos transitando por México de forma ilegal[19] en el 2014.

La mayoría de los residentes nacidos en el exterior ha vivido en México por, al menos, cinco años, y los que están en edad de trabajar, en general, han estudiado mínimo hasta preparatoria. Aunque muchos de los que participan en el mercado laboral son profesionales, existe una relación entre el tipo de ocupación y el país de origen[20] vinculada a las desigualdades educativas, las capacidades productivas y los estereotipos nacionales: con frecuencia, los residentes nacidos en el exterior tienen mejores ocupaciones en comparación con los nacidos en México. Por otra parte, con los datos de la ENADID se puede observar que la tasa de empleo se incrementa para aquellos individuos nacidos en el exterior que permanecen más de un año en el país (pasa de 87 a 96%).

Los nacidos en México, en el extranjero con nacionalidad mexicana y los extranjeros se distribuyen de manera distinta en las diferentes actividades económicas. En los servicios de alojamiento y alimentación, así como en las actividades profesionales, científicas y técnicas, estos últimos tienen mayor presencia relativa. Los mexicanos nacidos en el exterior están relativamente más concentrados en las actividades inmobiliarias y de enseñanza, así como en las de atención a la salud y asistencia social. Por su parte, los nacidos en México están sobrerrepresentados en las actividades manufacturera y agropecuaria (ver gráfica 1).

En cuanto a las ocupaciones, hay una mayor proporción de artesanos y trabajadores agropecuarios, así como operadores de maquinaria y ensambladores entre los nacidos en México, mientras que los extranjeros y los mexicanos nacidos en el exterior ocupan relativamente más posiciones profesionales y técnicas, directivas y gerenciales (ver gráfica 2).

En relación con el flujo de trabajadores migrantes internacionales, solo en el 2016, alrededor de 90 mil extranjeros recibieron tarjetas de residentes temporales o permanentes, o visitantes por razones humanitarias,[21] categorías migratorias en las que se permite trabajar a cambio de una remuneración.

Trabajadores fronterizos en el sur

En el 2016, el INM emitió 15 130 tarjetas de trabajador fronterizo. La EMIF-Sur muestra que 92% de los trabajadores extranjeros no residentes provenien-

16 La diferencia en la estructura etaria de las dos subpoblaciones de nacidos en el extranjero es notable y reveladora: los nacidos en EE. UU. tienen una edad mediana de 11 años, mientras que quienes nacieron en el resto del mundo tienen medianas de 38 años en el caso de los hombres y 36 en el de las mujeres (estimaciones del Consejo Nacional de Población y UPM-SEGOB basadas en la EI).

17 Según información del UNHCR, la tasa de solicitudes aceptadas pasó de 70 a 50% entre el 2015 y el 2017.

18 Fernández y Rodríguez (2016).

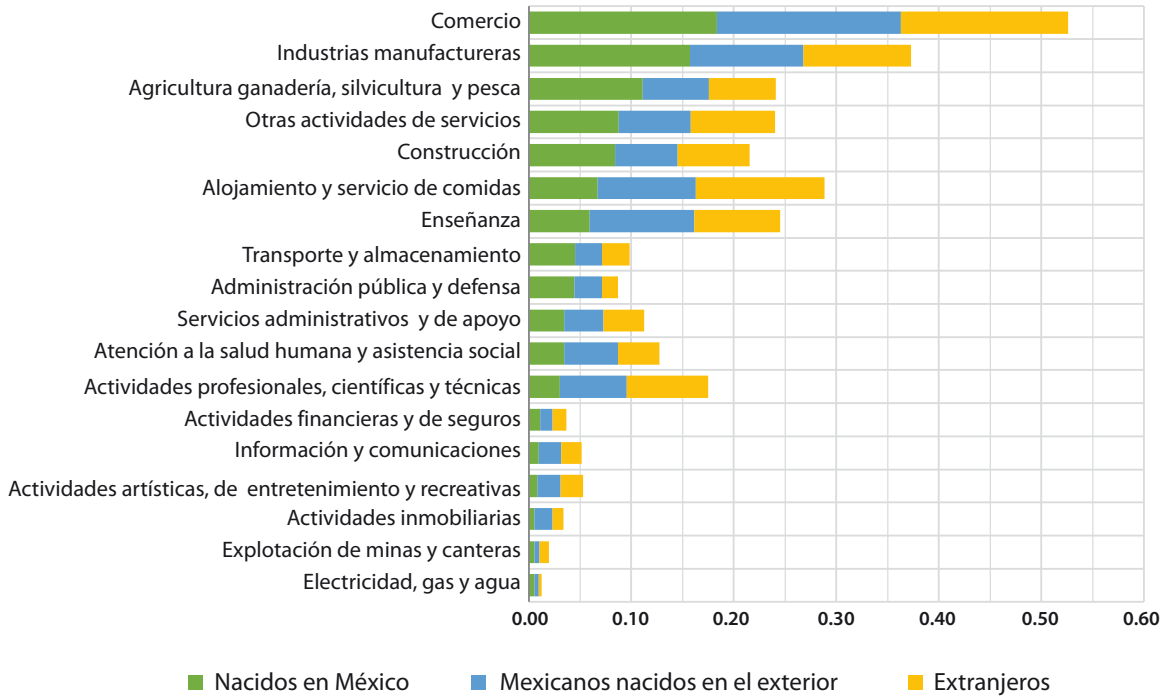
19 Es decir, sin cumplir con la regulación en materia migratoria.

20 Según las estimaciones del CONAPO basadas en datos de la EI.

21 La cifra corresponde a la suma de tarjetas de visitante por razones humanitarias, de residente temporal y de residente permanente emitidas ese año, menos las tarjetas de residente permanente por regularización.

Gráfica 1

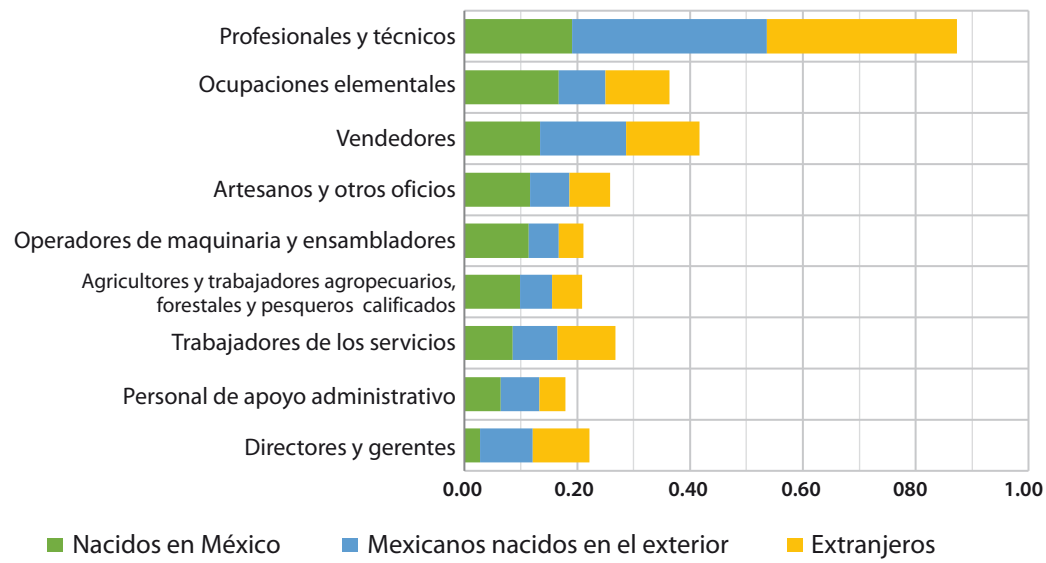
Distribución de personas empleadas por industria[a] según país de nacimiento y nacionalidad



[a] Según la *Clasificación Industrial Internacional Uniforme de Todas las Actividades Económicas* (*CIIU*).
Fuente: estimaciones de la autora basadas en: INEGI. *Encuesta Intercensal 2015*.

Gráfica 2

Distribución de personas empleadas por ocupación[a] según país de nacimiento y nacionalidad



[a] Según la *Clasificación Internacional Uniforme de Ocupaciones (CIUO)*.
Fuente: estimaciones de la autora basadas en: INEGI. *Encuesta Intercensal 2015*.

tes del sur son hombres con una edad mediana de 33 años. Usualmente, tienen nivel educativo básico o ninguna educación formal y 51% habla una lengua indígena. Los principales destinos en la actualidad son Frontera Comalapa y Tapachula, Chiapas. Su distribución por duración de estancia se presenta en la gráfica 3.

Uno de cada cuatro ingresa a México con una tarjeta de visitante regional que les permite permanecer un máximo de siete días, como se indica en el cuadro 1, pero al menos 18% de ellos excede ese tiempo. Por otra parte, más de 20% cruza sin ninguna documentación.[22]

Suelen trabajar en la agricultura, silvicultura y pesca; algunos son acompañados por sus familias. Entre los trabajadores extranjeros residentes que permanecen máximo 24 horas, muchos se emplean en ocupaciones elementales, como: vendedores ambulantes de servicios y afines, peones de la minería, limpiadores, asistentes, o en la construcción, la industria manufacturera y el transporte.

Solo uno de cada 10 obtiene un ingreso mayor a 2 salarios mínimos en México. Los hogares con ese nivel de ingreso se clasifican en los dos primeros deciles de la distribución nacional.[23]

Movilidad laboral internacional desde México

Residentes trabajando en el exterior

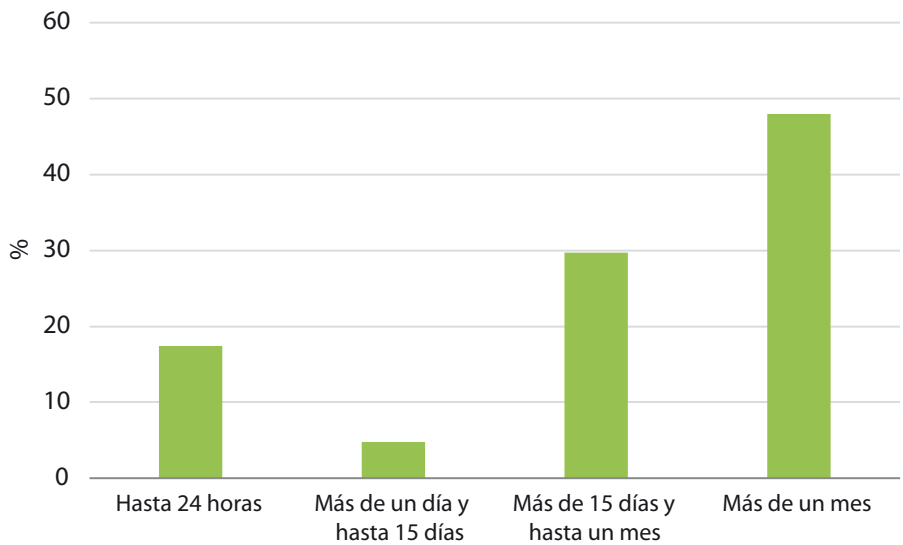
Las personas que habitualmente residen en México y trabajan en el exterior pueden identificarse con la pregunta de la Encuesta Intercensal: "¿En qué estado o país [está el negocio, empresa o lugar donde trabajó (NOMBRE) la semana pasada]?". La estimación corresponde al volumen o proporción dados en cierta semana del año, por lo que puede estar sesgada por efectos estacionales. Es importante anotar, también, que los residentes de nuestro país trabajando en el exterior pueden distinguirse en: nacidos en México; nacidos en el extranjero, pero con nacionalidad mexicana; y residentes extranjeros.

22 Según el reporte de *Indicadores anuales 2017* (COLEF, 2019).

23 Estimaciones de la autora con base en la ENIGH 2014.

Gráfica 3

Distribución por duración de estancia en México de los trabajadores fronterizos provenientes del sur



Fuente: estimaciones del CONAPO y la UPM-SEGOB basadas en la EMIF-Sur 2015.

Asimismo, la ENOE ofrece alguna información relevante sobre las características demográficas y ocupacionales de los residentes en México trabajando en el exterior, quienes pueden ser identificados con la pregunta: "¿Cuál es el nombre de la empresa, negocio o institución para la que trabaja o ayuda?". Las opciones de respuesta son: 1) anotar el nombre, 2) reportar que el negocio no tiene nombre, 3) declarar ser empleado de una unidad doméstica o trabajador de otro trabajador y 4) ser un trabajador en el extranjero. De esta y otras preguntas relacionadas se puede saber si estas personas son empleados o independientes, si tienen contrato y de qué tipo, si reciben o no prestaciones sociales, entre otras, así que es factible un estudio más profundo de los residentes trabajando en el exterior. La robustez de las estimaciones de la ENOE acerca de los residentes trabajando en el exterior se basa en el hecho de que el sesgo estacional es minimizado al tratarse de una encuesta trimestral llevada a cabo sobre un panel rotatorio, lo cual significa que una vivienda particular es entrevistada durante cinco trimestres consecutivos y 80% de la muestra es la misma de un trimestre al otro.

Una fuente complementaria para identificar residentes habituales que trabajan en el exterior es la ENIGH. Con el cuestionario para personas de 12 años y más de edad se puede detectar a quienes trabajaron en otros países, y es posible caracterizarlos según hayan sido empleados o trabajadores independientes, con o sin un contrato y de qué tipo, con prestaciones sociales o sin ellas, según el número de horas trabajadas, así como el tipo y tamaño de la organización para la que trabajaron. En la ENIGH, las preguntas sobre la fuerza de trabajo se refieren al mes previo, así que las estimaciones acerca de los residentes trabajando en el exterior podrían ser más robustas que las resultantes de la EI en relación con el sesgo estacional. No obstante, la muestra es más pequeña y el foco no es ni la migración ni la movilidad laboral, así que las estimaciones podrían, de hecho, ser menos precisas que las de la EI.

En suma, menos de 1% de la población ocupada nacida en México trabajó fuera del país en la semana de referencia en comparación con 14% de los nacidos en el exterior (aquí incluidos los que tienen nacionalidad mexicana). El porcentaje fue, incluso, mayor en el caso de los extranjeros: (16).[24]

Los hogares con, al menos, una persona trabajando fuera del país obtienen un ingreso promedio 50% mayor al de aquellos en los que nadie trabaja fuera de la nación (99.4% del total). Aunque la proporción de hogares con residentes trabajando en el exterior es pequeña, estos obtienen 1% del ingreso total de todos los hogares mexicanos. Este grupo de población tiene en su mayoría un empleo subordinado sin contrato ni prestaciones sociales: en 2014, 74% de ellos trabajaron de 30 a 60 horas a la semana en establecimientos de menos de 20 empleados.[25]

Trabajadores fronterizos en el norte

El porcentaje de la población residente en Baja California, Sonora, Chihuahua, Coahuila de Zaragoza, Nuevo León y Tamaulipas que va a EE. UU. para trabajar o estudiar es de alrededor de 1%; de ellos, un tercio reside en Tijuana, Baja California. La mayoría de los trabajadores fronterizos son hombres, pero también hay un número significativo de mujeres (80 710 y 43 918 con edades medianas de 33 y 24 años, respectivamente);[26] tres de cada cuatro hombres y seis de cada 10 mujeres se desplazan por razones de trabajo (UPM-CONAPO, 2016).

Emigrantes

La población mexicana tiene una larga e intensa historia de migración hacia los EE. UU.[27] y, en diferente escala, hacia otros países desarrollados en búsqueda de mejorar la calidad de su vida. Las es-

24 Estimaciones de la autora con base en la EI.

25 Estimaciones de la autora basadas en microdatos de la ENIGH 2014. Un perfil semejante para los residentes trabajando en el exterior se puede delinear con información de la ENOE.

26 Estimaciones del CONAPO y UPM-SEGOB basadas en la EI.

27 Entre el 2009 y 2014, 87% de los emigrantes mexicanos se desplazaron hacia EE. UU. (estimaciones del CONAPO y UPM-SEGOB basadas en la ENADID 2014).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

trategias seguidas para conseguirlo van desde matricularse en programas educativos, hasta migrar para conseguir mejores empleos, con o sin autorización legal.

Con datos de la *Current Population Survey (CPS)* y de la *American Community Survey (ACS)*, ambas de la Oficina del Censo de los EE. UU., se estima que el volumen de los residentes nacidos en México ascendía a 12 millones de personas en el 2015, de las cuales 70% eran económicamente activas.[28]

No obstante, desde finales del siglo pasado, la migración al país del norte ha descendido de forma gradual, tanto en términos de volumen como en porcentaje de la población mexicana (UPM-CONAPO, 2016). Para algunos analistas, esto se debe en parte a la aplicación más severa de la regulación migratoria que ha forzado a la gente (en particular, a los migrantes no autorizados) a permanecer más tiempo en EE. UU.[29]

Los datos de la ENADID muestran que 78% de los migrantes a ese país son hombres con edad mediana de 30 años, en tanto que la de las mujeres es de 28. La mitad de los varones entraron sin autorización, en comparación con solo una cuarta parte de las personas de sexo femenino. En el caso de las mujeres, reunirse con la familia y buscar trabajo fueron las causas más mencionadas como razones del desplazamiento.

La mayoría de los trabajadores mexicanos en EE. UU. se emplean de tiempo completo en el sector terciario de la economía, en establecimientos con menos de 100 empleados. En el 2014 obtenían un salario de 19 229 dólares anuales que, al tipo de cambio vigente, equivalía al ingreso de un hogar mexicano del decil más alto de la distribución.[30]

Migrantes en tránsito irregular

Su flujo por el país parece haber aumentado desde el 2011. En el 2014 se estimó una cifra de 389 600 personas en esta condición en México, 88% de ellas provenientes de Centroamérica.[31] En el 2015, los datos de la EMIF-Sur permitieron estimar un total de 86 700 devoluciones de centroamericanos que iban rumbo a EE. UU. por parte de las autoridades mexicanas, al igual que 42 200 llevadas a cabo por las estadounidenses después de haber estado en aquel país por un mes o menos. Se trata, con frecuencia, de hombres en el rango de 24-26 años de edad. La mitad de los devueltos por las autoridades mexicanas fueron detenidos en Chiapas o Veracruz de Ignacio de la Llave y 67% de los que consiguieron llegar a EE. UU. había cruzado por Reynosa, Tamaulipas.[32]

Conclusiones

Estudiar la MLI como un asunto en la intersección de la migración internacional y el mercado de trabajo global, alineando conceptos y definiciones de los ámbitos poblacional, laboral y comercial, en el marco de la legislación nacional, significó enfrentar el primer reto de compilar información de fuentes variadas con objetivos diferentes para construir un relato del estado y el curso del fenómeno en México.

El trabajo de revisión exhaustiva de las fuentes disponibles para su descripción fue relevante y útil para comprender la diversidad de información que se produce, la dispersión de la misma e, incluso, los límites de la correspondencia conceptual y metodológica con los criterios legales y las estadísticas producidas.

En el caso mexicano, debido a su volumen y patrón histórico, el estudio de la migración de com-

28 Estimaciones del CONAPO basadas en datos del *Integrated Public Use Microdata Series (IPUMS)*.

29 Passel, Cohn y Gonzalez-Barrera reportan que la propensión a vivir en EE. UU. por al menos un año fue más alta para los migrantes devueltos en el 2010 que para los aprehendidos cinco o 10 años antes.

30 Estimaciones de la autora con base en la ENIGH 2014.

31 Estimaciones de UPM-SEGOB basadas en sus boletines estadísticos, información del Departamento de Seguridad Nacional, el *Anuario estadístico del Servicio de Inmigración y Naturalización*, de la Patrulla Fronteriza y de la Oficina del Censo (*American Community Survey*), todos de los EE. UU.

32 Estimaciones del CONAPO y UPM-SEGOB basadas en la EMIF-Sur 2015.

patriotas a Estados Unidos de América (en una variedad de aristas) domina la literatura de lo que hemos identificado en este artículo con el nombre de *movilidad laboral internacional*. La medición de la migración interna e internacional ha sido, por ello, el foco de la recolección de información y de los esfuerzos de política pública. Aun así, otras formas de movilidad están ganando visibilidad y generando interés debido a su impacto económico; por ejemplo, ha habido un creciente interés por estudiar la movilidad laboral en México en su dimensión interna, lo que ha derivado en la compilación e integración de datos de fuentes diversas (tradicionales o no) para describir con suficiencia a la población que se mueve de un municipio a otro, en ocasiones, cruzando límites estatales.

De igual forma, también se comprende la relevancia de estudiar los movimientos por razones laborales en las fronteras con EE. UU. y Guatemala; sin embargo, poca atención se ha dado a la movilidad de personas residiendo en México, no necesariamente en las fronteras, y ocupándose en el exterior.

Un desafío interesante en esta investigación fue el de mover el reflector de la migración dirigida al país del norte (que tiene, sin duda, retos pendientes en cuanto a las estrategias de medición) para enfocarse en ciertos grupos demográficos que no representan una parte importante de la población de un país en términos de efectivos, pero que merecen atención debido, precisamente, a su especificidad. Quienes se desplazan por motivos laborales sin cambiar su residencia habitual son, en realidad, pocos entre los millones de habitantes en México, pero sus características y las de sus hogares son particulares y significativas.

Por otra parte, el estudio de otros grupos muy pequeños —difíciles de detectar en las encuestas, como el de los migrantes no autorizados en tránsito por el país— tiene una importancia enorme para la comprensión de asuntos de suma relevancia en materia de derechos humanos. Para esta dimensión de la movilidad internacional, tenemos aún grandes limitaciones en la generación de estadísticas fidedignas, a pesar del notable esfuerzo que se hace a través de las encuestas en las fronteras norte y sur del país. Aquí, el esfuerzo de integración de información de fuentes variadas debe ser mayúsculo, pasando por la recolección de registros en albergues a lo largo del país, hasta el contraste con las cifras estimadas de migrantes no autorizados en EE. UU., principalmente.

De los proyectos del INEGI revisados en el marco de esta investigación se obtiene información para describir a la población que reside en México, pero que nació en otro país y forma parte de la fuerza laboral; también, es posible una estimación de la proporción de residentes en México que trabaja fuera del país, así como delinear algunas de sus características. Además, en algunos de estos proyectos se recogen datos para estimar el volumen de emigrantes en hogares dentro del territorio nacional. Acerca de la población no residente en México que ingresa al territorio nacional por trabajo, los registros del INM y la EMIF-Sur de México son las principales fuentes de datos. Como resultado del análisis de estas, se enuncia enseguida una serie de recomendaciones para mejorar tanto la captación estadística como la disponibilidad de los datos.

La información sobre entradas y salidas del territorio nacional es indispensable para entender la dinámica de la movilidad de los trabajadores extranjeros no residentes y de los residentes trabajando en el exterior, pero es importante desarrollar los instrumentos que permitan distinguir la tipología de la MLI y relacionarla con los patrones de movilidad de cada caso (de los trabajadores fronterizos o estacionales, por ejemplo).

Por otra parte, a los instrumentos de recolección podrían añadírseles preguntas que informen sobre aspectos muy relevantes para un estudio integral del tema. En el caso de las encuestas en la frontera podría preguntarse acerca de la frecuencia o periodicidad del cruce de frontera por motivos laborales o, en otro tenor, sobre la exposición a factores de riesgo para la salud física o mental.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

En el caso de los censos y las encuestas en hogares, es imprescindible la inclusión de la pregunta sobre nacionalidad, además del lugar de nacimiento y, también, es indispensable preguntar acerca del año de llegada al país. De otra forma, cualquier análisis que considere la duración de la estancia de los migrantes internacionales sigue descansando sobre la extremadamente débil hipótesis de nula movilidad entre las fechas referidas en los proyectos estadísticos (un año atrás/cinco años atrás).

Si fuera posible el levantamiento de un módulo en los hogares para la medición de la MLI, podrían incluirse en él preguntas retrospectivas sobre el empleo en el exterior. [33] En caso contrario, al menos debería considerarse el ajuste del diseño estadístico de la ENOE para asegurar un suficiente número de trabajadores extranjeros no residentes y de residentes habituales trabajando en el exterior. En las encuestas en establecimientos levantadas en México no se tiene la experiencia de incluir alguna pregunta que pueda abonar al estudio de la MLI. Indagar en los Censos Económicos sobre la nacionalidad del personal empleado sería de enorme impacto analítico.

Es preciso seguir avanzando en la integración de información de fuentes distintas, como registros administrativos, censos, encuestas en hogares y en establecimientos, aprovechando el potencial de cada una de ellas. Asimismo, el intercambio con otros países es de gran importancia para llenar los vacíos de información o verificar la consistencia de las estimaciones.

33 Si bien el INEGI levantó la Encuesta Demográfica Retrospectiva (EDER) en 1998, 2011 y 2017, incluyendo, entre otras, la historia migratoria de residentes en México, los eventos registrados tienen duración mínima de un año, así que la movilidad sin cambio de residencia no es captada.

Fuentes

Cámara de Diputados del H. Congreso de la Unión. "Ley de Migración", en: *Diario Oficial de la Federación*. México, 2017 (DE) http://www.ordenjuridico.gob.mx/Documentos/Federal/html/wo83139.html

Courgeau, D. *Méthodes de mesure de la mobilité spatiale*. Paris, INED, 1988.

El Colegio de la Frontera Norte (COLEF), Secretaría del Trabajo y Previsión Social, Consejo Nacional de Población, Unidad de Política Migratoria, Secretaría de Relaciones Exteriores y Consejo Nacional para Prevenir la Discriminación. *Encuestas sobre migración en las fronteras norte y sur de México*. México, COLEF, 2019 (DE) https://www.colef.mx/emif/

Fernández, C. y M. Rodríguez. "Hondureños migrantes en México: del tránsito al asentamiento", en: *CANAMID Policy Brief Series*. PB011. Guadalajara, CIESAS, 2016 (DE) http://www.canamid.org/publication?id=PB011

Instituto Nacional de Estadística y Geografía (INEGI). *Censo de Población y Vivienda 2010*. México, INEGI, 2010 (DE) https://www.inegi.org.mx/programas/ccpv/2010/default.html

_____ *Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad*. México, INEGI, 2018 (DE) http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/enoe/default.html

_____ *Encuesta Nacional de Ingresos y Gastos de los Hogares 2016*. Nueva serie. México, INEGI, 2016 (DE) http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/enigh/nc/2016/default.html

_____ *Encuesta Intercensal 2015*. México, INEGI, 2015 (DE) http://www.beta.inegi.org.mx/proyectos/enchogares/especiales/intercensal/default.html

_____ *Encuesta Nacional de la Dinámica Demográfica 2014*. México, INEGI, 2014 (DE) http://www.beta.inegi.org.mx/proyectos/enchogares/especiales/enadid/2014/default.html

Instituto Nacional de Migración (INM). *Visa por razones humanitarias*. México, INM, 2016 (DE) https://www.gob.mx/inm/documentos/preguntas-frecuentes-para-solicitar-visa-por-razones-humanitarias

International Labour Office. *Resolución sobre la actualización de la Clasificación Internacional Uniforme de Ocupaciones*. Geneva, ILO, 2007 (DE) https://www.ilo.org/public/spanish/bureau/stat/isco/docs/resol08.pdf

_____ *The Thirteenth International Conference of Labour Statisticians*. Geneva, ILO, 1982 (DE) https://www.ilo.org/public/libdoc/ilo/1982/82B09_651_engl.pdf

Meza, L. "Visitantes y residentes. Trabajadores guatemaltecos, salvadoreños y hondureños en México", en: *CANAMID Policy Brief Series*. PB04. Guadalajara, CIESAS, 2015 (DE) http://www.canamid.org/publication?id=PB04

Nájera, J. "Formas de movilidad laboral transfronteriza de las y los guatemaltecos a Chiapas, una visión desde la familia", en: *Revista Latinoamericana de Estudios de la Familia*. Vol. 3, 2011, pp. 177-198 (DE) https://www.colef.mx/emif/resultados/articulos/2011%20-%20

Formas%20de%20movilidad%20laboral%20transfronteriza%20de%20las%20y%20los%20guatemaltecos%20a%20Chiapas,%20una%20vision%20desde%20la%20familia.pdf

Passel, J., D'V. Cohn y A. Gonzalez-Barrera. *Net Migration from Mexico Falls to zero–and Perhaps Less*. Washington, D.C., Pew Hispanic Center, 2012 (DE) http://www.pewhispanic.org/2012/04/23/net-migration-from-mexico-falls-to-zero-and-perhaps-less/

Rodríguez-Chávez, E. y S. Cobo. *Extranjeros residentes en México. Una aproximación cuantitativa con base en los registros administrativos del INM*. México, Centro de Estudios Migratorios, INM, SEGOB, 2012 (DE) http://www.politicamigratoria.gob.mx/work/models/SEGOB/CEM/PDF/Estadisticas/Poblacion_Extranjera/ExtranjerosResMex.pdf

Secretaría de Relaciones Exteriores (SRE). *Lineamientos para atender solicitudes de asilo y refugio*. México, SRE, 2016 (DE) https://extranet.sre.gob.mx/images/stories/asilo/asilo2016.pdf

Somohano, K. y P. Yankelevich. *El refugio en México. Entre la historia y los desafíos contemporáneos*. México, COMAR-SEGOB, 2011.

Statistics Austria. *Labour mobility*. 2. UNECE Working Paper Series on Statistics. Geneva, 2017 (DE) https://www.unece.org/fileadmin/DAM/stats/publications/2017/Issue2_Labour.pdf

UNECE. *Measuring International Labour Mobility*. Geneva, United Nations, 2018 (DE) https://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECSSTAT20187_WEB.pdf

Unidad de Política Migratoria-Secretaría de Gobernación (UPM-SEGOB). *Boletines Estadísticos*. México, UPM-SEGOB, 2018 (DE) http://www.politicamigratoria.gob.mx/es_mx/SEGOB/Boletines_Estadisticos

Unidad de Política Migratoria y Consejo Nacional de Población. *Prontuario sobre movilidad y migración internacional. Dimensiones del fenómeno en México*. México, SEGOB, 2016 (DE) http://www.politicamigratoria.gob.mx/es_mx/SEGOB/Prontuario

United Nations. *Clasificación Industrial Internacional Uniforme de Todas las Actividades Económicas (CIIU). Revisión 4*. New York, United Nations, 2009 (DE) https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev4s.pdf

_____ *International Recommendations for Tourism Statistics (IRTS) 2008*. New York, United Nations, 2010 (DE) https://unstats.un.org/unsd/publication/seriesm/seriesm_83rev1e.pdf

_____ *Recommendations on Statistics of International Migration (RSIM). Revision 1*. New York, United Nations, 1998 (DE) https://unstats.un.org/unsd/publication/SeriesM/SeriesM_58rev1e.pdf

United Nations, International Monetary Fund, Organisation for Economic Co-operation and Development, Statistical Office of the European Union, United Nations Conference on Trade and Development, World Tourism Organization, World Trade Organization. *Manual on Statistics of International Trade in Services 2010 (MSITS 2010)*. New York, United Nations, 2011 (DE) https://unstats.un.org/unsd/publication/seriesm/seriesm_86rev1e.pdf

United Nations High Commissioner for Refugees (UNHCR). *Statistical Yearbook 2015*. Geneva, UNHCR, 2017 (DE) https://www.unhcr.org/statistics/country/59b294387/unhcr-statistical-yearbook-2015-15th-edition.html#_ga=2.168125732.462805140.1550871504-1954695127.1550871504

_____ *Population Statistics*. Geneva, UNHCR, 2019 (DE) http://popstats.unhcr.org/en/overview#_ga=2.125782927.462805140.1550871504-1954695127.1550871504

Actualización de la matriz total de **insumo-producto de México del 2003.** Aplicación de los métodos de doble deflación y RAS

An Update of the Mexican **Input-Output Table of 2003.** An Application of the RAS and the Double Deflation Methods

Brenda Murillo-Villanueva, Martín Puchet Anyul y Gerardo Fujii-Gambero*

El análisis de insumo-producto muestra el aparato productivo de un país en un año determinado. Cuando las matrices de insumo-producto (MIP) de diferentes años están valuadas a los mismos precios es posible analizar la evolución de la estructura productiva. Para el caso de México, las MIP más recientes (2008 y 2012) están valuadas a precios del 2008; sin embargo, la del 2003 está a los precios de ese año. En este trabajo se estima la MIP del 2003 a precios del 2008 mediante dos métodos: el de la doble deflación y el RAS. Las estimaciones obtenidas se comparan según criterios de distancia entre matrices. Los resultados sugieren que la del 2003 a precios del 2008 estimada con el método RAS es mejor, ya que utiliza datos actualizados por el Instituto Nacional de Estadística y Geografía y mantiene el sentido económico.

Palabras clave: método de doble deflación; método RAS; criterio de distancia rectilínea.

The input-output analysis shows the productive structure of a country in a given year. When several input-output tables of different years are measured at the same prices, it becomes possible to analyze the evolution of the productive structure. In the Mexican case, the most recent tables (2008 and 2012) are measured at 2008 basic prices; however, the table of 2003 is measured at basic prices of 2003. Therefore, this paper estimates the input-output table of 2003 at 2008 basic prices with the RAS and the double deflation methods. The estimations obtained are compared with distance criteria between tables. The results suggest that the estimation with the RAS method of the input-output table of 2003 at prices of 2008 is better since it uses data from INEGI and keeps up the economic sense.

Key words: double deflation method; RAS method; distance criteria.

Recibido: 14 de noviembre de 2018.
Aceptado: 7 de marzo de 2019.

Nota: agradecemos los comentarios de dos dictaminadores anónimos que contribuyeron a mejorar el documento.

* Profesora-investigadora de la Facultad de Economía de la UAEM, bmurillov@uaemex.mx; profesor titular de la Facultad de Economía de la UNAM, anyul@unam.mx; y profesor titular de la Facultad de Economía de la UNAM, fujii@unam.mx; respectivamente.

The fruit seller or Market stall, 1951 painting by Olga Costa (1913-1993)/ DEA / G. DAGLI ORTI/Getty Images

1. Introducción

La matriz de insumo-producto (MIP) es un instrumento que tiene como objetivo principal examinar la interdependencia de las ramas de actividad económica que forman el aparato productivo de un país en un determinado año, el cual fue planteado y desarrollado por el economista W. Leontief entre 1928 y 1941; sin embargo, tiene sus raíces en trabajos de diversos autores, de entre los que destacan tres (ver Kurz, H. D. y N. Salvadori, 2000): en orden cronológico, tenemos que la *Tableau Economique* del fisiócrata François Quesnay, publicada en 1758, representó el primer intento por esquematizar el funcionamiento de un sistema económico integrado por tres tipos de agentes económicos a través de los cuales fluye la producción; este instrumento fue utilizado tiempo después por Marx para representar su estructura de *reproducción simple* en el capitalismo; y, por último, el economista y estadista von Bortkiewicz, quien fuera profesor de W. Leontief, utilizó el esquema de reproducción de Marx para resolver el problema de la transformación (ver Baumol, W. J. y T. Ten Raa, 2009).

Los aportes de estos tres economistas representaron las bases para que años más tarde Leontief construyera las primeras MIP: estuvo motivado por el análisis de la producción a partir de la consideración de que la economía es un sistema que transforma recursos en bienes y servicios finales; le interesaba distinguir las diferentes etapas del proceso productivo y resaltar la idea de que todos los productos, además de ser usados para consumo final, también son utilizados como insumos (ver Leontief, 1936); finalmente, en 1941, publicó las primeras MIP en el libro *Structure of the American Economy*. La matriz de insumo-producto se deriva del aparato conceptual-metodológico del sistema de cuentas nacionales y respeta la noción del equilibrio walrasiano.

El análisis basado en las MIP permite conocer la estructura productiva de un país en un determinado año, evidenciando las relaciones entre las ramas del aparato productivo. Muestra el producto de cada rama que es necesario para satisfacer la demanda intermedia y la final, así como los requerimientos de cada sector de insumos intermedios y primarios para la producción individual. La información detallada contenida en las matrices de insumo-producto permite realizar diversos ejercicios para probar el efecto que variaciones en las variables del modelo, o exógenas a este, tienen sobre la estructura productiva.

En ese sentido, el análisis de insumo-producto resulta fundamental para estudiar la evolución de la estructura productiva de un país en el tiempo. Sin embargo, para ello, se requiere lo siguiente: a) que los flujos que registran los intercambios entre ramas económicas se descompongan en la cantidad intercambiada y su precio o b) que los valores de los intercambios estén valuados a los mismos precios para cada año analizado. De no ser así, la evolución de una variable se explicaría, por un lado, por el cambio en los determinantes de la cantidad misma y, por el otro, debido al cambio en el precio, sin poder discriminar las respectivas magnitudes.

Por ello, lo adecuado en el análisis comparativo de MIP es utilizar información valuada al mismo nivel de precios, toda vez que la matriz de coeficientes (MC) a precios corrientes presenta las estructuras de costos de cada sector de la economía; cada componente del respectivo vector columna muestra cuál fue el gasto en cada insumo que se hizo para producir el valor bruto de la producción (VBP) de una mercancía sea esta un bien o un servicio no factorial. La MC a precios constantes, al eliminar el efecto de las variaciones tanto de los precios de los insumos como del precio del producto, acerca estos coeficientes en volúmenes demandados y ofrecidos a los coeficientes técnicos en cantidades físicas que están en el origen del análisis de insumo-producto.

Cuando se requiere estudiar el efecto de la técnica de producción y no el de la composición de los costos sobre otra variable, lo adecuado es usar la MC que resulta de estimar las transacciones a precios de un año base o a los constantes; por ejemplo, para estudiar los cambios en el empleo que genera el cambio en la técnica de producción, es imprescindible valuar las matrices a precios constantes (ver Murillo-Villanueva, 2018).

Como se detalla más adelante, las MIP del 2008 y 2012 se encuentran valuadas a precios del 2008. Considerando que la matriz del 2003 muestra diferencias menores en términos de su sistema de clasificación y nivel de desagregación con respecto a las del 2008 y 2012, la actualización de esta permitirá abarcar un periodo relativamente amplio con información homogénea, lo que posibilita analizar un sinnúmero de problemas económicos basados en el estudio de la evolución de la estructura productiva, el cambio en la técnica de producción de los distintos subsectores o la modificación en la distribución de los productos, por ejemplo: el efecto del cambio técnico sobre el empleo (ver Murillo-Villanueva, 2018), el crecimiento económico o el pago a los factores productivos. Por ello, este documento tiene como objetivo obtener la MIP de México del 2003 a precios del 2008 y, para tal fin, se utilizan dos métodos de deflación y se identifica el que mejor la actualiza.

El artículo se estructura de la siguiente manera: en el apartado dos se presentan las principales características de las matrices de insumo-producto disponibles para México, se identifican aquellas con características similares y se mencionan los trabajos de homologación realizados a la matriz del 2003; en el tercero se describen y aplican ambos métodos para el caso de la MIP del 2003; en el cuarto se muestran los resultados obtenidos y se realiza una comparación basada en el criterio de la distancia rectilínea para determinar cuál de las dos estimaciones se desvía menos de la matriz de insumo-producto del 2003 a precios corrientes; y en el último se presentan las conclusiones.

2. Matrices de insumo-producto en México

Las MIP pueden ser valuadas de distintas maneras: a precios de comprador,[1] de productor[2] o a básicos[3]; la diferencia entre ellas se encuentra en el tratamiento que se da a los impuestos, subsidios y márgenes del comercio y transporte. En el análisis de insumo-producto conviene trabajar con precios básicos, ya que permiten medir las transacciones sin incluir impuestos, subsidios, costos de transporte ni márgenes de comercialización.

En México, actualmente, el Instituto Nacional de Estadística y Geografía (INEGI) calcula y publica las MIP con base en la información del Sistema de Cuentas Nacionales de México (SCNM). A la fecha, dispone de 10 referidas a los años: 1950, 1960, 1970, 1975, 1978, 1980, 2003, 2008, 2012 y 2013. Las primeras dos fueron elaboradas por el Banco de México y permitieron integrar el primer conjunto de cuentas consolidadas del país; las siguientes cuatro las realizó la Dirección General de Estadística (DGE) del hoy INEGI (SPP, 1980); las correspondientes a 1975 y 1978 fueron parte de una actualización de la de 1970, en tanto que la de 1980 presentó características innovadoras en relación con las precedentes. Estas seis matrices (1950, 1960, 1970, 1975, 1978, 1980) se calcularon a precios de productor y en millones de pesos corrientes.

Después de casi un cuarto de siglo, el INEGI calcula y publica la MIP del 2003 valuada a precios básicos y en millones de pesos del 2003. Las siguientes dos matrices (2008 y 2012) tienen la misma estructura que la del 2003, se calculan ambas a precios básicos, pero en miles de millones de pesos a precios del 2008.

Cabe mencionar que, aunque a nivel de subsector, las matrices del 2003, 2008 y 2012 presentan el mismo número de subsectores (79), los de la MIP del 2003 difieren de aquellos incluidos en las matrices del 2008 y 2012 porque la primera utiliza el Sistema de Clasificación de América del Norte (SCIAN) del 2002, mientras que las del 2008 y 2012 usan el SCIAN del 2007. Por último, la matriz del 2013, que fue publicada por el INEGI a inicios del 2018, está valuada en millones de pesos a precios básicos del 2013 y contempla el cambio de año base del 2008 al 2013.

Como se puede notar, el hecho de que las MIP del 2008 y 2012 estén ambas valuadas al mismo

1 *Precio de comprador = precio de mercado - IVA.*
2 *Precio de producto = precio de comprador- márgenes comerciales y de transporte y fletes.*
3 *Precio básico = precio de productor - impuestos indirectos, a las ventas o IVA no deducible + subvenciones de productos.*

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Cuadro 1

Homologación SCIAN

SCIAN 2003	SCIAN 2007	Homologación
516 Creación y difusión de contenido exclusivamente a través de internet **517** Otras telecomunicaciones	**517** Otras telecomunicaciones	**517** Otras telecomunicaciones
931 Actividades del gobierno	**931** Actividades legislativas, gubernamentales y de impartición de justicia **932** Organismos internacionales y extraterritoriales	**931** Actividades del gobierno y organismos internacionales

Fuente: elaboración propia con base en las MIP del 2003, 2008 y 2012.

nivel de precios y consideran el mismo sistema de clasificación, permite la comparación entre ellas. Sin embargo, esta comparación podría enriquecerse con la homologación de la matriz del 2003. Aunque son varias las diferencias que existen entre esta y las del 2008 y 2012, la más importante de ellas reside en el nivel de precios básicos al que fueron valuadas. Otra diferencia relevante entre la MIP del 2003 y las del 2008 y 2012 es el sistema de clasificación industrial con el que fueron construidas. El cuadro 1 resume la diferencia entre las ediciones del SCIAN y la homologación realizada.

Por otro lado, introducir la matriz del 2013 para un futuro análisis comparativo resulta también muy atractiva. Sin embargo, el hecho de que solo se encuentra un año alejada de la MIP del 2012 y que considere el cambio de año base dificulta su inclusión ya que, para realizar una comparación a precios constantes, sería necesario no solo actualizar la matriz del 2003, sino las del 2008 y 2012 para un análisis a precios básicos del 2013. En ese sentido, en este trabajo se estima la matriz de insumo-producto del 2003 a precios básicos del 2008: esto requiere actualizarla. Dado que entre el 2003 y 2008 se registró un alza en los precios de los productos de los subsectores, las entradas de la MIP del 2003 se deben acrecentar (o inflar). La actualización o deflación de una matriz de insumo-producto puede realizarse de distintas maneras (ver Jackson and Murray, 2004).

En esta investigación se revisaron y aplicaron los dos métodos más utilizados: el de la doble deflación y el RAS. El primero se retomó debido a su simplicidad y validez teórica, mientras que el segundo, muy usado en la bibliografía de análisis de insumo-producto, tiene hoy la facilidad de instrumentarse gracias a la disponibilidad de los datos que requiere.

Al valuar con otros precios matrices de insumo-producto nos podemos enfrentar a dos inconvenientes: por un lado, es probable que las entradas de la MIP cambien significativamente (Wiebe y Lenzen, 2016) y, por el otro, se puede observar una modificación en la magnitud del valor agregado (VA). Dietzenbacher y Hoen (1998 y 1999) y Jackson y Murray (2004), entre otros, argumentan que los resultados de deflactar una matriz mediante RAS son mejores en ambos sentidos que los obtenidos por el método de la doble deflación. Respecto a este último, Durand (1994), Rajakumar y Shetty (2015), entre otros, han advertido de las deficiencias del método para estimar el valor agregado.

3. Dos métodos para deflactar matrices: doble deflación y RAS

Las relaciones de interdependencia que se muestran en las MIP se expresan en una serie de identidades contables, de las cuales aquí se destacan las dos más importantes (ver Schuschny, 2005, cap. 1):

$$VBP_i = X_i = X_{i1} + X_{i2} + \cdots + X_{in} + C_i + I_i + G_i + H_i + E_i \quad (1)$$

$$VBP_j = X_j = X_{1j} + X_{2j} + \cdots + X_{nj} + M_{1j} + M_{2j} + \cdots + M_{nj} + W_j + P_j + NT_j \quad (2)$$

La ecuación (1) muestra que la producción de cada sector puede venderse en los mercados de productos intermedios o como producto final. Así, el destino de la producción del sector $i(VBP_i)$ se encuentra en otras industrias ($X_{ij}; j=1, \ldots, n$), en el consumo de los hogares (C_i), empresas (I_i), gobierno (G_i), inventarios (H_i) o el resto del mundo (E_i). Por su parte, la ecuación (2) muestra que el valor de la producción de cada sector j se utiliza para comprar aquellos factores productivos necesarios para que cada sector produzca; da cuenta de la adquisición de insumos intermedios nacionales (X_{nj}) e importados (M_{nj}), de los insumos primarios de trabajo (W_j) y capital (P_j), así como del pago de impuestos netos (NT_j).

La información de las identidades (1) y (2) se presenta de dos formas: una es a través de la matriz interna de insumo-producto y la otra, de la total. La diferencia entre ellas reside en el tratamiento que se da a las importaciones. Mientras que en la interna las importaciones se registran mediante un vector fila que muestra la adquisición de insumos intermedios importados, en la total las importaciones de insumos intermedios y bienes finales se suman a los componentes de cada una de las identidades contables de la siguiente manera:

$$VBP_i = X_i = X_{i1} + M_{i1} + X_{i2} + M_{i2} + \cdots + X_{in} + M_{in} + C_i + C_i^m + I_i + I_i^m + G_i + G_i^m + H_i + H_i^m + E_i + E_i^m - M_i \quad (3)$$

$$VBP_j = X_j = X_{1j} + \cdots + X_{nj} + M_{1j} + \cdots + M_{nj} + W_j + P_j + NT_j \quad (4)$$

donde $M_i = M_{i1} + M_{i2} + \cdots + M_{in} + C_i^m + I_i^m + G_i^m + H_i^m + E_i^m$ y donde los componentes M_{nj} representan los insumos intermedios importados y los componentes con el supra-índice m ($C_i^m + I_i^m + G_i^m + H_i^m + E_i^m$), los bienes importados destinados a abastecer la demanda final. La matriz total de insumo-producto resulta de la suma de la matriz interna y la de importaciones. La decisión de usar la matriz interna o la total dependerá de los objetivos del investigador;[4] este trabajo plantea la actualización de la matriz total de insumo-producto del 2003.

La actualización o deflación de cualquier variable, en este caso de las transacciones que componen las MIP, se realiza mediante la aplicación de uno o varios índices de precios implícitos (IPI) (ver Díaz-Calleja, 2003). Por lo tanto, al comparar variables valuadas a un mismo precio, el investigador se asegura de captar únicamente los cambios en las cantidades producidas. El método de la doble deflación utiliza IPI, mientras que el RAS los estima.

3.1 Método de doble deflación

Se le llama así porque se desarrolla en dos etapas: la primera consiste en deflactar los insumos intermedios (Z_{ij}),[5] la demanda final (f_i)[6] y el VBP (X_i) valuados a precios corrientes utilizando, respectivamente, los IPI de los insumos intermedios, la demanda final y el VBP; la segunda etapa es para obtener por diferencia el valor agregado (v_j)[7] y, por lo tanto, un índice de precios del VA que iguale la identidad fundamental entre el VBP total por la oferta y por la demanda (Miller y Blair, 2009, cap. 4.8).

En este caso, el índice de precios es la razón entre el precio de una mercancía o producto a precios del 2003 y el precio en el nuevo año base (2008). Si definimos $\pi_i = p_i^b / p_i^t$ como el índice de precios o deflactor para la industria i donde p_i^b son los precios del año base (2008) y p_i^t representan los del año corriente (2003), al multiplicar cada

4 Por ejemplo, si el investigador desea conocer el impacto del cambio en alguna de las variables exógenas sobre la economía nacional, se recomienda hacer uso de las matrices internas; pero si desea conocer las técnicas de producción de los distintos sectores será mejor utilizar las matrices totales, ya que también consideran los insumos importados utilizados.

5 La matriz (Z_{ij}) representa los valores de las transacciones X_{ij}.

6 $f_i = C_i + I_i + G_i + V_i + E_i - M_i$.

7 $v_j = W_j + P_j + NT_j$.

índice de precios por los insumos intermedios, la demanda final y el VBP obtenemos los valores de cada variable valuados a precios del 2008:

$$z^b = \hat{\pi}_z z^t, \quad f^b = \hat{\pi}_f f^t, \quad x^b = \hat{\pi}_x x^t, \tag{5}$$

donde z^t representa la matriz de transacciones intersectoriales; $\hat{\pi}_z$, el vector diagonalizado del IPI de insumos intermedios de la industria i; f^t es el vector de la demanda final de la industria i; $\hat{\pi}_f$, el vector diagonalizado del deflactor de la demanda final de la industria i; x^t, el vector del VBP de la industria i; $\hat{\pi}_x$, el vector diagonalizado del deflactor del producto de la industria i; el supra-índice b denota precios del año base y el supra-índice t, precios corrientes. De manera clara, la actualización de una matriz de insumo-producto requiere que el importe que se desea actualizar sea multiplicado por el IPI, mientas que la deflación requeriría dividir al monto por el IPI. A partir de la identidad (2) es posible obtener el valor agregado(v^b) actualizado y su índice de precios para cada industria ($\hat{r}$.) El VA a precios del año base es la diferencia entre el VBP a precios del año base (x^b) y el consumo intermedio a precios del año base ($i'z^b$):

$$(v^b)' = (x^b) - i'z^b. \tag{6}$$

El índice de precios del VA resulta del cociente del VA de un año a precios básicos y el VA a los corrientes, matricialmente:

$$\hat{v}^b = \hat{r}\hat{v}^t, \ \hat{r} = \hat{v}^b (\hat{v}^t)^{-1}. \tag{7}$$

El método de la doble deflación ha sido muy criticado por dos cuestiones: la primera, porque la utilización de un índice de precios de la producción para deflactar una fila completa solo se justifica si el sector produce un único bien y, por lo general, los sectores producen más de uno; la segunda crítica reside en que el valor agregado es obtenido por diferencia, de manera que el error en su medición es igual a la suma de los errores en la medición de la demanda intermedia y la final, así como el producto.[8] En este sentido, las limitaciones del método de doble deflación se encuentran en que supone que hay un único IPI para todas las relaciones de entrega o distribución que un subsector tiene con el resto del aparato productivo y en que su funcionamiento enfatiza, principalmente, en las relaciones de demanda (intermedia y final) dejando de lado el cambio en los precios relativos causados por las relaciones de oferta (consumo intermedio y valor agregado); sin lugar a dudas, esto último ocasiona que las estimaciones de los agregados de consumo intermedio y VA se encuentren alejados de sus valores reales en una magnitud igual a la suma de las desviaciones de todas las relaciones intersectoriales que se explican por considerar solo el ajuste por la demanda.

Aplicación del método de doble deflación a la MIP del 2003

Se utilizaron los IPI de la demanda intermedia y de la producción (información disponible en el *Anexo A*) obtenidos del Banco de Información Económica del INEGI (2016a). Para la estimación de la matriz de insumo-producto del 2003 a precios del 2008, se realizó lo siguiente:[9]

- Se estimó la matriz total de transacciones intersectoriales deflactada (z^b) mediante la multiplicación del IPI de la demanda intermedia inter-industrial del 2003 base 2008 ($\hat{\pi}_z$) y la matriz de transacciones a precios del 2003 (z^t) (ecuación 5).
- Se obtuvo la demanda final y el VBP a precios del 2008 multiplicando la demanda final (f^t) y el VBP a precios del 2003 (x^t) por el IPI de la producción $\hat{\pi}_x$.[10]
- Por diferencia, se obtuvo el VA a precios del 2008 (v^b) (ecuación 6). Para obtener el desglose de los componentes del VA, se multiplicó la participación de cada componente del VA del 2003 y el total del

8 Durand (1994) propone como alternativa la estimación el valor agregado inter-industrial a través de los cuadros de oferta y utilización en lugar de por diferencia, como en el método de doble deflación.

9 Disponible bajo solicitud a los autores.

10 No se utilizó el IPI de la demanda final para deflactarla debido a que solo se encontró a un nivel de desagregación de 20 subsectores y la matriz que se pretende deflactar está desgregada en 78 subsectores.

VA a precios del 2008. La estimación de los componentes del VA a partir de la estructura observada en el 2003 implica que no hubo cambios importantes en los precios relativos de los insumos primarios, trabajo y capital y, por lo tanto, excluye el cambio en los requerimientos de estos insumos a causa de los precios de estos factores. Sin lugar a dudas, sería deseable estimar los componentes del VA considerando los IPI de los factores, sin embargo, en la práctica, la falta de este tipo de información limita su realización. La información de las participaciones se obtuvo del cuadro de oferta y utilización del 2003 y se encuentra disponible en el *Anexo B*.

3.2 Método RAS

Es una forma de ajuste biproporcional aplicada a las MIP. Fue introducido por Stone (1961), Stone y Brown (1962) y Bacharach (1970). Se le cataloga como un método que, sin utilizar o utilizando parcialmente información censal o de encuestas, actualiza matrices de insumo-producto. Permite estimar n^2 datos usando solo $3n$ datos (Jackson y Murray, 2004).

Sean $a_{ij} \in A$ los coeficientes técnicos de la MIP conocida y $q_{ij} \in Q$ los coeficientes de la matriz que se desea encontrar. Los $3n$ datos requeridos son los siguientes: a) el vector columna de la demanda intermedia del año objetivo (dn_i^{obj}), b) el vector fila de consumo intermedio del año objetivo (cn_{ji}^{obj}) y c) el vector columna del VBP del año objetivo (x_i^{obj}). El método consiste en obtener la demanda intermedia que se observa si la estructura de insumo-producto no experimentara ningún cambio,[11] es decir, con la estructura productiva (A) y el valor de la producción (x) originales: $dn^1 = Ax$. Las modificaciones del periodo explican las diferencias entre dn^{obj} y dn^1, de manera que si:

$$\boldsymbol{r_1 = \widehat{dn}^{obj} (\widehat{dn}^1)^{-1}}, \text{ entonces } \boldsymbol{Q^1 = r_1 A}. \quad (8)$$

Esta es la primera estimación de la nueva estructura de insumo-producto. Las sumas por fila de $Q^1 x$ ahora igualan los valores conocidos de dn^{obj}. Sin embargo, las sumas por columna no serán iguales a los valores conocidos de consumo intermedio (cn_{ji}^{obj}). A continuación se calcula $cn^1 = i' Q^1 \hat{x}$ donde $i' = (1, \ldots, 1)$ es el vector suma. Sea:

$$\boldsymbol{s_1 = \widehat{cn}^{obj} (\widehat{cn}^1)^{-1}}, \text{ entonces } \boldsymbol{Q^2 = Q^1 s_1}. \quad (9)$$

donde Q^2 es la matriz cuya suma por columna es igual a cn_{ji}^{obj}, pero cuya suma por fila es diferente a dn_i^{obj}. El método RAS consiste en obtener estimaciones sucesivas de r_i y s_i, hasta que la diferencia entre dn_i^{obj} y cn_{ji}^{obj} con sus respectivas estimaciones (r_i y s_i) sea la menor posible, lo cual sucede cuando la suma por columna es igual a la suma por fila. Por lo normal, el procedimiento converge hacia una estimación estable de Q después de un número relativamente pequeño de iteraciones.

El problema que resuelve RAS también puede plantearse como uno de optimización (Bacharach, 1970) en el que la matriz estimada Q se desvía lo menos posible de la observada alcanzando los valores dn_i^{obj} y cn_{ji}^{obj}. Este método, a diferencia del de doble deflación, no tiene como finalidad proveer un valor estimado de los respectivos componentes de las demandas intermedia y final, del consumo intermedio, del VA ni del VBP, más bien supone que se cuenta de manera exógena con los valores de dichas variables.

Aplicación del método RAS a la MIP del 2003

Los datos disponibles para esta estimación fueron el consumo intermedio, el VA, el VBP y las importaciones del 2003 a precios básicos del 2008 desagregados en 78 subsectores (información disponible en el *Anexo C*). Por medio del método RAS, la MIP se consiguió de la siguiente manera:[12]

- Con los datos disponibles y haciendo uso de las identidades contables[13] se obtuvo

11 Para mayor detalle, ver Jackson y Murray (2004) o Miller y Blair (2009, cap. 7.4).

12 Disponible bajo solicitud a los autores.

13 *Consumo intermedico + valor agregado + importaciones = demanda intermedia + demanda final.*

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

el monto de la suma de la demanda intermedia y la final.

- Del cuadro de utilización del 2003 del SCNM se obtuvo la participación de la demanda intermedia y de cada componente de la demanda final en el total del VBP (información disponible en el *Anexo D*). Las participaciones se multiplicaron por el monto obtenido arriba y se obtuvo la demanda intermedia y la final del 2003 a precios constantes del 2008.
- Con los vectores obtenidos de demanda intermedia y consumo intermedio objetivo (del 2003 a precios del 2008), dn_i^{obj} y cn_{ji}^{obj}, y con la matriz A de coeficientes técnicos del 2003 a precios del 2003 fue posible estimar la matriz Q de coeficientes técnicos del 2003 a precios del 2008. Sin embargo, la condición para que el método RAS converja es que el monto total de dn_i^{obj} debe ser idéntico al monto total de cn_{ji}^{obj}. En este caso, los valores discreparon en 5%, por lo que el valor total de dn_i^{obj} se ajustó hacia arriba.
- Por último, a través del cuadro de utilización del 2003 a precios del 2003, se obtuvo el valor de los componentes del VA en el total del 2008 a precios del 2003. Para ello, se multiplicó la participación de cada componente del VA y el VA deflactado (información disponible en el *Anexo B)*.

4. Distancia entre la MIP del 2003 a precios corrientes y sus dos actualizaciones

Se definen tres matrices: la del 2003 a precios del 2003 (A03), la del 2003 a precios del 2008 actualizada por RAS (A03RAS) y la del 2003 a precios del 2008 actualizada por el método de doble deflación (A03DD). Ahora bien, reconsiderando que cuando se obtienen las dos actualizaciones (A03RAS y A03DD), la única información disponible de las razones entre lo que se usa para producir y lo que se produce son las estructuras de costos que genera la matriz a precios corrientes; es claro que esta referencia observada puede cambiar radicalmente mediante los efectos conjuntos de los precios relativos de insumos a productos. No obstante, el único asidero empírico observado de los coeficientes resultantes de las razones en volúmenes (o a precios constantes) entre insumos y productos son los coeficientes de costos a producto que genera la matriz a precios corrientes. Por ese motivo, en este trabajo se eligió el criterio de la mínima distancia entre dicha matriz observada y las resultantes de ambos métodos de estimación. Este criterio de elección, como resulta claro, es endeble excepto por un simple hecho contable: la única información disponible es la de costos y, por lo general, la composición del ingreso por los gastos necesarios para su consecución no está radicalmente alejada de la composición técnica de insumos para la producción (ver Wiebe y Lenzen, 2016).

En ese sentido, una medida de similitud entre dos observaciones, vectores o matrices es la distancia; cuanto más parecidas sean las observaciones comparadas, menor será la distancia entre ellas y viceversa (Puchet, 1987). Existen diversas formas de medirla (Cuadras, 1989 y Gower y Legendre, 1986): las más utilizadas son la euclídea[14] y la rectilínea.[15]

El primer punto a considerar para medir la similitud entre las matrices está relacionado con el número de celdas iguales a cero: las que en la matriz A03 fueron cero deberán permanecer cero en las matrices deflactadas A03DD y A03RAS; en la A03DD, el número de celdas cuyo valor fue igual a cero se mantuvo;[16] en la A03RAS, el número de celdas igual a cero se incrementó en 81 (la A03 tiene 1 836 celdas en cero, mientras que la A03RAS cuenta con 1 917). Esto se debe a que los cuadros de oferta y utilización del 2003 a precios del 2008 registran que la demanda intermedia de los subsectores (213) *Servicios de la minería* y (483) *Transporte por ferrocarril* fue cero, mientras que en la matriz A03 son diferente de cero. El valor individual de 77 de las 81 celdas eliminadas representó menos de 1%

14 $d_{(i,j)} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$.

15 $d_{(i,j)} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$.

16 Esto se debe a que la matriz A03DD es el resultado de la multiplicación de A03 por el índice de precios de la demanda intermedia.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

del valor total del consumo intermedio del subsector columna al que pertenecen (solo en el subsector 212, el valor de las celdas que se eliminaron representó 1.5% del valor del consumo intermedio). Por su parte, el valor del consumo intermedio de las 81 celdas eliminadas representa 0.15% del monto de consumo intermedio total de la economía. Esto quiere decir que la importancia relativa de las celdas eliminadas es relativamente baja.

El segundo punto a considerar al comparar las matrices se refiere a la similitud entre los vectores columna que representan parte de la técnica de producción de cada subsector de la matriz. Con base en lo explicado arriba, se buscará que las distancias entre los vectores de coeficientes técnicos de A03 y sus estimaciones respectivas sean menores, de manera que las estimaciones no estén radicalmente alejadas de la composición observada. Para medir la distancia entre coeficientes, se calcularon las diferencias porcentuales relativas entre los coeficientes de las matrices estimadas A03DD y A03RAS y los de la observada A03 según la definición de distancia rectilínea (ecuación 10):

$$D(a_{ij}) = \left(\frac{|a_{ij}^{03DD} - a_{ij}^{03}|}{a_{ij}^{03}} \right) \times 100. \qquad (10)$$

La razón por la cual se utilizó el método de la distancia rectilínea se encuentra en que los coeficientes técnicos de las MIP, al no representar puntos de un plano cartesiano sino observaciones, conducen a que la diferencia entre observaciones homólogas resulte la misma con el criterio de distancia euclídea que con el de la rectilínea. Sin embargo, los resultados obtenidos de la euclídea no están acotados, mientras que los conseguidos de la rectilínea (ecuación 10) sí lo están.

Los cuadros 2 y 3 muestran las distancias entre los coeficientes de la matriz A03 y las A03DD y A03RAS, respectivamente. En la primera columna se presentan las distintas categorías en las que se clasifica la diferencia porcentual relativa (ecuación 10) entre coeficientes y que va desde cero hasta más de 100%; en este trabajo, se consideró que los coeficientes más similares serán aquellos que estén alejados de su valor en la matriz A03 en un porcentaje no mayor a 30%; la segunda muestra el número de coeficientes que se encuentran en cada categoría; la tercera indica el porcentaje que representan respecto al total de coeficientes; y en la cuarta se presenta la participación del valor de los coeficientes de cada categoría $k(c(k))$ en el valor del coeficiente de insumos totales ($coef.\ IT$) resultante de las transacciones intersectoriales del 2003 a precios del 2003 y que se define como:

$$Particip.\ en\ el\ coef.\ IT^{03} = \frac{\sum_i \sum_j a_{ij}^{03DD};\ (i,j)\ c(k), k = 0, \ldots, 10}{\sum_j^{78} \sum_i^{78} a_{ij}^{03}} \qquad (11)$$

donde el valor del denominador es 34.09.

El cuadro 2 muestra las diferencias relativas entre las matrices A03 y A03DD. Como se observa, aquellas celdas que constituyen 79% del coeficiente de insumos intermedios totales de la matriz A03 se captaron con un error porcentual que no excede a 30%, las cuales representan, de manera conjunta, a 82% del total de las celdas (3 491). Por su parte, aquellas celdas cuyo error relativo fue mayor a 30% tienen un peso relativo en las transacciones intermedias muy bajo. Además, se observa que la frecuencia disminuye considerablemente conforme aumenta el error representado por cada clase. Los resultados sugieren una buena aproximación individual de los coeficientes técnicos de la matriz actualizada por doble deflación.

Por su parte, el cuadro 3 presenta las diferencias relativas entre las matrices A03 y A03RAS. Se observa que de las 4 167 celdas diferentes de cero, solo 44% (1 840) muestra una diferencia relativa porcentual respecto a sus valores en A03 que oscila entre 0 y 30% y representa casi 60% del valor del coeficiente de consumo intermedio total, es decir, poco más de la mitad de los coeficientes técnicos estimados se desvían más de 30% respecto a su valor en A03 y su participación en el coeficiente de insumos intermedios totales es menor (59.6%); esto sugiere que, de acuerdo con este criterio, la

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Cuadro 2

Diferencias relativas porcentuales entre las matrices A03 y A03DD, así como participación en el coeficiente de insumos intermedios totales a precios del 2003

Clase	Frecuencia	Frecuencia relativa acumulada	Participación de los coeficientes de cada clase en el *coef. IT* (acum.)
0-10%	2 025	47.67	47.3
11-20%	1 046	72.29	69.1
21-30%	420	82.18	79.0
31-40%	266	88.44	83.9
41-50%	174	92.54	88.5
51-60%	99	94.87	91.0
61-70%	94	97.08	92.7
71-80%	32	97.83	93.5
81-90%	15	98.18	93.8
91-100%	7	98.34	94.3
Más de 100%	70	100.00	100.0
Total	**4 248**		

Fuente: elaboración de los autores con base en la MIP del 2003 deflactada con el método de doble deflación.

Cuadro 3

Diferencias relativas porcentuales entre las matrices A03 y A03RAS, así como participación en el coeficiente de insumos intermedios totales a precios del 2003

Clase	Frecuencia	Frecuencia relativa acumulada	Participación de los coeficientes de cada clase en el *coef. IT* (acum.)
0-10%	645	15.48	21.2
11-20%	657	31.25	47.1
21-30%	539	44.18	59.6
31-40%	520	56.66	71.0
41-50%	444	67.32	80.2
51-60%	421	77.42	86.0
61-70%	272	83.95	90.3
71-80%	148	87.55	92.2
81-90%	65	89.06	93.3
91-100%	183	93.45	94.8
Más de 100%	273	100.00	99.8*
Total	**4 167**		

* El total no es igual a 100.00 debido a las celdas que la actualización por RAS convirtió en cero.
Fuente: elaboración de los autores con base en la MIP del 2003 deflactada con el método RAS.

estimación de la matriz A03DD es mejor que la obtenida por el método RAS.

El tercer punto a considerar está relacionado con la evaluación de las estimaciones de los bordes de la matriz, es decir, de los coeficientes de demanda intermedia,[17] final,[18] consumo intermedio[19] y valor agregado.[20] Las diferencias relativas se obtuvieron a través de la distancia rectilínea; por ejemplo, la diferencia entre el valor estimado y el observado del coeficiente de demanda intermedia se define como en la ecuación (12), mientras que la participación de los coeficientes de cada categoría en el total del coeficiente del 2003 a precios del 2003 se define en la ecuación (13), y lo mismo se hace para el resto de variables:

$$D(dn_i) = \frac{|dn_i^{03DD} - dn_i^{03}|}{dn_i^{03}} \qquad (12)$$

$$P.\ Cdn^{03} = \frac{\sum_i dn_i^{03DD}(i)\, c(k), k = 0, \ldots, 10}{\sum_i^{78} dn_i^{03}}. \qquad (13)$$

Los cuadros 4 y 5 muestran las diferencias relativas entre los coeficientes observados (A03) y los estimados (A03DD y A03RAS) para las cuatro variables: la columna F presenta la frecuencia de las diferencias según la categoría, la FRA denota la frecuencia relativa acumulada y la columna PT_ representa la participación acumulada de cada clase en el total para la matriz A03 de la variable de interés correspondiente.[21]

Los resultados muestran que la matriz A03DD reporta los coeficientes de demandas intermedia y final y consumo intermedio más similares a los de la A03, ya que el porcentaje de los coeficientes estimados con menor desviación es mayor; en concreto, 62, 93 y 81% de los respectivos coeficientes fueron estimados con una desviación no mayor a 30% respecto a su valor en la matriz A03.

17 *Demanda intermedia:* $dn_i = \sum_j^{78} X_{ij} / X_i = \sum_j^{78} a_{ij}$.

18 *Demanda final:* $f_i = F_i / X_i$, *donde* $F_i = C_i + I_i + G_i + H_i + E_i - M_i$.

19 *Consumo intermedio:* $cn_j = \sum_i^{78} X_{ij} / X_j = \sum_i^{78} a_{ij}$.

20 *Valor agregado:* $v_j = V_j / X_j$, *donde* $V_j = W_j + P_j + NT_j$.

21 El espacio en blanco _ es ocupado, según el caso, por las variables demanda intermedia, demanda final, consumo intermedio o valor agregado.

Cuadro 4

Diferencias relativas porcentuales de los coeficientes de las demandas intermedia y final, el consumo intermedio y el valor agregado de las matrices A03 y A03DD, así como participación en el total de las variables respectivas

Clase	Demanda intermedia			Demanda final			Consumo intermedio			Valor agregado		
	F	FRA	PTDI	F	FRA	PTDF	F	FRA	PTCI	F	FRA	PTVA
0.0-0.1	33	42.3	61.2	53	70.7	87.3	31	40.3	39.8	42	53.8	62.8
0.1-0.2	8	52.6	76.5	8	81.3	93.4	23	70.1	68.1	14	71.8	81.3
0.2-0.3	7	**61.5**	85.0	9	**93.3**	97.8	8	**80.5**	79.5	8	82.1	89.2
0.3-0.4	8	71.8	87.9	2	96.0	99.3	8	90.9	91.6	3	85.9	91.8
0.4-0.5	0	71.8	92.6	0	96.0	99.3	3	94.8	95.1	1	87.2	92.8
0.5-0.6	4	76.9	93.1	1	97.3	99.5	3	98.7	99.2	3	91.0	95.3
0.6-0.7	3	80.8	93.1	0	97.3	99.5	1	100.0	100.0	2	93.6	96.9
0.7-0.8	4	85.9	93.1	1	98.7	100.4	0			1	94.9	97.7
0.8-0.9	1	87.2	93.1	0	98.7	100.4	0			1	96.2	98.4
0.9-1.0	2	89.7	93.1	0	98.7	100.4	0			2	98.7	99.3
Más de 1.0	8	100.0	100.0	1	100.0	100.0	0			1	100.0	100.0
Total	**78**			**75**			**77**			**78**		

Fuente: elaboración de los autores con base en la MIP del 2003 deflactada con el método de doble deflación.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Cuadro 5

Diferencias relativas porcentuales de los coeficientes de las demandas intermedia y final, el consumo intermedio y el valor agregado de las matrices A03 y A03RAS, así como participación en el total de las variables respectivas

Clase	Demanda intermedia			Demanda final			Consumo intermedio			Valor agregado		
	F	FRA	PTDI	F	FRA	PTDF	F	FRA	PTCI	F	FRA	PTVA
0.0-0.1	31	39.7	29.8	38	50.7	77.4	36	46.8	54.9	41	52.6	58.2
0.1-0.2	11	53.8	64.0	4	56.0	83.2	16	67.5	74.0	21	79.5	83.4
0.2-0.3	2	**56.4**	68.8	3	**60.0**	85.4	8	**77.9**	83.9	8	**89.7**	92.0
0.3-0.4	7	65.4	77.3	2	62.7	86.7	5	84.4	88.1	4	94.9	96.3
0.4-0.5	3	69.2	84.2	5	69.3	91.9	8	94.8	94.8	1	96.2	97.5
0.5-0.6	4	74.4	93.3	2	72.0	93.6	2	97.4	97.6	1	97.4	98.3
0.6-0.7	2	76.9	94.6	7	81.3	95.9	1	98.7	99.3	0	97.4	98.3
0.7-0.8	2	79.5	94.7	2	84.0	98.8	0	98.7	99.3	0	97.4	98.3
0.8-0.9	4	84.6	94.7	0	84.0	98.8	1	100.0	100.0	1	98.7	99.3
0.9-1.0	3	88.5	95.9	0	84.0	98.8	0			0	98.7	99.3
Más de 1.0	9	100.0	100.0	12	100.0	100.0	0			1	100.0	100.0
Total	**78**			**75**			**77**			**78**		

Fuente: elaboración de los autores con base en la MIP del 2003 deflactada con el método RAS.

Asimismo, se observa que la participación de los coeficientes en sus totales es de 85, 90 y 80%, respectivamente. En cambio, se ve que los coeficientes de VA fueron mejor estimados por el método RAS que por el de doble deflación, ya que 90% de los coeficientes de VA obtenidos en A03RAS muestran una desviación no mayor a 30% respecto a sus valores originales, con una participación en su total de 92 por ciento.

Como diversos autores señalan, la estimación del valor agregado y sus respectivos coeficientes a través del método de doble deflación muestra algunas deficiencias. Sin embargo, aunque en este ejercicio las estimaciones del VA por RAS resultan ser mejores que aquellas por el método de doble deflación, la diferencia entre ellas es menor. Mientras que el porcentaje de coeficientes de VA con una desviación menor a 30% respecto a A03 es 8% mayor en A03RAS que en A03DD, la participación de estos en el coeficiente de valor agregado total es solo 3% mayor.

El cuarto criterio para la elección del método que mejor actualiza la matriz A03 consiste en que los resultados tengan sentido económico. Este criterio toma relevancia porque en la estimación del VA de A03DD pueden surgir valores negativos, es decir, como el valor agregado se obtiene por diferencia $[(v^b)' = (x^b) - i'z^b]$, es posible que para algún sector el monto actualizado del consumo intermedio $(i'z^b)$ sea mayor al del VBP (x^b), ocasionando que el VA sea negativo, lo cual no tiene ningún sentido económico porque indica destrucción de valor; dicho de otro modo, una situación de VA negativo sugiere un déficit, deuda u obligación adquirida por el pago de los factores productivos, de manera que la producción observada, en lugar de generar ingreso para los hogares y las empresas, requirió de ingreso proveniente de estos sectores institucionales. Para el caso de la actualización de la matriz de A03DD, los resultados arrojan que el VA del subsector SCIAN (326) *Industria del plástico y del hule* toma un valor de -9 104 (en miles de millones de pesos a precios del 2008). Este resultado respalda con evidencia la deficiencia del método de doble deflación para la estimación del valor agregado.

Por último, los resultados de los cuatro criterios considerados para evaluar la calidad de las esti-

maciones de la matriz del 2003 a precios del 2008 sugieren algunas diferencias importantes y otras menores. Primero, se observa que la matriz estimada por el método de doble deflación mantiene el número de celdas igual a cero, mientras que la estimación por RAS lo altera, aunque se puede demostrar que la importancia relativa de las celdas eliminadas es mínima. Segundo, se pudo ver que la A03DD estimó mejor la matriz de coeficientes técnicos, ya que es la que menos se aleja de la A03, esto se debe a que 82% de los coeficientes técnicos se estimó con una desviación menor a 30% respecto a su valor en la matriz A03 y representan 79% del coeficiente total de insumos totales. Tercero, respecto a los coeficientes de las variables borde (como la demanda intermedia y la final, el consumo intermedio y el valor agregado), se observa que la estimación de los primeros tres es ligeramente mejor en A03DD, pero que la estimación de los coeficientes de VA en A03RAS es mejor. Además, se comprobó que el método de doble deflación es deficiente en la estimación del valor agregado, ocasionando que para el caso de un sector se obtuviera un coeficiente sin sentido económico. Sin lugar a dudas, este último punto limita el uso de la matriz A03DD.

5. Conclusiones

La comparación de diversas matrices de insumo-producto permite conocer la evolución de la estructura productiva de una economía durante un periodo determinado. Sin embargo, requiere del uso de matrices valuadas a los mismos precios. De las MIP disponibles para México, se encontró que la del 2003 puede homologarse para ser comparada con las del 2008 y 2012, que están valuadas al mismo nivel de precios y que consideran el mismo sistema de clasificación industrial, y ello requiere la estimación de la matriz del 2003 a precios del 2008.

Este trabajo describe detalladamente los pasos realizados para la obtención de la MIP del 2003 a precios del 2008 a partir de los métodos de doble deflación y RAS y presenta la información utilizada en ambos ejercicios. El criterio que se utiliza para la determinación de la mejor estimación es el de la distancia rectilínea. Se considera que el ajuste en los precios de referencia no debe alterar demasiado la relación que existe entre el uso y la distribución de los insumos y el producto total ni la relación entre la demanda intermedia y la final, el consumo intermedio, el VA y el VBP. Por ello, la matriz actualizada menos distante de la original será la mejor estimada. Para ello, se consideraron cuatro criterios: el primero incluye el número de celdas iguales a cero; el segundo, la distancia entre las entradas de la matriz que registran las relaciones intersectoriales; el tercero revisa la distancia entre las variables borde de la matriz; y el cuarto considera la congruencia económica de las estimaciones.

Los resultados muestran diferencias importantes entre las dos actualizaciones. Contrario a lo sugerido por Dietzenbacher y Hoen (1998 y 1999) y Jackson y Murray (2004), el ejercicio realizado para el caso de la MIP de México muestra que la matriz de coeficientes técnicos y los coeficientes de demandas intermedia y final y consumo intermedio se estiman mejor con el método de doble deflación, pero en concordancia con los mismos autores, este método es deficiente en la estimación del valor agregado. De manera puntual, se encontró que, debido a que el método de doble deflación obtiene por diferencia el VA, este último resultó negativo para el caso de un subsector, lo cual indica errores importantes en la medición y, por ello, falta de congruencia económica. Por lo tanto, considerando lo anterior, los resultados sugieren que la actualización por el método RAS es superior en el criterio de la congruencia económica y que la distancia entre los coeficientes estimados por RAS y observados de las variables borde (consumo intermedio, valor agregado, demandas intermedia y final, así como valor bruto de la producción) es aceptable. Por eso, y por el hecho de que los datos de las variables borde del 2003 a precios del 2008 fueron generados y actualizados por el propio INEGI, se concluye que la mejor estimación es la obtenida por el método RAS.

La actualización de la matriz de insumo-producto del 2003 representa un ejercicio importante

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

para la elaboración de información que eventualmente permite incorporar una nueva dimensión a la comparación y análisis de la evolución de la estructura productiva de México en el periodo 2003-2012, lo que, a su vez, facilita dar respuesta a diversos problemas económicos relacionados con el cambio en la técnica de producción y su efecto en otras variables de interés.

Fuentes

Bacharach, M. *Biproportional Matrices and Input-Output Change*. Cambridge, Cambridge University Press, 1970.

Baumol, W. J. & T. Ten Raa. "Wassily Leontief: In appreciation", en: *The European Journal of the History of Economic Thought*. 16:3, 2009, pp. 511-522.

Cuadras, C. M. "Distancias estadísticas", en: *Estadística Española*. Vol. 30, Núm. 119, 1989, pp. 295-378.

Díaz Calleja, E. "Deflactores y precios implícitos: índices de precio y volumen en la contabilidad nacional", en: *Revista de Economía Crítica*. 1, 2003, pp. 113-127.

Dietzenbacher, E., & A. R. Hoen, "Double deflation and aggregation", en: *Environment and Planning A*. 31(9), 1999, pp. 1695-1704.

Dietzenbacher, E. & Hoen, A. R. "Deflation of input-output tables from the user's point of view: a heuristic approach", en: *Review of Income and Wealth*. 44, 1998, pp. 111-122.

Durand, R. "An Alternative to Double Deflation to Measuring Real Industry Value-Added", en: *Review of Income and Wealth*. 40, 1994, pp. 303-316.

Gower, J. C. and P. Legendre. "Metric and Euclidean Properties of Dissimilarity Coefficients", en: *Journal of Classification*. 3, 1986, pp. 5-48.

INEGI. *Matrices de insumo-producto*. INEGI, 2017 (DE) http://www.inegi.org.mx/

_____ *Banco de Información Económica*. *Índices de precios implícitos base 2008*. INEGI, 2016a (DE) http://www.inegi.org.mx/

_____ *Banco de Información Económica*. *Cuentas nacionales base 2008*. INEGI, 2016b (DE) http://www.inegi.org.mx/

_____ *Sistema de Cuentas Nacionales de México. Matriz de insumo-producto del 2003*. INEGI, 2016c (DE) http://www.inegi.org.mx/

Jackson, R. W. and A. T. Murray. "Alternative Input-Output Matrix Updating Formulas", en: *Economic Systems Research*.16, 2004, pp. 135-148.

Kurz, D. H. & N. Salvadori. "'Classical' roots of input-output analysis: A short account of its long prehistory", en: *Economic Systems Research*. 12:2, 2000, pp. 153-179.

Leontief, W. "Quantitative input and output relations in the economic system of the United States", en: *Review of Economics and Statistics*. 18:3, 1936, pp. 105-125.

Miller, R. and P. Blair. *Input-Output Analysis*. *Foundations and Extensions*. 2nd edition. Cambridge University Press, 2009.

Murillo-Villanueva, B. *El desempleo tecnológico en la industria manufacturera en México, 2003-2012: el efecto del cambio técnico en el empleo*. Tesis doctoral. México, Universidad Nacional Autónoma de México, 2018.

Puchet, M. "Experimentos de actualización de matrices de insumo-producto de México" , en: P. Alonzo Quiroz et al. (editores). *Análisis aplicado de insumo-producto*. México, CIDE, 1987, pp. 90-122.

Rajakumar, J. Dennis & S. L. Shetty. "Gross value added: Why not the double deflation method for estimation", en: *Economic and Political Weekly*. Vol. I, No. 33, August 15th, 2015, pp. 78-81.

Schuschny, A. *Tópicos sobre el modelo de insumo-producto*. *Teoría y aplicaciones*. Serie de Estudios Estadísticos y Prospectivos. Santiago, CEPAL, División de Estadística y Proyecciones Económicas, 2005.

Secretaría de Programación y Presupuesto (SPP). *Bases informativas para la utilización del modelo de insumo-producto*. Tomo I: *Homogeneización de las matrices 1950-1960-1970*. México, SPP, 1980.

Stone, R. *Input-Output and National Accounts*. Paris, Organization for Economic Cooperation and Development, 1961.

Stone, R. & A. Brown. *A Computable Model of Economic Growth*. Vol. 1, *A Programme for Growth*. London, Chapman and Hall, 1962.

Wiebe, K. y M. Lenzen. "To RAS or not to RAS? What's is the difference in outcomes in the multi-regional input-output models?", en: *Economic Systems Research*. 28, 2016, pp. 383-402.

Anexo A

Índices de precios implícitos de la producción y la demanda intermedia del 2003, índice base 2008 = 100

SCIAN	IPI producción	IPI demanda intermedia	SCIAN	IPI producción	IPI demanda intermedia
111	68.0	72.5	485	79.1	69.8
112	77.2	70.0	486	77.7	89.4
113	84.3	82.5	487	75.5	74.3
114	88.1	80.2	488	77.9	73.2
115	72.3	74.7	491	84.7	79.2
211	35.9	66.5	492	33.4	78.9
212	54.0	63.8	493	80.6	85.2
213	71.5	71.7	511	83.1	78.8
221	71.5	52.7	512	80.8	82.1
222	76.4	50.7	515	71.9	82.0
236	74.1	73.8	517	100.5	78.6
237	70.3	73.6	518	81.4	89.1
238	73.8	74.5	519	79.1	94.9
311	71.4	67.6	521	104.8	81.7
312	89.4	73.0	522	84.3	80.8
313	82.6	70.2	523	122.6	80.2
314	81.5	84.7	524	98.1	86.0
315	87.3	77.4	531	81.2	81.5
316	86.4	87.4	532	66.2	85.4
321	80.6	80.4	533	81.1	78.6
322	78.3	88.9	541	79.8	82.4
323	84.0	82.5	551	74.1	83.5
324	30.4	82.3	561	79.1	77.6
325	61.6	29.2	562	80.2	82.7
326	73.6	74.5	611	72.1	79.1
327	83.8	73.9	621	72.4	81.8
331	42.5	88.4	622	86.2	82.5
332	66.1	47.8	623	82.0	82.3
333	71.7	71.4	624	75.1	77.9
334	101.7	64.9	711	81.1	78.2
335	72.0	105.3	712	78.0	80.6
336	86.3	63.1	713	76.7	81.2
337	80.3	76.7	721	87.9	80.3
339	79.2	85.3	722	80.2	76.2
43-46	74.6	78.2	811	81.8	78.0
481	66.8	78.2	812	81.3	83.2
482	69.8	66.1	813	82.9	79.8
483	94.8	70.9	814	73.6	79.7
484	76.3	133.1	931	75.6	79.6

Fuente: INEGI. *Banco de Información Económica. Índices de precios implícitos base 2008.* INEGI, 2016a.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Anexo B

Participación de los componentes del valor agregado en el valor agregado del 2003 a precios del 2003

SCIAN	VA	*t*1	*W*	*t*2	S	SCIAN	VA	*t*1	*W*	*t*2	*S*
111	1.0000	0.0060	0.1260	0.0000	0.8680	485	1.0000	0.0287	0.3181	-0.0036	0.6568
112	1.0000	0.0021	0.2599	0.0002	0.7378	486	1.0000	0.0035	0.4473	0.0034	0.5458
113	1.0000	0.0030	0.1526	0.0000	0.8444	487	1.0000	0.0527	0.3503	0.0199	0.5771
114	1.0000	0.0309	0.2518	0.0125	0.7048	488	1.0000	0.0039	0.1859	0.0070	0.8032
115	1.0000	0.0049	0.4011	0.0011	0.5930	491	1.0000	0.0022	0.9794	-0.0635	0.0819
211	1.0000	0.0011	0.0406	0.6658	0.2924	492	1.0000	0.0285	0.3976	0.0106	0.5634
212	1.0000	0.0081	0.2165	0.0105	0.7648	493	1.0000	0.0044	0.5557	0.0162	0.4237
213	1.0000	0.0115	0.4478	0.0143	0.5264	511	1.0000	0.0007	0.2309	0.0041	0.7644
221	1.0000	0.0283	0.3261	0.0053	0.6403	512	1.0000	0.0003	0.0599	0.0148	0.9250
222	1.0000	-0.0010	0.5499	0.0124	0.4387	515	1.0000	-0.0014	0.1462	0.0026	0.8526
236	1.0000	0.0057	0.3943	0.0025	0.5975	517	1.0000	0.0067	0.2802	0.0066	0.7065
237	1.0000	0.0102	0.4147	0.0071	0.5680	518	1.0000	0.0013	0.6485	0.0146	0.3356
238	1.0000	0.0069	0.4375	0.0015	0.5540	519	1.0000	0.0004	0.6934	0.0161	0.2901
311	1.0000	0.0032	0.2255	0.0033	0.7680	521	1.0000	0.0002	0.1810	0.0056	0.8131
312	1.0000	0.0068	0.2839	0.0110	0.6983	522	1.0000	-0.0003	0.3304	0.0229	0.6470
313	1.0000	0.0109	0.4716	0.0147	0.5028	523	1.0000	-0.0002	0.4914	0.0106	0.4981
314	1.0000	0.0188	0.4369	0.0070	0.5373	524	1.0000	-0.0003	0.2062	0.0252	0.7688
315	1.0000	0.0134	0.4325	0.0077	0.5464	531	1.0000	0.0001	0.0096	0.0103	0.9800
316	1.0000	0.0064	0.3914	0.0074	0.5948	532	1.0000	0.0031	0.0778	0.0054	0.9137
321	1.0000	0.0080	0.3237	0.0027	0.6656	533	1.0000	0.0000	0.0025	0.0000	0.9975
322	1.0000	0.0105	0.3290	0.0097	0.6508	541	1.0000	0.0022	0.2702	0.0036	0.7240
323	1.0000	0.0051	0.4841	0.0177	0.4930	551	1.0000	0.0021	0.4739	0.1506	0.3734
324	1.0000	0.0179	0.4103	0.0011	0.5706	561	1.0000	0.0011	0.5342	0.0053	0.4594
325	1.0000	0.0115	0.3446	0.0148	0.6291	562	1.0000	0.0096	0.2784	0.0093	0.7027
326	1.0000	0.0076	0.4443	0.0127	0.5354	611	1.0000	0.0002	0.7560	0.0026	0.2412
327	1.0000	0.0099	0.2724	0.0039	0.7138	621	1.0000	0.0015	0.4126	0.0020	0.5839
331	1.0000	0.0168	0.2357	0.0048	0.7428	622	1.0000	0.0017	0.7994	0.0039	0.1950
332	1.0000	0.0072	0.4361	0.0121	0.5446	623	1.0000	0.0019	0.7331	0.0057	0.2593
333	1.0000	0.0078	0.7054	0.0127	0.2741	624	1.0000	0.0041	0.6253	0.0050	0.3655
334	1.0000	0.0028	0.4135	0.0099	0.5738	711	1.0000	0.0004	0.0734	0.0032	0.9230
335	1.0000	0.0067	0.4582	0.0111	0.5239	712	1.0000	0.0001	0.9352	0.0162	0.0484
336	1.0000	0.0059	0.3398	0.0045	0.6498	713	1.0000	0.0021	0.3678	0.0203	0.6097
337	1.0000	0.0088	0.3919	0.0078	0.5915	721	1.0000	0.0012	0.1117	0.0080	0.8791
339	1.0000	0.0081	0.4905	0.0097	0.4918	722	1.0000	0.0018	0.3945	0.0034	0.6003
43-46	1.0000	0.0005	0.2411	0.0101	0.7484	811	1.0000	0.0045	0.2034	0.0056	0.7864
481	1.0000	0.0791	0.6648	0.0466	0.2096	812	1.0000	0.0009	0.0746	0.0019	0.9226
482	1.0000	0.0285	0.3301	0.0192	0.6221	813	1.0000	0.0010	0.4208	0.0029	0.5753
483	1.0000	0.0211	0.5386	0.0041	0.4362	814	1.0000	0.0000	1.0000	0.0000	0.0000
484	1.0000	0.0258	0.3367	0.0000	0.6374	931	1.0000	0.0026	0.9837	0.0075	0.0062

Donde *t*1 es impuestos netos sobre bienes y servicios; *W*, salarios; *t*2, impuestos netos a la producción; y P, excedente bruto de operación.

Fuente: INEGI. *Sistema de Cuentas Nacionales de México. Matriz de insumo-producto del 2003*. INEGI, 2016c.

Anexo C

Consumo intermedio, valor agregado, valor bruto de la producción e importaciones del 2003 a precios del 2008

SCIAN	CI	VA	X	M	SCIAN	CI	VA	X	M
111	73 804	227 745	301 550	92 019	**485**	176 072	220 719	396 791	0
112	112 712	105 032	217 745	2 586	**486**	3 211	6 356	9 568	0
113	4 185	15 478	19 663	701	**487**	1 720	1 883	3 603	0
114	5 380	8 022	13 403	61	**488**	23 894	43 446	67 340	600
115	1 978	4 012	5 989	9 801	**491**	482	2 298	2 780	0
211	76 055	980 154	1 056 209	0	**492**	5 915	27 725	33 640	0
212	33 515	76 228	109 742	16 916	**493**	7 692	4 700	12 393	0
213	40 251	44 969	85 220	0	**511**	11 706	16 431	28 137	0
221	177 144	135 871	313 015	196	**512**	8 629	6 174	14 802	0
222	14 822	35 385	50 207	0	**515**	16 989	11 832	28 822	2 341
236	429 231	568 622	997 853	0	**517**	96 249	119 206	215 455	4 062
237	180 595	134 183	314 778	0	**518**	1 309	1 845	3 154	0
238	41 604	84 334	125 938	0	**519**	708	687	1 395	0
311	679 950	405 815	1 085 765	111 118	**521**	1 333	11 516	12 849	0
312	103 353	82 524	185 877	6 515	**522**	78 809	148 948	227 757	416
313	43 391	18 280	61 672	65 109	**523**	7 791	4 495	12 286	0
314	17 400	12 426	29 826	7 744	**524**	69 775	25 101	94 876	28 969
315	89 653	59 821	149 474	36 418	**531**	99 278	1 199 506	1 298 784	0
316	28 612	17 093	45 705	23 069	**532**	8 447	16 342	24 789	3 766
321	28 971	20 037	49 009	15 420	**533**	376	14 430	14 806	0
322	81 234	34 182	115 416	64 146	**541**	83 437	247 027	330 463	23 936
323	23 483	14 251	37 733	14 663	**551**	17 370	56 110	73 480	0
324	705 872	87 579	793 451	77 656	**561**	65 087	336 249	401 336	939
325	540 042	247 941	787 983	325 973	**562**	2 373	3 425	5 798	0
326	134 583	52 511	187 095	128 683	**611**	52 406	442 838	495 244	0
327	92 453	97 916	190 368	27 842	**621**	29 962	103 659	133 621	0
331	283 090	138 256	421 346	114 007	**622**	55 804	101 981	157 785	0
332	132 348	58 344	190 691	141 380	**623**	753	1 279	2 032	0
333	75 727	55 111	130 838	235 906	**624**	6 229	7 761	13 991	0
334	396 978	92 841	489 819	301 636	**711**	3 187	19 593	22 780	0
335	156 716	60 623	217 339	205 082	**712**	1 230	4 095	5 326	0
336	563 722	200 668	764 390	311 297	**713**	12 733	27 813	40 545	0
337	36 038	28 029	64 068	9 752	**721**	44 524	91 498	136 022	0
339	82 080	40 172	122 251	58 352	**722**	85 317	159 427	244 744	101
43-46	396 190	1 431 039	1 827 229	0	**811**	40 735	51 633	92 368	1 899
481	44 904	16 500	61 404	12 898	**812**	25 664	89 449	115 112	0
482	9 634	15 174	24 807	0	**813**	18 585	31 503	50 087	0
483	6 850	7 869	14 719	0	**814**	ND	50 081	50 081	0
484	121 079	268 197	389 276	0	**931**	135 180	427 605	562 785	0

ND: no disponible.

Fuente: INEGI. *Banco de Información Económica. Cuentas nacionales base 2008*. INEGI, 2016b.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Anexo D

Continúa

Participación de la demanda intermedia y de los componentes de la demanda final del 2003 en el total de la demanda global a precios básicos del 2008 (*consumo intermedio + valor agregado + importaciones = demanda intermedia + demanda final*)

SCIAN	DG	DI	DF	DF	C	G	K	V	E
111	1	0.5159	0.4841	1	0.5354	0.0000	0.0130	0.1645	0.2871
112	1	0.7662	0.2338	1	0.6917	0.0000	0.0891	0.1463	0.0730
113	1	0.8429	0.1571	1	0.7329	0.0000	0.0002	0.1316	0.1353
114	1	0.1518	0.8482	1	0.8888	0.0000	0.0000	0.0000	0.1112
115	1	0.8150	0.1850	1	0.1233	0.0000	0.0000	0.5844	0.2922
211	1	0.5396	0.4604	1	0.0000	0.0000	0.0000	-0.0101	1.0101
212	1	0.7987	0.2013	1	0.0000	0.0000	0.0000	0.0661	0.9339
213	1	0.0000	1.0000	1	0.0000	0.0000	1.0000	0.0000	0.0000
221	1	0.6725	0.3275	1	0.9455	0.0000	0.0000	0.0000	0.0545
222	1	0.7022	0.2978	1	1.0000	0.0000	0.0000	0.0000	0.0000
236	1	0.0037	0.9963	1	0.0000	0.0000	1.0000	0.0000	0.0000
237	1	0.0385	0.9615	1	0.0000	0.0000	1.0000	0.0000	0.0000
238	1	0.8967	0.1033	1	0.0000	0.0000	1.0000	0.0000	0.0000
311	1	0.1847	0.8153	1	0.9377	0.0000	0.0000	0.0182	0.0441
312	1	0.0367	0.9633	1	0.9187	0.0000	0.0000	0.0040	0.0772
313	1	0.7500	0.2500	1	0.5905	0.0000	0.0000	0.0849	0.3247
314	1	0.2910	0.7090	1	0.6664	0.0000	0.0105	0.0132	0.3099
315	1	0.0910	0.9090	1	0.7349	0.0000	0.0000	-0.0011	0.2662
316	1	0.2448	0.7552	1	0.8748	0.0000	0.0000	0.0092	0.1160
321	1	0.8432	0.1568	1	0.6078	0.0000	0.0065	0.0729	0.3127
322	1	0.7436	0.2564	1	0.6956	0.0000	0.0000	0.0769	0.2276
323	1	0.6607	0.3393	1	0.6814	0.0753	0.0000	0.0012	0.2422
324	1	0.6671	0.3329	1	0.8249	0.0000	0.0000	-0.0426	0.2177
325	1	0.5753	0.4247	1	0.7901	0.0000	0.0001	0.0371	0.1694
326	1	0.6777	0.3223	1	0.5761	0.0000	0.0000	0.0678	0.3561
327	1	0.6403	0.3597	1	0.6780	0.0000	0.0018	0.0358	0.2844
331	1	0.7848	0.2152	1	0.0090	0.0000	0.0619	0.0834	0.8132
332	1	0.6043	0.3957	1	0.3757	0.0000	0.1354	0.0206	0.4683
333	1	0.2942	0.7058	1	0.0297	0.0000	0.6416	0.0070	0.3217
334	1	0.3647	0.6353	1	0.1269	0.0000	0.1326	0.0029	0.7376
335	1	0.4075	0.5925	1	0.2258	0.0000	0.0954	0.0091	0.6696
336	1	0.2086	0.7914	1	0.3931	0.0000	0.1581	0.0096	0.4391
337	1	0.1129	0.8871	1	0.5416	0.0000	0.2993	0.0034	0.1557
339	1	0.2862	0.7138	1	0.5079	0.0000	0.0422	0.0087	0.4411
43-46	1	0.3680	0.6320	1	0.7216	0.0000	0.0944	0.0000	0.1840
481	1	0.2936	0.7064	1	0.7682	0.0000	0.0000	0.0000	0.2318
482	1	0.0000	1.0000	1	1.0000	0.0000	0.0000	0.0000	0.0000
483	1	0.0282	0.9718	1	1.0000	0.0000	0.0000	0.0000	0.0000
484	1	0.4268	0.5732	1	0.7007	0.0000	0.0893	0.0000	0.2100
485	1	0.0343	0.9657	1	1.0000	0.0000	0.0000	0.0000	0.0000

Anexo D

Concluye

Participación de la demanda intermedia y de los componentes de la demanda final del 2003 en el total de la demanda global a precios básicos del 2008 (*consumo intermedio + valor agregado + importaciones = demanda intermedia + demanda final*)

SCIAN	DG	DI	DF	DF	C	G	K	V	E
486	1	0.4334	0.5666	1	0.6990	0.0000	0.0898	0.0000	0.2112
487	1	0.0000	1.0000	1	1.0000	0.0000	0.0000	0.0000	0.0000
488	1	0.4440	0.5560	1	0.9345	0.0000	0.0000	0.0000	0.0655
491	1	0.9140	0.0860	1	1.0000	0.0000	0.0000	0.0000	0.0000
492	1	0.6406	0.3594	1	1.0000	0.0000	0.0000	0.0000	0.0000
493	1	1.0112	-0.0112	1	0.7216	0.0000	0.0944	0.0000	0.1840
511	1	0.5899	0.4101	1	0.6498	0.0000	0.3502	0.0000	0.0000
512	1	0.4011	0.5989	1	0.9784	0.0000	0.0216	0.0000	0.0000
515	1	0.0595	0.9405	1	0.9775	0.0018	0.0000	0.0000	0.0206
517	1	0.3330	0.6670	1	0.9871	0.0000	0.0000	0.0000	0.0129
518	1	1.0112	-0.0112	1	0.7007	0.0000	0.0893	0.0000	0.2100
519	1	0.5400	0.4600	1	1.0000	0.0000	0.0000	0.0000	0.0000
521	1	0.9177	0.0823	1	0.0000	1.0000	0.0000	0.0000	0.0000
522	1	0.3318	0.6682	1	0.9946	0.0000	0.0000	0.0000	0.0054
523	1	0.5244	0.4756	1	1.0000	0.0000	0.0000	0.0000	0.0000
524	1	0.3995	0.6005	1	0.7635	0.0000	0.0000	0.0000	0.2365
531	1	0.1433	0.8567	1	0.9999	0.0001	0.0000	0.0000	0.0000
532	1	0.8735	0.1265	1	0.9994	0.0000	0.0000	0.0000	0.0006
533	1	0.1910	0.8090	1	0.0000	0.0000	1.0000	0.0000	0.0000
541	1	0.8305	0.1695	1	0.7054	0.2068	0.0035	0.0000	0.0843
551	1	1.0112	-0.0112	1	0.6990	0.0000	0.0898	0.0000	0.2112
561	1	0.9479	0.0521	1	1.0000	0.0000	0.0000	0.0000	0.0000
562	1	0.2371	0.7629	1	1.0000	0.0000	0.0000	0.0000	0.0000
611	1	0.0056	0.9944	1	0.2338	0.7662	0.0000	0.0000	0.0000
621	1	0.0000	1.0000	1	0.4524	0.5476	0.0000	0.0000	0.0000
622	1	0.0000	1.0000	1	0.2468	0.7532	0.0000	0.0000	0.0000
623	1	0.0000	1.0000	1	0.7221	0.2779	0.0000	0.0000	0.0000
624	1	0.0000	1.0000	1	0.3205	0.6795	0.0000	0.0000	0.0000
711	1	0.0592	0.9408	1	0.9726	0.0274	0.0000	0.0000	0.0000
712	1	0.0048	0.9952	1	0.1485	0.8515	0.0000	0.0000	0.0000
713	1	0.0000	1.0000	1	0.9886	0.0114	0.0000	0.0000	0.0000
721	1	0.2113	0.7887	1	1.0000	0.0000	0.0000	0.0000	0.0000
722	1	0.1028	0.8972	1	1.0000	0.0000	0.0000	0.0000	0.0000
811	1	0.5080	0.4920	1	1.0000	0.0000	0.0000	0.0000	0.0000
812	1	0.0536	0.9464	1	1.0000	0.0000	0.0000	0.0000	0.0000
813	1	0.1869	0.8131	1	1.0000	0.0000	0.0000	0.0000	0.0000
814	1	0.0000	1.0000	1	1.0000	0.0000	0.0000	0.0000	0.0000
931	1	0.0004	0.9996	1	0.0029	0.9971	0.0000	0.0000	0.0000

Fuente: INEGI. *Banco de Información Económica. Cuentas nacionales base 2008*. INEGI, 2016b.

Registro de los nacimientos en México.
Una mirada crítica
de su evolución en las últimas tres décadas

The Birth Registry in Mexico.
A Critical View
of its Evolution During the Last Three Decades

Marta Mier y Terán Rocha* y Víctor Manuel García Guerrero**

El objetivo del artículo es llevar a cabo una evaluación de las modificaciones en las estadísticas de los nacimientos registrados en México y sus entidades federativas en las últimas tres décadas, la cual se realiza mediante el análisis de la temporalidad del registro en las estadísticas vitales y el contraste con otras fuentes de información. En los últimos años, la diferencia entre el número total de nacimientos registrados y el de los que se han inscrito de forma oportuna se ha reducido de manera sustancial como consecuencia de un menor registro tardío. A partir del 2012, el descenso en el número de nacimientos registrados totales y de menores de un año es reflejo, también, de una disminución en la fecundidad. En el periodo analizado, la mejora en las estadísticas se caracterizó por la baja paulatina del tiempo entre el nacimiento y el registro; este avance fue logrado en todas las entidades fede-

The aim of this paper is to evaluate the evolution of birth statistics in Mexico during the last three decades. This evaluation is carried out by analyzing the timing of the registration in the vital statistics and comparing it with other sources of information. In recent years, the difference between the total number of births and the number of births registered in a timely manner has decreased, as a consequence of declining late registration. By 2012, the decrease in the number of total births registered, and of those aged less than one year, is a result of fertility decline. During the period of time analyzed, the improvement in the registry has been characterized by the reduction between time at birth and time at registration. Such improvements have included all of the Mexican States, although timing deficiencies remain in some of them. In general, census data

Nota: proyecto apoyado por el Fondo Sectorial CONACYT-INEGI; agradecemos el apoyo invaluable de Isaías Ramírez Bañales, Yazmín Berenice González Mayorga y Nayeli Rodríguez Mendieta en el manejo de las bases de datos y la elaboración del material gráfico.

* Universidad Nacional Autónoma de México, martamyt@sociales.unam.mx
** El Colegio de México, vmgarcia@colmex.mx

rativas del país, aunque en algunas de ellas subsisten deficiencias en la anotación oportuna; en general, las cifras censales son consistentes con las de las estadísticas vitales. De la evaluación de la cobertura del registro mediante información de las encuestas se observa la creciente inscripción en el tiempo de los menores de cero a cinco años, aunque el aumento no ha sido constante.

Palabras clave: estadísticas vitales; registro de nacimientos; evaluación de la información; fecundidad.

Recibido: 4 de marzo de 2019.
Aceptado: 9 de mayo de 2019.

are consistent with vital statistics. The evaluation of the registry coverage using surveys shows the registration of children under five is increasing over time, although such increase has not been constant.

Key words: Vital Statistics; Birth Registry; Data Assessment; Fertility.

Sunshine Babies/Fox Photos/Getty Images

Introducción

Las estadísticas vitales son el recuento de los hechos que acontecen en la población de un país, como nacimientos, defunciones, matrimonios y divorcios.

La estadística de nacimientos proporciona la frecuencia con la que ocurre este hecho en cierta unidad territorial; además, contiene datos sobre ciertas características, como lugar en el que ocurrió el nacimiento, la condición de sobrevivencia, el sexo del registrado, su orden y las características demográficas y socioeconómicas de los padres. La importancia de este tipo de información radica en que, a partir de ella, es posible analizar la dinámica de la población y proponer políticas que permitan el desarrollo económico y social (INEGI, 1997). Por su naturaleza, dichas estadísticas son la fuente de información más adecuada para la medición y el estudio de la fecundidad, siempre y cuando sean oportunas y precisas.

En México, el Instituto Nacional de Estadística y Geografía (INEGI) es el encargado de la generación de las estadísticas vitales, mientras que el Registro Civil es la institución del Estado responsable de dar constancia de los hechos vitales a los ciudadanos e informar al INEGI sobre los mismos; también se encarga de "…la inscripción continua, permanente y obligatoria de los hechos vitales…" (INEGI, 1997), lo cual permite asegurar que las personas tengan una identidad legal y puedan acceder a los beneficios y protecciones del Estado (Naciones Unidas, 2016).

El derecho a la identidad, que se obtiene mediante la inscripción, se reconoce como un derecho humano en la *Declaración universal de derechos humanos*; en la Convención sobre los Derechos del Niño se establece que "…el niño será inscrito inmediatamente después de su nacimiento y tendrá derecho a un nombre, a una nacionalidad, a conocer a sus padres y a ser cuidado por ellos…" (UNICEF, 2002; UNICEF-INEGI, 2012). De igual manera, en la *Agenda 2030 para el desarrollo sostenible* se enfatiza la necesidad de tener información precisa y oportuna de las estadísticas vitales; en los *Objetivos de Desarrollo Sostenible*, el número 16 reconoce una relación directa entre los sistemas de registro civil y de estadísticas vitales, la identidad legal y el desarrollo sostenible (Naciones Unidas, 2016).

Como organismo recolector de información estadística, el Registro Civil ha tenido limitaciones por su complejo funcionamiento y su vinculación con otras instituciones de la administración pública, pero se ha reconocido la importancia de sus funciones y sus aportes al conocimiento de la dinámica de la población (Gaete-Darbó, 1965). En México, una de las principales deficiencias es que la inscripción de los nacimientos y el trámite del acta pueden ocurrir meses e, incluso, años después de que ocurre el evento. Además, la complejidad del registro se acentúa al ser responsabilidad tanto de las autoridades federales como de las estatales y, en ocasiones, municipales; los reglamentos y la organización pueden variar de una entidad a otra (RENAPO, 2012).

Los organismos internacionales han jugado un papel relevante en la promoción de una mejora en esta actividad en México (BID, 2006; Gaete Darbó, 1974; Harbitz y Arcos, 2013; Naciones Unidas, 1974, 1985 y 2003; OEA, 2010a y 2010b; UNICEF-INEGI, 2012); se ha trabajado para homogeneizar la información captada en los registros civiles y se ha promovido la eficiencia y exactitud de sus sistemas y las estadísticas vitales (INEGI, 2015).

Con el objetivo de coordinar las actividades y reducir el número de ciudadanos sin registrar, los organismos internacionales han propuesto un programa de modernización, con énfasis en las poblaciones más vulnerables: indígenas, migrantes y marginadas (UNICEF, 2002).

En México, con anterioridad, se habían elaborado diagnósticos y diseños para corregir las anomalías y las carencias en el sistema de registro (López, 1988). Al final de la década de los 90, en el seno del Registro Nacional de Población

(RENAPO), se propuso la modernización del sistema por medio de la revisión del marco jurídico, ya que los códigos civiles de las entidades federativas trataban de manera distinta desde los formatos para captar la información hasta los requisitos para celebrar registros extemporáneos (Herrero, 1998). Se planteó la revisión de las estructuras administrativas porque no coincidían las jerarquías de los empleados entre los estados, y en las oficialías los salarios no estaban contemplados, lo que propiciaba comportamientos irregulares en cuanto a dádivas e inscripciones fraudulentas. También, se señaló la necesidad de dignificar las oficinas, así como la formación y la capacitación del personal permanente mediante programas continuos. Se planteó la necesidad de la automatización, ya que en muchas entidades el procedimiento seguía siendo manual con los consecuentes errores operativos y el nulo control de dobles inscripciones, además de que ninguna de las entidades federativas tenía estudios sobre la calidad del registro.

En años recientes, después de grandes esfuerzos por atender la problemática de la mejora del Registro Civil, el discurso ha cambiado: se plantea que garantizar el registro universal, oportuno y gratuito de los nacimientos es una obligación del Estado (INEGI, 2015).[1] Se mencionan, también, las modificaciones de las prácticas que se han llevado a cabo en las entidades federativas y que han tenido resultados muy positivos en la inscripción oportuna de los niños (SEGOB, DIF y UNICEF, s/f; UNICEF-INEGI, 2012).

En este marco, en el presente trabajo se evalúa de manera sistemática la evolución de las estadísticas de los nacimientos registrados en el país. Dicho análisis se lleva a cabo utilizando la temporalidad del registro y el contraste con otras fuentes de información.

Fuentes de datos y principales trabajos sobre la enumeración de los nacimientos en México

En la actualidad, las estadísticas vitales no son la única fuente de datos para la enumeración continua de los nacimientos en el país. Desde el 2008 se cuenta, también, con el certificado de nacimiento del Subsistema de Información sobre Nacimientos (SINAC) recabado por la Secretaría de Salud. El objetivo de esta nueva fuente es integrar información de los nacidos vivos ocurridos en el país a partir de este documento, el cual certifica el hecho cuando acontece a la vez que garantiza la identidad del niño y proporciona certeza legal de la relación madre-hijo. Sin embargo, como las mismas autoridades lo reconocen, su principal limitación es la dificultad para captar los nacimientos que ocurren fuera de alguna unidad médica, por lo que sufren de subestimación, sobre todo en las regiones con menor acceso a los servicios de salud.[2]

La información censal también proporciona el número de nacimientos que ocurren en el país. Esta fuente no es continua, pero, a partir del dato sobre la fecha de alumbramiento del último hijo nacido vivo, permite obtener estimaciones de los ocurridos en los 12 meses anteriores a los levantamientos del 2000, 2010 y 2015. Los datos censales adolecen de limitaciones como: omisiones de mujeres con hijos pequeños y, en particular, de niños pequeños, así como distorsiones en la ubicación temporal de los nacimientos; lo anterior, más el hecho de que quien proporciona la información en los censos puede ser una tercera persona, no necesariamente la madre, han llevado a que se ignore esta fuente en las estimaciones oficiales de la fecundidad (Hernández *et al*., 2015). No obstante, varios trabajos la han empleado con éxito (CONAPO, 2005; UNICEF, 2002; UNICEF-INEGI, 2012), como el que se presenta ahora, donde se obtuvieron resultados valiosos y consistentes, lo cual se verá más adelante.

1 El registro oportuno del nacimiento es el que se realiza ante la autoridad del Registro Civil durante el periodo establecido por la ley. En México, el plazo difiere entre las entidades federativas; en la mayoría de ellas se define como 180 días a partir del nacimiento, pero hay otras en las que es de un año (UNICEF-INEGI, 2012).

2 http://www.dgis.salud.gob.mx/contenidos/sinais/s_sinac.html, consultado el 28 de septiembre de 2018.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Otras fuentes para evaluar la cobertura de las estadísticas vitales son la Encuesta Nacional de la Dinámica Demográfica (ENADID), realizada en 1992, 1997 y el 2014, así como la Encuesta Intercensal (EI) 2015, las cuales proporcionan datos sobre la población que cuenta con acta de nacimiento. Con ellas se han realizado evaluaciones de la calidad de las estadísticas vitales.

En algunas investigaciones se han señalado las limitaciones que subsisten y deben subsanarse pero, en general, se coincide en reconocer mejoras notables en la calidad de las estadísticas de nacimientos en el país (García, 2016; Hernández *et al*., 2015; Mier y Terán, 2013; UNICEF-INEGI, 2012; Galindo y Ordorica, 2007).

De acuerdo con Figueroa (1998), la información más antigua sobre la edad en el registro de los nacimientos en el país es de 1933, cuando de los 737 mil captados, 94.2% correspondía a menores de 1 año. Con el tiempo, al aumentar la inscripción de las personas que no habían sido registradas previamente de manera oportuna, la autora afirma que la proporción de nacimientos anotados de niños menores de 1 año en las décadas de los 60 y 70 tiende a reducirse hasta alcanzar un valor mínimo de 67.4% en 1974, que coincide con la puesta en marcha de una campaña para propiciar el registro. También señala cúspides en el asiento de los nacimientos en las edades de ingreso al sistema escolar y salida de la primaria, así como cerca de los 18 años; menciona que esta última pudiera estar motivada por el servicio militar o el matrimonio.

El registro múltiple es una de las deficiencias de las estadísticas vitales más difícil de evaluar. De manera directa, solo en dos ediciones de la ENADID (1992 y 1997) se ha indagado sobre la edad al momento del primer registro y el número de veces que la persona ha sido registrada. Los resultados son difíciles de evaluar porque, en ambos casos, la inscripción múltiple es menor a 1% y la falta de respuesta es cerca del doble. De acuerdo con el Consejo Nacional de Población (CONAPO, 2005:66) "…los nacimientos de las estadísticas vitales adolecen de registro múltiple y se deben de tener precauciones al usar esos datos para estimar los niveles y tendencias de la natalidad y la fecundidad…".[3] Con la información disponible no es posible obtener estimaciones de la magnitud de este hecho para cohortes más recientes. No obstante, se espera que esta práctica se haya reducido de forma sustancial en el tiempo con la modernización del Registro Civil, que facilita la obtención de las actas en entidades distintas a las de nacimiento y el incremento de requisitos para el registro extemporáneo, en especial, la constancia de inexistencia y, en fecha más reciente, la puesta en marcha del certificado de nacimiento, que es un documento requerido para el registro (INEGI, 2015; 15).

Galindo y Ordorica (2007) proponen una serie histórica de nacimientos que cubre la segunda mitad del siglo pasado; afirman que la subenumeración de los ocurridos en ese periodo se debe, sobre todo, a la mortalidad y la migración, y plantean el problema del registro múltiple.[4] También, aplican una función cuadrática a datos de las estadísticas vitales para estimar los nacimientos registrados de niños menores de 1 año, lo que podría asimilarse con los ocurridos, y las cifras que obtienen van de 1.2 millones en 1950 a 2.4 millones en 1985 y a 2.5 millones en el 2000.

Con el objetivo de obtener estimaciones de la fecundidad para el país con datos de las estadísticas vitales (1990-2013), Hernández y colaboradoras (2015) realizaron un cálculo del número de los nacimientos anuales mediante la reconstrucción de las cohortes con un seguimiento del registro hasta antes de que cumplan 8 años de edad. Señalan que la proporción del registro en el primer año de vida varía entre 70 y 75% en las cohortes nacidas hasta el 2006; a partir del siguiente año, esta proporción tiende a aumentar hasta ser cercana a 90% en la cohorte nacida en el 2013. Las autoras contrastan sus estimaciones con datos del SINAC (2008-2014) y observan una creciente cobertura de este último,

3 El autor del trabajo menciona ciertas cifras, pero advierte que lo hace solo con fines ilustrativos: el registro múltiple originaría una sobreestimación de los nacimientos de la cohorte 1985 del orden de 6.5 a 10.6% (CONAPO, 2005:66).

4 Para el periodo 1950-1980, por ejemplo, plantean una subenumeración de los nacimientos por mortalidad y migración de 8.1% (Galindo y Ordorica, 2007).

de manera que la diferencia entre las dos fuentes es 4.5% en el año más reciente; como conclusión, mencionan que se requiere contrastar los resultados de la reconstrucción de los nacimientos a partir de las estadísticas vitales con los de otras fuentes para tener mayor certeza sobre su enumeración y las estimaciones de la fecundidad.

Investigaciones recientes evalúan la cobertura del registro mediante la confrontación de los datos de las estadísticas vitales con los de los censos de población (UNICEF-INEGI, 2012; García, 2016). Se compara el número de nacimientos registrados de menores de 1 año de edad en 1999, 2009 y 2014 de las estadísticas vitales con el de los calculados con la información censal de la fecha de nacimiento del último hijo nacido vivo en México y las entidades federativas (García, 2016). Se señala que, en el país, el registro oportuno de los nacimientos en el Registro Civil aumentó de forma notable: de 78.7% en 1999, a 93.4% en el 2009 y a 94.5%,[5] en el 2014. Una limitación de estas estimaciones estriba en que los nacimientos captados en el censo para el año anterior al del levantamiento corresponden al total del año, lo que sería correcto si los censos tuvieran lugar en enero, pues la técnica es pertinente para los 12 meses inmediatos anteriores al levantamiento censal (Moultrie *et al.*, 2013; Spoorenberg, 2015). Alejarse de esta ventana temporal propicia que algunos de los nacimientos ocurridos no sean de últimos hijos, de manera que no son contabilizados. Dado que el Censo del 2000 se llevó a cabo en febrero, el del 2010 en junio y la EI 2015 en marzo, la estimación de los nacimientos para el 2009 es la más afectada, lo que origina una sobreestimación del registro oportuno para este año.

En la EI 2015, así como en otras encuestas, se indaga acerca de la situación del registro de las personas, lo que permite evaluar el grado de cobertura de los nacimientos en las estadísticas vitales,[6] por ejemplo, en la ENADID 1997, para todos los hijos de las mujeres en edades reproductivas, se indagó sobre su condición de registro, la edad y el número de veces que la persona había sido registrada; en la del 2014, además, se preguntó sobre la obtención del certificado de nacimiento, lo que permite hacer la vinculación de la cobertura de las dos fuentes. A diferencia de estas, en la EI 2015 se incluyó para todas las personas una pregunta sobre si contaban con acta de nacimiento o estaban inscritas en el Registro Civil.

Con base en los datos de la EI 2015, García (2016) señala que una alta proporción de niños menores de 6 años cuenta con acta de nacimiento (96.4%), la cual varía entre las entidades federativas (de 90.3% en Chiapas a 98.7% en Querétaro), de manera que en todos los estados el registro de los niños de estas edades sería superior a 90%; no obstante, cabe mencionar que la falta de respuesta en esta fuente no es despreciable, ya que se desconoce la condición de registro de 2% de los niños menores de 6 años en el país.

Más adelante analizamos con mayor detalle estas cifras, pero antes pasamos a evaluar los datos de los nacimientos registrados mediante el análisis de la coherencia interna de las estadísticas vitales. De esta manera, se analizan patrones en los datos para conocer las prácticas del registro de los nacimientos y evaluar en qué medida se aproximan a la situación deseada de un registro universal y oportuno que provea de datos confiables sobre los nacimientos que ocurren en el país. En este trabajo, nos referimos al registro oportuno como aquel que acontece antes del primer aniversario.[7]

Resultados. El registro de los nacimientos en México

La calidad de los datos puede evaluarse con varios criterios, entre los cuales están la relevancia, precisión (cobertura, ausencia de respuesta, etc.), oportunidad y accesibilidad (Naciones Unidas, 2016). La evaluación de la coherencia en el tiempo y el contraste con otras fuentes son formas que permiten atender especialmente al de la precisión.

5 Los límites de confianza al 90, 91.8 y 97.4 por ciento.

6 INEGI. *Tabulados de la Encuesta Intercensal 2015*. // INEGI. *Encuesta Nacional de la Dinámica Demográfica*. Ediciones 1997 y 2014.

7 Ver nota 1.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Ausencia de respuesta en el registro de los nacimientos

El análisis de la ausencia de respuesta, o de valores faltantes, se llevó a cabo de acuerdo con el año en el que se registraron los nacimientos, con la ventana temporal de tres décadas (1985 al 2016) que reporta el INEGI de los registros administrativos correspondientes (ver gráfica 1).[8] Evaluamos los valores perdidos en 13 de las variables sociodemográficas más relevantes para el estudio de la fecundidad captadas en el registro de los nacimientos y presentadas por el INEGI (2015). De ellas, las fundamentales para la medición de la fecundidad (como año del nacimiento, edad al registro y sexo), prácticamente no tienen valores perdidos. En ningún caso, después de 1987 se alcanzó 0.1% de valores no especificados; de hecho, a partir del 2014, la proporción de estos en las tres variables es inferior a 0.02 por ciento. Por otro lado, para la variable que señala si el registrado fue presentado vivo, solo se cuenta con información a partir de 1992 y, después de este año, se observa una mejoría en su registro, salvo en 1994 cuando se desconoce si estaban vivos o no 3% de los anotados.[9] En los últimos años, dicha variable tiene muy pocos casos no especificados. Lo anterior indica que estas cuatro variables fundamentales para el análisis de la fecundidad están bien captadas en la fuente y son las mejor recabadas.

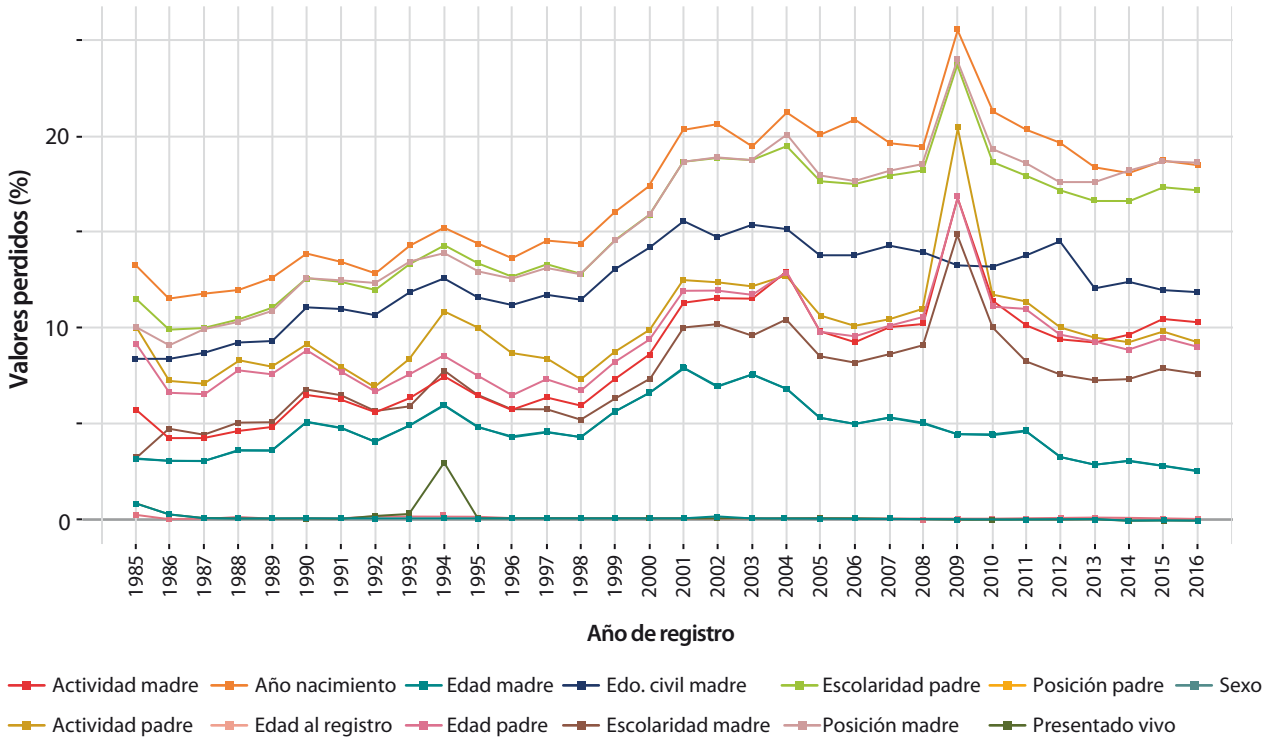
La edad de la madre al nacimiento también es una variable clave para la estimación de la fecundidad y, en general, está bien captada. En la década

8 En 1985, el RENAPO y el INEGI establecieron un convenio para que el Instituto tuviera acceso a una copia del acta para el proceso de producción estadística. Esto, y el empleo del formato único para captar los nacimientos, permitió mayor facilidad en la crítica y la validación de información, así como la inclusión de datos acerca de las características socioeconómicas de los padres del nacido vivo. La copia del acta fue la base para la elaboración de las estadísticas a partir de 1986 (INEGI, 2015, p. 18).

9 En 1994 hubo un cambio en el instrumento de captación, lo que pudo haber ocasionado esta alta proporción. El formato adoptado en este año sigue vigente hasta ahora (INEGI, 2015, p. 18).

Gráfica 1

Proporción de nacimientos registrados con ausencia de respuesta (valores perdidos) en variables utilizadas en el estudio de la fecundidad, 1985-2016



Fuente: elaboración propia con base en INEGI. *Estadísticas vitales. Nacimientos.*

de los 80, los valores perdidos eran inferiores a 5%, pero la captación empezó a deteriorarse en la década de los 90 y alcanzó los valores más altos de omisión al inicio del siglo XXI, cuando la proporción aumentó a casi 8 por ciento. A partir del 2004, el registro de la edad de la madre tendió a mejorar hasta llegar a tener en el 2016 solo 2.5% de valores perdidos. De otras variables de la madre, la escolaridad tuvo la misma tendencia que la edad, aunque en niveles más elevados, alrededor de 10% en los primeros años del presente siglo y varió en torno a 8% en los últimos años observados.

El estado civil y la condición laboral de la madre, así como las variables del padre, tuvieron una captación más deficiente. El periodo en el que este aspecto del registro adoleció de mayores problemas fue a partir de 1999 y continuó durante el siglo actual. En el 2009 hubo un aumento notable de valores faltantes en la escolaridad y actividad de la madre, así como en todas las características del padre.[10] Si eliminamos el dato atípico del 2009, la omisión de la información de la actividad económica alcanza 13% de valores faltantes; de las variables del padre, la edad llega a tener 16% de no especificados; la escolaridad y la actividad económica, 20%; y la posición en el trabajo, 21 por ciento.

Número total de nacimientos registrados y número de nacimientos con registro oportuno

La cantidad de nacimientos que se registraron a partir de 1985 varió entre 2.3 millones y 2.9 millones, con una tendencia creciente en los primeros 10 años observados y, en general, una al descenso a partir de entonces hasta el 2016 (ver gráfica 2). Se observa que la magnitud de los nacimientos registrados fue muy superior a la de los menores de 1 año, es decir, los que tuvieron un asentamiento oportuno, oscilaron en torno a los 2 millones; en varios años, la diferencia fue de más de medio millón de nacimientos, lo cual equivalió a entre una quinta y una cuarta parte de los registros anuales. En los últimos años, la diferencia entre las dos cifras se ha reducido de manera sustancial como consecuencia de un menor registro tardío. Esta tendencia a converger se observa con claridad en la proporción de nacimientos con registro oportuno, que parte de valores algo inferiores a 80%, pero que en la década de los 90 y hasta el 2003 tenía valores más bajos como consecuencia del creciente registro de personas que no se habían inscrito en años anteriores. A partir del 2004, la proporción empezó a aumentar hasta alcanzar valores de 87% en los últimos años, lo que representó una mejora de 10% en el registro oportuno entre 1985 y el 2016.

Una posible explicación del número elevado de los nacimientos es el registro múltiple. En la propuesta de modernización del sistema del registro de la población en el país (Herrero, 1998), se mencionaba que las deficiencias en el Registro Civil podrían estar dando lugar a este problema. Como se mencionó, esta práctica es difícil de evaluar, solo se tienen evidencias de que ha existido (CONAPO, 2005) y que debe haber afectado durante el periodo aquí analizado, en particular antes de dicha modernización y en ciertas entidades federativas de México. La reducción en la brecha entre el número total de nacimientos anotados y el de los de menores de 1 año que se observa en años recientes apoya esta hipótesis.[11]

Temporalidad del registro: una aproximación transversal

El principal error de cobertura en las estadísticas vitales del país ha sido el registro de las personas después de su primer año de vida. Aunque hay quienes no son anotados en toda su existencia, una proporción importante se lleva a cabo después

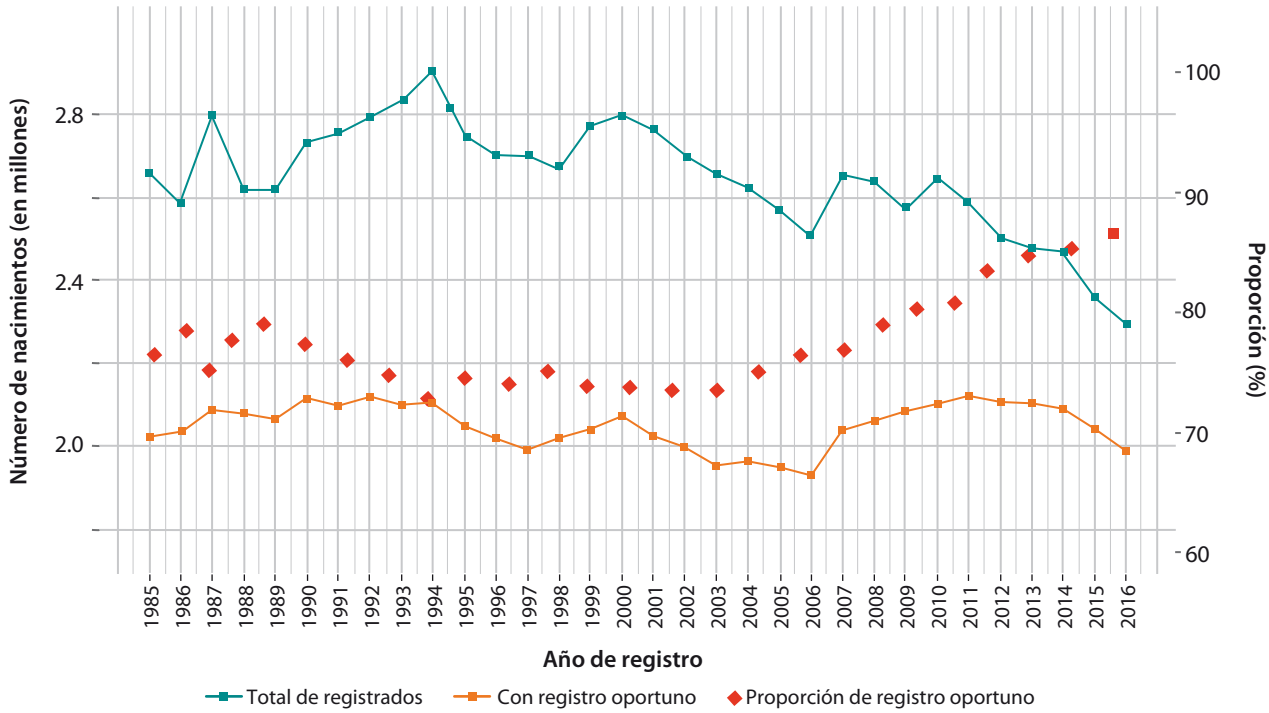
10 En el 2001, 2002, 2003, 2008, 2009 y 2017 han sucedido cambios en los requerimientos del INEGI para la automatización del Registro Civil, en los que se especifican las variables y las categorías en cada una de ellas —INEGI (DE) https://www.inegi.org.mx/programas/natalidad/ e INEGI, 2015, pp. 32-35—; es probable que su implementación haya afectado la calidad de la información porque estos años coinciden con el aumento en los valores faltantes de algunas de las variables.

11 Una posible causa de cierta sobreenumeración de los nacimientos es el registro de personas que no nacieron en el país ni son hijos de padres mexicanos; sin embargo, no es posible conocer su magnitud ni su tendencia en el tiempo porque obedece a múltiples factores.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Gráfica 2

Nacimientos registrados totales y con registro oportuno, así como proporción de registro oportuno, 1985-2016



Fuente: elaboración propia con base en INEGI. *Estadísticas vitales. Nacimientos.*

del primer aniversario. La captación inmediata después de acontecido el nacimiento no ha sido una práctica común, por lo que el contacto con las instituciones gubernamentales que requieren el acta de nacimiento ha sido la motivación para realizar esta acción de una gran parte de la población. El ingreso a la escuela es uno de los motivos principales, de manera que la expansión del sistema educativo ha propiciado la mayor cobertura y el registro en edades cada vez más tempranas, en particular con la obligatoriedad del nivel preescolar.[12] La participación en programas sociales de gobierno y otros eventos (como una posible migración o el matrimonio) también han favorecido que se lleve a cabo el registro.

Los cambios en el tiempo pueden ser de orden coyuntural (campañas para el registro o inicio de la obligatoriedad del preescolar) o no coyunturales (como el proceso de modernización del sistema del registro o la creciente necesidad de tener un acta, que ha sido cada vez más temprana). Los primeros ocasionan modificaciones que pueden ser reversibles, mientras que los segundos implican tendencias que persisten; ambos son de interés para tener más elementos sobre los patrones del registro de los nacimientos.

En cuanto a la temporalidad, una primera aproximación a su evolución se obtiene mediante el número de nacimientos registrados, de acuerdo con su edad en el momento del registro y año en que fueron anotados. Así, en la gráfica 3 se muestra el monto de los nacimientos con registro oportuno y se desglosan las edades en las que ocurren las inscripciones de los demás. Se observa que el primer caso se ha mantenido con leves variaciones a lo largo del tiempo, oscilando en alrededor de 2 millones. Después de un máximo relativo en el 2000, ocurre un leve descenso hasta el 2006, pero en el año siguiente da co-

12 El tercer año de preescolar fue obligatorio a partir del ciclo 2004-2005; el segundo, desde el ciclo siguiente; y el primero, en el ciclo 2008-2009.

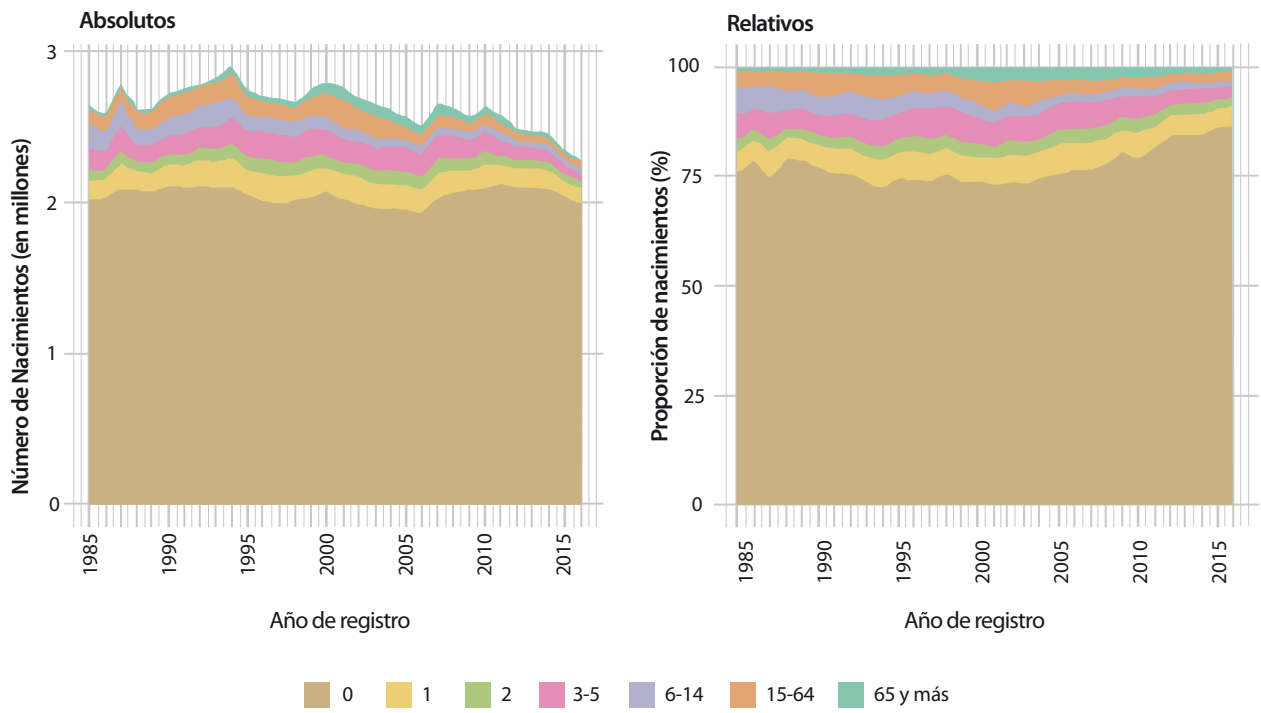
mienzo un repunte que termina en el 2011, para iniciar otra caída en los últimos años.

De esta manera, las variaciones en el monto total de nacimientos captados en el periodo tienen su origen sobre todo en los registros después del primer año de vida; por ejemplo, la reducción entre los dos primeros años observados se debe exclusivamente a la disminución del registro de mayores de 1 año y el máximo relativo en 1987 es, en especial, por el mayor asiento de niños de 1 a 14 años. La tendencia creciente en los siguientes periodos tampoco está relacionada con los registros en el primer año de vida, sino con el de las demás edades. La cúspide en 1994 se debió a un mayor asentamiento de niños de 1 a 5 años y de edades mayores a 14 años. Un caso distinto es el máximo relativo en el 2007, en el que tanto el registro de los de más edad como de los menores de 1 año tuvo un repunte; en estos últimos el alza continuó unos años, probablemente favorecida por la puesta en marcha del certificado de nacimiento. En el 2010 hubo un máximo relativo que se debió a un mayor registro generalizado, pero más notable en los mayores de 1 año. Un hecho que destaca tanto en términos absolutos como relativos es el incremento de registros del grupo etario 15-64 entre 1999 y el 2003, lo que coincidió con el incremento de la emigración mexicana en ese mismo periodo y con la puesta en marcha de algunos programas sociales.[13] A partir del 2012 (y sobre todo en los dos últimos años), el descenso en el número de nacimientos registrados totales y de menores de 1 año es reflejo de una baja de la fecundidad, una inercia demográfica que empieza a ceder a un crecimiento más lento y una población que, en su gran mayoría, cuenta con acta de nacimiento.

13 Por ejemplo, el Programa Nacional de Desarrollo Social 2001-2006, *Superación de la pobreza: una tarea contigo* de la Secretaría de Desarrollo Social consultado el 24 de noviembre de 2018 en https://www.gob.mx/cms/uploads/attachment/file/13832/PNDS_2001_2006rr.pdf

Gráfica 3

Nacimientos según edad en el registro, 1985-2016



Fuente: elaboración propia con base en INEGI. *Estadísticas vitales. Nacimientos.*

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

El número de personas registradas en las edades de 1, 2 y 3 a 5 años varió a lo largo del periodo en torno a 5, 3 y 5%, respectivamente; ahí se encontraban quienes se inscribirían para su ingreso a la educación preescolar. Las personas anotadas en edades escolares de 6 a 14 años iniciaron el lapso con valores alrededor de 5%, proporción que bajó muy rápido hasta ser cercana a 2% en el 2004 y disminuyó aún más en los años recientes; el alto porcentaje del inicio está vinculado al ingreso a la primaria y su tendencia refleja el ingreso cada vez más temprano de los niños al sistema educativo. Las personas que se registraron en las edades laborales de 15 a 64 variaron de forma considerable: de valores cercanos a 4% en los primeros años analizados aumentaron a 6% al principio del siglo XXI y, después, disminuyeron hasta ser solo 1% en el 2016; en este amplio grupo de edades pudo haber motivaciones distintas, desde el ingreso al trabajo y el servicio militar o algún programa social, hasta contraer matrimonio o migrar, y su reducción se debió, sobre todo, a que las personas ya habían sido registradas en edades más tempranas. En las mayores (de 65 años o más), el registro estuvo vinculado principalmente a los programas sociales dirigidos a este grupo de población y a la necesidad de realizar algunos trámites, como el de las herencias; fueron pocas las personas que se registraron en estas edades, pero su número aumentó a partir de 1999 hasta el 2001, cuando alcanzó su máximo (3%) y permaneció más o menos elevado hasta el 2011; en los últimos años es inferior a 1 por ciento.[14]

Temporalidad del registro en las cohortes de nacimiento

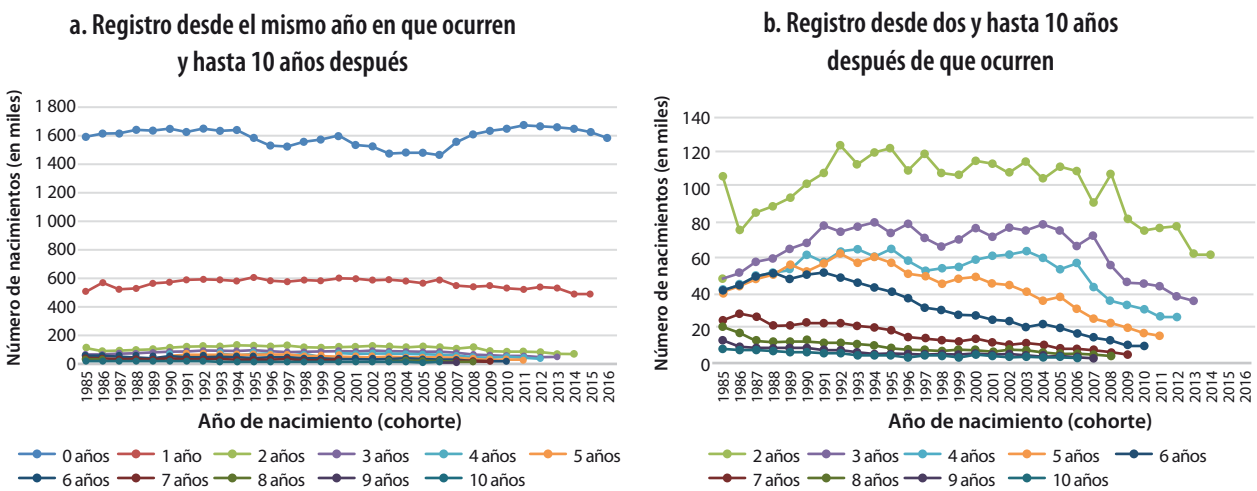
Otra forma de evaluar la temporalidad del registro de los nacimientos ocurridos es mediante el seguimiento del registro de las cohortes, ya sea que se lleve a cabo en el mismo año que nacen o en los subsecuentes, en edades más tardías (ver gráfica 4). Los resultados de esta visión longitudinal corroboran lo que se señaló en la gráfica anterior, pero sugieren nuevos elementos en cuanto a patrones de registro, ya que las cohortes viven en distintos momentos de su vida los cambios que ocurren en determinados años.

Los nacimientos registrados el mismo año fueron más numerosos en las primeras cohortes ob-

14 El máximo en el 2001 coincide con el inicio del Programa de Apoyo Alimentario, Atención Médica y Medicamentos Gratuitos para adultos de 70 años de edad y más del gobierno de la Ciudad de México, el que en el 2007 tenía una cobertura de más de 400 mil personas (IAAM-DF, 2008).

Gráfica 4

Número de nacimientos según año en el que ocurren (cohorte) y número de años que transcurren hasta su registro, 1985-2016



Fuente: elaboración propia con base en INEGI. *Estadísticas vitales. Nacimientos.*

servadas (cercanos a 1.6 millones), al igual que en la del 2000 y, a partir de la cohorte nacida en el 2007 (ver gráfica 4a). Destaca que el repunte en la de los nacidos en el 2000 ocurrió casi de manera exclusiva en esta curva y no en la de los registrados más tarde, mientras que el aumento a partir de la del 2007 coincidió con la puesta en marcha del certificado de nacimiento y pareciera ser reflejo del registro cada vez más cercano al nacimiento, ya que en las curvas de los asentados en años subsecuentes no se observa repunte alguno en estas cohortes, más bien lo contrario. El número de nacimientos registrados el mismo año alcanzó un máximo en la cohorte 2011 que correspondió a 1.7 millones. A partir de la del 2013 hubo una leve reducción que se acentuó en las del 2015 y 2016, ya como resultado de una inercia demográfica asociada a un crecimiento más lento y al descenso de la fecundidad.

Durante todo el periodo observado, algo más de 500 mil nacimientos se registraron al año siguiente de que ocurrieron; entre ellos, algunos eran menores de 1 año y otros ya lo habían cumplido (ver gráfica 4a). Esta cifra tuvo leves repuntes y, a partir de la cohorte 2007, descendió de forma pausada, baja que se acentuó en las cohortes 2014 y 2015. La tendencia de los últimos años se explica por registros cada vez más cercanos al nacimiento y por la reducción en su número.

Los nacimientos registrados dos o más años después fueron menos numerosos, pero tuvieron patrones interesantes, por lo que en la gráfica 4b presentamos la amplificación de las curvas de los nacimientos asentados de dos y hasta 10 años después de acontecidos. Los captados dos años después fueron del orden de 100 mil en las cohortes más alejadas, con un descenso sustancial en la de 1986, asociado a un aumento de los registrados un año antes y, en fecha más reciente, un aumento hasta alcanzar un máximo en la cohorte 1992 (registrados en 1994) cuando, con irregularidades, empezó una leve tendencia al descenso que se acentuó a partir de la cohorte 2009.

En todos los casos, pero en particular en los nacimientos asentados dos o más años después, persiste una tendencia en las distintas curvas que consiste en un mayor número de registros en 1994, patrón que ya habíamos mencionado en el análisis transversal, pero que llama la atención cómo va afectando el registro de las distintas cohortes y que al siguiente año se captan en menor medida, mostrando el carácter coyuntural de la mejora en 1994. Otra pauta que se repite es el menor número de registros de las diferentes cohortes en el 2009 y la mayor cantidad en el 2010, cuando hay un repunte de los nacimientos registrados totales.

Los nacimientos asentados tres años después de ocurridos son del orden de 50 mil y tienen un sensible aumento hasta ser de cerca de 80 mil, pero, a partir de la cohorte 2008, descienden hasta ser cercanos a 35 mil nacimientos en la del 2013. Con similar tendencia a la de los registrados tres años después, la curva de los de cuatro años después de ocurridos aumenta en las cohortes nacidas del 2000 al 2003, cuando se inscriben para ingresar a la educación preescolar; en la cohorte 2004 inicia un descenso que se acentúa en las más recientes. El registro cinco años después, asociado al ingreso a la primaria se reduce a partir de la cohorte nacida en 1995, mientras que los de seis años después descienden de manera continua desde la cohorte nacida en 1992.

Resulta interesante observar cómo el registro es cada vez menos desfasado conforme las cohortes son más recientes; por ejemplo, a partir de la de 1986, disminuye el registro ocho y nueve años después de ocurrido; en la de 1991, inicia un descenso casi sin interrupción del registro siete años después, es decir, de niños de 6 y 7 años de edad. En las primeras cohortes observadas, el número de registros cuatro, cinco o seis años después de ocurridos es el mismo en cada caso; a partir de la cohorte nacida en 1989, la inscripción seis años después empieza a ser menos común y hasta la del 2001, el registro cinco años después inicia su reducción. Los inscritos cuatro años después descienden de manera casi ininterrumpida a partir de la cohorte 2007, que fue la primera registrada por el SINAC.

Así, en el periodo analizado, la mejora se caracterizó por la reducción paulatina del tiempo entre

el nacimiento y el registro y, a partir de la cohorte 2007, el que se realiza en el mismo año empieza a aumentar de manera sensible. Los registros uno o más años después de ocurridos inician un descenso casi ininterrumpido a partir de la cohorte 2011.

Enumeración de los nacimientos ocurridos mediante el seguimiento del registro de las cohortes

Uno de los aspectos cruciales para utilizar los datos de las estadísticas vitales en la estimación de la fecundidad es definir hasta cuántos años después de ocurridos se da seguimiento a las cohortes para conocer su magnitud, ya que se desprenden niveles un tanto distintos según el tiempo que dure este seguimiento. En investigaciones anteriores se ha optado por duraciones que van de los 3 años de edad (Pérez y Meneses, s/f) a los 4 años después de acontecido el nacimiento (Mier y Terán, 2013) y hasta los 7 años de edad (Hernández *et al*., 2015). En este trabajo, nuestro interés es abonar a la discusión sobre la pertinencia de estas duraciones de seguimiento; proponer cifras del número de nacimientos ocurridos está fuera de sus alcances.

En la gráfica 5 acumulamos las cifras del seguimiento de las cohortes hasta su registro en las distintas duraciones después de acontecidos; en cada línea se representa a los registrados hasta cierto número de años después de acontecidos, desde el siguiente año en que ocurren hasta 10 años después.[15 y 16] Se observan dos grandes tendencias: en las primeras cohortes y hasta la de 1992, un número creciente de nacimientos y, a partir de entonces, una de largo aliento al descenso que se ve interrumpida en las cohortes en torno al 2000 y, más tarde, a partir de la del 2007 y hasta la del 2012. Estas tendencias se observan de manera más o menos acentuada en todas las curvas. El aumento de las cifras de las primeras cohortes se explicaría por un mayor impacto de la inercia demográfica que del descenso en los niveles de fecundidad, situación que tendería a invertirse en las cohortes siguientes. En general, es probable que estas variaciones se deban a cambios en los efectivos de mujeres; el máximo relativo del 2000 puede originarse, además, en la atracción de este primer año del siglo XXI para tener un hijo, o bien para registrar como nacidos en ese año a algunos de los ocurridos en años anteriores.

Con algunas excepciones, la curva del registro hasta un año después varía entre 2.1 millones y 2.2 millones, mientras que la de hasta 10 años después oscila entre 2.4 millones y 2.6 millones, de manera que, dependiendo de la duración del seguimiento, las diferencias en las magnitudes de las cohortes varían entre 300 mil y 400 mil nacimientos, lo cual significa que entre 13 y 15% de los nacimientos se registraron del segundo al décimo año después de ocurridos. Cabe señalar que después de alcanzar máximos en las cohortes 1992 y 1994, las diferencias en términos absolutos se reducen de forma ininterrumpida a partir de la cohorte 2001, como resultado de la mejora en la temporalidad del registro. De seguir la tendencia observada, los nacimientos que se registran entre el segundo y el décimo año después de ocurridos seguirán reduciéndose de manera significativa en las cohortes más recientes.

Enumeración de los nacimientos ocurridos con datos de las estadísticas vitales y otras fuentes

Una forma diferente de valorar la cobertura y temporalidad del registro de los nacimientos en las estadísticas vitales es mediante el contraste con los resultados de otras fuentes de información. Para hacerlo, comparamos las cifras de las cohortes con los resultados de los instrumentos censales para los 12 meses anteriores al levantamiento y con las cifras del SINAC para el periodo 2008-2016.

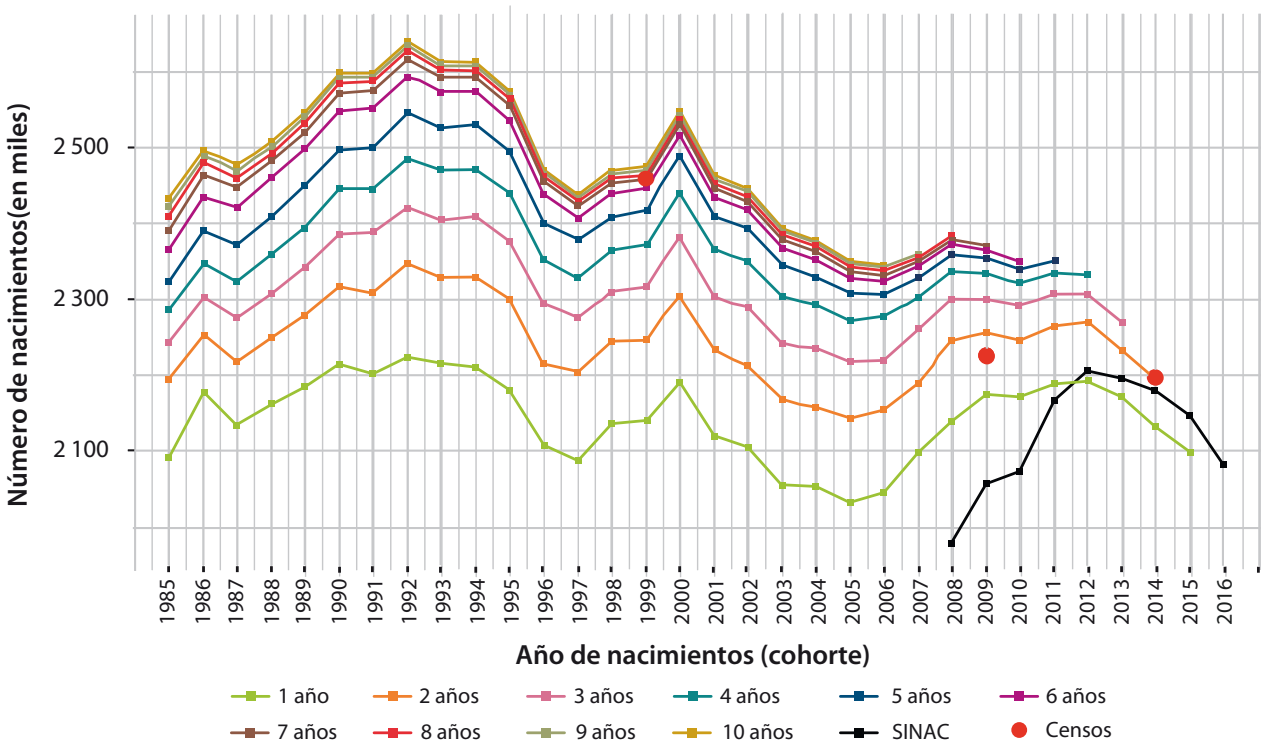
A partir de la pregunta sobre la fecha de nacimiento del ultimo hijo nacido vivo planteada a

15 No incluimos la curva de los registrados el mismo año en que ocurren que son la gran mayoría, como se observa en la gráfica 4, porque impiden observar las diferencias entre las magnitudes más pequeñas de las otras curvas.

16 La última cohorte observada hasta los 10 años después es la nacida en el 2006. En las siguientes cohortes, el truncamiento va siendo cada vez más temprano, hasta llegar a la nacida en el 2016 para la que solo se tiene a quienes fueron registrados el mismo año de su nacimiento.

Gráfica 5

Reconstrucción de las cohortes hasta 10 años después de acontecido el nacimiento (1985-2015) y nacimientos con datos del SINAC (2008-2016) y estimados con los censales (1999, 2009 y 2014)



Fuente: elaboración propia con base en INEGI. *Estadísticas vitales. Nacimientos*; censos de población. Ediciones 2000 y 2010 y *Encuesta Intercensal 2015*. // SSA. *Datos de nacimientos*.

las mujeres de 12 años o más en los ejercicios del 2000, 2010 y 2015, obtenemos estimaciones de los nacimientos ocurridos en los 12 meses anteriores al levantamiento. Si las cifras censales fueran precisas, podríamos definir hasta qué duración habría que hacer el seguimiento de las cohortes en las estadísticas vitales. Sin embargo, como ya mencionamos, estos datos censales pueden adolecer de omisiones de mujeres con hijos pequeños y, en particular, de los pequeños, así como distorsiones en la ubicación temporal de los últimos nacimientos, además del hecho de que quien proporciona la información en los censos es una tercera persona, no necesariamente la madre (Hernández *et al*., 2015).

No obstante, coincidimos con otros trabajos que han empleado con éxito esta información censal (CONAPO, 2005; UNICEF-INEGI, 2012; García, 2016) en cuanto a sus valiosos resultados, coincidentes con los de otras fuentes. Como se muestra en la gráfica 5, con el dato sobre los nacimientos en los 12 meses anteriores al Censo del 2000, se obtiene una cifra de 2.46 millones, 35% superior a la de los nacimientos de la cohorte 1999 registrados en el mismo año en que acontecieron, pero solo 12% más elevada que la de los asentados hasta un año después; la cifra censal coincide con el seguimiento de la cohorte hasta el registro entre el sexto y el séptimo año después de acontecidos. El número de nacimientos que se obtiene del Censo del 2010 es de 2.23 millones, cifra que se encuentra para la cohorte 2009 entre los registrados hasta uno y dos años después de ocurridos. En el caso de la EI 2015, el número de alumbramientos de los 12 meses anteriores es de 2.20 millones, y coincide con los de la cohorte 2014 inscritos hasta dos años después de acontecidos. Así, las cifras censales coinciden con las de las estadísticas vitales y sugieren una mejora notable en la temporalidad del registro en los nacimientos ocurridos en el siglo actual.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Por otro lado, los datos del SINAC tienen valores bajos en los primeros años —pero con una tendencia a aumentar— como resultado de una cobertura creciente; en el 2012 alcanzaron un valor máximo de 2.19 millones, algo más que los de esa cohorte registrados hasta un año después de acontecidos. A partir del 2013, las cifras de los nacimientos del SINAC tendieron al descenso, es decir, que el incremento en la cobertura cedió ante el descenso de la fecundidad; esta tendencia de la reducción en los alumbramientos también la sugieren las cifras de las estadísticas vitales. En estos últimos años, los datos del SINAC se ubican entre las curvas de las cohortes registradas hasta uno y dos años después de acontecidos.

De esta manera, se valida que las estadísticas vitales —como fuente de datos continua— permiten la enumeración de los nacimientos, aunque con ciertas limitaciones en cuanto a la falta de registro oportuno de una parte de ellos, la que tiende a reducirse en el tiempo. Los datos del SINAC aumentan su cobertura desde su inicio en el 2008, de tal forma que en años recientes tienden a converger con las cifras de las estadísticas vitales. En los últimos años, la cifra del SINAC es superior a la de los registrados hasta un año después, lo cual refleja el retraso del registro de algunos niños que ya cuentan con el certificado de nacimiento. Las estimaciones puntuales obtenidas a partir de los instrumentos censales son valiosas porque sirven de referencia en la validación de las estadísticas continuas.

Evaluación de la cobertura del registro mediante información de las encuestas

Una forma directa para evaluar la cobertura de las estadísticas vitales es mediante los datos que proporcionan algunas encuestas sobre la condición de registro de nacimiento de las personas. Esta información es de suma utilidad porque, a través del seguimiento de las cohortes que realizamos, pudimos conocer la temporalidad del registro, mas no su cobertura a ciencia cierta, la cual, en principio, sí se obtiene con estos datos transversales en los que se conoce la situación de registro del total de la población susceptible de ser asentada. Una limitación de los datos de las encuestas es su posible imprecisión debido a que se trata de información relativamente sensible, de manera que algunos entrevistados pueden responder lo que consideran debería de ser, por lo que es probable que la cobertura de la inscripción resulte un tanto sobreestimada.

En las ediciones 1992, 1997 y 2014 de la ENADID se preguntó sobre la condición de registro de los hijos de las mujeres en edades reproductivas; en la EI 2015 fue a toda la población.[17] En esta última puede ser una tercera persona quien responde y es posible que desconozca la situación de registro de los miembros del hogar, lo cual se refleja en una mayor frecuencia de no respuesta en esta fuente.[18]

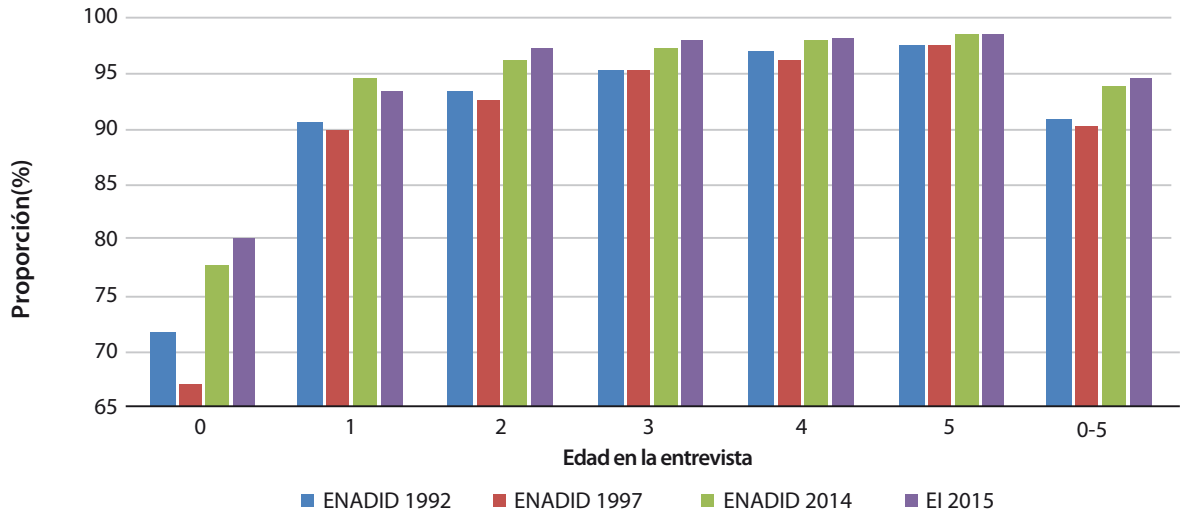
Los resultados son consistentes con lo esperado: reflejan el creciente registro en el tiempo de los niños de 0 a 5 años, aunque el aumento no ha sido lineal (ver gráfica 6). Entre las dos primeras encuestas (1992 y 1997), el registro decreció, en especial entre los menores de 1 año: de 71.6 a 67.1 por ciento. Es posible que una mejor captación en la segunda encuesta estuviera en el origen de este decrecimiento de la cobertura del registro. No obstante, en el periodo que las separa, el número de nacimientos con registro oportuno de las estadísticas vitales disminuyó y la proporción de registro oportuno se redujo de 1992 a 1994 (ver gráficas 2 y 3), lo que puede estar relacionado con una reducción real en la cobertura de los menores de 1 año entre 1992 y 1997.

17 En el 2015, a toda la población se le preguntó si contaba con acta de nacimiento o estaba inscrita en el Registro Civil del país. En la ENADID se indagó sobre la situación de los hijos nacidos vivos de las mujeres entrevistadas; en las de 1992 y 1997 se preguntó si el nacimiento del hijo se registró en una oficina del Registro Civil; en la del 2014, para todos los hijos nacidos vivos a partir del 2008, se preguntó si habían obtenido el certificado de nacimiento que otorga el sector salud y si registraron el nacimiento en el Registro Civil (http://www.beta.inegi.org.mx/contenidos/programas/enadid/2014/doc/mujer_enadid14.pdf).

18 En el 2015 la falta de respuesta no es despreciable: 0.8% para la población en su conjunto, pero alcanza a ser de 5.4% entre los menores de 1 año. Nosotros supusimos que la no respuesta en el 2015 correspondía a valores perdidos y calculamos la proporción de registro solo para quienes se tenía respuesta válida.

Gráfica 6

Condición de registro en la ENADID (ediciones 1992, 1997, 2014) y Encuesta Intercensal 2015. Proporción de la población de 0 a 5 años con acta según edad en la entrevista y año de la encuesta



Fuente: elaboración propia con base en INEGI. *Encuesta Nacional de la Dinámica Demográfica* y *Encuesta Intercensal 2015*.

Las mayores diferencias en la cobertura se observan entre las dos primeras encuestas y las dos últimas, cuando el registro fue más común, en particular en el 2015; en este año, por ejemplo, 80% de los niños menores de 1 año, 93.5% de los de 1 año cumplido y 97.1% de los de 2 años estaban registrados; en los de 5 años, el registro declarado fue de 98.6%, es decir, que solo 1.4% de los que estaban por ingresar a la escuela primaria no contaban con acta.

En esta fuente más reciente, 98.4% de la población de todas las edades declaró tener acta de nacimiento (ver gráfica 7). Al considerar en su conjunto a los niños que no habían empezado la primaria (0 a 5 años), la cobertura fue elevada: 94.5% tenía acta.[19] En las edades normativas de asistir a primaria y secundaria, esta proporción aumentó a 98.8%, casi igual a la del siguiente grupo (15 a 64 años), en el que el valor alcanzó su máximo (98.9%). En las personas de edades más avanzadas (65 años y más), al tratarse de cohortes de nacimiento más antiguas, la proporción que tenía acta fue levemente menor (98.4%).

Así, la información disponible sugiere una alta y creciente cobertura del Registro Civil, aunque la inscripción oportuna no es aún universal. En el 2015, al menos 6.5% de los niños no fueron registrados en su primer año de vida y casi 2% de la población total no contaba con acta de nacimiento.

Enumeración de los nacimientos en las entidades federativas

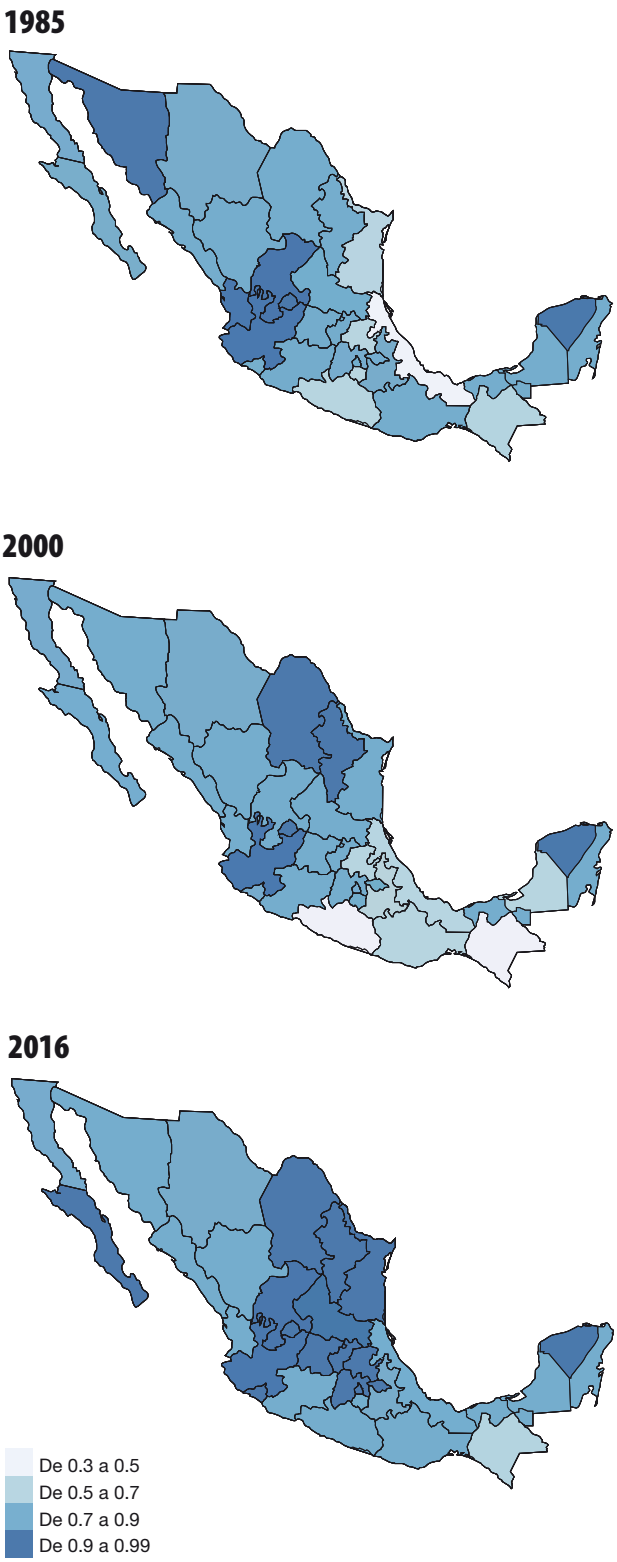
La situación del registro en el conjunto del país es resultado de particularidades muy distintas en los estados. Al desagregar el análisis del registro oportuno a este nivel, se observan condiciones muy dispares en un inicio, pero una generalizada mejora en el tiempo (ver mapas). En 1985 solo seis de las entidades (Aguascalientes, Jalisco, Nayarit, Sonora, Yucatán y Zacatecas) tenían una frecuencia del registro oportuno alta, superior a 90%; en el 2000 se les sumaron Coahuila de Zaragoza y

19 Según los resultados de la Encuesta Nacional de los Niños, Niñas y Mujeres en México (ENIM, 2015) realizada por el Instituto Nacional de Salud Pública y UNICEF-México (2016), 95.1% de los niños menores de 5 años se reportó como registrado. Esta cifra es muy cercana a la obtenida con la EIC 2015 para los menores de 6 años (94.5%), lo que sugiere resultados aceptables en esta última. No obstante, con el objetivo de verificar la información, en la ENIM se pidió a las madres mostrar el acta de sus hijos y solo la mostró 72.9 por ciento.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

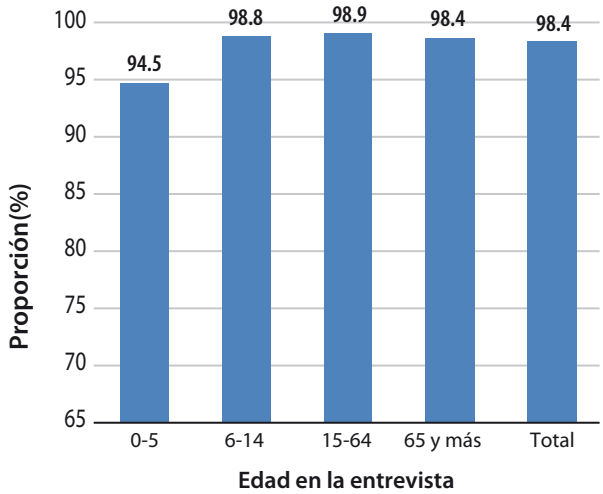
Mapas

Registro oportuno en las entidades federativas, 1985, 2000 y 2016 (proporciones)



Gráfica 7

Condición de registro en la Encuesta Intercensal 2015. Proporción de la población que cuenta con acta de nacimiento según edad en la entrevista



Fuente: elaboración propia con base en INEGI. *Encuesta Intercensal 2015.*

Nuevo León con buenos resultados, pero dejaron el grupo Nayarit, Sonora y Zacatecas. En el 2016, Baja California Sur, Colima, la Ciudad de México, Guanajuato, Hidalgo, el estado de México, Querétaro, San Luis Potosí, Tamaulipas, Tlaxcala y Zacatecas se agregaron también a este grupo con las mejores inscripciones. En ese periodo, se pasó de seis entidades a 16, de manera que en la mitad del país el registro oportuno fue generalizado en este último año. Es de señalar que a partir del 2000, una vez que un estado llega a este nivel, no desciende, lo que refleja la persistencia de las mejoras.

En una situación opuesta, las entidades con el registro oportuno más limitado (menor a la mitad de los registros) resultaron ser Veracruz de Ignacio de la Llave en 1985 y Chiapas y Guerrero en el 2000; en el 2016 ninguna de las entidades tuvo valores tan bajos. El siguiente grupo, con una situación menos deficiente, donde algo más de la mitad (entre 50 y 70%) de los registros fueron de niños menores de 1 año en 1985 estuvo formado por Chiapas, Guerrero, Hidalgo, Morelos y Tamaulipas; en el 2000, dejaron este grupo Tamaulipas y Morelos porque mejoraron, y se añadió Veracruz de Ignacio de la Llave que mejoró, además de Puebla y Oaxaca

donde se redujo el registro oportuno; en el 2016, solo Chiapas se encontraba en esta situación.

Se afirma que el entorno geográfico y los medios de comunicación y transporte, además de otras cuestiones sociales, económicas, culturales y administrativas afectan el registro de los nacimientos (Pérez y Meneses, s/f, 2; INEGI, 2015). Nosotros observamos que el registro oportuno limitado en las entidades federativas está relacionado con condiciones de alta marginación; en el 2000, los estados donde este rubro se encontraba en el rango más bajo fueron seis de los siete con el índice de marginación más elevado.[20] No obstante, en el lado opuesto del espectro, la relación entre el registro y la marginación no fue lineal, ya que las entidades con los índices de marginación más bajos no siempre tenían el mejor registro, y algunas con alto grado de marginación contaban con un registro excelente. El caso de Yucatán es un claro ejemplo donde, con un alto grado de marginación en el 2000, se tenía una tradición de buen registro que se reflejó en un nivel elevado a lo largo de todo el periodo.

Conclusiones

El objetivo de este trabajo es evaluar de manera sistemática la evolución de las estadísticas de los nacimientos registrados en México en las últimas tres décadas. La justificación radica en que las estadísticas vitales constituyen información estratégica para la planeación del país. La evaluación se realizó mediante el análisis de la temporalidad del registro y el contraste con otras fuentes de datos.

Como organismo recolector de información estadística, el Registro Civil ha tenido limitaciones por su complejo funcionamiento en los distintos niveles de gobierno y su vinculación con otras instituciones de la administración pública. No obstante, a pesar de esta complejidad, la evolución hacia un mejor registro de los nacimientos durante el periodo observado es innegable. En una primera etapa, la mejora consistió sobre todo en una creciente cobertura del registro, con frecuencia relacionada con cambios coyunturales propiciados por campañas de registro o la puesta en marcha de programas sociales, aunque también con la expansión del sistema educativo en los niveles de primaria y secundaria. En los años centrales del periodo, esta creciente cobertura ocasionó un deterioro en la proporción de nacimientos con registro oportuno. A partir del inicio del presente siglo, tanto la cobertura como la inscripción oportuna en las estadísticas vitales han progresado de manera notable y han coincidido en el tiempo con la obligatoriedad de la educación preescolar y, unos años más tarde, la puesta en marcha del SINAC, lo que es probable haya generado una sinergia con efectos positivos en la precisión de los datos.

En la comparación con otras fuentes, es de resaltar la pertinencia de la información censal y del SINAC como referencia en la validación de las estadísticas vitales. A pesar de las limitaciones de los datos censales para obtener estimaciones del número de nacimientos, los resultados señalan la plausibilidad de las cifras, resultado del seguimiento de las cohortes de nacimiento, y coinciden en señalar la mejora en la temporalidad del registro. Las cifras del SINAC han tendido a converger con los de las estadísticas vitales en los últimos años. La información de las encuestas sobre la cobertura del registro también tiene limitaciones, pero es relevante, en particular, porque ratifica los adelantos en la cobertura y la temporalidad del registro a partir del cambio de siglo; sugiere que aún en años recientes subsisten ciertas deficiencias en la inscripción oportuna.

A pesar de que la evolución en el periodo ha sido muy desigual en las entidades federativas, se observa una tendencia generalizada a la mejora en la temporalidad del registro de los nacimientos; en el 2016, ninguno de los estados tenía valores tan bajos (menos de la mitad) en su registro oportuno. Algunas entidades (como Aguascalientes, Jalisco y Yucatán) han tenido un progreso generalizado desde 1985, mientras que otras (como Guerrero y

20 CONAPO. *Índices de marginación en el 2000*. http://www.conapo.gob.mx/es/CONAPO/Indices_de_marginacion_2000_

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Veracruz de Ignacio de la Llave) han sufrido grandes deficiencias; Chiapas se ha mantenido con valores bajos, aun en el último año observado.

En suma, para que las estadísticas vitales se consoliden como fuente de información precisa y oportuna de los nacimientos que ocurren en el país y sean de utilidad en las estimaciones de la fecundidad, se requiere continuar con los esfuerzos para lograr la universalidad del registro oportuno en todas las entidades federativas, en especial de manera inmediata después del acontecimiento y en los estados con mayores rezagos.

La calidad de la información básica sobre la persona que se registra es adecuada, pero es necesario atender también la calidad de los datos de las variables vinculadas con el cálculo de la fecundidad —en particular la edad de la madre— y otras sociodemográficas de la madre y del padre. Información precisa y oportuna sobre los nacimientos y las características de los padres en todas las regiones de México permitiría un buen conocimiento de los patrones de natalidad y fecundidad en el país, las entidades federativas y otros niveles de desagregación y, con ello, proporcionar elementos para la planeación eficiente del desarrollo.

Fuentes

Banco Interamericano de Desarrollo (BID). *El registro de nacimientos: consecuencias en relación al acceso a derechos y servicios sociales y a la implementación de programas de reducción de pobreza en 6 países de Latinoamérica*. Washington, DC, EE. UU., BID, 2006.

CEPAL-UNICEF. "Desafíos", en: *Boletín de la infancia y adolescencia sobre el avance de los Objetivos de Desarrollo del Milenio. El derecho a la identidad: los registros de nacimientos en América Latina y el Caribe*. Núm. 13, CEPAL-UNICEF, 2011.

Cody, C. *Count every child: The right to birth registration*. Working Plan Ltd., 2009.

CONAPO. *La fecundidad en México. Niveles y tendencias recientes*. Serie *Documentos técnicos*. México, CONAPO-Secretaría de Gobernación, 2005.

Figueroa, Beatriz. "El registro extemporáneo de los nacimientos. Una fuente de información desatendida", en: *DEMOS. Carta demográfica de México*. México, 1998, pp. 35-37.

Gaete-Darbó, Adolfo. "Evaluación de las estadísticas vitales en América Latina", en: *Boletín de la Oficina Sanitaria Panamericana*. Noviembre de 1965, pp. 431-443.

_____ "The population register as an agency for collecting statistical data", en: *Reunión del Comité de Expertos para el Mejoramiento de las Fuentes de Estadísticas Demográficas*. CEPAL. ST/ECLA/Conf. 47/L.1. Naciones Unidas, Economic and Social Council, 1974.

Galindo, Carlos y Manuel Ordorica. "Estimación de nacimientos ocurridos y registrados, México 1950-2000", en: *Papeles de Población*. 13(54). México, 2007, pp. 39-86.

García, Juan Enrique. "La cobertura oportuna y sub-cobertura de los nacimientos en México", en: *Coyuntura Demográfica. Revista sobre los Procesos Demográficos en México Hoy*. Núm. 10. México, SOMEDE, 2016, pp. 87-97.

Harbitz, Mia e Iván Arcos Axt. *Diccionario para registros civiles e identificación*. Washington, DC., EE. UU., BID, 2013.

Hernández, María Felipa, Graciela Tapia, Xóchitl Alarcón y María de la Cruz Muradás. "Aproximaciones al nivel de la fecundidad en México 1990-2014", en: *La Situación Demográfica en México 2015*. México, CONAPO, 2015, pp. 17-42 (DE) https://www.gob.mx/cms/uploads/attachment/file/88780/02_Aproximaciones_al_nivel_de_la_fecundidad_en_Mexico.pdf, consultado el 27 de octubre de 2017.

Hernández, Juan Eugenio, Lina Sofía Palacio, Leonel González, Concepción García, Diana Molina, Armando David Quezada y Ana Lidia Salgado. *Desarrollo de un modelo que combina métodos probabilísticos, geográficos y demográficos para estimar y corregir el subregistro de las defunciones en México* (DE) http://www.beta.inegi.org.mx/eventos/2018/conacyt/doc/p_JuanHdz.pdf, consultado el 24 de noviembre de 2018.

Herrero, J. M. "El registro de población. Problema añejo", en: *DEMOS. Carta Demográfica de México*. Núm. 11. Ciudad de México, Demos, 1998, pp. 38-39.

INEGI. Censos de población y vivienda. Ediciones 2000 y 2010 (DE) https://www.inegi.org.mx/programas/ccpv

_____ *El ABC de las estadísticas vitales*. Aguascalientes, México, INEGI, 1997.

_____ *Encuesta Intercensal 2015* (DE) https://www.inegi.org.mx/programas/intercensal/2015/default.html

_____ *Encuesta Nacional de la Dinámica Demográfica*. Ediciones 1992, 1997, 2006, 2009 y 2014 (DE) https://www.inegi.org.mx/programas/enadid/2014/default.html

_____ *Estadísticas vitales*. *Nacimientos*. Bases de datos de 1985 al 2016 (DE) http://www.beta.inegi.org.mx/proyectos/registros/vitales/natalidad/default.html

_____ *Estadística de nacimientos. Marco metodológico*. México, INEGI, 2015 (DE) http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825075026.pdf, consultado el 1 de noviembre de 2018.

_____ INEGI (DE) https://www.inegi.org.mx/programas/natalidad/

Instituto Nacional de Salud Pública (INSP) y UNICEF-México. *Encuesta Nacional de Niños, Niñas y Mujeres 2015-Encuesta de Indicadores*

Múltiples por Conglomerados 2015. *Informe final*. INSP y UNICEF-México, Ciudad de México, 2016.

Instituto para la Atención de los Adultos Mayores en el DF (IAAM-DF). *Programa Institucional 2008-2012*. México, Secretaría de Desarrollo Social, Dirección General. IAAM-DF, 2005 (DE) http://www.adultomayor.cdmx.gob.mx/images/pdf/Programa_Institucional_2008-2012.pdf, consultado el 25 de noviembre de 2018.

López, Guadalupe. "Problemas de la estadística demográfica desde el punto de vista de los productores", en: Bronfman y Gómez de León (comps.). *La mortalidad en México: niveles, tendencias y determinantes*. México, El Colegio de México, 1988, pp. 57-67.

Mier y Terán, Marta. "Los nacimientos que ocurren en el país. ¿Qué revelan las fuentes sobre su número en años recientes?", en: *Coyuntura Demográfica. Revista sobre los Procesos Demográficos en México Hoy*. Núm. 3. México, SOMEDE, 2013, pp. 53-59.

Moultrie, T. A. "Evaluation of data on recent fertility from censuses", en: Moultrie, T. A., R. E. Dorrington, A. G. Hill, K. Hill, I. M. Timæus & B. Zaba (eds.). *Tools for Demographic Estimation*. Paris, International Union for the Scientific Study of Population (DE) https://researchonline.lshtm.ac.uk/25321/1/Tools%20for%20Demographic%20Estimation_GOLD%20VoR.pdf

Naciones Unidas. *United Nations Expert Group Meeting on the Methodology and lessons learned to evaluate the completeness and quality of vital statistics data from civil registration*. UN/POP/EGM-CRVS/2016/INF.1. Nueva York, EE. UU., Population Division, Department of Economic and Social Affairs, United Nations Secretariat, 2016.

_____ *Principios y recomendaciones para un sistema de estadísticas vitales*. (ST/ESA/STAT/SER.M/19/Rev.2). Nueva York, EE. UU., Departamento de Asuntos Económicos y Revisión 2 División de Estadísticas Sociales, 2003 (DE) https://unstats.un.org/unsd/publication/SeriesM/SeriesM_19rev2s.pdf

_____ "Manual de sistemas y métodos sobre estadísticas vitales. Volumen II. Examen de las prácticas nacionales", en: *Estudio de Métodos*. Serie F, núm. 35 (ST/ESA/STAT/SER.F/35). Nueva York, EE. UU., Departamento de Asuntos Económicos y Sociales Internacionales, Oficina de Estadística, 1985 (DE) http://unstats.un.org/unsd/demographic/standmeth/handbooks/Series_F35/Series_F35es_v2.pdf

_____ *Principios y recomendaciones para un sistema de estadísticas vitales. Informes estadísticos*. Serie M, núm. 19 (ST/STAT/SER.M/19/Rev.I). Nueva York, EE. UU., Departamento de Asuntos Económicos y Sociales, Oficina de Estadística, 1974 (DE) http://unstats.un.org/unsd/demographic/standmeth/principles/Series_M19Rev1es.pdf

Organización de los Estados Americanos. *Diagnóstico del marco jurídico institucional y administrativo de los sistemas de registro civil en América Latina*. Washington, DC, EE. UU., OEA, 2010a.

_____ *Manual de prácticas exitosas para el registro civil*. Washington, DC, EE. UU., Departamento de Modernización del Estado y Gobernabilidad, Programa de Universalización de la Identidad Civil de las Américas, OEA, 2010b.

Pérez Paredes, Elsa y Eloina Meneses Mendoza. *El registro de los nacimientos en México*. s/f (DE) unstats.un.org/unsd/vitalstatkb/Attachment1110.aspx?AttachmentType=1, consultado el 1 de noviembre de 2018.

RENAPO. *Programa de Modernización Integral del Registro Civil: Conceptos y Estructura*. México, 2012 (DE) http://www.renapo.gob.mx/swb/work/models/RENAPO/Resource/317/ConceptosMIRC.pdf

Secretaría de Desarrollo Social. *Programa Nacional de Desarrollo Social 2001-2006, Superación de la pobreza*: *una tarea contigo*. (DE) https://www.gob.mx/cms/uploads/attachment/file/13832/PNDS_2001_2006rr.pdf, consultado el 24 de noviembre de 2018.

SEGOB, DIF y UNICEF. *Derecho a la identidad. Buenas prácticas del registro de nacimiento de niñas y niños en México* (informe elaborado por Leonardo Mier Bueno y Martín Álvarez Gutiérrez). UNICEF-México, s/f.

SSA, Dirección General de Información en Salud. *Datos de nacimientos*. Bases de datos del 2007 al 2016 (DE) https://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_nacimientos_gobmx.html

Spoorenberg, T. *Evaluation and Analysis of Fertility Data, Regional Workshop on the Production of Population Estimates and Demographic Indicators*. Adís Abeba, Etiopía, Population Division, United Nations, Department of Economic and Social Affairs, 5-9 de octubre, 2015 (DE) http://www.un.org/en/development/desa/population/events/pdf/other/11/ppt_Fertility.pdf

UNICEF. "El registro de nacimiento. El derecho a tener derechos", en: *Innocenti Digest*. 9. Florencia, Italia, Centro de Investigaciones Innocenti, 2002, pp. 1-16.

UNICEF-INEGI. *Derecho a la identidad. La cobertura del registro de nacimiento en México en 1999 y 2009*. México, UNICEF- INEGI, 2012.

UNICEF-INSP. *México. Encuesta Nacional de Niños, Niñas y Mujeres 2015. Informe final. Programa de la Encuesta de Indicadores Múltiples por Conglomerados (MICS)*. México, 2017 (DE) https://www.insp.mx/enim2015/informe-final-enim.html, consultado el 7 de diciembre de 2018.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Funcionamiento en muestras finitas de técnicas de imputación y retropolación: caso de las series de encuestas económicas nacionales del INEGI

Finite Sample Performance of Imputation and Retropolation Techniques: the INEGI's National Economic Surveys' case

Francisco de Jesús Corona Villavicencio,* Jesús López-Pérez y Nelson Omar Muriel Torrero*****

Nota: se agradecen los comentarios y sugerencias realizadas por el personal de la Dirección General Adjunta de Encuestas Económicas del INEGI, en particular a Araceli Martínez Gama, Santiago Ávila Ávila, Juan José Ríos Franco, Francisco Reyes Piña, Ramón Bravo Cepeda, Ramón Sánchez Trujano, Roberto Tovar Soria, Diana Gabriela Cedeño Robles, Rodrigo G. Carranza Trinidad y Zelic Marco Antonio Rosa Vara; también, nuestro agradecimiento a Gerardo Leyva Parra, director general adjunto de Investigación por sus valiosas aportaciones realizadas.

* Instituto Nacional de Estadística y Geografía (INEGI), franciscoj.corona@inegi.org.mx.
** INEGI, jesus.lopezp@inegi.org.mx
*** Universidad Iberoamericana CDMX, nelson.muriel@ibero.mx

Snow and Rosss Geese, Bosque Del Apache, New Mexico, USA, Winter/Education Images/Getty Images

El objetivo de este trabajo es analizar el funcionamiento en muestras finitas de diferentes técnicas de imputación y retropolación en el contexto de series de tiempo. Para estos fines, se realiza un experimento Monte Carlo simulando diversas series de tiempo a través de modelos de factores dinámicos estacionales, considerando factores estacionarios, no estacionarios y la combinación de ambos; lo anterior, bajo distintas especificaciones en el componente idiosincrático. Este proceso generador de datos es validado ajustando dichos modelos a cinco bases de datos de las encuestas económicas nacionales del INEGI. Los principales resultados indican que, para imputar datos, es conveniente usar información adicional; de otra forma, los métodos basados en el filtro de Kalman son una buena alternativa. En retropolación, los mejores resultados se obtienen incluyendo restricciones sobre el pasado de la serie de tiempo a retropolar; caso contrario, empalmar series de tiempo ofrece una solución útil en la práctica.

Palabras clave: cointegración; encuestas económicas nacionales; modelos de factores dinámicos; experimento Monte Carlo; raíz del error cuadrático medio.

Recibido: 4 de marzo de 2019.
Aceptado: 9 de mayo de 2019.

The purpose of this research is to analyze the finite sample performance of frequently used imputation and retropolation time series techniques. In this way, we carry out a Monte Carlo experiment simulating several time series through seasonal Dynamic Factor Models (DFMs) considering stationary factors, non-stationary common factors and the combination of both, along with different specifications in the idiosyncratic component. This data-generating process is validated through DFMs estimation on five database of National Economic Surveys (EEN) from INEGI. The main results indicate that for data imputation, using information of correlated time series helps to minimize the estimation error; on the other hand, the methods based on the Kalman filter are a useful alternative. For retropolation, the better results are obtained using restrictions over the past of the time series; otherwise, the splice method is a useful alternative in practice.

Key words: Cointegration; National Economic Surveys; Dynamic Factor Models; Monte Carlo Experiment; Root mean squared error.

1. Introducción

Las instituciones encargadas de generar series de tiempo económicas de carácter oficial se enfrentan a diversos retos metodológicos para construir las que reflejen un concepto económico, ofrecer al público las que cubran periodos mayores, incorporar nuevos aspectos metodológicos ante los cambios de año base, etcétera. Todos estos ejemplos requieren el uso de técnicas estadísticas o econométricas que deben tener sustento teórico y empírico para garantizar la calidad de la información oficial publicada.

Centrándonos en el problema de generar indicadores económicos con temporalidad larga, observamos que, con frecuencia, estos dependen de series de tiempo que, por diversas causas (problemas técnicos ocurridos al momento de realizar la recolección de la información, falta de recursos económicos para observarla, inexistencia de la misma, etc.), tienen datos faltantes. Por otro lado, para construir indicadores económicos que tengan las características de pertinencia y utilidad al público en general, es necesario, en ocasiones, recurrir a información de series de tiempo complementarias para poder construir otras más largas. En estos casos, es necesario implementar técnicas ya sea de imputación de datos, de retropolación de series de tiempo o ambas para construir indicadores adecuados.

La generación de los indicadores económicos disponibles en el Banco de Información Económica (BIE) del Instituto Nacional de Estadística y Geografía (INEGI) depende de los datos disponibles en series de tiempo de estudios de dominios menores. En ellas, por razones múltiples, suele haber datos faltantes e incompatibilidades, producto —estas últimas— de los cambios metodológicos a

los que su registro se ha sujetado. En particular, las encuestas económicas nacionales (EEN) —entre ellas, la Mensual sobre Empresas Comerciales (EMEC), la Mensual de la Industria Manufacturera (EMIM), la Mensual de Servicios (EMS), la Nacional de Empresas Constructoras (ENEC) y la Mensual de Opinión Empresarial (EMOE)— presentan estas características, lo cual hace que su integración en los índices del BIE sea un reto y que la búsqueda de indicadores económicos con temporalidades más largas se beneficie de las técnicas de retropolación e imputación.

Existen algunos trabajos previos que proponen y/o evalúan el funcionamiento de diversas técnicas de imputación en diferentes áreas del conocimiento. En cuanto a propuestas metodológicas tenemos, por ejemplo, a Harvey & Pierse (1984), quienes recomiendan un método basado en la representación espacio-estado del Modelo Autorregresivo Integrado de Media Móvil (ARIMA, por su nombre en inglés) y la aplicación del filtro de Kalman. Peña & Tiao (1989), por otro lado, plantean que se puede hacer uso de series de tiempo complementarias para imputar la información faltante utilizando esperanzas condicionales o promedios como estimadores. En el ámbito de series de tiempo múltiples, Guerrero & Gaspar (2010) proponen una metodología que considera la edición e imputación de la serie de forma conjunta basándose en procesos de la familia de vectores autorregresivos (VAR). En cuanto a la evaluación de técnicas de imputación, destacan los trabajos de Pfeffermann & Nathan (2001) concentrados en encuestas; Schafer & Graham (2002), quienes presentan una evolución histórica de las técnicas de imputación de datos faltantes en cortes transversales; Moritz *et al*. (2015), donde el problema se aborda desde el ámbito computacional; Schmitt *et al*. (2015), concentrados en aplicaciones en las ciencias naturales y Pratama *et al*. (2016), quienes basan su análisis en la causa de la falta de datos. En el caso particular de las técnicas de retropolación —como se explica en De la Fuente-Moreno (2014)—, la técnica usada con más frecuencia por organismos internacionales se basa en el encadenamiento o empalme de series de tiempo. De manera alternativa, Guerrero & Corona (2018) sugieren el concepto de retropolación restringida para *llevar hacia atrás* el Producto Interno Bruto (PIB) trimestral por entidad federativa de México, base 2008.

De esta forma, aunque existen trabajos previos que evalúan el funcionamiento de diversas técnicas de imputación y retropolación, no hay uno que sea integral y que valore a fondo el funcionamiento en muestras finitas de diversas técnicas en el contexto de series de tiempo y bajo diferentes estructuras estocásticas, como lo pueden ser la estacionariedad y la variabilidad. En este trabajo nos proponemos cubrir esa laguna en la literatura con el propósito particular de que nuestra propuesta sea de utilidad en el contexto de las EEN del INEGI, pero rescatando la aplicabilidad de las técnicas en escenarios más generales. En definitiva, el objetivo fundamental de esta investigación es evaluar el funcionamiento de diversas técnicas de imputación y retropolación usadas de manera frecuente en el contexto de series de tiempo. Partimos del hecho de que existe información que puede servir para imputarlas y/o retropolarlas; capturamos esta relación proponiendo como proceso generador de datos uno de factores dinámicos, en el cual las series de tiempo comparten factores comunes y se relacionan a través de ellos pero tienen, a su vez, una dinámica individual, también estocástica. Para el caso de imputación, usamos la especificación espacio-estado aplicando el suavizamiento de Kalman como método para imputar datos. También, se ponen a prueba diferentes tipos de *splines*, medias móviles y métodos de reemplazo basados en la información disponible dentro de la muestra. Asimismo, se considera la técnica basada en los *vecinos más cercanos*, la cual es usada con frecuencia en el caso de corte transversal, pero aplicable también a series de tiempo. Para el caso de retropolación, se utiliza el enfoque del empalme de series de tiempo, un algoritmo fácil de implementar y muy usado por distintos organismos internacionales generadores de información. Se especifican, además, dos técnicas novedosas de retropolación, ambas basadas en la estimación de modelos VAR, la primera realizando *backasting* o pronósticos *hacia atrás* y la segunda utilizando

dichos pronósticos, pero imponiendo restricciones lineales en la línea de Guerrero & Corona (2018).

El resto del artículo se organiza como sigue: en la segunda sección se describen los métodos de imputación y retropolación usados en este trabajo; en la siguiente se detalla el diseño del experimento Monte Carlo implementado para evaluar el rendimiento de los diferentes métodos de imputación y retropolación en muestras finitas; en la cuarta se evalúan las condiciones del experimento Monte Carlo en el contexto de las EEN; en la posterior se presentan los principales resultados de este experimento y, por último, se llega a las conclusiones y se hacen algunas recomendaciones.

2. Métodos de imputación y retropolación

2.1. Técnicas de imputación

2.1.1 Filtro de Kalman

La especificación general está dada por la representación gaussiana espacio-estado, donde la serie observada, denotada por y_t, viene dada por:

$$y_t = Z_t \alpha_t + u_t,$$
$$\alpha_{t+1} = T_t \alpha_t + R_t e_t, \quad (1)$$

para $t = 1,\ldots,T$. En la ecuación (1), $u_t \sim N(0,H_t)$, $e_t \sim N(0,Q_t)$ y $\alpha_1 \sim N(a_1,P_1)$. Para consultar más detalles sobre el modelo espacio-estado dado por la expresión (1) ver, entre otros, a Helske (2017). Usando el suavizamiento de Kalman, podemos estimar de forma recursiva, con toda la información disponible, la ecuación de estado como:

$$\hat{\alpha}_t = E(\alpha_t \mid y_1,\ldots, y_T). \quad (2)$$

El primer método de imputación, llamado TKS, supone que $Z_t = T_t = R_t = 1$, $\alpha_t = t$ y $H_t = Q_t = 0.01$; es decir, es un modelo donde las observaciones se mueven alrededor de una tendencia determinística y donde existe poca variabilidad en los términos de error en la representación espacio-estado. El segundo, denominado LLKS, supone que α_t es una caminata aleatoria y que H_t y Q_t son estimadas por máxima verosimilitud (ver Moritz & Bartz-Beielstein, 2017). Nótese que este modelo, a diferencia de TKS, supone una tendencia estocástica.

2.1.2 *Splines*

El segundo grupo de métodos de imputación está basado en el uso de *splines* que, en esencia, representan una serie de interpoladores lineales y no-lineales para unir puntos entre datos faltantes. Matemáticamente diríamos que una función $s(y)$ es una *spline* en el intervalo $[a,b]$, si existe una partición del intervalo,

$$P = \{a = y_0 < y_1 < \ldots < y_T = b\} \quad (3)$$

de tal forma que $s(y)$ es un polinomio, tal vez distinto, en cada intervalo $[y_i, y_{i+1}]$, $i = 0,1,\ldots,T$ - 1. En este caso, los puntos y_t se denominan los nodos del *spline*.

En este trabajo se utiliza la *spline* lineal, donde se unen los puntos a través de una proporcionalidad lineal entre los datos faltantes. También, se usan los procedimientos sugeridos por Forsythe *et al*. (1977), donde se ajusta una posición cúbica exacta a través de los cuatro puntos en cada extremo de los datos y el algoritmo de Stineman (1980), donde se computan pendientes más bajas cerca de pasos o picos abruptos en la secuencia de puntos. Al primero se le denota como LIN; al segundo, SPL; y al tercero, STI.

2.1.3 Medias móviles

Otro grupo de métodos de imputación está basado en la estimación de promedios o medias móviles basadas en el promedio de un número determinado, n, de observaciones centrales; en él se encuentran la simple (SMA), la ponderada de forma lineal

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

(LWMA) y la ponderada de manera exponencial (EWMA). En la SMA se pondera con el mismo peso a todas las observaciones que forman el promedio. En la LWMA, los pesos decrecen aritméticamente, es decir, las observaciones junto a un valor central i tienen un peso$1/2$; las observaciones siguientes (i - 2, i + 2), $1/3$; las que siguen (i - 3, i + 3), $1/4$ y así, en lo sucesivo. Por último, para el caso de la EWMA, los pesos van decreciendo de manera exponencial de la forma $(1/2)^i$.

2.1.4 Imputación basada en la información disponible en la muestra

Este grupo de procedimientos está sustentado en argumentos netamente heurísticos y se basan solo en utilizar el dato anterior o el siguiente disponible para rellenar los faltantes. En términos estadísticos, el método de arrastre de la última observación realizada (LOCF) está dado por:

$$\hat{y}_t = y_{t-J}, \quad J = \min(i | y_{t-i} \neq NA). \tag{4}$$

Por otro lado, el método de arrastre de la primera observación realizada (NOCB):

$$\hat{y}_t = y_{t+J}, \quad J = \max(i | y_{t+i} \neq NA). \tag{5}$$

En ambos casos, $\hat{y}_t$ es el estimador de y_t usando la información disponible en el momento i.

2.1.5 Método basado en los k vecinos más cercados (KNN)

Utiliza la información de una serie de tiempo, x_{jt}, para imputar los datos faltantes en la serie de tiempo de interés, y_t. De forma estadística, y para simplificar, si definimos k =1, el estimador KNN está definido de la siguiente manera:

$$\tilde{y}_t = \{x_t \mid \delta(x_{it}, y_t) < \delta(x_{jt}, y_t) \forall j\} \tag{6}$$

donde $\hat{y}_t$ representa el estimador KNN, $\delta(x_{it}, y_t) = | y_t - x_{jt} |$ es la distancia entre las series de tiempo y_j y x_{jt} al tiempo t para $j = 1, \ldots, K$, siendo K el número de series de tiempo disponibles como candidatos a vecinos más cercanos. De manera adicional, se puede reformular δ_j usando la información de $K \geq k > 1$ vecinos más cercanos. También, nótese que diversas formas funcionales de x_{jt} pueden ser usadas para imputar y_t. Para mayor detalle sobre este método de imputación, ver Kowarik & Templ (2016).

2.2. Técnicas de retropolación

2.2.1 Empalme de series de tiempo

Consiste en heredar el comportamiento pasado de una serie de tiempo a otra; por ejemplo, cuando una está desactualizada, pero tiene una temporalidad hacia atrás mayor que una más reciente, puede *llevarse hacia atrás* la más reciente con la temporalidad pasada de la desactualizada. Esto, claro está, resulta en la herencia de comportamientos específicos de la serie de tiempo desactualizada sin perder el nivel de la de interés y es una práctica común en los organismos encargados de generar información estadística oficial.

De manera más general, podemos explotar la correlación instantánea entre dos series de tiempo siempre que estas compartan un periodo común y que una de ellas tenga mayor información temporal en el pasado. En este contexto, se puede definir el estimador por empalme $\tilde{y}_h$ (SPLICE) de la siguiente manera:

$$\tilde{y}_h = f(x_h), \tag{7}$$

donde x_h se obtiene de maximizar la correlación de y_{t^*} con alguna variable del vector $X_{t^*} = (x_{1t^*}, \ldots, x_{Kt^*})'$, para $t^* = H + 1, \ldots, T$, donde $h = 1, \ldots, H$. Vale la pena destacar que es necesario que el vector (y_{t^*}, x_{jt^*}) esté compuesto por series de tiempo cointegradas para que la estimación de la correlación sea estadísticamente válida.

2.2.2 Retropolación: VAR

Denotemos las observaciones como el vector $Y_{t^*} = (Y_{t^*}, X_{t^*})'$ de dimensión N x 1, y supongamos que siguen un proceso VAR(p), es decir, existen matrices de coeficientes de dimensiones N x N y un vector de errores tales que:

$$Y_{t^*} = A_1 Y_{t^*-1} + \ldots + A_p Y_{t^*-p} + v_t \quad (8)$$

Es conocido que, ya que cada variable dependiente tiene las mismas variables independientes, la estimación de las A_j puede realizarse a través de mínimos cuadrados ordinarios (MCO). Asimismo, puede mostrarse que el mejor estimador lineal insesgado (MELI) de Y_{t^*+h} es $Y_{t^*+h|T}$ o, sin pérdida de generalidad, $Y_{h|T}$ es el MELI de Y_h. De esta forma, podemos utilizar los coeficientes de MCO y realizar la cadena de pronóstico para obtener:

$$Y_{h|T} = \hat{A}_1 Y_{h+1|T} + \ldots + \hat{A}_p Y_{h+p|T}. \quad (9)$$

Por consiguiente, pueden obtenerse las retropolaciones para Y_h usando el estimador $Y_{h|T}$ que se denota en este trabajo como VAR. Es importante considerar las características estocásticas de Y_{t^*}, pues si el modelo expresado en la ecuación (8) no es estacionario, estaremos ignorando relaciones de cointegración. La prueba más utilizada en el contexto de modelos multivariados viene dada por Johansen (1991); sin embargo, cuando las series son cointegradas y el objetivo es pronosticar, Lütkepohl (2005) argumenta que una representación VAR(p) en niveles es apropiada.

2.2.3 Retropolación restringida

Utilizando la regla general de combinación de Guerrero & Peña (2003) se pueden utilizar las estimaciones dadas por la expresión (9) e incorporar restricciones lineales para obtener predicciones restringidas de Y_h. Dichas restricciones cumplen la siguiente relación:

$$z = C y_{D(h-1)+d}, \quad (10)$$

donde $y_{D(h-1)+d} = (y_{D(h-1)+1}, \ldots, y_{D(h-1)+D})'$ y $C = \frac{1}{D}\mathbf{1}_D$, siendo D la frecuencia de la serie (por ejemplo, D =12 para datos mensuales) y $\mathbf{1}_D$ un vector de unos de dimensión D x 1. En particular, sean $Y_{h|T}$ las estimaciones preliminares de y_h, de forma que $Y_h = Y_{h|T} + E_h$, donde E_h es un proceso estacionario vectorial que admite la misma representación VAR(p), el MELI de y_h basado en $Y_{h|T}$ y en z está dado por:

$$\hat{y}_h = y_{h|t} + \hat{a}(z - C y_{D(h-1)+d}), \quad (11)$$

donde $\hat{a}$ se obtiene al expresar el VAR(p) en su forma de medias móviles, estimándose los coeficientes de manera recursiva. Para más detalles sobre la estimación de los coeficientes ver Guerrero & Corona (2018). Dicho estimador lo denotamos como RVAR.

Nótese que esta representación es multivariada, por lo que se obtienen todos los pronósticos restringidos para Y_h, de tal forma que las expresiones (10) y (11) están dadas para representaciones de series de tiempo múltiples, es decir, modelos VAR(p); sin embargo, decidimos enfocarnos en la variable que se desea retropolar con el objetivo de aligerar la notación.

3. Diseño del experimento Monte Carlo

El proceso generador de datos para evaluar el funcionamiento empírico de los métodos de imputación y retropolación está basado en una estructura de factores dinámicos. Una de las implicaciones más importantes de este diseño experimental es que las series de tiempo están relacionadas por uno o más factores comunes. Por ello, al imputar o retropolar, se podría estar realizando con series relacionadas entre sí. En este estudio, consideramos diferentes estructuras de dependencia y variabilidad en el componente idiosincrático de cada serie de tiempo. De esta manera, las observaciones, Y_t, siguen el proceso de factores dinámicos:

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

$$
\begin{aligned}
Y_t &= \mu + \lambda F_t + \mathcal{E}_t, \\
(\mathrm{I} - \phi L)(I - \Phi L^D) F_t &= (I + \Theta L^D)\eta_t, \\
\varepsilon_t &= \Gamma \varepsilon_{t-1} + a_t,
\end{aligned} \tag{12}
$$

donde μ es el vector de medias de las observaciones de dimensión N x 1, F_t y η_t son los factores comunes y sus respectivas innovaciones de dimensión r x 1, λ es una matriz de dimensión N x r que denota la contribución de los factores sobre las observaciones, ε_t y a_t son vectores de dimensión N x 1 que representan al componente idiosincrático y sus errores, respectivamente. Por otra parte, ϕ, Φ, Θ y Γ son matrices de coeficientes que, por simplicidad, se suponen diagonales. Las dimensiones de las primeras tres son r x r y de la última, N x N. Finalmente, L es el operador de diferencias tal que $LY_t = Y_{t-1}$. Conocemos al sistema de ecuaciones de la expresión (12) como un modelo de factores dinámicos (DFM, por sus siglas en inglés).

Asumiendo estacionariedad en la parte estacional, algunos resultados importantes recaen sobre las matrices de coeficientes, en específico sobre ϕ y Γ; por ejemplo, si algún elemento de la diagonal de ϕ es 1, entonces los factores no son estacionarios, por lo que Y_t no es estacionaria (si $\lambda \neq 0$). Más aún, si el componente idiosincrático es estacionario, cada par de series de tiempo del vector Y_t es cointegrado, es decir, los factores comunes son las tendencias comunes de las observaciones. Por otra parte, si el componente idiosincrático no es estacionario, habrá relaciones espurias en Y_t, es decir, cada serie de tiempo es una caminata aleatoria independiente.

Nuestro experimento Monte Carlo considera los siguientes modelos generales:

$$
\begin{aligned}
&\mathrm{M1}: \phi = 1, \\
&\mathrm{M2}: \phi = 0.5, \\
&\mathrm{M3}: \phi = \begin{bmatrix} 1 & 0 \\ 1 & 0.5 \end{bmatrix}.
\end{aligned} \tag{13}
$$

El primero (M1) corresponde al caso en el que las observaciones son generadas por una caminata aleatoria; el segundo (M2), a donde el factor es estacionario; y el tercero (M3) es la combinación de los dos anteriores. Asimismo, se consideran tres casos en la matriz de coeficientes VAR del componente idiosincrático, a saber, $\Gamma_1 = diag(-0.8)$, $\Gamma_2 = 0$ y $\Gamma_3 = diag(0.99)$, es decir, solo se consideran modelos cointegrados o estacionarios, aunque al asumir elementos de la diagonal en la matriz de coeficientes del componente idiosincrático cercanos a la unidad, nos acercamos al caso espurio. En todos se toma $\mu = 100$, $\lambda \sim U(0,1)$, $\Sigma_n = diag(1)$, $\Phi = diag(0.7)$ y $\Theta = diag(0.3)$, para $D = 12$.

Consideramos tres distintas matrices de covarianzas para las innovaciones del componente idiosincrático cuando es homocedástico, a saber $\Sigma_{a1} = diag(0.1)$, $\Sigma_{a2} = diag(1)$ y $\Sigma_{a3} = diag(10)$. Por otra parte, bajo heterocedasticidad especificamos las matrices de covarianzas como $\Sigma_{a1} = diag(U\sim[0.05, 0.15])$, $\Sigma_{a2} = diag(U\sim[0.5, 1.5])$ y $\Sigma_{a3} = diag(U\sim[5, 15])$. Por último, también consideramos el caso de errores correlacionados transversal y contemporáneamente siguiendo la sugerencia de Corona *et al.* (2017), donde se propone utilizar una estructura válida de correlación débil.

Los tamaños de muestra seleccionados son $N_1 = 15$, $T_1 = 100$ y $N_2 = 30$, $T_2 = 200$, considerando paneles de dimensiones similares a los de las EEN, tanto si fuesen por subsectores económicos como por entidad federativa. De esta forma, considerando tres dinámicas en el componente idiosincrático, tres variabilidades, tres tipos de error y dos tamaños de muestra, tenemos 3 x 3 x 3 x 2 = 54 distintas especificaciones por cada uno de los tres modelos en el experimento. Consideramos $R = 500$ réplicas y para cada diferente especificación eliminamos de forma aleatoria 24 o 48 observaciones, alrededor de 25% de la muestra. Para el estudio de la imputación, estas observaciones se eliminan de posiciones seleccionadas de manera aleatoria y para el de la retropolación del inicio de la muestra. En cada réplica, cada especificación y para cada serie de tiempo del DFM se aplican las técnicas de imputación y retropolación. La estimación del VAR(p) es bivariada, seleccionando la variable a retropolar y aquella que tenga más correlación con la variable objetivo. Cuando se incorporan restricciones lineales se asume que estas son conocidas y, asimismo,

de carácter temporal, es decir, que el promedio de cada tres periodos es conocido.

Para evaluar el funcionamiento de cada técnica, se considera como estadístico a la raíz cuadrada del error cuadrático medio (RMSE, por sus siglas en inglés). En cada réplica, y para cada panel de series de tiempo, se estima el valor promedio de los RMSE y su desviación estándar. Al finalizar las 500 réplicas se toman los promedios globales tanto para la media como para la desviación estándar para cada especificación.

Las medias móviles tienen un horizonte de $n = 4$ y usamos $k = 5$ para el método de KNN. El experimento Monte Carlo es programado y estimado con ayuda del programa R, usando las librerías de KFAS para la implementación del método TKS e imputeTS para la estimación de LLKS, LIN, SPL, STI, LOCF y NOCB. Para la aplicación del KSS se utiliza la librería VIM y, por último, para la estimación de los modelos VAR se usa la librería vars.

4. Estimación del Modelo de Factores Dinámicos: EEN

Con el objetivo de establecer que las condiciones en las que se genera el experimento Monte Carlo son válidas para los datos de las EEN, se estiman los componentes de diferentes DFM y se ajustan posibles modelos ARIMA estacionales (SARIMA, por su nombre en inglés) a los factores y modelos autorregresivos de orden 1, a los componentes idiosincráticos con el fin de estimar los valores de los parámetros autorregresivos y las varianzas del error del componente idiosincrático. La estimación de los factores y sus cargas asociadas se realizan a través de componentes principales (PC, por sus siglas en inglés) como lo propone Bai (2004), es decir, se asume *a priori* que el componente idiosincrático es estacionario. También, el número de factores, $\hat{r}$, es determinado por el criterio de Ahn & Horenstein (2013). Para la implementación de este ejercicio, se centran las observaciones en 100, de tal forma que tengan la misma media que las series simuladas en el experimento.

Vale la pena destacar que, si los factores comunes son estacionales, es conveniente usar otros métodos de estimación tal como Nieto *et al*. (2016); no obstante, nuestro objetivo no es desentrañar las características particulares de los DFM, sino solo investigar si la estimación tradicional de un DFM a través de un método no paramétrico como lo es PC captura características similares a las condiciones en las que es generado el experimento. Nótese que el supuesto fundamental para la estimación consistente de los factores y sus cargas a través de PC es que conforme N tiende a infinito, el efecto de los factores sobre las observaciones es lo que permanece, convergiendo a cero el efecto del componente idiosincrático. Corona *et al*. (2017) prueban que, en muestras finitas y con factores estacionarios y no estacionarios, esto ocurre con N alrededor de 15. Así, la estimación del modelo a través de PC es considerada suficiente para nuestros fines.

En consecuencia, para la estimación de los parámetros del DFM, se utilizaron las siguientes bases de datos, todas con año base = 2008:

- EMEC. Personal ocupado total de comercio al por mayor por entidad federativa, 2008:01-2018:07 ($N = 32$, $T = 127$).
- EMIM. Índice de personal ocupado por rama de la industria manufacturera, 2007:01-2018:07 ($N = 21$, $T = 139$).
- EMS. Ingresos totales por la prestación de servicios según sector y dominio, transportes, correos y almacenamiento, 2008:01-2018:07 ($N = 13$, $T = 127$).
- ENEC. Personal ocupado total por entidad federativa, 2006:01-2018:07 ($N = 132$, $T = 151$).
- EMOE. Indicador de confianza empresarial del sector, 2008:01-2018:09 ($N = 7$, T = 129).

El cuadro presenta los resultados de las estimaciones de los parámetros asociados al DFM, mostrándose el número estimado de factores y la especificación de los modelos SARIMA ajustados a los factores. Es importante señalar que dicha especificación a los factores es realizada a través de

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Cuadro

Resumen de la estimación de los parámetros del DFM

Encuesta	$\hat{r}$	Modelo SARIMA $(p,d,q)(P,D,Q)$: F_t	Cuantiles 2.5, 50 y 97.5% de $diag(\Gamma)$	Cuantiles 2.5, 50 y 97.5% de $diag(\Sigma_a)$
EMEC	1	SARIMA(0,1,0)(1,0,0)[12]	0.54 0.91 0.98	0.08 0.64 5.78
EMIM	1	SARIMA(2,2,1)(2,0,0)[12]	0.93 0.99 1.00	0.00 0.01 0.51
EMS	1	SARIMA(0,1,1)(0,1,2)[12]	0.53 0.81 0.89	8.59 15.24 112.49
ENEC	1	SARIMA(1,1,0)(1,0,0)[12]	0.39 0.77 0.92	13.95 71.29 241.22
EMOE	1	ARIMA(1,1,1)	0.48 0.77 0.84	0.84 2.05 2.91

Nota: p = orden autorregresivo de la serie observada, d = orden de integración, q = orden autorregresivo de medias móviles, P = orden autorregresivo del componente estacional, D = orden de integración del componente estacional, Q = orden autorregresivo de medias móviles y 12 = estacionalidad mensual.

la selección automática que otorga la librería *forecast* del programa R, presentándose también los cuantiles 2.5, 50 y 97.5% tanto de $diag(\Gamma)$ como de $diag(\Sigma_a)$.

Se puede apreciar que, en todos los casos, de acuerdo con el criterio de Ahn & Horenstein (2013), un factor es suficiente para resumir la parte común de las series de tiempo para cada encuesta, aunque tal vez existan más factores si utilizamos otros criterios, por ejemplo, el de Onatski (2010). En todos los casos, con excepción de la EMOE, hay evidencia para afirmar que el factor es estacional y todos los factores tienen, al menos, una raíz unitaria en la parte no estacional. Solo para el caso de las EMS se detectó una raíz unitaria en la parte estacional. Los cuantiles asociados a la diagonal de Γ son positivos y sus valores varían entre 0.4 y 1, es decir, todas las autocorrelaciones de los componentes idiosincráticos son positivas y, en algunos casos, da lugar a DFM espurios. Para finalizar, los valores de $diag(\Sigma_a)$ llegan hasta 241.22, los cuales pueden considerarse muy grandes aunque para la EMEC, EMIM y EMOE los valores máximos de varianza se alcanzan en 5.78, siendo el máximo límite inferior de 13.95 para las cinco encuestas. De esta forma, el M1 parece adaptarse de buena forma a las bases de datos de las EEN tomadas como ejemplo; es decir, son paneles de series de tiempo cointegrados, con errores heterocedásticos y de muy distinta variabilidad, donde en algunos casos podemos encontrar relaciones espurias entre las variables.

5. Resultados del experimento Monte Carlo

5.1 Imputación

En esta subsección se presentan los RMSE para los métodos de imputación y retropolación descritos de forma previa. En ambos casos se hace énfasis en modelos con errores heterocedásticos dado que no se encontraron diferencias significativas entre modelos con errores homocedásticos, heterocedásticos y correlacionados de manera transversal, pero sí en términos de la estructura de autocorrelación en los componentes idiosincráticos y en los tamaños de la varianza de su respectivo error y de las muestras. En todas las gráficas se presentan los intervalos de confianza aproximados del RMSE a través de las 500 réplicas. En ellas, las columnas muestran los tres tamaños de la varianza del error del componente idiosincrático y los renglones, las dos medidas de muestra consideradas. Se presentan los resultados para los casos más representativos; no obstante, el resto está disponible bajo petición.

La gráfica 1 presenta los resultados para el M1 considerando $\Gamma = diag(-0.8)$, es decir, cuando los errores idiosincráticos tienen correlación serial negativa. En este caso, las observaciones son generadas por un factor común no estacionario y, dada la estructura del error, las series de tiempo están

cointegradas para cada par de estas y, también, los factores comunes representan la tendencia común de las observaciones.

Se puede apreciar que, cuando la varianza del error del componente idiosincrático es pequeña, todos los criterios presentan errores menores a la unidad y que SPL es el método cuyo RMSE se incrementa conforme la varianza del componente idiosincrático lo hace. Nótese que cuando la varianza del error del componente idiosincrático alcanza su máximo valor, $\sigma_a^2 = 10$, todos los criterios incrementan su variabilidad, siendo TKS el método con menor RMSE, tanto en el promedio de sus medias como de sus desviaciones estándar a través de las réplicas. El efecto del tamaño de la muestra es casi imperceptible, excepto cuando $\sigma_a^2 = 10$, caso en el que los intervalos de confianza son más pequeños cuando la muestra es más grande, i.e., cuando $N = 30$ y $T = 200$.

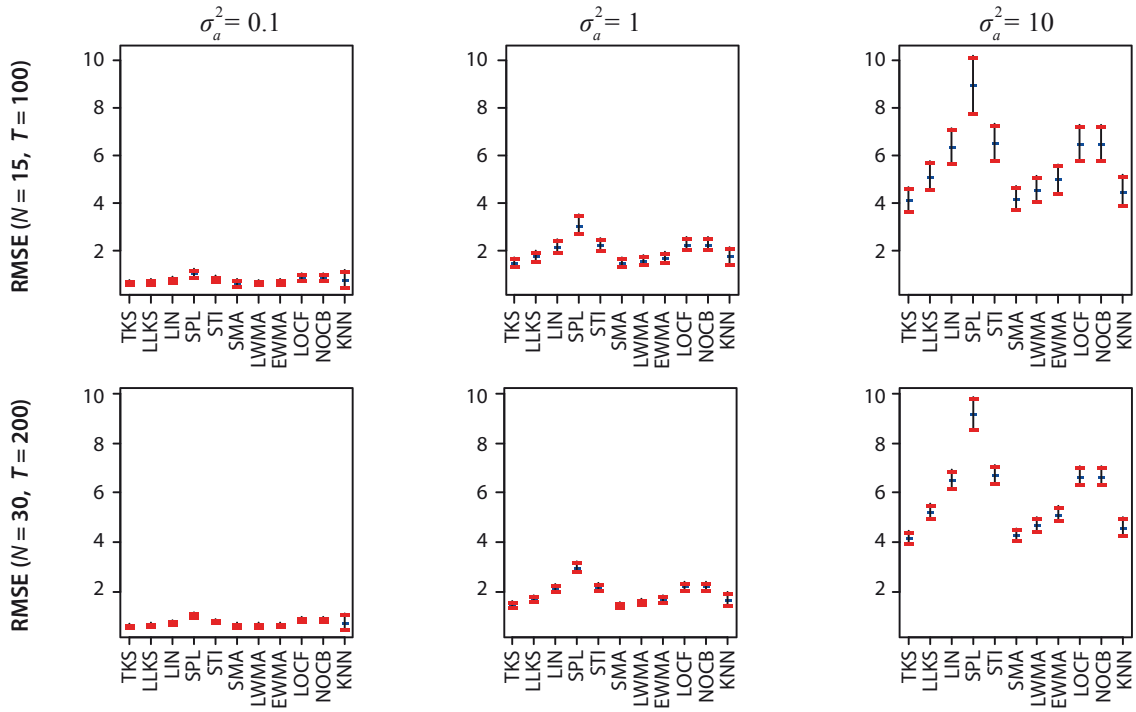
La gráfica 2 muestra los mismos resultados que la 1 cuando $\Gamma = diag(0)$, es decir, cuando los componentes idiosincráticos son ruido blanco. En esta situación, al igual que el caso anterior, las series de tiempo son cointegradas.

En la gráfica 2 se puede denotar que KNN presenta una mayor variabilidad respecto a los otros métodos. Llama la atención que su variabilidad es más grande que la del resto de los procedimientos cuando $\sigma_a^2 = 0.01$. Asimismo, es interesante observar que conforme σ_a^2 aumenta, SPL es el procedimiento que tiende a resentirlo más, incrementándose la media del RMSE a través de las réplicas. De forma similar a lo obtenido en la gráfica 1, el tamaño de muestra tiene el mismo efecto en todos los procedimientos, siendo los intervalos de confianza menores en el tamaño de muestra más grande cuando $\sigma_a^2 = 10$.

En la gráfica 3 se presentan los resultados de los intervalos de confianza de los RMSE estimados a través de las réplicas para cada uno de los métodos de imputación para el M1 cuando $\Gamma = diag(0.99)$.

Gráfica 1

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M1 con errores heterocedásticos y $\Gamma = diag(-0.8)$



INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

Nótese que esta parametrización representa el caso cuando las variables están siendo simuladas casi como unas caminatas aleatorias independientes, aunque en términos teóricos, las series de tiempo siguen siendo cointegradas.

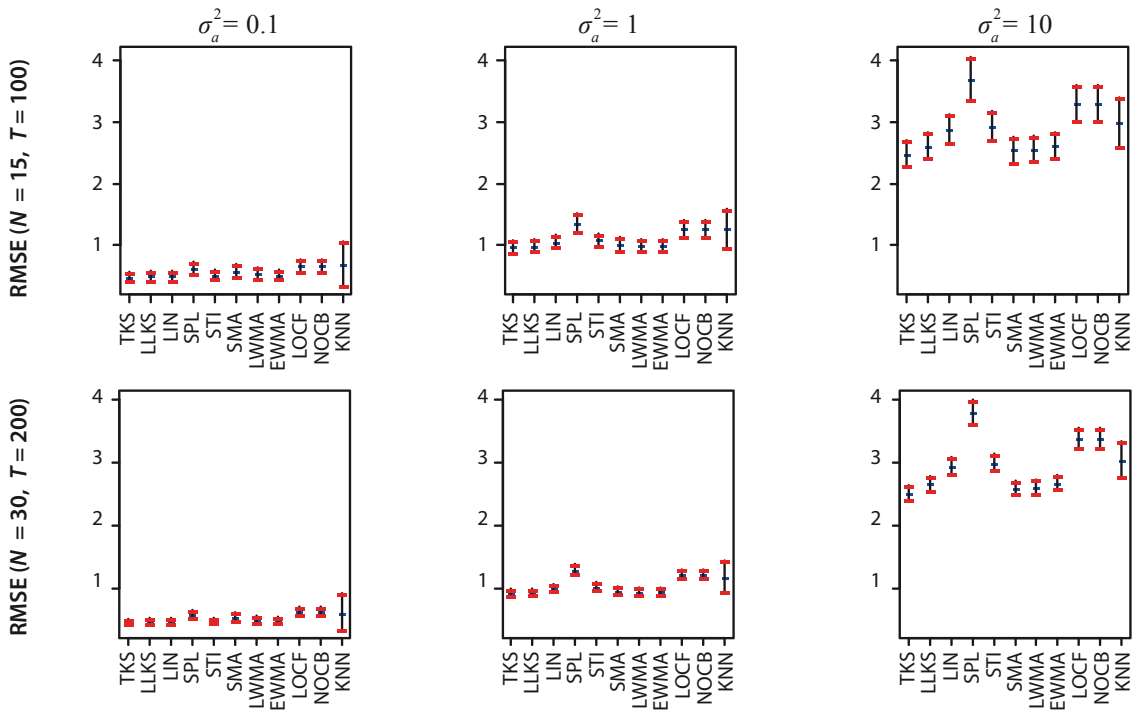
Puede observarse que cuando $\sigma_a^2 = 0.1$ o $\sigma_a^2 = 1$, todos los procedimientos de imputación, salvo el KNN y los métodos basados en la información disponible en la muestra, tienden a funcionar de manera similar mostrando un límite superior del intervalo confianza que no rebasa a 1. Por otra parte, cuando $\sigma_a^2 = 10$, los intervalos superiores se incrementan, pero lo realizan apenas sobrepasando el 1 en algunos casos, con excepción del KKN, que llega ahora a rondar las tres unidades. Este resultado es esperado, dado que KNN funciona bien conforme *los vecinos* aportan información relevante para imputar datos faltantes y esto se da en el caso de paneles cointegrados, es decir, cuando la diagonal de Γ sea menor a la unidad en valor absoluto. Asimismo, nótese que la varianza del componente idiosincrático es también una función de σ_a^2 y conforme esta sea más grande, más variabilidad habrá entre las series que forman al DFM, lo que puede afectar el funcionamiento de este método.

Para M2 y M3 solo presentamos los casos cuando $\Gamma = diag(0)$ y $\Gamma = diag(0.99)$, respectivamente, ya que, como se denotó en la subsección anterior, este tipo de dinámicas autorregresivas son las situaciones que más se presentan en la práctica. La gráfica 4 muestra el funcionamiento de los procedimientos de imputación para el M2 cuando $\Gamma = diag(0)$. En este caso, el factor común que genera a las observaciones es estacionario, por lo que las series de tiempo se mueven alrededor de una variable que tanto su media como su varianza no dependen del tiempo.

Se puede apreciar que el método del KNN presenta tanto valores medios como intervalos de confianza un poco más pequeños en los RMSE que los otros procedimientos, sobre todo cuando

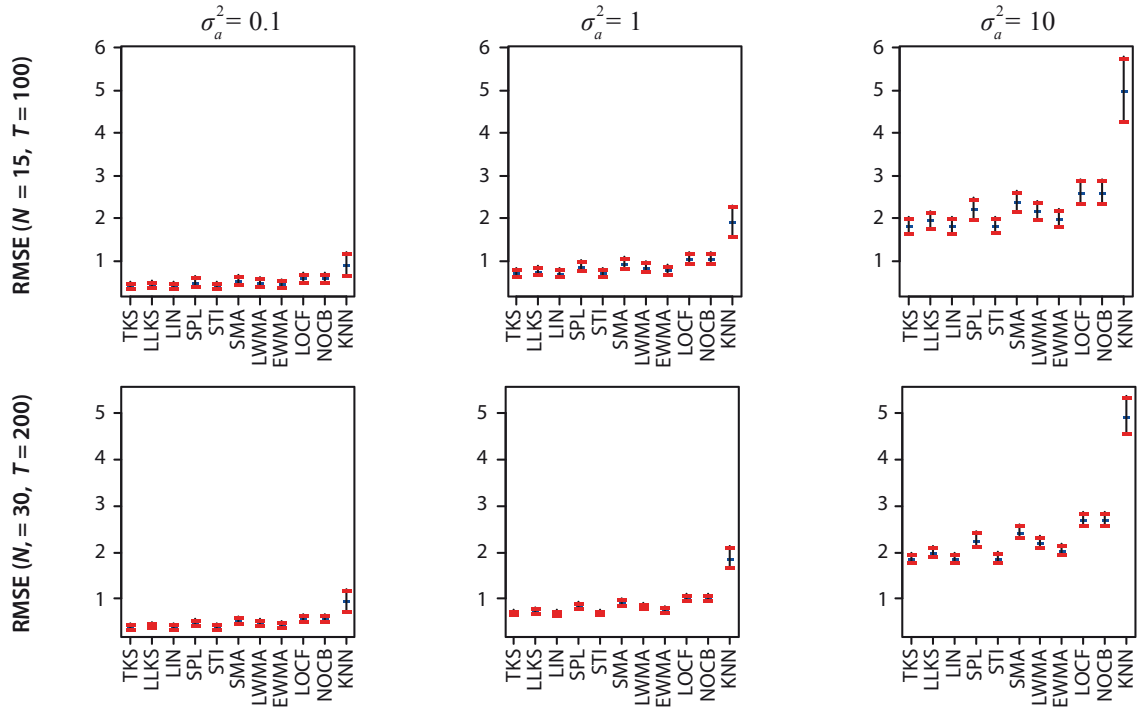
Gráfica 2

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M2 con errores heterocedásticos y $\Gamma = diag(0)$



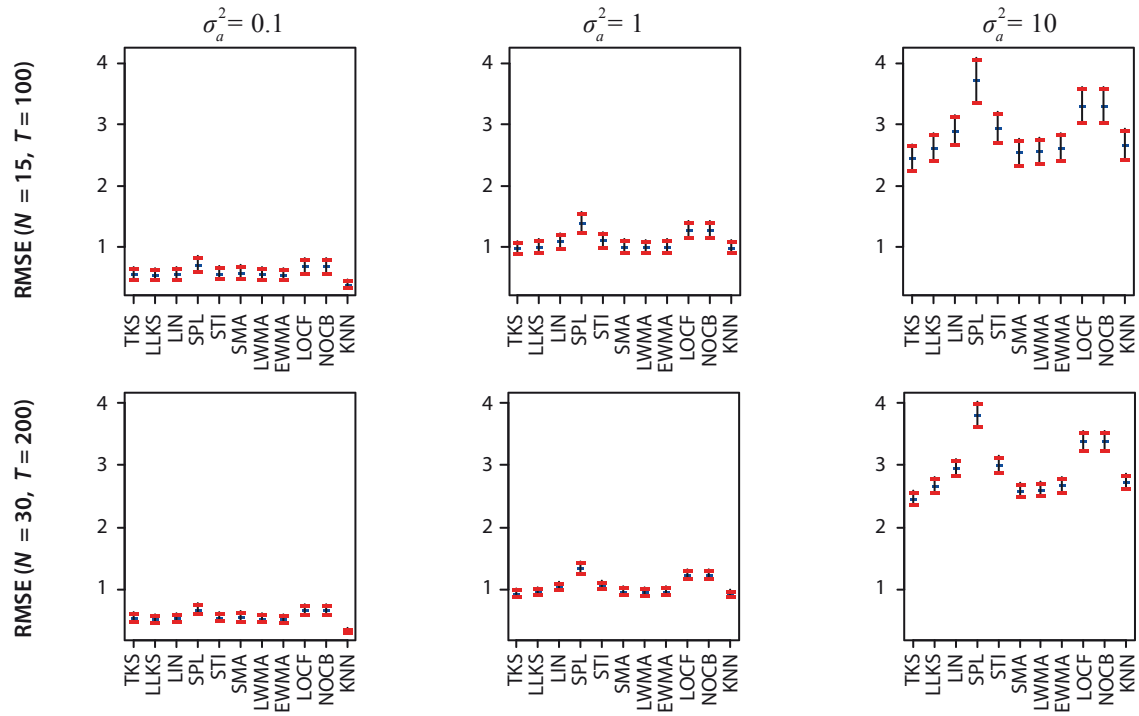
Gráfica 3

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M1 con errores heterocedásticos y $\Gamma = diag(0.99)$



Gráfica 4

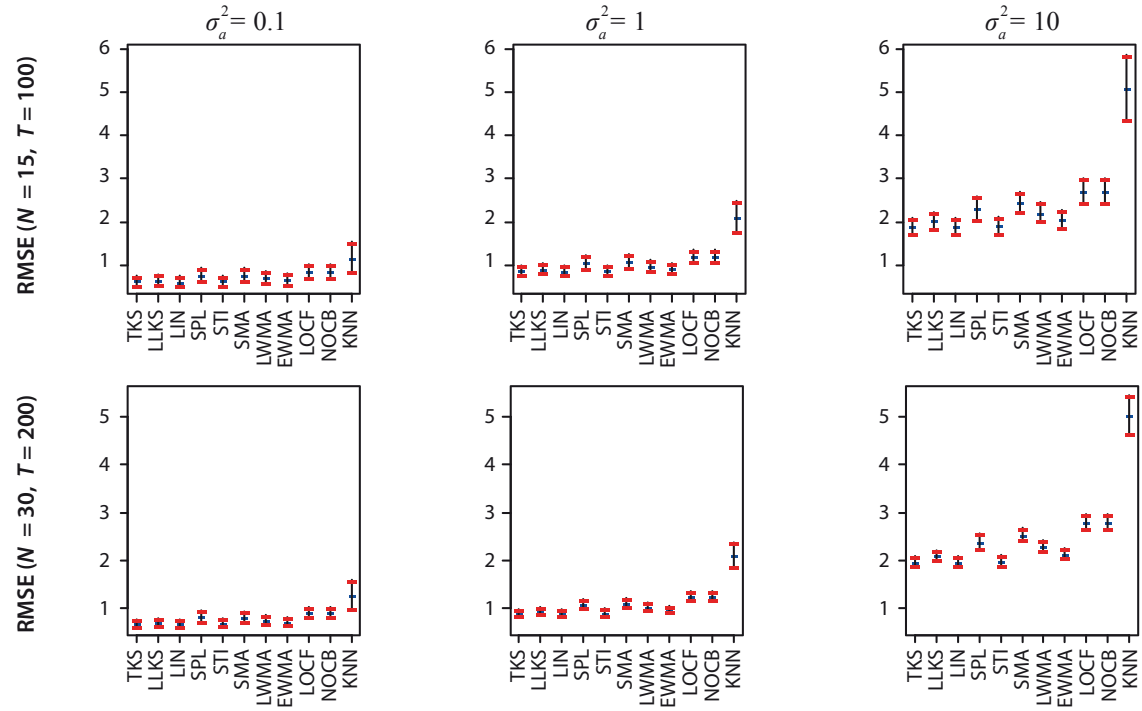
Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M2 con errores heterocedásticos y $\Gamma = diag(0)$



INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

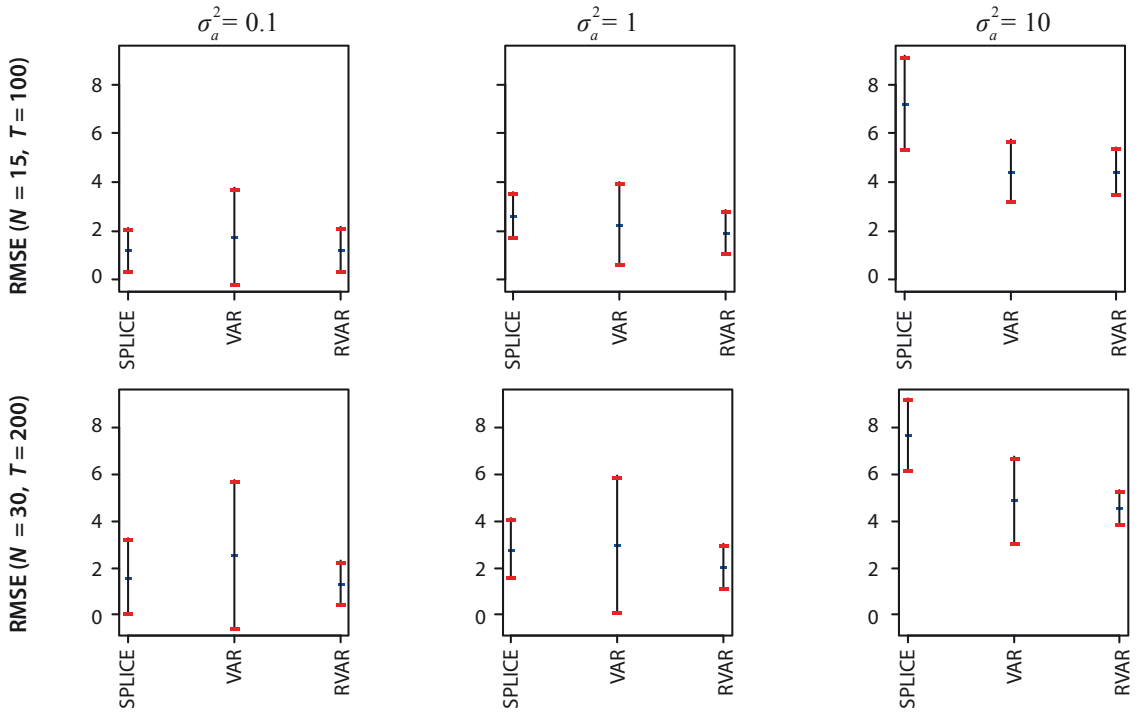
Gráfica 5

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M3 con errores heterocedásticos y $\Gamma = diag(0.99)$



Gráfica 6

Intervalos de confianza (95%) de los RMSE para los distintos métodos de retropolación para M1 con errores heterocedásticos y $\Gamma = diag(-0.8)$



$\sigma_a^2 < 10$. Sin embargo, cuando $\sigma_a^2 = 10$, los métodos presentan una variabiliad similar y, más aún, los intervalos de los RMSE tienden a intersectarse con excepción de SPL, LOCF y NOCB, por lo que, en la práctica, el resto de los métodos de imputación que no son estos tres tienden a funcionar casi igual cuando la varianza del error del componente idiosincrático no es grande.

Para finalizar, la gráfica 5 muestra los mismos resultados que la 4, pero ahora para el M3 considerando el caso cuando el componente idiosincrático tiene una matriz de coeficientes autorregresivos $\Gamma = diag(0.99)$. Este caso es de interés porque, aunque en teoría, las series de tiempo son cointegradas, notemos que, dependiendo de las cargas de λ, las series pueden estar al límite de ser caminatas aleatorias independientes, por ejemplo cuando $y_{it} = \lambda_1 F_{1t} + \varepsilon_{it} (\lambda_2 = 0)$ o bien, el caso de cuando comparten un factor estacionario, pero domina el comportamiento individual no estacionario, es decir, cuando $y_{it} = \lambda_2 F_{2t} + \varepsilon_{it}$ $(\lambda_1 = 0)$, bajo el argumento de que ε_t tiende a ser una serie de tiempo no estacionaria.

Se puede apreciar que KNN funciona de manera deficiente. Esto tiene que ver con que estamos utilizando información de *vecinos* con los cuales la relación puede considerarse, casi, espuria. Este efecto se vuelve aún más importante conforme σ_a^2 crece. La presencia del factor estacionario es irrelevante y esto está relacionado con el hecho de que la variación en las observaciones es dominada por la variación del factor común no estacionario.

Como conclusión, y dadas las características consideradas en este experimento, los métodos basados en el filtro de Kalman tienden a funcionar, en promedio, un poco mejor que el resto de los procedimientos, sobre todo el TKS. Otro resultado importante es que, ya que la variabilidad alrededor de la media es similar entre los basados en medias móviles y el filtro de Kalman, podemos concluir que, ya que los intervalos se intersectan, dichas familias de métodos tienden a funcionar de manera similar. También, el utilizar un *spline* lineal funciona, incluso, mejor que el de la misma familia, es decir, el interpolador denotado como STI. Asimismo, los métodos más débiles para imputar datos resultaron ser el SPL y los basados en la información disponible en la muestra. Otra conclusión muy importante es que, aunque el KNN tiene un pobre funcionamiento cuando no hay *buenos vecinos,* funciona bien cuando los vecinos utilizados aportan información relevante y esto ocurre con frecuencia en la práctica.

5.2 Retropolación

En este contexto, como comentamos en la sección 3, nos interesa saber qué método *retropola* de forma más acertada las observaciones faltantes. La gráfica 6 presenta los resultados de los métodos de retropolación para el M1 cuando $\Gamma = diag(-0.8)$.

Resulta interesante observar que, sin considerar el tamaño de muestra y de la varianza asociada al error del componente idiosincrático, el RVAR tiende a funcionar hasta cierto punto mejor, dado que obtenemos medias y varianzas más pequeñas en los RMSE calculadas a partir de las réplicas. En otras palabras, tenemos estimaciones con una mayor precisión y una menor incertidumbre. No obstante, cuando la varianza del error del componente idiosincrático es pequeña, SPLICE es muy competitivo. Esto es relevante si consideramos lo flexible que resulta implementar este método respecto al VAR y al RVAR, los cuales requieren una mayor cantidad de información y supuestos para su respectiva implementación. Solo en el caso cuando $\sigma_a^2 = 10$, el método RVAR —que supone la existencia de restricciones, en este caso temporales— es claramente el mejor.

Los resultados obtenidos para el M1 al ir variando el parámetro de autocorrelación en los componentes idiosincráticos no muestran cambios relevantes respecto a lo obtenido en la gráfica 6, motivo por el cual solo presentamos este caso para el M1.

La gráfica 7 muestra los resultados encontrados para el M2 cuando $\Gamma = diag(0)$, es decir, cuando

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

las observaciones tienen un factor común estacionario y los errores individuales son ruido blanco.

Podemos apreciar resultados parecidos a los mostrados en la gráfica 6, aunque los intervalos para RVAR se hacen más grandes, lo cual nos muestra que las predicciones tienen una incertidumbre más alta en el caso estacionario que en el no estacionario. Nótese que al generar series de tiempo cointegradas, aprovechamos su tendencia común al pronosticarlas, motivo por el cual las predicciones son más certeras en el caso no estacionario. En este modelo también se observa que cuando $\sigma_a^2 = 10$, el método RVAR es radicalmente mejor tanto respecto al SPLICE como al VAR.

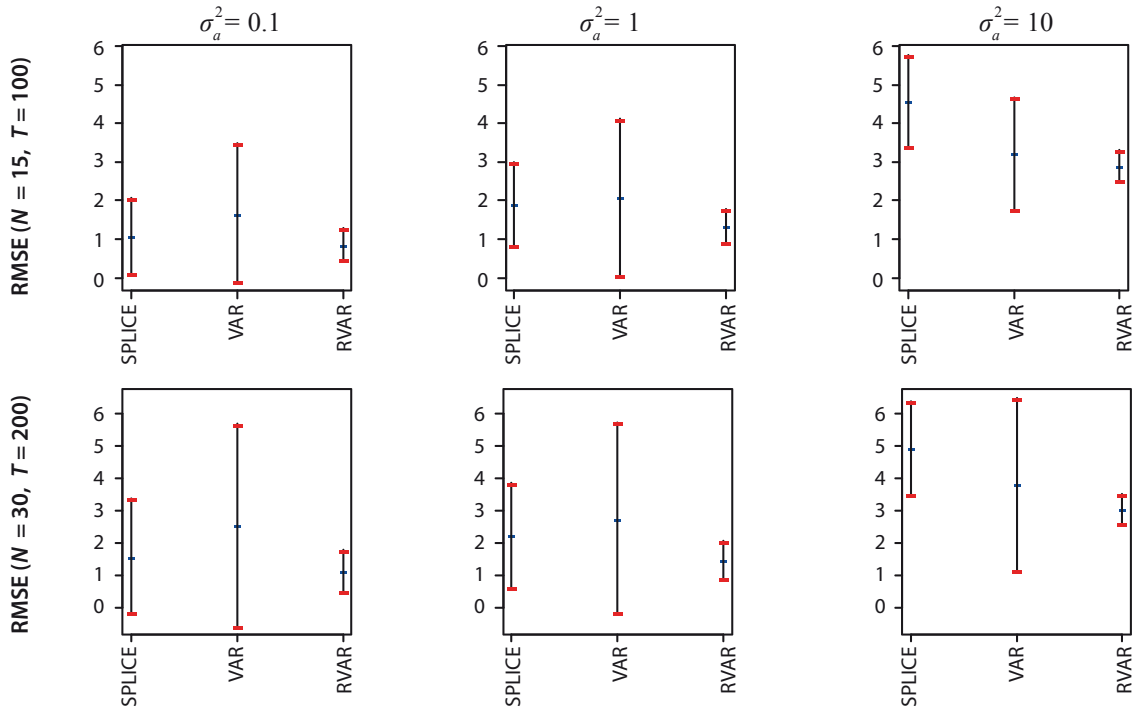
Por último, la gráfica 8 presenta lo obtenido para el M3 cuando $\Gamma = diag(0.99)$.

Solo se pueden apreciar algunos cambios respecto a la gráfica 6, es decir, y al igual que en el caso de imputación, el factor no estacionario domina sobre el estacionario; no obstante, podemos apreciar ahora que, cuando $\sigma_a^2 = 1$ y el tamaño de muestra es más grande, el RVAR presenta intervalos de confianza más pequeños que el SPLICE y el VAR.

Un resultado importante es que, aun cuando el RVAR da resultados más certeros y con menor incertidumbre, el SPLICE puede considerarse estadísticamente competitivo, sobre todo en los casos cuando la variabilidad individual de las observaciones es pequeña, incluso en algunos casos funciona mejor que el VAR y esto resulta de utilidad porque su implementación requiere solo del supuesto que la variable de empalme esté correlacionada con la variable a retropolar. Nótese que el VAR y el RVAR requieren de un número suficiente de observaciones para poder ser implementados (se estiman $N + Np^2$ parámetros), además de que, de manera empírica, es necesario validar que las series son cointegradas (pronóstico en niveles) y que los residuales están libres de autocorrelación.

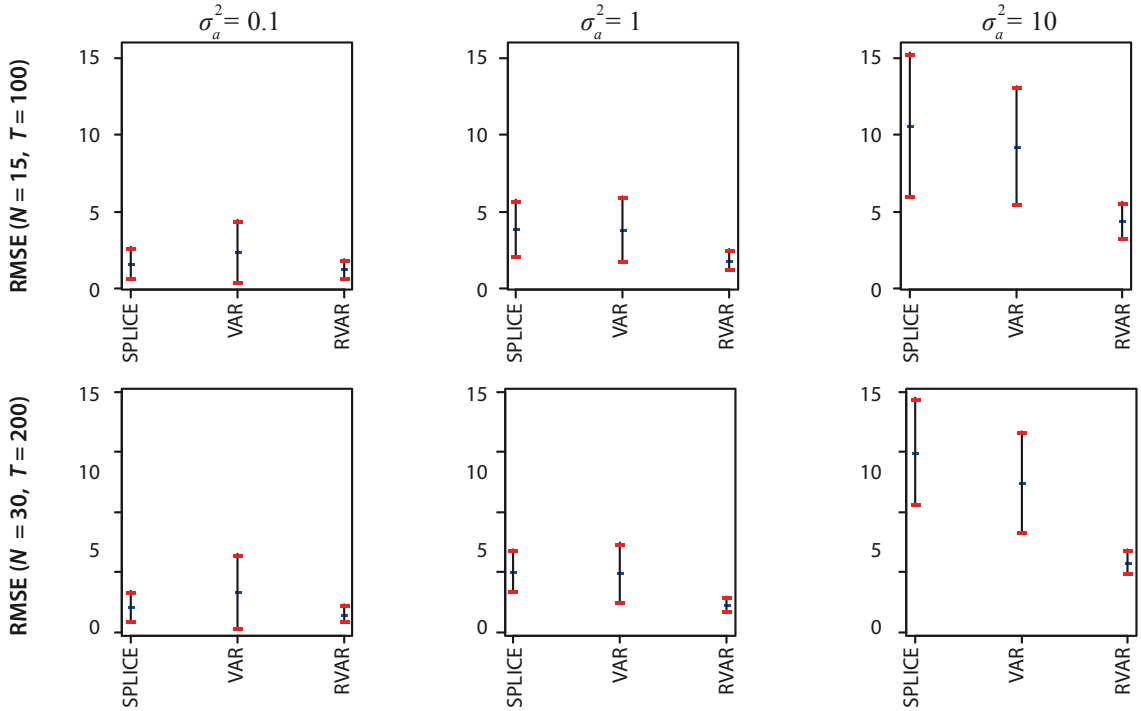
Gráfica 7

Intervalos de confianza (95%) de los RMSE para los distintos métodos de retropolación para M1 con errores heterocedásticos y $\Gamma = diag(0)$



Gráfica 8

Intervalos de confianza (95%) de los RMSE para los distintos métodos de retropolación para M3 con errores heterocedásticos y $\Gamma = diag(0.99)$



6. Conclusiones y recomendaciones

Este trabajo se orientó en evaluar el funcionamiento en muestras finitas de diferentes métodos de imputación y retropolación en el contexto de series de tiempo, cuyos resultados puedan ser útiles para los generadores de información oficial. Nos enfocamos en simular paneles de series de tiempo como DFM, cuyas características fueron validadas para cinco bases de datos de las EEN del INEGI. En este sentido, se consideraron DFM estacionales con factores no estacionarios, estacionarios y la combinación de ambos, variando sobre todo el tamaño de muestra, los parámetros de autocorrelación en el componente idiosincrático y diferentes estructuras de dependencias en estos, como la homocedasticidad, heterocedasticidad y la autocorrelación cruzada. En otras palabras, se simularon series de tiempo que se observan con frecuencia en la práctica. Se pudo ver que, en el contexto de imputación, los métodos que utilizan en el filtro de Kalman son los más competitivos, aunque su rendimiento es similar a los que se basan en medias móviles. También, el KNN es el más competitivo cuando la información de los *vecinos más cercanos* es de calidad, situación relacionada con el caso de DFM cointegrados. Los métodos basados en la información disponible en la muestra y *splines* cúbicas resultaron ser los menos precisos. En lo que respecta a la retropolación, el RVAR es de forma clara el mejor de los tres analizados; no obstante, se requiere conocer, en cada caso particular, las restricciones lineales adecuadas y necesarias para su implementación. En caso de no conocer estas restricciones, se observa que el SPLICE otorga con frecuencia mejores resultados que los obtenidos a través de un VAR.

De esta forma, se recomienda a los generadores de información que, para imputar, es útil usar in-

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

formación de variables la cual esté correlacionada en algún sentido, estadístico o económico, con la serie de tiempo de interés. En caso de no contar con este tipo de información, se recomienda preferir los métodos basados en el filtro de Kalman o en medias móviles. Para retropolar, es claro que tener más información del pasado de la serie de tiempo permite minimizar el error atribuible a la estimación, lo anterior restringiendo los pronósticos obtenidos en una fase anterior a través de un modelo VAR; no obstante, si no existen dichas restricciones, empalmar series de tiempo correlacionadas entre sí resulta una alternativa flexible y útil.

Fuentes

Ahn, Seung C. y Alex R. Horenstein. "Eigenvalue ratio test for the number of factors", en: *Econometrica*. Vol. 81, núm. 3. 2013, pp. 1203-1227.

Bai, Jushan. "Estimating cross-section common stochastic trends in nonstationary panel data", en: *Journal of Econometrics*. Vol. 122, núm. 1. 2004, pp. 137-183.

Corona, Francisco, Pilar Poncela y Esther Ruiz. "Estimating non-stationary common factors: Implications for risk sharing", en: *DES-Working Papers. Statistics and Econometrics*. WS 24585. Departamento de Estadística, Universidad Carlos III de Madrid, 2017.

De la Fuente Moreno, Ángel. "A mixed splicing procedure for economic time series", en: *Estadística Española*. Vol. 56, núm. 183. 2014, pp. 107-121.

Forsythe, George. E., Cleve B. Moler y Michael A. Malcolm. *Computer Methods for Mathematical Computations*. Prentice-Hall, 1977.

Guerrero, Víctor M. y Blanca I. Gaspar. "Edition and Imputation of Multiple Time Series Data Generated by Repetitive Surveys", en: *Journal of Data Science*. Vol. 8, núm. 4. 2010, pp. 555-577.

Guerrero, Víctor M. y Daniel Peña. "Combining multiple time series predictors: a useful inferential procedure", en: *Journal of Statistical Planning and Inference*. Vol. 116, núm 1. 2003, pp. 249-276.

Guerrero, Víctor M. y Francisco Corona. "Retropolating some relevant series of Mexico's System of National Accounts at constant prices: The case of Mexico City's GDP", en: *Statistica Neerlandica*. Vol. 72, núm. 4. 2018, pp. 495-519.

Harvey, Andrew C. y Richard G. Pierse. "Estimating missing observations in economic time series", en: *Journal of the American Statistical Association*. Vol. 79, núm. 385. 1984, pp. 125-131.

Helske, Jouni. "KFAS: Exponential Family State Space Models in R", en: *Journal of Statistical Software*. Vol. 78, núm 10. 2016, pp. 1-39.

Hyndman, Robert J. y Yeasmine Khandakar. "Automatic time series forecasting: the forecast package for R", en: *Journal of Statistical Software*. Vol. 26, núm. 3. 2007, pp. 1-22.

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

Kowarik, Alexander and Matthias Templ. "Imputation with R package VIM", en: *Journal of Statistical Software*. Vol. 74, núm. 7. 2016, pp. 1-16.

Lütkepohl, Helmut. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Luzi, Orietta., *et al*. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. ISTAT, CBS, SFSO, Eurostat, 2007.

Moritz, Steffen *et al. Comparison of different methods for univariate time series imputation in R*. arXiv preprint arXiv:1510.03924, 2015.

Moritz, Steffen y Thomas Bartz-Beielstein. "ImputeTS: time series missing value imputation in R", en: *The R Journal*. Vol. 9, núm. 1. 2017, pp. 207-218.

Nieto, Fabio. H., Daniel Peña y Dagoberto Saboyá. "Common seasonality in multivariate time series", en: *Statistica Sinica*. Vol 26, núm. 4. 2016, pp. 1389-1410.

Onatski, Alexei. "Determining the number of factors from empirical distribution of eigenvalues", en: *The Review of Economics and Statistics*. Vol. 92, núm 4. 2010, pp. 1004-1016.

Peña, Daniel & George C. Tiao. "A note on likelihood estimation of missing values in time series", en: *The American Statistician*. Vol. 45, núm. 3. 1991, pp. 212-213.

Pfaff, Bernhard. "VAR, SVAR and SVEC Models: Implementation Within R Package vars", en: *Journal of Statistical Software*. Vol. 27, núm. 4. 2008, pp. 1-32.

Pfeffermann, Danny y Gad Nathan. "Imputation for wave nonresponse: Existing methods and a time series approach", en: Groves, Robert *et al*. (eds.). *Survey Nonresponse*. New York, Wiley, 2001, pp. 417-429.

Pratama, Irfan, *et al*. "A review of missing values handling methods on time-series data", en: *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*. Bandung, Indonesia, octubre de 2016, pp. 1-6.

R Core Team. "R: A language and environment for statistical computing", en: *R Foundation for Statistical Computing*. Vienna, Austria, 2017 (DE) https://www.R-project.org/

Schafer, Joseph L. y John W. Graham. "Missing data: our view of the state of the art", en: *Psychological Methods*. Vol. 7, núm. 2. 2002, pp. 147-177.

Schmitt, Peter, Jonas Mandel y Mickael Guedj. "A comparison of six methods for missing data imputation", en: *Journal of Biometrics & Biostatistics*. Vol. 6, núm. 1. 2015, pp. 1-6.

Stineman, Russel W. "A Consistently Well Behaved Method of Interpolation", en: *Creative Computing*. Vol. 6, núm. 7. 1980, pp. 54-57.

Colaboran en este número

Benito Durán Romo

Es licenciado en Informática por el Instituto Tecnológico de Aguascalientes, con una primera maestría en esa materia por la Universidad Autónoma de Aguascalientes y una segunda en Análisis Estadístico por el Centro de Investigación en Matemáticas (CIMAT), AC. Desde 1989, labora en el Instituto Nacional de Estadística y Geografía (INEGI), donde inició en el procesamiento de encuestas especiales en la entonces Dirección de Estadísticas de Corto Plazo; después, colaboró en el procesamiento de la Encuesta Nacional de Ingresos y Gastos de los Hogares para luego fungir como responsable del área de procesamiento de esta; actualmente, es subdirector de Investigación en Indicadores Sociales, Demográficos y Económicos en la Dirección General Adjunta de Investigación del INEGI. Sus temas de interés son la imputación automática, el aprendizaje máquina y los *ninis*.

Contacto: benito.duran@inegi.org.mx

Carlos Samuel Pérez Pérez

De nacionalidad mexicana. Es licenciado en Matemáticas Aplicadas y Economía por el Instituto Tecnológico Autónomo de México (ITAM) y matemático por la Universidad Nacional Autónoma de México (UNAM); en fecha reciente, terminó los estudios de la Maestría en Ciencias de Datos en el ITAM. En su carrera profesional, ha participado en conferencias internacionales de estadística y sistemas complejos; ha formado parte del personal docente de la UNAM y el ITAM y ha colaborado en decenas de proyectos de investigación y consultoría de índole pública y privada. En la actualidad, se desempeña como consultor en proyectos de ciencias de datos y como colaborador de la División Académica de Actuaría, Estadística y Matemáticas del ITAM.

Contacto: carlos.perez@itam.mx

Luis Enrique Nieto Barajas

De nacionalidad mexicana. Es doctor en Estadística por la Universidad de Bath en Inglaterra. Pertenece al Sistema Nacional de Investigadores (SNI) con nivel II y ha publicado más de 40 artículos de investigación en revistas internacionales arbitradas. Es presidente de la Asociación Mexicana de Estadística para el periodo 2019-2021. En la actualidad, se desempeña como profesor de tiempo completo en el Departamento de Estadística del ITAM.

Contacto: lnieto@itam.mx

Olinca Páez

Mexicana. Licenciada en Economía por la Universidad Veracruzana y maestra en Demografía por El Colegio de México (El COLMEX); cursó dos diplomados en el Centro de Investigación y Docencia Económicas (CIDE) Región Centro, uno sobre Género, Sexualidad y Derecho y otro en Gobierno, Gestión y Políticas Públicas. Es investigadora desde el 2003 y ha publicado sobre temas económicos y demográficos para un público diverso, apoyándose de técnicas estadísticas y econométricas. En el 2012

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 10, Num. 3, septiembre-diciembre, 2019. 2022

obtuvo el segundo lugar del Premio Nacional de Investigación Social y de Opinión Pública que otorga el Centro de Estudios Sociales y de Opinión Pública de la Cámara de Diputados. En el 2018 fue beneficiaria del *Trust Fund for Statistical Capacity Building* del Banco Mundial. Del 2015 al 2018 formó parte del grupo internacional encargado de desarrollar el reporte *Measuring international labour mobility*. En la actualidad, es subdirectora de Investigación de Información Econométrica en el INEGI.

Contacto: olinca.paez@inegi.org.mx

Brenda Murillo-Villanueva

Nació en México. Es doctora en Economía por la Facultad de Economía de la UNAM. Actualmente, es profesora-investigadora del Centro de Investigación en Ciencias Económicas de la Universidad Autónoma del Estado de México (UAEM). Es candidata a investigadora nacional del SNI.

Contacto: bmurillo@uaemex.mx

Martín Puchet Anyul

Nació en Durazno, Uruguay. Estudió Economía en la Universidad de la República (Montevideo) y obtuvo su maestría y doctorado en esa disciplina en el CIDE y en la UNAM en México, respectivamente. En esa misma casa de estudios se desempeña como profesor titular de métodos cuantitativos desde 1990; además, es tutor de maestría y doctorado de los programas de posgrado en Economía (desde el 2001) y en Filosofía de la Ciencia (desde el 2010). Es miembro del SNI con nivel III.

Contacto: anyul@unam.mx

Gerardo Fujii-Gambero

Nació en Chile. Obtuvo su doctorado en Economía en la Universidad de Lomonósov de Moscú, Rusia. Es tutor del programa de doctorado en Economía e investigador en la Facultad de Economía de la UNAM. Pertenece al SNI con nivel II.

Contacto: fujii@unam.mx

Marta Mier y Terán Rocha

Mexicana de nacimiento. Obtuvo su doctorado en Demografía en la Universidad de Montreal, Canadá. Es investigadora titular en el Instituto de Investigaciones Sociales de la UNAM y docente en la Maestría en Demografía Social del Posgrado en Ciencias Políticas y Sociales de la misma casa de estudios. Pertenece al SNI del Consejo Nacional de Ciencia y Tecnología con el nivel II, y sus líneas de investigación son: estimación y análisis de la fecundidad; transiciones a la vida adulta: escuela, trabajo y formación de familias entre los jóvenes; así como curso de vida y trayectorias familiares y laborales. En la actualidad, desarrolla los proyectos *Evaluación de las fuentes de datos para el estudio de la fecundidad en México* y *Trayectorias familiares y laborales durante el proceso de transición de la fecundidad en México*. Cuenta con numerosos trabajos publicados en libros y revistas nacionales y del extranjero. Ha impartido cursos y dirigido tesis en la UNAM y otras instituciones de educación superior en el campo de los estudios de la población.

Contacto: martamyt@sociales.unam.mx

Víctor Manuel García Guerrero

Nació en México. Es doctor en Estudios de Población por El COLMEX. Se desempeña como profesor-investigador de tiempo completo del Centro de Estudios Demográficos, Urbanos y Ambientales de El COLMEX. Ha sido asesor en métodos demográficos del Fondo de Población de Naciones Unidas, el Consejo Nacional de Población, la Secretaría de Desarrollo Social, el Banco Interamericano de Desarrollo y distintos despachos de consultoría, bancos y aseguradoras. Es investigador nacional nivel II por el SNI; sus temas de investigación son: estimaciones y proyecciones de población y su uso en la política pública y toma de decisiones, así como modelación matemática, estadística y computacional de la mortalidad, fecundidad y migraciones.

Contacto: vmgarcia@colmex.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). 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; también, tiene una en *Sport Analytics*; de ambas líneas ha publicado trabajos en revistas arbitradas de circulación internacional. Pertenece al SNI, nivel candidato.

Contacto: franciscoj.corona@inegi.org.mx

Jesús López-Pérez

De nacionalidad mexicana. Es licenciado en Economía y maestro en Estadística Aplicada por el ITESM. En la actualidad, es investigador invitado 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

Nelson Omar Muriel Torrero

De nacionalidad mexicana, es actuario, maestro y doctor en Ciencias Matemáticas por la UNAM. Es profesor de tiempo completo en el Departamento de Física y Matemáticas de la Universidad Iberoamericana CDMX. Ha trabajado para distintas instituciones educativas, como: la Facultad de Ciencias de la UNAM, el CIMAT, en Guanajuato, el Departamento de Economía de la UC3M, en España, y el Centro Universitario de Ciencias Económico-Administrativas de la Universidad de Guadalajara impartiendo cursos para licenciaturas, maestrías y programas de doctorado. Su principal área de investigación es la teoría econométrica con énfasis en el análisis de series de tiempo, y su aplicación en áreas como las finanzas, el crecimiento económico y el estudio de la criminalidad.

Contacto: nelson.muriel@ibero.mx

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**

¿DÓNDE NACISTE?

¡PREGÚÚÚNTAME!

YA VIENE EL CENSO
MARZO 2020

INEGI

2020 censo de Población y Vivienda

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

MÉXICO