Vol. 6, Núm. 3, septiembre-diciembre 2015.

ISSN 2007-2961

REALIDAD, DATOS Y ESPACIO. REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

INTERNATIONAL JOURNAL OF STATISTICS AND GEOGRAPHY

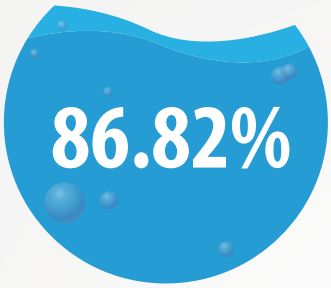
REALITY, DATA AND SPACE.

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

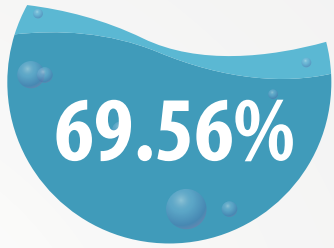
Agua y desarrollo sostenible en México

Cobertura a nivel nacional

Agua potable y alcantarillado



Tratamiento y disposición de aguas residuales

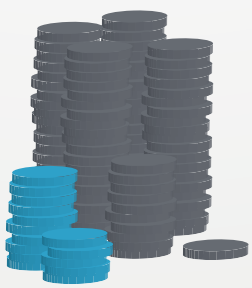


Extracción y recarga de acuíferos
(millones de metros cúbicos)

Fuente: INEGI. *Sistema de Cuentas Nacionales de México. Cuentas Económicas y Ecológicas de México, 2013.* Preliminar, año base 2008, datos del Banco Mundial.

La **actividad agropecuaria** fue la que más consumió el recurso hídrico subterráneo, con 70 por ciento.

Los gastos en protección ambiental fueron de **148 699 millones de pesos.**



Los costos totales por agotamiento y degradación ambiental (CTADA) fueron de **909 968 millones de pesos.**

Fuentes: INEGI. *Censo Nacional de Gobiernos Municipales y Delegacionales 2013.*
——— *Sistema de Cuentas Nacionales de México. Cuentas Económicas y Ecológicas de México, 2013.* Preliminar, año base 2008, datos del Banco Mundial.

Conociendo México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa **@INEGI_INFORMA**

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Vol. 6, Núm. 3, septiembre-diciembre 2015.

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

Contenido

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

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 Plataforma Open Access de Revistas Científicas Electrónicas Españolas y Latinoamericanas *e-Revist@s.*

REALIDAD, DATOS Y ESPACIO. REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA, Vol. 6, Núm. 3, septiembre-diciembre 2015, 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: Enrique Jesús Ordaz López. 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.

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

Editorial

Mantener una revista durante cinco años no ha sido fácil; el trabajo alrededor de Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía ha sido arduo y la experiencia, enriquecedora. Hemos salido en busca de nuevas formas de producción editorial, que van de la mano de las actuales tecnologías digitales de transmitir el conocimiento, a través de los estándares que nos permitan tener mayores oportunidades de visibilidad en índices y bases de datos internacionales, elementos indispensables para certificar la calidad de una publicación.

Además de *Latindex Catálogo* y *CLASE*, obtuvimos el registro en la Plataforma *Open Access* de Revistas Científicas Electrónicas Españolas y Latinoamericanas *e-Revist@s*.

Para cerrar este 2015, el primer artículo, *Mismos individuos en diferentes bases de datos sin identificador común: la unión de las bases de datos de beneficiarios de la SAGARPA con una base universal*, presenta la metodología que permita obtener tal identificador, usualmente carente en las bases de datos de interés público —como aquéllas—, para lograr su vinculación.

El siguiente, *Análisis de los microdatos del censo de 1930: a 80 años del México posrevolucionario*, es una aportación que los autores hacen para contribuir a completar la historia demográfica de México mediante estimaciones con microdatos recuperados del Quinto Censo de Población.

Estimación del ingreso por trabajo en los municipios y las delegaciones de México utilizando técnicas de estimación para áreas pequeñas presenta un ejercicio con modelos lineales mixtos que compara los resultados de la estimación directa en áreas grandes —a partir de los datos recabados por la ENIGH 2010— con los obtenidos mediante dichas técnicas.

Más adelante, en el artículo *Clasificación de cultivos agrícolas utilizando técnicas clásicas de procesamiento de imágenes y redes neuronales artificiales*, se muestra una metodología basada en técnicas de procesamiento de imágenes satelitales de baja resolución y reconocimiento de patrones que permite realizar la adecuada clasificación de áreas agrícolas en una región de prueba en Sinaloa, México.

Cambios recientes en la esperanza de vida en México, análisis por medio de su descomposición lleva a la reflexión de conocer sus modificaciones negativas en distintos ámbitos y niveles del país para estar en posición de apoyar en la producción de medidas que aumenten el bienestar de la población.

Por último, *Notes on the Relationship between Trade and Employment in U. S. Manufacturing Sector, 1998-2008* es el resultado de una investigación que incorpora elementos relevantes para analizar la relación entre el comercio (exportaciones e importaciones) y el empleo en ese sector del vecino del norte.

Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía ha contado con la participación y colaboración de numerosas personas en el quehacer de cada número: para cada una de ellas, va el reconocimiento del INEGI.

http://rde.inegi.org.mx

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

A quienes han hecho **Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía**

Con este último número del 2015 de la Revista, el Instituto Nacional de Estadística y Geografía reconoce la labor de las personas que han hecho posible estos más de cinco años de publicación de un extenso proceso que involucra, en primera instancia, tanto los textos que han enviado los autores como el trabajo de selección y dictaminación para hacer posible la integración de cada edición, seguida por la corrección de estilo, la traducción, el diseño y la edición, así como la fase de preprensa e impresión final y la preparación de la alternativa digital para internet. No se olvida también la importancia que conlleva su difusión y distribución a distintos usuarios.

Es así que durante este lustro se han gestado importantes cambios que han dado la pauta para continuar con la realidad de este proyecto. Todavía hay mucho trabajo por hacer y mucho que aportar.

Nuestro sincero agradecimiento, no sin antes pedir una disculpa por alguna omisión involuntaria, a:

Abbdel Camargo Martínez
Abel Gómez Gutiérrez
Abel López Pacheco
Abelardo Aníbal Gutiérrez Lara
Abigail Álvarez Torres
Abigail Rojas Huerta
Abraham Aparicio Cabrera
Abraham Toriz Cruz
Ada Borjas López
Adalberto Tejeda-Martínez
Adela Angoa
Adolfo Alberto Laborde Carranco
Adolfo Sánchez Almanza
Adrián Antonio Silva Jiménez
Adrián Franco Barrios
Adriana del Carmen Riveroll Arellano
Adriana Sletza Ortega Ramírez
Adriana Verónica Hinojosa Cruz
Agustín Escobar Latapí
Aída Díaz-Tendero Bollain
Alan Peugnet Núñez
Alberto Contreras Cristán
Alberto Moritz Cruz
Alberto Ortega y Venzor
Alberto Palloni
Alejandra Arroyo Martínez Sotomayor
Alejandra López Caloca
Alejandro Cervantes Mancilla
Alejandro Contreras Lugo
Alejandro Federico Alva Martínez
Alejandro González Martínez
Alejandro González Roque
Alejandro Mina Valdés
Alejandro Tapia Vargas
Alenka Guzmán Chávez
Alfonso Mejía Modesto
Alfonso Mendoza Velázquez
Alfonso Mercado García
Alfredo Amador García
Alfredo Bustos y de la Tijera
Alfredo Luna Arrioja
Alicia Girón González
Álvaro Ramírez Alujas
Amando Ramos González
Amarela Varela Huerta
Ana Delia Contreras Hernández
Ana García de Fuentes
Ana Karen Díaz
Ana Luisa González Arévalo
Ana Magdalena Castellanos López
Ana Melisa Pardo Montaño
Ana Sojo
Ana Villafuerte Islas
Anastasio Reyes Pérez
Andrés Flores Montalvo
Andrés Reyes Pérez
Ángel Fernando Argüello Ortiz
Ángel Manuel Ortiz Marín
Angélica Álvarez
Angélica Hernández Quintero
Angélica Rivera Olvera
Antonina Galván Fernández
Antonina Ivanova Boncheva
Antonio Baldemar Méndez
Araceli Damián González
Ariadna García García
Armando Aguiar Rodríguez
Armando Gómez Guerrero
Armando Ramos González
Armando Sánchez Vargas
Arsenio Merlos Rodríguez
Artemio Gallegos García
Arturo Antón Sarabia
Arturo Blancas Espejo
Arturo Colín Cruz
Arturo Federico Belmonte Preza
Atocha Aliseda Llera
Baldemar Méndez Antonio
Beatriz Meneses Aguirre
Beatriz Mondragón Hernández
Beatriz Romero Sánchez
Belém Trejo Valdivia
Benjamín Martínez López
Benjamín Ortiz Espejel
Benjamín Temkin Yedwab
Berenice Patricia Ramírez López
Bernardo Hernández Prado
Bernardo Olmedo Carranza
Brígida García Guzmán
Bruno López-Videla Mostajo
Carla Cano Álvarez
Carla Pederzini Villarreal
Carlos Alberto Francisco Cruz
Carlos Alberto Mercado Salvador

Carlos Álvarez Jiménez
Carlos Eduardo Canfield Rivera
Carlos Félix Garrocho Rangel
Carlos Gay García
Carlos Guerrero de Lizardi
Carlos Guerrero Élemen
Carlos Javier Cervantes
Carlos López Portillo Tostado
Carlos Luque Ancona
Carlos Manuel Welsh Rodríguez
Carlos Patiño Gómez
Carlos Roberto López Pérez
Carlos Rodolfo Torres Navarrete
Carole Farell Baril
Carolina Andrea Ochoa Martínez
Carolina Martínez Salgado
Catalina Pérez Correa González
Cecilia Andrea Rabell Romero
Cecilia Guadarrama Tavira
Celia Romero Islas
César Bistrain Coronado
César García Callejas
Christian Eduardo Díaz Sosa
Claudia Marisol Serna Martínez
Claudia Mireya Ramírez Núñez
Claudia Paloma Delgado Sánchez
Claudia Paloma Salas Esparza
Claudia Patricia Pardo Hernández
Claudio Alberto Dávila Cervantes
Consuelo Mañón Salas
Covadonga Escandón Martínez
Cristina Guirette Saldaña
Cristina Gutiérrez Delgado
Cristinne Leo Martel
Cuauhtémoc Ruiz Esparza Pérez
Daniel Ventosa-Santaulària
Daniela Francisca Díaz Echeverría
David Arellano Gault
David Castro Lugo
David Madrigal González
David Márquez Molina
David Moctezuma Navarro
David Mulato Martínez
David Rocha Romero
Delfino Vargas Chanes
Delia Margarita Vergara Reyes
Diana Betancourt Ocampo
Diego Hernández
Dolores Edwiges Luna Reyes
Domitilo Pereyra Díaz
Dora Elvira García
Edna Jaime Treviño
Eduardo Gutiérrez Peña
Eduardo Javier Ramírez Espino
Eduardo Rodríguez-Oreggia †
Eduardo Sojo Garza Aldape
Efraín Martínez Pérez
Elena Azaola Garrido
Eliana Rocío González Molano
Elías Lagunas Canchola
Eliseo Cantellano de Rosas
Eliseo Gerardo Guerrero Eufrasio
Elizabeth Bello Hernández
Eloísa Domínguez Mariani
Elsa Leticia Flores Márquez
Emma Liliana Navarrete
Emma Mendoza Bremauntz
Enoch Bonilla Jiménez
Enrique Cabrero Mendoza
Enrique Cortés Martínez
Enrique de Alba Guerra
Enrique de la Garza Toledo
Enrique Dussel Peters
Enrique Fernando Nava López
Enrique García Ramírez
Enrique Hernández Laos
Enrique Lora Toro
Enrique Minor Campa
Enrique Ordaz López
Enrique Ortiz González
Eric Manuel Rodríguez Herrera
Erick Sánchez Flores
Erick Thorbecke
Érika Gómez Olivares
Ernesto Carlos Leyva Pedrosa
Ernesto Espíndola Advis
Esperanza Vargas Álvarez
Estela Rivero Fuentes
Eugenio Corpus Moreno
Eugenio Gómez Reyes
Eunice Danitza Vargas Valle
Eva Castillo Navarrete
Fabiola Sosa Rodríguez
Federico Arellano Segura

Federico Horacio Dickinson Bannack
Felipe Máximo Bello
Felipe Pérez Gachuz
Felipe Roboam Vázquez Palacios
Félix Acosta Díaz
Félix Vélez Fernández Varela
Fernando Chávez Gutiérrez
Fernando Cortés Cáceres
Fernando Estrada Hernández
Fernando Filgueira Prates
Fernando Javier Chávez Gutiérrez
Fernando Morales Gómez
Fernando Nava
Fernando Riosmena
Fernando Toriz
Ferrán Padrós Blázquez
Fidel Aroche Reyes
Fidel Pérez Moreno
Fiorella Mancini
Florina Arredondo Trapero
Francisco Estrada Porrúa
Francisco González Munive
Francisco Gurri
Francisco Guzmán López Figueroa
Francisco Hansen Albites
Francisco Javier Moreno Núñez
Francisco Javier Rivas Rodríguez
Francisco Javier Solís Delgado
Francisco Javier Soto Acosta
Francisco José Zamudio Sánchez
Francisco Marroquín Figueroa
Francisco Moreno Sánchez
Francisco Venegas Martínez
Freddy Alberto Domínguez Ortiz
Froylán Hernández Lara
Froylán Vladimir Enciso Higuera
Gabriel Darío Uribe Guerra
Gabriel González König
Gabriel Martínez Frausto
Gabriel Núñez Antonio
Gabriel Purón Cid
Gabriel Ramírez Sandoval
Gabriel Soto Cortés
Gabriela Muñoz Meléndez
Gail Mummert Fulmer
Gardy Augusto Bolívar Espinoza
Gerardo Bocco Verdinelli
Gerardo Castillo Ramos
Gerardo Espejo Abelar
Gerardo Esquivel Hernández
Gerardo González Chávez
Gerardo Leyva Parra
Gerardo Oliva Gutiérrez
Gerardo Piña
Gerardo Rosales Macías
Germán Vázquez Sandrín
Gian Carlo Delgado Ramos
Gilberto Francisco Pérez Aquino
Giulia Mugellini
Gladys Stella Rodríguez
Gloria Clementina Zúñiga Juárez
Gloria Moreno Álvarez
Gloria Soto Montes de Oca
Gonzalo Alonso Jiménez Alegría
Gonzalo Hernández Licona
Graciela Bensusán Aerous
Graciela Freyermuth
Graciela Teruel Belismelis
Graciela Tonon de Toscano
Grisel Ayllón Aragón
Griselda Luz María Acosta Castañeda
Guadalupe de los Ángeles Cano Meléndez
Guadalupe Vázquez Rodríguez
Guido Pinto Aguirre
Guillermo Alberto Sandoval Sánchez
Guillermo Sierra Juárez
Gustavo Aguilar Mata
Gustavo Alarcón Martínez
Gustavo Rivas Mendoza
Hector Édgar Buenrostro Mercado
Héctor Manuel Pedraza Rosales
Héctor Mauricio Núñez Amortegui
Héctor Mayagoitia Domínguez
Héctor Rodríguez Ramírez
Héctor Santiago Vélez Muñoz
Herberto Rodríguez Regordosa
Hiram Beltrán Sánchez
Homero Alonso Sánchez
Honorio Juárez Hernández
Hubert C. de Grammont
Hugo Hidalgo Silva
Hugo Montes de Oca Vargas
Ignacio Ibarra López
Ignacio Méndez Ramírez

Ingrid Fabiola Sada Correa
Irene Casique Rodríguez
Irma Leticia Martínez Sánchez
Irma Patricia Bárcenas Valtierra
Isabel de la Asunción Valadez Figueroa
Isalia Nava Bolaños
Ismael Aguilar Benítez
Ismael del Carmen Sandoval Montes
Israel Estrada Contreras
Itzi Segundo Métay
Jacinto Elías Sedeño Diaz
Jacob Ryten
Jacobo Ramírez Huerta
Jaime de Ávila González
Jason Schachter
Javier Guerra Estrella
Javier Reyes Ruiz
Javier Santibáñez Cortés
Javier Zermeño Duardo
Jesús Elías Fermín
Jesús Medina Vázquez
Johan Van Horebeek
Jonathan Heath Constable
Jorge Alberto Montejano Escamilla
Jorge Armando Morales Novelo
Jorge Arturo Meave del Castillo
Jorge Caro Oceguera
Jorge Daudé Balmer
Jorge Eduardo Mendoza Cota
Jorge Lara Álvarez
Jorge Lira Chávez
Jorge Luis Vázquez Aguirre
Jorge Martínez Pizarro
Jorge Prado Molina
Jorge Rodríguez Vignoli
Jorge Torres Rodríguez
Jorge Yamamoto Suda
José Alberto Incera Diéguez
José Alfredo Sosa Licona
José Ambrosio Bastián
José Andrés Ortiz Domínguez
José Antonio Cardona Pérez
José Antonio de la Peña Mena
José Antonio Mejía Guerra
José de Jesús Brambila Paz
José de Jesús García Vega
José de Jesús Romo Ramírez
José Édgar Villalobos Enciso
José Eduardo Ibarra Olivo
José Elías Rodríguez Muñoz
José Eliud Silva Urrutia
José Gasca Zamora
José Luis Ángel Rodríguez Silva
José Luis Briseño Cervantes
José Luis Durán Fernández
José Luis López Santana
José Luis Maya Cruz
José Luis Olmos Estrada
José Luis Pérez Garmendia
José Luis Silván Cárdenas
José María Duarte Cruz
José Pablo Covarrubias Ordiales
José Paúl Carrasco Escobar
José Pedro Olmedo Cornejo
José Ramón Arámbula García
José Ramón Narro Robles
José Raúl Santa Cruz Pérez
José Rodríguez Rocha
José Vences Rivera
José Víctor Tamaríz Flores
Juan Andrés Burgueño Ferreira
Juan Carlos Chávez Martín del Campo
Juan Carlos Espinoza Reyes
Juan Carlos Martínez Méndez
Juan Carlos Moreno Brid
Juan Carlos Ramírez Rodríguez
Juan Carlos Rodríguez Sierra
Juan Felipe Nieto Romero
Juan Francisco Islas Aguirre
Juan José Ambríz García
Juan José Campos Estévez
Juan Martín Mendoza
Juan Rivera Dommarco
Juan Sergio Salvador Flores Ponce
Juan Trejo Castro
Judith Domínguez Serrano
Judith López Peñaloza
Julia Guadalupe Pacheco Ávila
Julio Adolfo Cruz Reséndiz
Julio Baca Del Moral
Julio Campo Alves
Julio Pérez Díaz
Jürgen Weller
Kara Sinbernagel

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Keisuke Kondo
Landy Lizbeth Sánchez Peña
Laura Alicia Ibáñez Castillo
Laura Georgina Ahuactzin Pérez
Laura López Ortiz
Laura Luna González
Laura Nuño Gómez
León Rafael Garduño Estrada
Leonardo Roa
Leopoldo Villarruel Sahagún
Leslie Solís Saravia
Leticia Gracia Medrano
Leticia Ramírez Ramírez
Leticia Robles Silva
Leticia Ruiz Mendoza
Lidia Bratanova
Lilia Domínguez Villalobos
Lilia Granillo Vázquez
Lilia Rodríguez Tapia
Lilia Susana Padilla y Sotelo
Liliana Meza González
Luis Ángel Monroy Gómez Franco
Luis Ayala Alvarado
Luis Bermúdez
Luis Brito Castillo
Luis Chías Becerril
Luis Daniel Torres
Luis David Ramírez de Garay
Luis Enrique Nieto Barajas
Luis Ernesto Derbez Bautista
Luis Fernando Sánchez Martínez
Luis Foncerrada Pascal
Luis Huesca Reynoso
Luis Ignacio Román Morales
Luis Jaime Sobrino Figueroa
Luis Manuel Galván Ortiz
Luis Mariano Rojas Herrera
Luis Rubalcava Peñafiel
Manola Brunet India
Manuel Mendoza Ramírez
Manuel Molano Ruiz
Manuel Ordorica Mellado
Marcela Eternod Arámburu
Marco Antonio Berger García
Marco Antonio Gutiérrez Romero
Marco Aurelio Morales Martínez
Marcos Adrián Ortega Guerrero
Marcos B. González Tejeda
Marcos Esaú Domínguez Viera
Margarita Díaz Reyes
Margoth Laredo Rojas
María Adela Monreal Gómez
María Ángeles Durán Heras
María Cleotilde Báez García
María Concepción Arredondo García
María Cristina Gutiérrez Delgado
María del Carmen Reyes Guerrero
María del Consuelo Hernández Berriel
María del Rocío Ruiz Chávez
María del Rosario Cárdenas Elizalde
María Dolores Paris Pombo
María Edith Pacheco Gómez Muñoz
María Elena Diaz Trujillo
María Elena Zamora Dorantes
María Estela Rivero Fuentes
María Eugenia Cosío Zavala
María Eugenia González Ávila
María Graciela Freyermuth Enciso
María Guadalupe Villarreal Guevara
María Isabel Monterrubio Gómez
María Luisa Negrete
María Magdalena Rojas Rojas
María Margarita Parás Fernández
María Marlen Téllez Macías
María Matilde Gutiérrez López
María Mercedes Alaniz Ruvalcaba
María Mónica Vulling Garza
María Perevochtchikova
María Rosa Gudiño
María Salud Rangel Martínez
María Viridiana Sosa Márquez
Mariana Gabriela Cendejas Jáuregui
Mariana García
Marina Emilia Ariza Castillo
Marina Esperanza Covarrubias Amaya
Mario Alberto Hernández Hernández
Mario Arturo Ortiz Pérez
Mario Herrera Anguiano
Mario Miguel Ojeda Ramírez
Mario Palma Rojo
Mario Villalpando Benítez
Marisol Luna Contreras
Maritza Caicedo Riascos
Martha Mier y Terán

Martín Calderón Orduño
Martín Guadalupe Romero Morett
Martín Hermosillo Martínez
Martín Lima Velázquez
Mary Frances Rodríguez Van Gort
Mauricio Cecilio Domínguez Aguilar
Mauricio Pablo Cervantes Salas
Mercedes Pedrero Nieto
Mercedes Pedrosa Islas
Michel Séruzier
Miguel Adolfo Guajardo Mendoza
Miguel Ángel Alatorrre Mendieta
Miguel Ángel Mancera Gutiérrez
Miguel Ángel Martínez Damián
Miguel Ángel Suárez Campos
Miguel Cervera Flores
Miguel del Castillo Negrete Rovira
Miguel Moctezuma Flores
Milagros Alario Trigueros
Miriam Romo Anaya
Mónica Anzaldo Montoya
Mónica Toledo González
Natalia Volkow Fernández
Nayeli del Carmen González Novelo
Nicolás Heredia Delgado
Noé Aarón Fuentes
Noé Becerra Rodríguez
Noé Fuentes Flores
Nohora Beatriz Guzmán Ramírez
Norberto Roque Díaz de León
Norma Castillo
Norma Lorena Aguilar Monciváiz
Nydia Suppen Reynaga
Olga Lorena Rojas Martínez
Omar Ávila Flores
Orlandina de Oliveira
Óscar Adolfo Sánchez Valenzuela
Óscar Alonso Zamora Luna
Óscar González Muñoz
Óscar Jaimes Bello
Pablo Beltrán Ayala
Pablo Luis Covarrubias Toy
Pablo González
Pablo Hernández Ávila
Pablo Jasso Salas
Pablo Pérez Akaki
Pablo Ruiz Nápoles
Patricia Arias Rozas
Patricia Fernández Ham
Patricia Rodríguez López
Patricia Román Reyes
Paul Adolfo Taboada González
Paul Anand
Paul Cheung
Pedro Álvarez Icaza Longoria
Pedro Cayetano Martínez
Pedro Hancevic
Pedro Pablo Parra Díaz
Rafael Dávila Bugarín
Rafael de Jesús López Zamora
Rafael Garduño Rivera
Rafael López Vega
Rafael Moreno Sánchez
Ramón Arteaga Ramírez
Raúl Aguirre Gómez
Raúl Alcalá Campos
Raúl Alonso Chávez Reyes
Raúl Ángel Otero Díaz
Raúl Mejía González
Raúl Rueda Díaz del Campo
Raúl Sergio González Ramírez
Regina Hernández Franyuti
René Flores Arenales
René Millán Valenzuela
Ricardo César Aparicio Jiménez
Ricardo Orozco Zavala
Ricardo Villasís Keever
Roberto Antonio Ulloa Esquivel
Roberto Antonio Vázquez Espinoza de los Monteros
Roberto Bonifaz Alfonzo
Roberto Constantino Toto
Roberto Ham Chande
Roberto Ramírez Hernández
Roberto Rodríguez Rodríguez
Rodolfo Corona Vázquez
Rodolfo Cruz Piñeiro
Rodolfo de la Torre García
Rodolfo Nieves Martínez
Rodolfo Sánchez Sandoval
Rodrigo Martínez
Rodrigo Meneses Reyes
Rodrigo Negrete Prieto
Rodrigo Peláez López

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Rodrigo Sandoval Almazán
Rogelio Briseño Palomares
Rogelio Granguillhome Morfín
Rogelio Huerta Quintanilla
Rogelio Vázquez González
Rolando Ocampo Alcántar
Román Álvarez Béjar
Roque Quintanilla Montoya
Rosa María Jiménez Ambríz
Rosa Silvia Arciniega Arce
Rosalía Castelán Vega
Rosario Cárdenas Elizalde
Roxana Ivette Arana Ovalle
Rubén Barón Martell
Ryan Macdonald
Sagrario Garay Villegas
Salvador Marín Córdova
Santiago Soto Figueroa
Sara María Ochoa León
Saralina Ruiz Carús
Saúl Sandoval Merlos
Sergio Camarillo Calzada
Sergio Gaxiola Robles Linares
Sergio Gómez Anaya
Sergio Hernández González
Sergio Hernández Trejo
Sergio Llamas Mercado
Sergio Omar Frías
Sergio Velarde Villalobos
Silvia Castillo Argüero
Silvia Elena Giorguli Saucedo
Silvia Mejía Arango
Silvia Ruiz-Velasco Acosta
Simone Cecchini
Sofía Vázquez Herrera
Soledad Ramírez Villanueva
Sonia Frías Martínez
Sonia Tatiana Sánchez Quispe
Steven Vale
Susan W. Parker
Sylvia Irene Schmelkes
Sylvie Turpin Marion
Tanya Buxton Torres
Teodora Hurtado Saa
Teresa García Ramírez
Thérèse Lalor
Tito Alegría Olazabal
Tomás Ramírez Reynoso
Tonatiuh Guillén López
Ulises López Enríquez
Úrsula Oswald Spring
Vanesa Vansteenkiste
Vania Pérez Morales
Verónica Alejandra Olvera Estrada
Verónica Alonso Herrera
Verónica Irastorza Trejo
Verónica Villarespe Reyes
Víctor Alcocer Yamanaka
Víctor E. Tokman
Víctor García Vilchis
Víctor Gaspar Martín Hernández
Víctor Manuel García Guerrero
Víctor Manuel Guerrero Guzmán
Víctor Orlando Magaña Rueda
Víctor Pérez Hernández
Virgilio Partida Bush
Virginia Abrín Batule
Viseslav Simic
Vivian Milosavljevic
Vladimir Canudas Romo
Waldenia Cosmes Martínez
Walter Lepore
Walter Radermacher
Warren Jochem
Wilfrido Ruiz Ochoa
William F. Maloney
Wulfrano Castañeda Vidal
Ximena Fernández Martínez
Yedwab Benjamín Temkin
Yolanda Robles Escobedo
Yolanda Zavaleta Cortés
Yoloxóchitl Bustamante Díez
Yves Daniel Bussière
Zenón Cano Santana

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Mismos individuos en diferentes bases de datos sin identificador común: la unión de las bases de datos de beneficiarios de la SAGARPA con una base universal

Carlos Alberto Francisco Cruz, Jorge Lara Álvarez, Juan Francisco Islas Aguirre, Ana Karen Díaz Méndez y Felipe Pérez Gachuz

Mexican laborers cut broccoli stalks for Smith Farms' crew A as.../Portland Press Herald/Getty Images

Este trabajo muestra una metodología para unir varias bases de datos con información de individuos idénticos entre ellas, pero sin un identificador universal. Con un mínimo de información de cada persona, las podemos unir. Lo que proponemos requiere una interfaz inteligente que emplea un algoritmo de pareo por registro (*record matching*), y su finalidad es detectar a los mismos individuos en diferentes bases de datos. Demostramos nuestra metodología pareando una base universal con diversas bases de datos de beneficiarios de la Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA).

Palabras clave: unión de bases de datos, SAGARPA, pareo por registro.

This paper shows a methodology to join several databases which hold information on identical individuals but without a universal identifier. With minimum information on each individual, it is possible to combine different databases. The proposed methodology requires an intelligent interface using a matching algorithm for each record (record matching) in order to detect the same individuals among different databases. Its usefulness is demonstrated by matching a universal database with several beneficiaries' databases from Agriculture, Cattle, Rural Development, Food, and Fishing Ministry of Mexico (*SAGARPA* in Spanish).

Key words: Joined Databases, SAGARPA, record matching.

Recibido: 29 de diciembre de 2014.
Aceptado: 26 de marzo de 2015.

Nota: se agradecen los valiosos comentarios de Leonardo Pérez Sosa y Natalia Eugenia Volkow Fernández, así como de los participantes del *2014 Mexican Stata Users Group meeting*.

1. Introducción

Una característica suficiente para unir dos bases de datos es que ambas compartan una clave que identifique a los individuos; no obstante, en los países en vías de desarrollo, como México, es común la ausencia de ésta debido, principalmente, a la falta de calidad en la recolección de los datos y a la diversidad de instituciones que los colectan. Esta carencia de un identificador es frecuente en las bases de datos de interés público: beneficiarios de programas sociales, padrones de productores, personas en situación vulnerable, entre otras. Existe un gran potencial de información que se está desaprovechando porque no es sencillo unir las diversas bases de datos; por lo tanto, su vinculación representa un reto y una gran oportunidad para mejorar las políticas públicas.

Actualmente, en México se cuenta con una cantidad importante de datos disponibles referente a la situación social y económica del país, los cuales son generados por diversas instituciones y oficinas de gobierno de manera independiente, y no se tiene el cuidado de vincular las observaciones a través de un identificador común. Leicester (2001) señala cuatro beneficios de unir bases de datos provenientes de distintas fuentes: uno de ellos es que con la base de datos obtenida es posible generar nueva inferencia estadística sobre un fenómeno en particular; la segunda es que permite incorporar a los estudios variables adicionales; la tercera amplía las opciones para estudios longitudinales; y la cuarta reduce considerablemente los costos de una investigación al utilizar información ya recabada. No obstante, como lo señalan Hernández y Stolfo (1998), unir bases de datos sin un identificador común es un trabajo que puede resultar complejo y laborioso.

La vinculación y unión está sujeta a la disponibilidad de un identificador que cumpla con dos funciones (Christen, 2006): 1) reconocer cada una de las observaciones al interior de cada base de datos y 2) distinguir a cada observación para su posible unión con diversas bases de datos que compartan el mismo identificador. Note que

cuando existe el mismo individuo en dos bases de datos y logramos unir la información contenida en ambas se dice que se realizó un *pareo*. Si se tiene el mismo identificador que permita la unión de los individuos en las dos, la tarea es fácil, pero cuando no se cuenta con éste se vuelve una labor compleja y, en ocasiones, imposible. La pregunta que motiva este documento es la siguiente: ¿cómo unir bases de datos que comparten algunos individuos cuando no se cuenta con un identificador como el que se menciona?

En los últimos años se han desarrollado diferentes metodologías y técnicas computacionales que buscan resolver este problema; por ejemplo, hay algunas que se sustentan en exploraciones a través de permutaciones, tal es caso de Reif (2010) y Barker (2012), quienes proponen dos algoritmos (implementados en Stata[1]) que permiten realizar una búsqueda de texto en otra base de datos. Con la permutación se puede encontrar una combinación similar a la palabra que se esté rastreando. No obstante, estos métodos tienen dos limitantes importantes: la primera es que se debe señalar cuáles serían las posibles diferencias entre las palabras a buscar y la segunda es que, una vez concluido el procedimiento, se debe realizar una revisión para detectar posibles errores. Actualmente se tienen estudios que se han beneficiado de la unión de diferentes bases de datos. Machado (2004) describe investigaciones llevadas a cabo en el campo de la salud infantil a partir de la vinculación de bases de datos provenientes de diferentes fuentes y que no cuentan con un identificador. Este mismo autor destaca la importancia de la recopilación y unión de la información existente.

En ese sentido, el objetivo de este trabajo es mostrar una propuesta que permita vincular y unir dos[2] bases de datos que no cuenten con un identificador. Nuestra metodología consiste en comparar variables comunes en ambas y diagnosticar si tenemos al mismo individuo en éstas. Asimismo, intenta utilizar toda la información disponible secuencialmente; además, tiene una interfaz inteligente para corroborar que el pareo se realice de manera correcta. Este método hace posible unir bases de datos de millones de observaciones. Hasta donde tenemos conocimiento, es el primero que se propone para unirlas sin un identificador común. Para demostrar su valor empírico, se presenta el caso de diversas bases de datos de beneficiarios de la SAGARPA y una base universal,[3] la cual contiene a los productores agropecuarios beneficiarios y a los que no lo son, por eso la llamamos así.

El primer paso fue unir entre ellas las bases de la Secretaría en lo que llamamos *padrón de beneficiarios*. La conformación de éste permite comparar a los beneficiarios de los diversos apoyos otorgados a través de programas sociales de la SAGARPA. Posteriormente, la vinculación del padrón con la base universal dio la pauta para un diseño riguroso de evaluación de impacto, a cargo de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, por sus siglas en inglés)-México, para tres programas de la Secretaría: PROCAMPO, PROGAN y Fomento Productivo al Café.

El resto del documento está desarrollado de la siguiente manera: la sección 2 explica la metodología y los problemas para su implementación, la 3 muestra un ejemplo real de nuestra propuesta y la 4 presenta conclusiones.

2. Vinculación y unión de bases de datos sin identificador

Como se mencionó antes, la unión de bases de datos está condicionada a la existencia de un identificador que, en primer lugar, permita reconocer de manera individual cada una de las observaciones en una base de datos y que, después, se pueda identificar información de las mismas en otra.

1 Es un *software* estadístico que permite realizar análisis cuantitativo y cualitativo. Debido a sus características y a la facilidad de su interfaz, es uno de los programas más utilizados en la actualidad.

2 No obstante, nuestra metodología se puede ampliar; por ejemplo, si se desea unir tres bases de datos (A, B y C), una opción sería juntar A con B y después la unión de éstas con la C.

3 Esta base universal incluye información de productores agropecuarios.

En caso de que la base de datos sea a nivel persona, lo ideal sería contar con un número de identificación nacional[4] para así poder ligar a cada persona con cualquier otra que lo contenga. En México, la Clave Única de Registro de Población (CURP) es una opción de identificador a nivel individual; lamentablemente, no toda la gente cuenta con este registro; además, está la situación de que un individuo tiene más de una CURP. En el 2014, de acuerdo con datos de la Secretaría de Gobernación, en la Base de Datos Nacional de la Clave Única del Registro de Población había 177 892 081 registros, cifra que supera al total de habitantes —a mediados de ese año, con base en la proyección del Consejo Nacional de Población (CONAPO), en México había 119 713 203—; esto se debe, principalmente, a que hay un registro acumulado y sin depurar, y con algunos duplicados, por lo cual es una alternativa que se encuentra limitada. Otra opción es hacer uso del Registro Federal de Contribuyentes (RFC), pero este dato es aún menos común que la CURP, debido a que sólo se genera para personas mayores de 18 años de edad y que se encuentren registrados en la Servicio de Administración Tributaria (SAT). Entre los regímenes fiscales más comunes destacan las personas que prestan servicios profesionales subordinados, aquellas que ofrecen servicios profesionales, o bien, las que tienen una empresa o negocio, de tal manera que la gente que cuenta con RFC debe cumplir con las obligaciones que establece la ley en el pago de impuestos.

Una opción cuando las bases de datos carecen de un identificador común es unirlas a partir de la información contenida en cada una de ellas; de nuevo, si ésta es a nivel individuo, posiblemente las bases de datos contengan variables que nos permitan distinguir al individuo. Éstas pueden ser numéricas, como la edad o número del domicilio o en texto, por ejemplo nombre de la persona o calle del domicilio. Entonces, el reto es usarlas en conjunto para suplantar a un identificador, permitiendo así unir diferentes fuentes de información. La combinación hace que el trabajo de unión sea posible pero, nótese, se debe hacer una selección para sustituir al identificador; además, las variables empleadas pueden contener errores de captura o diferentes maneras de registro, por lo que buscar la información en otra base datos debe permitir que ésta pueda no ser idéntica. Un error de captura puede ser una falta de ortografía o un número erróneo, un ejemplo es: 01/01/1981 y 01/01/81.

Entonces, la primera interrogante que surge es la siguiente: ¿es posible comparar la información de una variable capturada en dos bases de datos diferentes con posibles errores y heterogeneidad de captura? Si éstas son pequeñas, se puede hacer de forma manual; no obstante, si se desea comparar dos con más de cien observaciones cada una, el proceso se torna laborioso. En caso de tener millones de observaciones, una comparación manual se vuelve imposible.

Una alternativa acorde para vincular dos variables de distintas bases de datos que potencialmente están capturadas de una manera diferente o con error es un pareo por registro (Fellegi y Sunter, 1969), el cual intenta identificar entradas en diferentes bases de datos que corresponden a la misma unidad de observación. Existe una rutina en Stata para esto nombrada *reclink*, que emplea una cadena comparadora a partir de un bigrama (Blasnik, 2010), el cual es utilizado como plataforma para el análisis estadístico de texto y permite realizar la comparación entre textos (Collins, 1996). El bigrama proporciona la probabilidad condicional de una palabra (W_n), dada una palabra precedente (W_{n-1}):

$$P(W_n | W_{n-1}) = \frac{P(W_n, W_{n-1})}{P(W_{n-1})}$$

Esta forma de vincular información entre dos bases de datos permite superar las variaciones de formato y los errores de captura. Sin embargo, su estructura y extensión pueden ocasionar que incluso la aplicación de este algoritmo lleve mucho

4 Por ejemplo, Documento Nacional de Identidad (DNI) en Argentina, Rol Único Nacional (RUN) en Chile y la Clave Única de Registro de la Población (CURP) en México.

tiempo y/o sea muy compleja. La complejidad radica en que puede resultar que la información para el mismo individuo no coincida exactamente en las dos bases de datos, sino que sólo sea muy semejante. Esto lo explicaremos más a detalle en la siguiente sección. Además, por lo general, debemos utilizar información contenida en más de una variable para asegurarnos de que se trata del mismo individuo. Esto dificulta la búsqueda.

Por estas razones, hemos diseñado una metodología que minimiza el tiempo computacional y maximiza la posibilidad de encontrar pareos. A continuación se describe y se presenta un ejemplo con el caso particular de los beneficiarios de la SAGARPA y nuestra base universal.

2.1 Etapas para unir bases de datos sin identificador a través de Stata

El objetivo es vincular aquellas que tienen diferentes fuentes de información a través de cinco etapas: 1) selección de variables a considerar en el pareo, 2) homogeneización, 3) división por grupos o bloques, 4) comparación, 5) interfaz inteligente y 6) pareo. Estos pasos son recomendables cuando se tienen bases de datos con un número importante de observaciones, en concreto más de 100 mil.[5] A continuación se describe cada una de las fases mencionadas.

Selección de variables

En esta parte del proceso se escogen las que se van a usar en cada base de datos para identificar a los individuos comunes entre ellas. Mientras más *única* sea la variable más valiosa va a ser; por ejemplo, la fecha de nacimiento (día, mes y año) es más útil que el estado de residencia. Si una base de datos sólo contiene variables muy *generales*, como sexo y edad en años, no va a ser posible unirla con otras bases usando este método.

Cabe señalar que las variables a comparar deben estar en ambas bases.[6] Por último, se recomienda depurar errores sistemáticos (caracteres adicionales, errores de ortografía comunes) en las que sean seleccionadas.

Homogeneización

Después de la selección, la siguiente etapa consiste en homogeneizar las variables de las dos bases de datos para el proceso de vinculación, es decir, el formato de captura en ambas bases. Esta fase depende, fundamentalmente, de la calidad de la información contenida en cada uno de los tabulados a vincular. Las variables de texto son las más complicadas de homogeneizar, esto debido a la diversidad de formatos en los que pueden ser capturadas y a las abreviaturas. Christen (2006) señala cinco posibles fuentes de heterogeneidad en los nombres propios en el momento de registrarse en una base de datos:

1. Variaciones por deletreo; ejemplo: Leydi *vs*. Laydi.
2. Variaciones por pronunciación; ejemplo: Paula *vs*. Paola.
3. Nombres compuestos; ejemplo: Marisol *vs*. María del Sol.
4. Nombres alternativos; ejemplo: Alfonso *vs*. Poncho.
5. Nombres sólo indicados con la primera letra; ejemplo: José Juan *vs*. J. J.

Además, podríamos agregar otro orden:

6. Nombre completo capturado en un solo campo, sin distinguir nombre de primer y segundo apellidos; ejemplo: Alfonso Rivera Gómez *vs.* Rivera Gómez Alfonso.[7]

5 Esto depende de la capacidad del procesador; 100 mil observaciones son suficientes en un procesador i9, a 8 Gb de RAM.

6 A manera de ejemplo: la base de datos A contiene las variables de sexo, nombre, apellido paterno y edad y la B, las de nombre, apellidos paterno y materno y edad. Entonces, no es posible emplear ni la variable apellido materno —dado que no está en la base A— ni tampoco la de sexo —pues no está en la B—; por lo tanto, sólo podríamos usar nombre, apellido paterno y edad.

7 Cabe señalar que todos los nombres aquí presentados son inventados; no obstante, ejemplifican situaciones reales.

Estas posibles variaciones son difíciles de detectar cuando se tienen muchas observaciones en las bases de datos. Una complicación adicional es que, debido a dichas variaciones y a la ausencia de un identificador, cada tabulado puede tener individuos duplicados capturados con una variación diferente. El cuadro 1 muestra un ejemplo de un registro duplicado en la misma base de datos; se observa que al interior de cada registro administrativo se pueden tener múltiples diferencias, es posible tener datos con letras mayúsculas y otros con minúsculas, además de las variaciones antes señaladas.

Cuadro 1

Ejemplo de un duplicado al interior de una base de datos

Nombre	Entidad	Municipio
María Guadalupe Luna Perea[a]	Chiapas	Arriaga
MA. G. LUNA PEREA	CHIAPAS	ARRIAGA

[a] Este nombre es sólo un ejemplo ilustrativo y no corresponde a una persona real.

Fuente: elaboración propia.

Ante esta situación, la etapa de homogeneización tiene dos objetivos específicos. El primero consiste en dejar todas las bases de datos en una condición similar; es recomendable substraer de las cadenas de texto los caracteres especiales —¡ , ” , # , $, % , & , (, = , ? , . , :—, además de los acentos y tener el texto en letras mayúsculas; otra recomendación, si se está trabajando con personas, es separar el nombre y los apellidos en celdas independientes (Herzog, Scheuren y Winkler, 2007). El segundo es revisar si una vez homogeneizadas las variables detectamos individuos duplicados dentro de cada base de datos. Alcanzar estos objetivos contribuye considerablemente a los resultados finales.

Una vez homogeneizadas las variables, podemos hacer un ejercicio de pareo perfecto, es decir, comparamos nuestras bases de datos para que coincidan con exactitud en todas las que seleccionamos en la primera fase. Dichos pareos perfectos los separaremos del grupo para reducir la carga computacional del pareo por registro que realizaremos en las siguientes etapas. Se relajará dicha exigencia hasta un nivel mínimo aceptable para considerar que dos observaciones corresponden a la misma persona. Esto se recomienda cuando: 1) las bases de datos contienen muchas observaciones y 2) se cuenta con varias variables útiles.

Grupos o bloques

Esta fase es importante cuando se van a unir tabulados con un número importante de observaciones. El método propuesto, basado en bigramas, puede ser muy tardado, sobre todo en bases de datos con miles o millones de observaciones, ya que cada una de las que contiene una base de datos es buscada en la otra. Una forma de reducir el esfuerzo computacional es generar grupos o bloques a partir de criterios comunes en ambas bases de datos (Baxter, Christen y Churches, 2003); por ejemplo, si en las dos se tiene información de las 32 entidades federativas del país, es recomendable generar una base de datos para cada una de ellas y realizar la vinculación con las bases de datos correspondientes a una entidad. Este planteamiento se puede realizar usando una sola variable o combinando varias; por ejemplo, podríamos agrupar por entidad federativa y sexo.

Comparación

En esta etapa se emplea la rutina de *reclink* (Barker, 2012). Comparamos las variables X_1, X_2, …, X_k de dos bases de datos al mismo tiempo. *Reclink* permite asignar un orden de relevancia a cada variable; mientras más *única* sea ésta más *peso* le adjudicaremos, por ejemplo, los apellidos paternos en México son más *únicos* que los nombres, por ello, les daríamos más relevancia a éstos. También, nos da la posibilidad de asignar un *peso* por no pareo, es decir, qué tanto nos importa si dos variables que estamos comparando no coinciden. Esta característica es clave en la rutina, pues se puede atribuir un gran peso de pareo a variables relevan-

tes, aunque tengan una alta probabilidad de ser capturadas con error, por ejemplo, la dirección del domicilio. A dicha variable le podríamos asignar un *alto peso* si coincide, porque es información única, pero un *bajo peso* en caso de no coincidir, porque puede tener mucha heterogeneidad en la captura.

Interfaz inteligente

A partir de los resultados de la comparación, es necesario establecer una medida o distancia que permita determinar qué tan similar es una observación en la primera base respecto a una observación en la otra (Bilenko, Mooney, Cohen, Ravikumar y Fienberg, 2003). En el caso de la rutina *reclink*, al basarse en un bigrama, es posible obtener la probabilidad de similitud entre una observación y otra. Lo que se obtenga depende, a su vez, de los pesos asignados a cada variable a considerar. Una probabilidad de 1 sugiere que los individuos de ambas bases de datos son idénticos; una de cero indica que no tienen relación.

Los registros con un valor inferior a 1 requieren de una interfaz inteligente, es decir, el discernimiento de una persona; es el usuario quien determinará si se trata del mismo individuo o no. El cuadro 2 muestra la sugerencia de los autores para clasificar las observaciones; se puede ver una posible clasificación de los registros comparados; cabe señalar que ésta se encuentra sujeta a una correcta elección de variables y sus respectivos pesos.

Pareo

Una vez finalizada la etapa de clasificación, es posible fusionar las bases de datos provenientes de fuentes de información distintas. El resultado es una base unificada confiable, a pesar de no contar con un identificador y de las dificultades de vinculación. Es importante destacar que mientras mejor sea la calidad de los datos y existan más variables en ambas bases de datos que permitan identificar a los individuos, más fácil será su vinculación.

3. Unión del padrón de beneficiarios de la SAGARPA y la base universal

En esta sección ejemplificaremos nuestra metodología con un caso práctico: la vinculación entre las bases de datos de beneficiarios de programas de la SAGARPA y de éstas con una base universal.[8] Una vez hecho el pareo, el análisis estadístico se hizo sobre variables que no nos permitían identificar a las personas.

En México, la SAGARPA es la principal entidad pública que se encarga de brindar apoyo a los sectores productivos agrícola, ganadero, de acuicultura y pesca, así como de postproducción. La población

8 Si bien, como explicamos arriba, es necesario usar variables como el nombre, sexo y edad, debemos dejar muy en claro que los investigadores que participamos en el presente estudio firmamos una carta compromiso para no extraer información personal de ninguna de las bases de datos utilizadas en esta investigación.

Cuadro 2

Clasificación de los registros con base en su probabilidad de similitud

Probabilidad	Clasificación
0.9500-0.9999	Se trata de la misma persona con pocas diferencias. Requiere de una revisión poco exhaustiva.
0.9000-0.9500	Puede tratarse de la misma persona, pero con varias diferencias. Requiere de una revisión exhaustiva.
0.8500-0.9000	Se tiene poca certeza de que sea la misma persona. Requiere de una revisión muy exhaustiva.
Menor a 0.8500	No se trata de la misma persona.

Fuente: elaboración propia.

objetivo a la que se orientan sus programas incluye personas físicas y morales, grupos, proyectos o cooperativas, dependiendo de las características del programa o componente en cuestión, de los montos de apoyo, así como de los criterios de elegibilidad. Hoy en día, en la Secretaría existen más de 50 programas sociales,[9] cada uno con su propia base de beneficiarios. Todos estos detalles e información adicional se modifican o conservan anualmente y se plasman en las reglas de operación, las cuales son publicadas en el *Diario Oficial de la Federación* en enero de cada año.

Por otro lado, en cuanto a la sistematización de la información de sus beneficiarios, no hay uniformidad, no se han establecido estándares de cómo llevar a cabo los registros. Cada subunidad de la SAGARPA opera su programa social; en algunos casos, ésta es la que enlista a los beneficiarios; en otros, la delegación estatal es la que lo hace. Cabe mencionar que no hay una vinculación entre las diferentes delegaciones estatales de la SAGARPA, así como con las subunidades responsables, las ventanillas de recepción de solicitudes, y otros agentes involucrados en la operación de los componentes. Entonces, la información recabada por cada componente no es homogénea y, por ende, resulta complicado para la Secretaría tener control respecto a quiénes se están destinando recursos públicos. En síntesis, esta instancia de gobierno no cuenta con un mecanismo preciso y eficiente para identificar a los beneficiarios entre sus componentes, pues no se vincula información entre los diferentes programas de apoyo.

Sus bases de datos cuentan con potenciales identificadores, como la CURP o el RFC. Sin embargo, existen muchos individuos sin observación para estas variables, además de que la calidad de captura no es buena. Un reto adicional es que cada una de ellas, en potencia, tiene individuos duplicados. Esto se debe a la mala calidad en la captura, puede ocurrir que una misma persona aparezca dos o más veces, pero no necesariamente la información capturada para cada uno de ellos es idéntica en sus duplicados, es decir, puede haber errores y heterogeneidad de captura para un mismo individuo.

Para identificar duplicados, depuramos los errores más comunes y, después, utilizamos las siguientes variables para comprobar si existieron errores sistemáticos de captura y contábamos con observaciones duplicadas:

1. Nombre, apellido paterno, apellido materno, entidad, municipio y localidad.[10]
2. Nombre y CURP.[11]
3. Nombre y RFC.

Una base de datos que unifique todos los listados de beneficiarios de la SAGARPA traería los siguientes beneficios:

- Un conteo preciso de esas personas.
- Un conteo del total de apoyos que recibe cada beneficiario.
- La relación entre los programas que recibe un beneficiario con varios apoyos.
- Un conteo del monto total de apoyo que recibe cada uno.

Como cada componente colecta datos sobre diferentes aspectos de un beneficiario, se podría lograr una caracterización más detallada de los apoyados por más de un componente. Entonces, el primer objetivo es unificar la información para 31 programas sociales de la SAGARPA, de los cuales pudimos tener acceso a sus bases de datos. Como regla general, usamos nuestro método de la siguiente manera:

a) Selección de variables a considerar en el pareo. Las que pudieran identificar al individuo, las cuales incluyeron, dependiendo de su disponibilidad, nombre, apellidos paterno y

9 Llamados *componentes*, divididos en nueve grandes ramas.

10 La información sobre la localidad permitiría detectar duplicados con facilidad. Sin embargo, en la mayoría de las bases siempre se reporta la localidad del beneficiario. Para ello, se aplicó el filtro a nivel municipal.

11 En algunos casos se usó la fecha de nacimiento obtenida de la CURP. Esto se hizo cuando de forma manual se detectaron errores de captura en la CURP. De manera similar, para el caso del RFC incorrectamente capturado, se utilizó la fecha de alta tributaria en el componente de la SAGARPA. Este procedimiento, además, permitió obtener la edad del beneficiario en el 2013 (o en cualquier otro año).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

materno, fecha de nacimiento (obtenida de la CURP), año de nacimiento/edad (CURP), RFC, entidad, municipio y localidad.

b) Homogeneización de formato de ciertos campos. Pusimos todas las variables de texto en mayúsculas, la fecha de nacimiento en el mismo formato, el año de nacimiento lo pasamos a edad en el 2013 y la entidad, municipio y localidad los convertimos a clave usando el catálogo de claves de localidades del Instituto Nacional de Estadística y Geografía (INEGI). Intentamos un *pareo perfecto* con las siguientes opciones:
   1. Nombre, apellidos paterno y materno, entidad, municipio y localidad.
   2. Nombre y CURP.
   3. Nombre y RFC.

Esto siempre y cuando los individuos contaran con toda la información, es decir, si las personas no tenían CURP no usábamos la opción 2.

Parte recursiva. Ronda 1

Una vez seleccionadas las observaciones con las que íbamos a realizar un pareo por registro, la primera ronda era la más exigente:

3) División por grupos o bloques. Exigíamos que la edad, la entidad, el municipio y la localidad coincidieran. Le dábamos más relevancia a los apellidos que a los nombres.
4) Comparación. Comparábamos los registros de apellidos y nombre.
5) Interfaz inteligente. Cada registro menor a 0.93 de probabilidad de coincidencia se revisó.
6) Pareo. Se unía la información de los individuos que coincidían en ambas bases de datos.

Parte recursiva. Ronda 2

Una vez seleccionadas las observaciones con las que íbamos a realizar un pareo por registro, esta segunda fase relajaba que la edad y la localidad coincidieran. Esto porque notamos que la edad, sobre todo cuando no era calculada de una CURP disponible, estaba mal capturada. También, notamos un problema de calidad en la captura de los datos de localidad:

3) División por grupos o bloques. Exigíamos que la entidad y el municipio coincidieran. Le dábamos más relevancia a los apellidos que a los nombres.
4) Comparación. Comparábamos los registros de apellidos y nombre.
5) Interfaz inteligente. Cada registro menor a 0.95 de probabilidad de coincidencia se procedía a revisar.
6) Pareo. Se unía la información de los individuos que coincidían en ambas bases de datos.

Parte recursiva. Ronda 3

Por último, relajamos la restricción de que el municipio debería de coincidir. Sin embargo, fuimos más estrictos al revisar el nombre y la edad de las personas:

3) División por grupos o bloques. Exigíamos que la entidad y la edad coincidieran. Le dábamos más relevancia a los apellidos que a los nombres.
4) Comparación. Comparábamos los registros de apellidos y nombre.
5) Interfaz inteligente. Cada registro menor a 0.999 de probabilidad de coincidencia se procedía a revisar.
6) Pareo. Se unía la información de los individuos que coincidían en ambas bases de datos.

Con este método pudimos formar una base única de beneficiarios de la SAGARPA que contiene información de 31 programas sociales que se encuentran actualmente en funcionamiento. Nuestro resultado fue que existe un total de 2 683 713 beneficiarios, de los cuales 75.3% son hombres y 24.7%, mujeres.

La conformación de una base única permite obtener información que no sería completa si

se tuviera la de un solo componente. El cuadro 3 muestra la distribución de beneficiarios por entidad federativa; se puede observar con mayor precisión que Oaxaca es la entidad con mayor número, ya que concentra a más de 11.24%, le sigue Chiapas (9.68%), Veracruz de Ignacio de la Llave (8.51%), Puebla (6.55%) y Guerrero (5.98%). Por su parte, las entidades con la menor cantidad son el Distrito Federal (0.08%), Baja California (0.20%) y Baja California Sur (0.09%).

Cuadro 3

Beneficiarios de la SAGARPA

Entidad	Beneficiarios	%	% Acum.
Oaxaca	301 761	11.24	11.24
Chiapas	259 752	9.68	20.92
Veracruz de Ignacio de la Llave	228 404	8.51	29.43
Puebla	175 775	6.55	35.98
Guerrero	160 602	5.98	41.96
México	144 796	5.40	47.36
Michoacán de Ocampo	132 234	4.93	52.29
Zacatecas	129 014	4.81	57.10
Hidalgo	117 839	4.39	61.49
Guanajuato	107 105	3.99	65.48
Jalisco	106 333	3.96	69.44
San Luis Potosí	99 336	3.70	73.14
Durango	91 171	3.40	76.54
Sinaloa	81 264	3.03	79.57
Chihuahua	74 068	2.76	82.33
Tamaulipas	68 621	2.56	84.89
Yucatán	50 952	1.90	86.79
Nayarit	45 180	1.68	88.47
Tabasco	40 717	1.52	89.99
Tlaxcala	39 205	1.46	91.45
Campeche	36 095	1.34	92.79
Coahuila de Zaragoza	34 688	1.29	94.08
Querétaro	28 751	1.07	95.15
Morelos	25 787	0.96	96.11
Quintana Roo	24 462	0.91	97.02
Sonora	23 024	0.86	97.88
Nuevo León	21 853	0.81	98.69
Aguascalientes	14 538	0.54	99.23
Colima	8 008	0.30	99.53
Baja California	6 603	0.25	99.78
Baja California Sur	2 441	0.09	99.87
Distrito Federal	2 224	0.08	99.95
Total	**2 683 713**	**100.00**	

Fuente: elaboración propia.

A su vez, la base de datos permite saber a cuántos programas sociales está inscrito cada uno de los beneficiarios. En el cuadro 4 se puede observar que más de 82.51% sólo recibe apoyos de un componente, mientras que 14.14% lo tiene de dos y 2.87%, de tres. También, es posible identificar el caso en el que se otorgan varios componentes por beneficiario. El mismo cuadro muestra que se tiene un beneficiario que recibe apoyo de ocho componentes; nueve, de siete y 139, de seis.

Cuadro 4

Número de programas sociales de la SAGARPA por beneficiario

Número de componentes	Beneficiarios	%
1	2 214 212	82.51
2	379 514	14.14
3	76 953	2.87
4	11 415	0.43
5	1 470	0.05
6	139	0.01
7	9	0.00
8	1	0.00
Total	**2 683 713**	**100.00**

Fuente: elaboración propia.

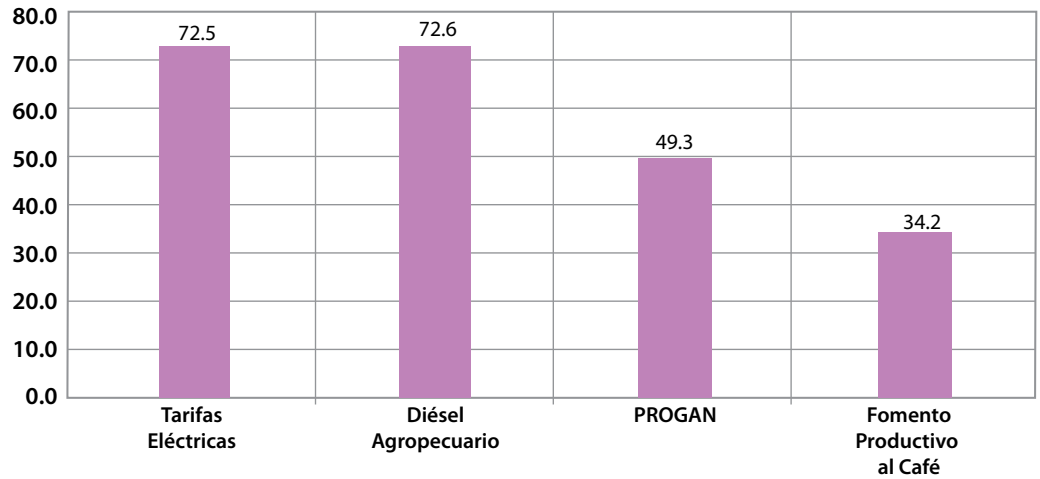
A partir de la información anterior es posible identificar el cruce que se puede tener entre dos componentes; por ejemplo, PROCAMPO es el que cuenta con mayor número de beneficiarios, por lo que se tiene un cruce importante de sus beneficiarios con otros componentes. La gráfica 1 muestra la cantidad de beneficiarios de Diésel Agropecuario, PROGAN, Fomento Productivo del Café y Tarifas Eléctricas que, a su vez, tienen PROCAMPO. En el caso de PROGAN, más de 150 mil (cerca de 50%) reciben, al mismo tiempo, PROCAMPO. A más de 34% de la población beneficiaria de Fomento Productivo del Café se le otorga, de manera simultánea, PROCAMPO. En el caso de Diésel Agropecuario y Tarifas Eléctricas son 72.6 y 72.5%, respectivamente.

3.1 Uniendo el padrón de beneficiarios de la SAGARPA y la base universal

Este padrón, que se mencionó al final de la introducción de este documento y que se formó mediante la unión de las bases de datos de la SAGARPA, dio la pauta para plantear un diseño de evaluación de impacto, por lo que, dadas las características de la información y la situación actual de los programas, se optó por un diseño cuasi-experimental. En este caso se requería de un marco muestral que permitiera ob-

Gráfica 1

Beneficiarios de PROCAMPO que simultáneamente son beneficiarios de otros programas (porcentajes)



Fuente: elaboración propia.

tener un grupo de control con el cual se pudiera establecer una comparación rigurosa con los beneficiarios del programa y, con ello, determinar el impacto de los programas de la SAGARPA en sus beneficiarios.

En ese sentido, una base universal que ofrece información detallada de los productores agropecuarios del país se determinó como la opción más viable para identificar un posible grupo de control. Para ello, fue necesario identificar en ésta a los beneficiarios de los programas de la SAGARPA. El reto era aún mayor, pues teníamos que vincular una base de datos de 2.6 millones de observaciones con una de cerca del doble.

La forma de vincular se llevó a cabo mediante el mismo método antes descrito, aunque es importante destacar que esta tarea se realizó sólo con los componentes de PROCAMPO, PROGAN y Fomento Productivo al Café.

Los resultados obtenidos son importantes (ver cuadro 5), ya que en el caso de PROCAMPO se identificaron 62% de los registros dentro de la base universal. En PROGAN se logró reconocer a 67% y en el de Fomento Productivo al Café, a 69 por ciento.

Cuadro 5

Beneficiarios de la SAGARPA identificados en una base universal

Componente	% en una base universal
PROCAMPO	62
PROGAN	67
Fomento Productivo al Café	69

Fuente: elaboración propia.

4. Conclusiones

La vinculación de la información proveniente de fuentes distintas permite un mayor aprovechamiento de los datos existentes, por lo que pueden disminuir considerablemente los costos de una investigación. La metodología descrita en este documento de trabajo brinda la posibilidad de unir dos bases de datos con individuos en común con un mínimo de información disponible e, incluso, con problemas en la calidad de captura de los datos. En ese sentido, el uso de este método, que emplea la rutina de Stata de *reclink*, hace posible la vinculación de dichas bases.

Nuestra metodología se ejemplificó con un caso real: la vinculación de información de la SAGARPA y una base universal, lo cual permitió realizar un diseño riguroso de una evaluación de impacto para los programas gestionados por la SAGARPA: PROCAMPO, PROGAN y Fomento Productivo al Café.

Fuentes

Barker, M. *Stata module to calculate the Levenshtein distance, or edit distance, between strings.* 2012. Consultado en: http://ideas.repec.org/c/boc/bocode/s457547.html

Baxter, R., P. Christen y T. Churches. "A Comparison of Fast Blocking Methods for Record Linkage", en: *CIMS Technical Report 03/139. CSIRO Mathematics, Information and Statistics*. 2003.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar y S. Fienberg. "Adaptive Name Matching in Information Integration", en: *IEEE Intelligent Systems 18(5)*, 2003, pp. 16-23.

Blasnik, M. *RECLINK: Stata module to probabilistically match records*. 2010. Consultado en: http://EconPapers.repec.org/RePEc:boc:bocode:s456876.

Christen, P. "A Comparisonof PersonalName Matching: Techniques and Practical Issues", en: *Joint Computer Science Technical Report Series*. The Australian National University, 2006.

Collins, M. "A nex statistical parser bases on bigram lexical dependencies", en: *In Proceeding of the 34th Annual Meeting of the Association of Computational Linguistic*. Santa Cruz, CA, 1996, pp. 184-191.

Elmagarmid, A. K., P. G. Ipeirotis y V. S. Verykios. "Duplicate Record Detection: A Survey", en: IEEE Transactions on Knowledge and dataEngineering. Vol. 19, Núm. 1, enero de 2007, pp. 1-16.

Fellegi, I. P. y A. B. Sunter. "A Theory for Record Linkage", en: *Journal of the American Statistical Association*. 1969, pp. 1183-1210.

Hernández, M. A. y S. J. Stolfo. "Real-World Data is Dirty: Data cleasing and the merge/purge problem", en: *Data Mining and Knowledge Discovery*. 2, 1998, pp. 9-37.

Herzog, T. N., F. J. Scheuren y W. E. Winkler. *Data Quality and Record Linkage Techniques*. United States of America, Springer, 2007.

Leicester, G. *Methods for Automatic Record Matching and Linkage and Their Use National Statistics. National Statistics Methodological Series*. Oxford, 2001.

Machado, C. J. "A literature review of record linkage procedures focusing on infant health outcomes", en: *Cad. Saú de Pública, Rio de Janeiro*. 20(2), 2004, pp. 362-371.

Reif, J. Stata module to match strings base on their Levenshtein edit distance. 2010. Consultado en: http://ideas.repec.org/c/boc/bocode/s457151.html

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Análisis de los microdatos del censo de 1930: a 80 años del México posrevolucionario

Francisco José Zamudio Sánchez, Roxana Ivette Arana Ovalle, Waldenia Cosmes Martínez, Javier Santibáñez Cortés y Margoth Laredo Rojas

Con la finalidad de hacer estimaciones que contribuyan a completar la historia demográfica de México, se recuperó 10% de los microdatos del Quinto Censo de Población. Para ello, las boletas originales fueron digitalizadas, capturadas y validadas. Posteriormente, se hicieron cálculos de todas las variables incluidas en el censo de 1930, las cuales fueron comparadas con las del 2010.

En estas ocho décadas se nota el impacto de las políticas públicas de salud, educación, economía y sociales (inclusión de género).

Los hallazgos más importantes son: la atípica pirámide de edades de la población de 1930, que el índice de masculinidad del 2010 está en el mismo nivel que en 1930, la pérdida de 10% de la población hablante de lengua indígena, la incorporación de 26% de la población femenina económicamente activa a las actividades productivas y la conservación del catolicismo.

La consistencia en los resultados de este trabajo sugiere que la base de microdatos puede ser usada en otros estudios.

Palabras clave: microdatos, indicadores demográficos, políticas públicas y censo.

Recibido: 27 de septiembre de 2014.
Aceptado: 5 de mayo de 2015.

In order to make estimations which help to complete the demographic history of Mexico, it has been recovered 10% of the micro data of the Fifth Population Census. To do this, the original documents were scanned, captured and validated. Subsequently, estimations were made of all variables included in the 1930 census, and they were compared with the variables of the 2010 census.

In these eight decades, is visible the impact of public policies on: health, education, economy, and on the social aspects (gender inclusion).

The most important findings are: the population pyramid in 1930, the sex ratio in 2010 is on the same level as in 1930, the loss of 10% of the population speaking indigenous language, the incorporation of 26 % of the economically active female population to productive activities, and the conservation of Catholicism.

Consistency in the results of this work suggests that the base of microdata recovered, can be used in other studies.

Key words: microdata, demographic indicators, public policy and population Census.

Mexiko, Ansichten/Ullstein bild/Getty Images

Presentación

Desde el punto de vista científico, las encuestas constituyen, sin duda alguna, un instrumento valioso para el conocimiento de la realidad social y política; en ese sentido, son ampliamente utilizadas en la investigación en Ciencias Sociales (Riba y Cuxart, 2004); también, los microdatos censales son un recurso de gran valor por un doble motivo: su propia condición (registros individuales que permiten explorar de manera simultánea las características de los individuos, familias, hogares y viviendas en que residen) y porque proceden de censos, fuente estadística sin igual, pues ninguna otra ofrece una densidad muestral, profundidad cronológica y cobertura geográfica comparables; no en vano, el censo conserva la mayor representatividad a escala nacional. Durante el último medio siglo, la mayor parte de los principales organismos de estadística han preparado los archivos de microdatos censales para el análisis por parte del personal y, en muchos casos, por investigadores externos; hoy en día, éstos son ampliamente utilizados por ellos y los encargados de formular políticas en los países desarrollados, pero son poco usados en el resto de las naciones (McCaa *et al*., 2005).

El presente trabajo describe a la población mexicana de 1930, haciendo estimaciones con una muestra de los microdatos del Quinto Censo de Población, los cuales fueron obtenidos del proyecto de investigación *Muestreo probabilístico para la recuperación de los microdatos del censo general de 1930*, financiado por el Fondo Sectorial CONACYT-INEGI.

Debe notarse que las boletas censales datan de 1930 y, por su valor histórico, se encuentran resguardadas en el Archivo General de la

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Nación (AGN) bajo las medidas de preservación necesarias para este tipo de documentos. Lo anterior, representó un reto importante, ya que se digitalizaron los originales en imágenes de alta resolución y se utilizó un sistema de captura en línea; el proyecto fue coordinado por el Departamento de Estadística Matemática y Cómputo de la Universidad Autónoma Chapingo (UACh).

El objetivo de la investigación fue recolectar los microdatos de 10% de los habitantes de 1930 que ascendía a 16 552 722 personas —Departamento de Estadística Nacional (DEN), 1932—; sin embargo, resultaron extraviadas las boletas de aproximadamente 24% de la población, es decir, en el AGN sólo se encontraban los registros correspondientes a 12 555 147 pobladores. A partir de la muestra, se logró estimar 99.9% de los habitantes registrados en el AGN, que representaba 75.8% de la población total de ese entonces. Se pudo medir la eficiencia de esta estimación, ya que existen los tabulados básicos publicados por el DEN en 1932.

La pérdida de datos más crítica se dio en el Distrito Federal (DF), pues no se encontró boleta alguna y, por ello, no se pudo hacer estimación alguna; le siguieron Quintana Roo (80%), Aguascalientes (77%), Colima (72%), Querétaro (66%), Baja California (Distrito Norte) (61%), Baja California (Distrito Sur) (58%) y Tlaxcala (51%).[1] Del resto de los estados sólo se perdieron, en promedio, 17% de las boletas. Los resultados anteriores son estimaciones de la población.

Es importante hacer notar que, aunque desde 1932 existen los tabulados básicos del censo, los microdatos recuperados en el presente estudio abren la posibilidad de realizar diferentes cruces importantes para una detallada caracterización de la población de la época a escala municipal; además, podrán ser usados a partir de ahora como otra fuente de información para realizar estimaciones más precisas, confiables y con mayor complejidad estadística. De manera fundamental, su recuperación proporcionará información a escala municipal, con cohortes de edad, sexo, estado civil, religión, etcétera.

El muestreo estratificado[2] permite tener una precisión controlada para cada estrato o subdivisión de la población; puede producir una ganancia en la precisión de las estimaciones de toda la población dependiendo de la homogeneidad de la información dentro de los estratos (Cochran, 1977). El muestreo por conglomerado es útil en la práctica porque usualmente es más económico y conveniente que hacerlo al azar en la población (Lohr, 1999). Además, el muestreo sistemático es, quizá, el procedimiento de selección más conocido, de uso común, simple de aplicar (Kish, 1965) y disminuye los errores de recolección (Cochran, 1977).

Dadas las propiedades teóricas comentadas, se propuso un diseño estratificado por conglomerados con selección sistemática, donde los estratos fueron los municipios y los conglomerados, las boletas censales que contenían, en promedio, 80 registros (habitantes); éste fue validado y enriquecido por el doctor Robert McCaa del Centro de Población de Minnesota.

Para disminuir los posibles factores que afectan la calidad de la información demográfica, en particular de un censo de población, es necesario realizar una revisión en las diferentes fases de obtención de los datos: precensal, de levantamiento y poscensal (Naciones Unidas *et al*., 2014). En la captura de 10% de la muestra se tuvieron en consideración los siguientes aspectos para asegurar la calidad de la información:

- La evaluación de la etapa precensal y el levantamiento del censo de 1930 se realizó bajo la supervisión del DEN; de acuerdo con los documentos históricos, se sabe que este operativo fue el primero realizado en México con altos estándares de calidad.

1 Los porcentajes que se indican entre paréntesis indican las pérdidas correspondientes.

2 "Es la muestra obtenida mediante la separación de los elementos de la población en grupos que no presenten traslapes, llamados estratos" (Scheaffer, R. *et al*., 1987).

Además, se conoce del trabajo presentado por McCaa y Gómez-Galvarriato (2013), en el cual se realizó un muestreo piloto de 1% de la información del Censo General de 1930, la recomendación de capturar 10% del mismo, pues consideraron que los documentos tenían la calidad suficiente para proporcionar información valiosa de la población mexicana de 1930. Con estos antecedentes, el INEGI decidió destinar recursos para la recuperación de estos datos.

- En este estudio, para determinar si la información que se estaba capturando tenía la calidad necesaria, se instrumentó la captura de 10 submuestras aleatorias. Al terminarla, se efectuó un proceso de validación para verificar la correcta inserción de los valores en las variables; además, se realizaron procesos de validación estadísticos manuales, semiautomáticos y automáticos con el objetivo de verificar la congruencia entre variables.
- Previo al análisis de los microdatos, se hizo una etapa de evaluación y correción de la información censal. Para llevarla a cabo, hay diversas técnicas y herramientas que se usaron de manera crítica acorde con la situación. Cabe destacar que, como lo mencionan Chackiel y Macció (1978): "No existe un método único capaz de proveer el mejor ajuste ni hay una técnica que reúna las cualidades de una receta universal. Por el contrario, el trabajo de corrección y evaluación se rige por los intentos sucesivos y las pruebas alternativas que conducen al 'ajuste más plausible' o el 'error más probable' determinado por el juicio y pericia del investigador".
- Para el manejo de los microdatos del censo de 1930, se usaron diversas técnicas de corrección, basadas sobre todo en la coherencia de la información y utilizadas en no más de 10% de la información por variable; estos procedimientos siguieron la pauta de algunos de los principios básicos generales, enunciados por Chackiel y Macció: serendipia, rehabilitación, consistencia o coherencia, robustez, ausencia de norma estricta y conocimiento de las circunstancias históricas y culturales del país.

De los resultados obtenidos y las comparaciones realizadas con los tabulados básicos del censo de 1930 se puede decir que, en general, la muestra representa de manera eficiente a la población y que los resultados en números relativos a escalas nacional y estatal tendrán un error de alrededor de 5%, salvo en los casos que se mencione.

Antecedentes

El ritmo de crecimiento de la población a nivel mundial aumentó de forma leve y se mantuvo en torno a 0.5% durante el siglo XIX, llegando a 2 mil millones de habitantes en 1927. Los últimos mil millones se acumularon en 127 años. Desde entonces, al 2012 se han sumado 5 mil millones más (UNFPA *et al*., 2012). Lo anterior, deja claro que el periodo comprendido entre 1927 y nuestros días ha sido de grandes cambios poblacionales a escala mundial. Por esto, varios países se han ocupado en recuperar datos históricos para la explicación de la dinámica de sus poblaciones, todo ello con el fin de proveer información relevante para el diseño de políticas públicas que se ocupen de brindar mayor calidad de vida a sus habitantes.

Como en el resto del mundo, la población mexicana ha cambiado de manera drástica en los últimos 80 años; el conflicto armado de 1910, las políticas poblacionales adoptadas por el gobierno y la industrialización, entre otros factores, han provocado grandes cambios en las tasas de crecimiento (ver gráfica 1).

En los primeros 30 años del siglo pasado, México sufrió tres acontecimientos importantes que diezmaron su población: 1) la Revolución Mexicana (1910-1921); 2) un brote de influenza (1918-1919), considerado como uno de los más críticos ocurridos en el país y el mundo y 3) la Guerra Cristera (1926-1929). Todos tuvieron incidencia en la población de 1930, cuando se realizó el censo bajo estudio, de

manera que las variables sociodemográficas tienen reflejado su impacto, entre otras: la tendencia del crecimiento poblacional, sexo por grupo quinquenal, estado civil y defectos físicos.

Para comprender la importancia que tiene el Quinto Censo de Población de México para las estadísticas nacionales, a continuación se hace una reseña de los principales eventos relacionados.[3]

La historia de los censos de población en México comienza en 1882 con la entonces Dirección General de Estadística (DGE) a cargo de don Antonio Peñafiel, quien sería el responsable de levantar los tres primeros: 1895, 1900 y 1910, todos bajo el régimen de Porfirio Díaz. Aunque en estos operativos se hizo un intenso trabajo para levantar la información, sirvieron muy poco en términos de la administración nacional del Estado.

Desde finales de 1910 hasta 1921 se desarrolló la Revolución Mexicana, la cual trajo consigo un importante retroceso en las estadísticas nacionales. Después de su periodo más crucial, Luis I. Mata fue el encargado de ejecutar el Censo General de Habitantes (1921), el cual se desarrolló en medio de diversas complicaciones logísticas. Los puntos más críticos en el levantamiento fueron: personal insuficiente y mal preparado, escasa participación por parte de varios estados y la falta de actualización de la cartografía municipal. Lo anterior, como es de suponerse, fue producto del proceso de reestabilización que atravesaba el país.

A partir de 1923, la estadística nacional empezó a tomar fuerza nuevamente con la desaparición de la DGE y la aparición del DEN, creado con la finalidad de centralizar y organizar los censos. Personaje clave en la nueva visión del manejo de las estadísticas nacionales fue el ingeniero Juan de Dios Bojorquez, director del DEN, cuya idea fundamental era crear un área de estadística que concentrara los registros nacionales, estatales y municipales; además de estar a cargo del Quinto Censo de Población, fue el responsable de aplicar el primer Censo Industrial y el Censo Agrícola Ganadero inicial, todos realizados en 1930.

A finales de 1929, el Censo de Población de 1930 ya había generado grandes expectativas en varios funcionarios del gobierno, quienes también encontraron importante recabar información del Estado desde 1923, como: estadísticas de crímenes, mortalidad, suicidios y divorcios, entre otras. Además, se habían realizado la primera y segunda reuniones nacionales de Estadística (1927 y 1929, respectivamente), donde Bojorquez y su equipo lograron aquilatar los conocimientos teóricos y prácticos adquiridos y difundir la importancia del papel de la estadística para el desarrollo de la República. El objetivo señalado para la Segunda Reunión Nacional de Estadística fue: "Encontrar los medios más prácticos y eficaces para hacer que los censos de 1930 sean lo mejor que México haya realizado en esta materia…", según el discurso final del ingeniero.

Por fortuna, no sólo se trató de un discurso, sino que la documentación del Quinto Censo de Población indica que se realizó cuidando de manera escrupulosa muchos detalles, desde la planeación hasta la consolidación de la información en los tabulados básicos presentados en 1932. Se puede decir que fue el primer censo realizado en México con los mejores criterios de calidad de la época, además de la utilización de una intensa campaña de difusión que pretendió informar a todos los habitantes acerca de la importancia de brindar datos fidedignos a los empadronadores quienes, a su vez, fueron seleccionados entre los habitantes más letrados de entonces. Uno de los aspectos más importantes en la ejecución del levantamiento fueron las boletas censales (en las que se podían registrar hasta 100 personas), las cuales, por primera vez, eran levantadas por un empadronador y no como en los censos anteriores, que se hicieron por autoempadronamiento.

Las temáticas de ese censo fueron: sexo, edad, estado civil (soltero, casado por la iglesia, casado por lo civil, divorciado y viudo), alfabetismo (sabe

3 Consultados en INEGI. *125 años de la Dirección General de Estadística 1882-2007* (INEGI, 2010a).

leer y sabe escribir), quehaceres domésticos, ocupación, lugar de nacimiento, lugar de residencia, nacionalidad, lengua indígena, bienes raíces (propiedad en la ciudad y propiedad en el campo), tenencia de la vivienda, defectos físicos y mentales, religión, desempleo, asistencia escolar y bienes raíces. Todas las variables de estas temáticas y las territoriales fueron digitalizadas para este proyecto.

Resultados

Lo que se presenta a continuación proviene de la Base de Microdatos (BMD).[4] En este documento se usan los resultados a escalas nacional y estatal, ya que se trata de la primera investigación realizada con estos datos, y los niveles de agregación son pertinentes para difundir la información recabada, así como la calidad de la misma.

Características de la población

A continuación, se presenta el análisis comparativo de la población mexicana de 1930 y el 2010 usando la BMD y, cuando se indique, los datos provenientes de los tabulados básicos de ambos censos.

La decisión acerca de tomar el año 2010 para hacer las comparaciones que se presentan se basó en el hecho de ser el censo más reciente con el que contamos y se deseaba observar los cambios experimentados en la población; aunque no es posible hacer comparaciones en algunas métricas a nivel nacional (debido a los datos extraviados), sí se puede hacer con otros donde se tiene suficiente muestra y, por supuesto, todas las estimaciones pertinentes a nivel municipal de los 2 019 municipios que se pudieron recuperar.

Para el levantamiento de la muestra censal de microdatos 2010, se diseñó un cuestionario ampliado con el que fueron censadas alrededor de 2.9 millones de viviendas en el país, que representaron 10% de las viviendas habitadas, la cual fue seleccionada con criterios probabilísticos especificados por el INEGI (2011). Como es de suponerse, la calidad y el nivel de profundidad de los procesos usados en este censo son mayores que los utilizados en 1930; además, las temáticas abordadas incluyen, también, características económicas y de las viviendas, al mismo tiempo que profundizan en las demográficas, culturales y sociales (INEGI, 2015).

Crecimiento de la población

Según McCaa (2003), las pérdidas asociadas a la Revolución se estiman en un rango de 1.9 a 3.5 millones de personas. Se han realizado diversos estudios en torno al número de vidas que se perdieron en la guerra civil de 1910 en México (Loyo, 1935; Collver, 1965; González Navarro, 1974; Mier y Terán, 1982; Ordorica y Lezama, 1993; Ibarra, 1996), sin embargo, ha sido complicada la estimación porque se cuenta con poca información histórica de la época o ésta es de poca calidad, además de que estos estudios tomaban en cuenta otros factores como: migración, mortalidad por epidemias, descenso de número de nacimientos, etcétera. Debido a esto, en la actualidad no existe consenso entre los investigadores respecto al impacto demográfico que tuvo la Revolución Mexicana ni en torno a los factores que deben tomarse en cuenta.

La gráfica 1 muestra el crecimiento poblacional en México; se identifica de forma clara el periodo comprendido entre 1910-1921, ya que es el único punto donde el crecimiento presenta un valor negativo; en consecuencia, el siguiente (1930) cobra particular relevancia en la historia sociodemográfica del país y es un aspecto importante de los microdatos recuperados por el proyecto. Con éstos, se tienen las estimaciones de las poblaciones de todos los municipios de la muestra, los cuales presentan errores menores a 5 por ciento.

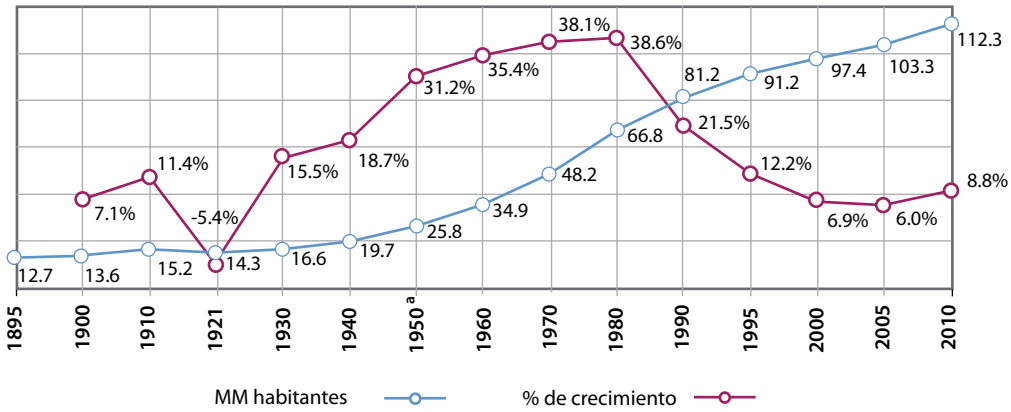
El acelerado crecimiento de la población mexicana se explica por las bajas tasas de mortalidad y los altos niveles de natalidad (Weeks, 1999). La

4 Es la entregada al INEGI como producto final del proyecto de investigación, la cual fue corregida como se indicó en la presentación. Los datos fueron expandidos según la metodología usada en los microdatos del censo del 2010 para obtener estimaciones poblacionales (INEGI, 2011).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 1

Crecimiento de la población mexicana, 1895-2010



[a] El total incluye 11 763 habitantes, dato registrado bajo el concepto de *Complementarios*, el cual no se presentó por entidad federativa.
Fuente: INEGI. Censos de población y vivienda de 1895 al 2010. Tabulados básicos.

disminución de la tasa de mortalidad infantil durante el siglo XX es notable, y aunque los registros estadísticos desde 1896 a 1923 son muy irregulares debido a las malas condiciones para recabarlos y al periodo revolucionario, se encuentran datos que indican que la mortalidad infantil disminuyó de 308 defunciones por cada mil nacimientos vivos en 1896 a 222 en 1923; este indicador siguió descendiendo hasta 131.6 en 1930 (Cordero, 1968) y para el 2010 se ubicó en 14.1 (CONAPO, 2014).

Las tasas de natalidad entre 1930 y 1970 se mantuvieron en el orden de 40-45 nacimientos por cada mil habitantes en un año (Alba, 1976). Después de la aplicación de la *Ley General de Población* de 1974, la tasa descendió de forma continua: a 35 en 1980, 27.9 en 1990, 23.4 en el 2000 y 19.7 en el 2010 (CONAPO, 2014).

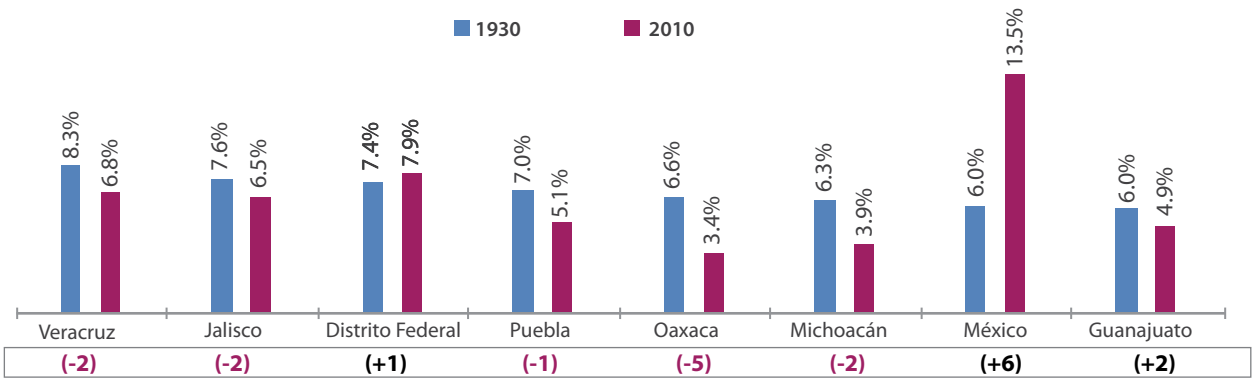
Concentración y densidad de la población

En 1930, 55.2% de la población se concentraba en sólo ocho entidades, composición que, prácticamente, no ha cambiado en los últimos 80 años, pues en la actualidad, en estos mismos estados, radica 52% de los habitantes (ver gráfica 2).

Explicaciones sobre la concentración de la población mexicana hay varias, a continuación se comentan dos:

Gráfica 2

Entidades que concentran más de 50% de la población en 1930 y el 2010



Nota: el número entre paréntesis indica cuántos lugares ha descendido (–) o aumentado (+) la entidad.
Fuente: INEGI. Censos de 1930 y 2010. Tabulados básicos.

- Gutiérrez (2003) indica que los factores geográficos, económicos, sociales y políticos, específicamente las vías de comunicación con las que cuenta la Ciudad de México, el desarrollo industrial que proporciona fuentes de trabajo, los centros de cultura, de arte y de diversión, así como la disponibilidad de los servicios públicos, han propiciado la concentración de habitantes en el centro del país.
- Asuad (2014), por su lado, comenta que: "A pesar de la reestructuración de la economía mexicana y su orientación a las exportaciones, el crecimiento económico de la Ciudad de México y los servicios que presta, siguieron atrayendo población y expandiendo su mancha urbana de manera explosiva".

Por otro lado, a lo largo de poco más de 80 años, la población mexicana ha cambiado de manera significativa en todas las variables demográficas. El principal indicador de densidad poblacional creció de 8.4 hab./km^2 en 1930 (DEN, 1932) a 57 hab./km^2 en 2010 (INEGI, 2010b).

Microdatos

Respecto al total poblacional, se capturaron 11.9% de las boletas (unidades muestrales), lo que, al expandir la muestra, corresponde a 75.8% de la población (96.9% de las entidades y 88.1% de los municipios).

División geográfica

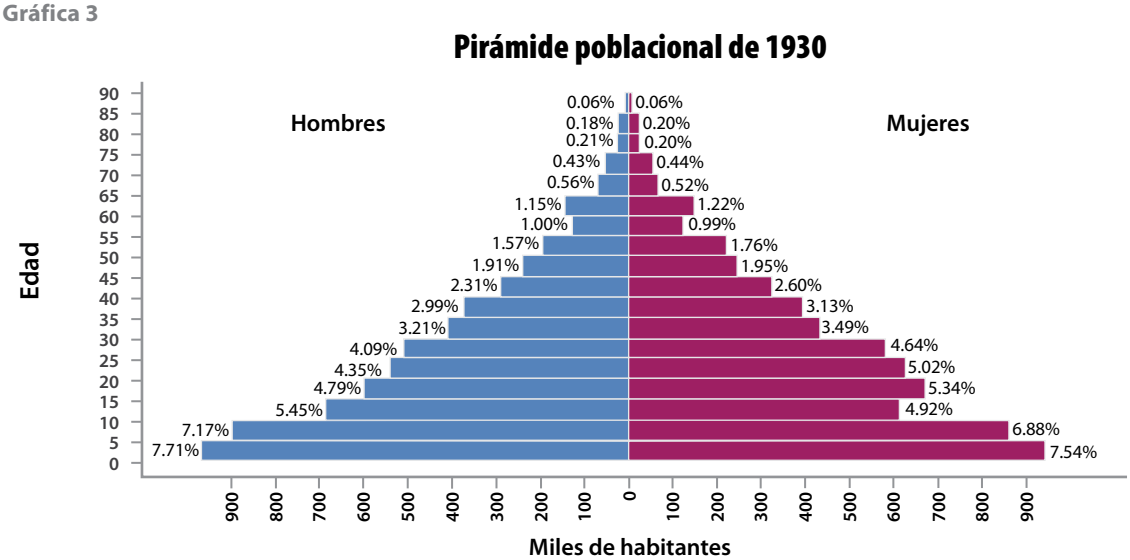
En 1930 existían dentro de las 30 entidades y el DF 2 293 municipios y 84 452 localidades.[5] En la muestra de microdatos se lograron recolectar 2 019 municipios y 11 719 localidades, las cuales se desagregan en ranchos (36.9%), pueblos (22.5%), rancherías (10.1%), haciendas (9.5%) y congregaciones (8.8%).

Edad y sexo

Su pirámide resume la historia de la población de al menos una generación y muestra el efecto de acontecimientos demográficos como el cambio en la fecundidad y en la mortalidad de una población. En la gráfica 3 se presenta la pirámide de 1930 usando los microdatos recolectados. Como se puede observar, tiene una base ancha que se estrecha rápidamente, lo que denota una población joven; también, se detectan grupos quinquenales con asimetrías entre hombres y mujeres. Un hecho

5 Representadas por ciudades, villas, pueblos, colonias, barrios, fábricas, congregaciones, haciendas, rancherías, ranchos, colonias agrícolas, minerales, plantas eléctricas, estaciones de ferrocarril, secciones de ferrocarril, campamentos y otras. Aunque en varios casos la categoría no parece pertenecer a una localidad, en este estudio se decidió apegarse a los resultados publicados por el DEN para hacer las validaciones de calidad correspondientes.

Gráfica 3



Fuente: microdatos expandidos del censo de 1930.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

que no tiene explicación inmediata es la marcada irregularidad en favor de los hombres en el grupo de 10-14 años (la cual es posible que sea el reflejo de la epidemia de influenza de 1918-1919, cuando parte de esta cohorte estaba entre los 0 y 2 años); sin embargo, de acuerdo con Márquez y Molina (2009), de los tres eventos que diezmaron a la población mexicana de 1910 a 1928, ninguno explica la asimetría mencionada, ya que la gripe española tomó, sobre todo, vidas de varones que para 1930 tendrían entre 32 y 42 años de edad.

Por otro lado, los movimientos armados afectaron principalmente a la población masculina.

Se deben hacer análisis que incorporen diferentes factores demográficos relevantes al fenómeno para explicar cuáles fueron los sucesos políticos, salubres, económicos y/o sociales que produjeron asimetrías tan marcadas en algunos grupos quinquenales.

Relación hombres/mujeres

La población mexicana estaba compuesta por 51% mujeres y 49% hombres, relación que se ha mantenido sin cambios desde entonces, exceptuando en los censos de 1960 y 1970, donde la composición fue de 50 y 50 por ciento.

Por otro lado, el índice de masculinidad (IM)[6] en 1930 era de 96.3, y en 1921 fue de 95.5, lo cual significa que por cada 100 mujeres había un hombre menos que en 1930; esto se puede explicar por el movimiento armado, el cual involucró, sobre todo, al sexo masculino.

Al estudiar la relación entre hombres y mujeres por grupo de edad a través del IM, es posible identificar patrones de diferenciación por sexo y edad de la mortalidad. Para dejar explícita esta relación, se debe señalar que mientras menor sea el valor del índice de masculinidad, se puede pensar que más han sido las pérdidas en la población masculina.

Los factores que afectan de manera negativa a una población son la emigración y la mortalidad. Ordorica y Lezama (1993), citados por Verduzco (1997), señalan que en el periodo de 1911 a 1921, del total de la población mexicana perdida, el impacto de la migración internacional fue de 13% —del que se atribuye 8% hacia Estados Unidos de América (EE.UU.) y 5% a otros países, principalmente Cuba y Guatemala—, lo cual deja paso a la mortalidad para explicar los cambios en la composición de la población.

En la gráfica 4 se presentan los índices de masculinidad calculados por grupos de edad de la población de 1930 y del 2010. Esto nos permite estudiar los efectos de la sobremortalidad masculina y comparar sus cambios a través del tiempo. Se espera que hombres jóvenes y adultos tengan tasas de mortalidad elevadas. Garduño (2001) señala que esto es debido a que los rasgos impuestos en la masculinidad están relacionados con la destructividad.

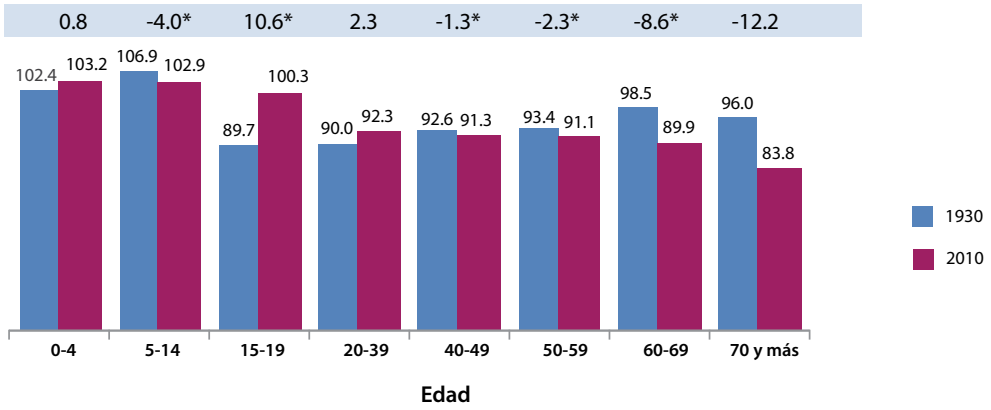
En primer lugar, se puede apreciar que la tendencia a la baja del IM con la información de 1930 no es tan fuerte como la que se observa en el 2010. También, se ve que en los rangos de 0-4 años de edad, 15-19 y 20-39 las diferencias son positivas entre el IM del 2010 frente al calculado en 1930. Por el contrario, éstas son negativas en los rangos de 5-14 años, y más de 40 años. Las diferencias positivas hablan de mejores condiciones para los hombres en el 2010 que las que hubo en 1930. Lo opuesto ocurre con las negativas, que se presentan de los 40 años en adelante.

Se debe resaltar que el indicador reportado en el 2010 fue 95.4, 0.1 menor que en 1921 y 0.9 menor que en 1930; es decir, en la actualidad presenta una situación disminuida para los hombres, como a finales del movimiento armado, sobre todo en los rangos mayores a 49 años. Aunque en todos los niveles existen diferencias, las más significativas están señaladas con asterisco; de éstas, la mayor se encuentra en el rango de 70 años de edad. Para explicar este fenómeno se debe observar la migración internacional y la mortalidad. En cuanto

6 Relación que muestra el número de hombres por cada 100 mujeres.

Gráfica 4

Índice de masculinidad por grupos de edad



Nota: el asterisco denota diferencia significativa entre las medias, con un nivel de significancia $\alpha = 0.05$.
Fuente: microdatos expandidos del censo de 1930 y el Censo de Población y Vivienda 2010.

a la primera, la Organización Internacional para las Migraciones (OIM) (2014) reportó que, en suma, alrededor de 600 mil mexicanos abandonan el país cada año (sobre todo hacia EE.UU.), con la intención de encontrar mejores oportunidades de empleo; de éstos, se estima que 75.4% son hombres (en promedio, mayores de 26 años). Por otro lado, durante el 2006 y el 2010, la mortalidad general anual aumentó entre 4 y 6%; en particular, la población masculina entre 25 y 44 años de edad fue disminuida principalmente por agresiones y accidentes; por último, los hombres de más de 45 años han sido afectados por enfermedades como la diabetes mellitus, las del corazón y las hepáticas, estas últimas relacionadas con el consumo de alcohol (INEGI, 2010c).

Condiciones tan desfavorables para los hombres del 2010 respecto a los de 1930 señalan que se debe trabajar en políticas económicas, de salud y seguridad para asegurar un índice de masculinidad que les brinde condiciones más equitativas (Zamudio F. *et al*., 2012).

Dependencia

El índice de dependencia demográfica (IDD)[7] expresa el número de personas en edades dependientes por cada 100 potencialmente activas, es decir, es la relación entre quienes tienen de 15 a 65 años de edad y los menores de 15 y mayores de 65 (INE, 2008).

La dependencia en 1930 era, en promedio, 72 por cada 100, sin embargo, en estados como Tabasco y Chiapas era mayor a 80; por el contrario, en Querétaro era de sólo 50. En el 2010, el IDD se situó alrededor de 55 a escala nacional, siendo Chiapas el que reportó el más alto (58) y el DF, el más bajo (32). En el 2010, en promedio había 17 personas dependientes menos que en 1930 por cada 100 personas activas. Ahora, la distribución del trabajo se encuentra mejor repartida por el cambio en la estructura poblacional.

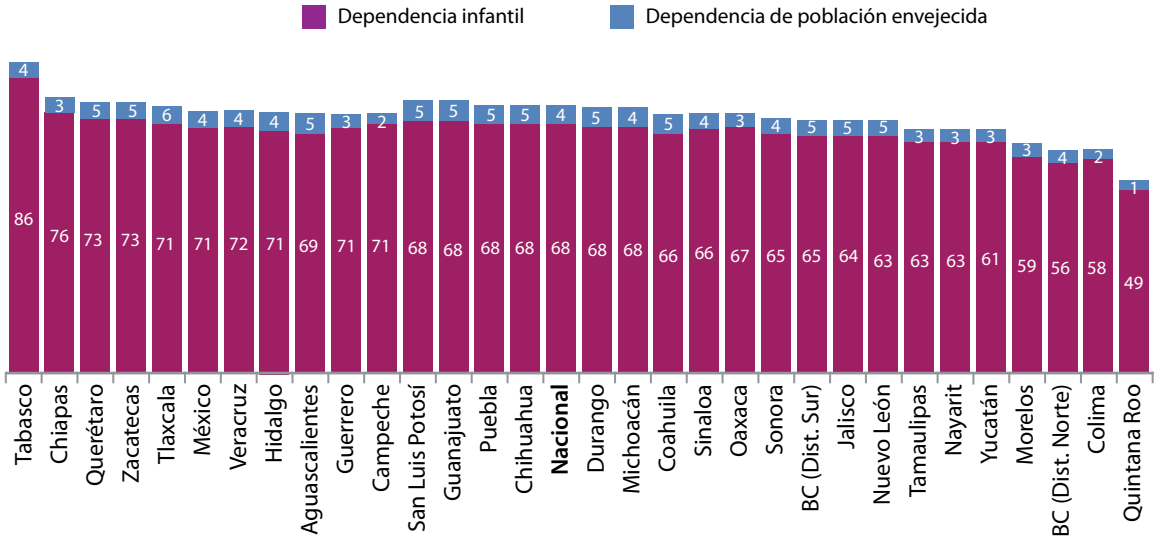
El proceso de envejecimiento demográfico trae consigo un periodo o ventana de oportunidad en el que se presentan las condiciones más favorables para el desarrollo debido al aumento de la población en edad laboral y a la reducción de la menor de 15 años; al mismo tiempo, la población adulta mayor todavía mantiene un peso, aunque pequeño. Las ventajas que ofrece esta situación serán mayores durante el lapso 2005-2030, cuando el índice de dependencia total será menor a 60 personas en edades dependientes por cada 100 en edad laboral. A partir de la tercera década de este siglo, el incremento pronunciado de la población adulta mayor cerrará este periodo de oportunidad demográfica (CONAPO, 2004).

7 Índice claramente económico que representa la medida relativa de la población potencialmente inactiva sobre la potencialmente activa (INE, 2014).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 5

Índice de dependencia por entidad, 1930



Fuente: microdatos expandidos del censo de 1930.

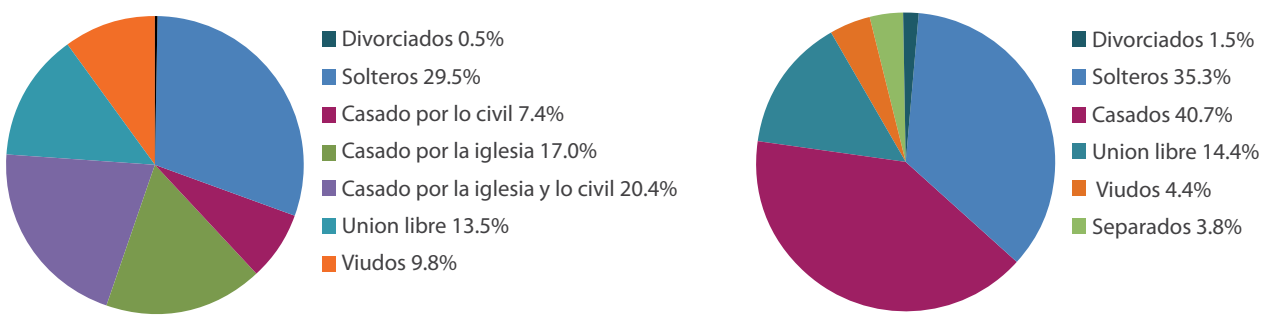
Nupcialidad

En cuanto a la variable estado civil, los microdatos indicaron que 44.8% de la población estaba casada por alguna de las tres modalidades (iglesia, civil o ambas), esto es cuatro puntos porcentuales más que el indicador en el 2010 (ver gráfica 6). La situación anterior se explica porque en ese año se tenían mayores porcentajes en las categorías de solteros, unión libre y divorciados (esta última triplicó el valor de 1930).

Por último, en la gráfica 6 se observa, también, que en la categoría de viudos el valor de 1930 es más del doble que en el 2010. Al respecto, en la gráfica 7 se puede ver que esta categoría está compuesta sobre todo por mujeres (9.1 puntos más), presumiblemente porque habían perdido a sus esposos durante el movimiento armado, ya que 59% de las viudas se encontraron en el rango de 40-65 años de edad. Aunque se podría pensar que este número es demasiado alto y que algunas

Gráfica 6

Situación conyugal de la población de 1930 (>14 años) y el 2010 (>12 años)



Fuente: microdatos expandidos del Quinto Censo de Población y el Censo de Población y Vivienda 2010.

de estas viudas *ficticias,* que se hacían llamar así por prejuicios sociales para no aceptar que eran madres solteras como ocurrió en la Colonia (McCaa, 1991), Arrom (1985) afirma que, aun cuando a su parecer no todas las viudas eran legítimas en ese periodo, las implicaciones demográficas y sociales son casi idénticas; de esta manera, deben hacerse análisis acerca de las implicaciones que tuvo esta variable en el desarrollo demográfico del México posrevolucionario.

Migración

Durante el Porfiriato se favoreció la inmigración, ya que se tenía la idea de que la población nativa era insuficiente para alcanzar el índice de progreso que otros países habían logrado. Después de debatir entre los beneficios económicos y los argumentos racistas, el gobierno decidió permitir la llegada de obreros chinos (por lo barato de su mano de obra) para trabajar en la agricultura y la industria, en especial la ferrocarrilera (Velázquez, 2005).

Posterior a 1914 sobrevino una gran inestabilidad en el tema migratorio, debido sobre todo a los conflictos tanto nacionales como internacionales, de manera que el Censo General de Habitantes de 1921 marcó un decremento en la cifra de ciudadanos extranjeros radicados en México; sin embargo, igual que durante la época del Porfiriato, las políticas demográficas de los regímenes de Álvaro Obregón (1920-1924) y Plutarco Elías Calles (1924-1928) favorecieron la inmigración extranjera, con lo que se registró una recuperación en 1930 (Salazar,1996).

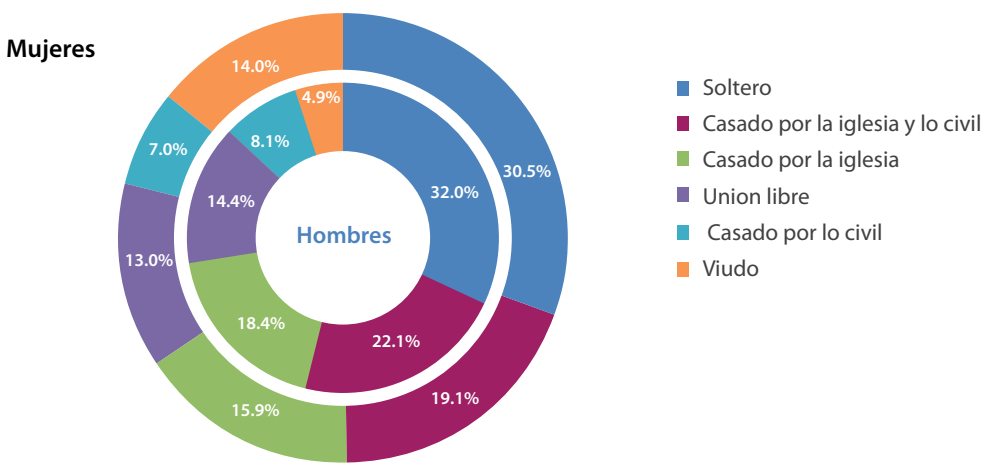
Por otro lado, EE.UU. repatrió en forma masiva población de origen mexicano después de 1921, estimándose que entre 1929 y 1935 regresaron a nuestro país más de medio millón de personas que tuvieron que ubicarse sobre todo en las ciudades fronterizas (Gutiérrez, 1995).

En 1930, 7.4% de la población que vivía en México no había nacido en el lugar de residencia reportado al momento del Censo, siendo Quintana Roo, Baja California (Distrito Norte), Baja California (Distrito Sur) y Tamaulipas las que reportaron mayor porcentaje de inmigración. Esto concuerda con lo señalado por Sobrino (2010), quien indica que los principales receptores en 1930 fueron los estados del norte del país y Quintana Roo en el sur, éste, desde entonces, se distingue por su flujo migratorio.

Por el contrario, las entidades que reportaron mayor emigración fueron Aguascalientes, Colima y Querétaro.

Gráfica 7

Nupcialidad por sexo, 1930



Fuente: microdatos expandidos del censo de 1930.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

En el 2000, las principales entidades receptoras de migrantes nacionales resultaron ser Quintana Roo (55.5% de su población total), Baja California (41.2%) y el estado de México (38.6%); para el 2010 los resultados fueron similares, excepto que esta última entidad fue desplazada por Baja California Sur. En cuanto a las entidades que tuvieron mayores porcentajes de expulsión en el 2000 se encuentran el DF (51.8% de su población), Zacatecas (38.6%) y Durango (30.9%), que para el 2010 conservaron valores similares.[8]

Población extranjera

La representada por los microdatos expandidos corresponde a 84 005 personas, lo cual equivale a 0.67% de la población que vivía en México en 1930.

En el DF se encontraba 31% de ella, por este hecho, la estimación nacional de los microdatos difiere del porcentaje que se reporta en los informes de 1932, donde se indica que 1% de la población que residía en México en 1930 era de origen extranjero.

Para dar una visión acerca de la población extranjera en México en 1930, se usaron los tabulados básicos de la DEN, ya que a nivel nacional no es adecuado reportar las estimaciones de los microdatos; sin embargo, a nivel municipal será correcto hacerlo.

Las entidades que albergaban más extranjeros eran el DF (31%), Chiapas (12%) y Veracruz (9%). La participación por nacionalidad se componía de la siguiente forma: estadounidense (26%), española (21%), china (11%), guatemalteca (9%), canadiense (4.2%) y alemana (3.2%), entre las de mayor representación.

En el 2010, la población extranjera radicada en México siguió significando 1% de la población, siendo Baja California (13%), Jalisco (9%), Chihuahua (8%) y el DF (7%) los que tenían mayor cantidad.

8 Cálculos realizados con los datos de los censos de población y vivienda del 2000 y 2010 (Romo, 2012).

En 1930, según los microdatos, sólo 1.1% de la población hablaba un idioma extranjero, entre los más frecuentes: inglés (62.4%), chino (12.3%) y alemán (6.1%). Para el 2010, esta situación ha cambiado de manera considerable, ya que se estima que 12% de la población habla alguna lengua extranjera, en particular, inglés (González *et al*., 2011).

Analfabetismo

Desde el México independiente se hablaba de la educación como el camino para progresar; sin embargo, los eventos armados, el brote de influenza y la Guerra Cristera crearon un ambiente de inestabilidad y un claro estancamiento en el desarrollo del país; en particular, el analfabetismo en los adultos era muy alto; por esta situación, en 1920, José Vasconcelos impulsó la primera campaña formal de alfabetización desde la rectoría de la Universidad Nacional de México, que buscó solventar la falta de instrucción que tenía la población (Bonilla, 2011).

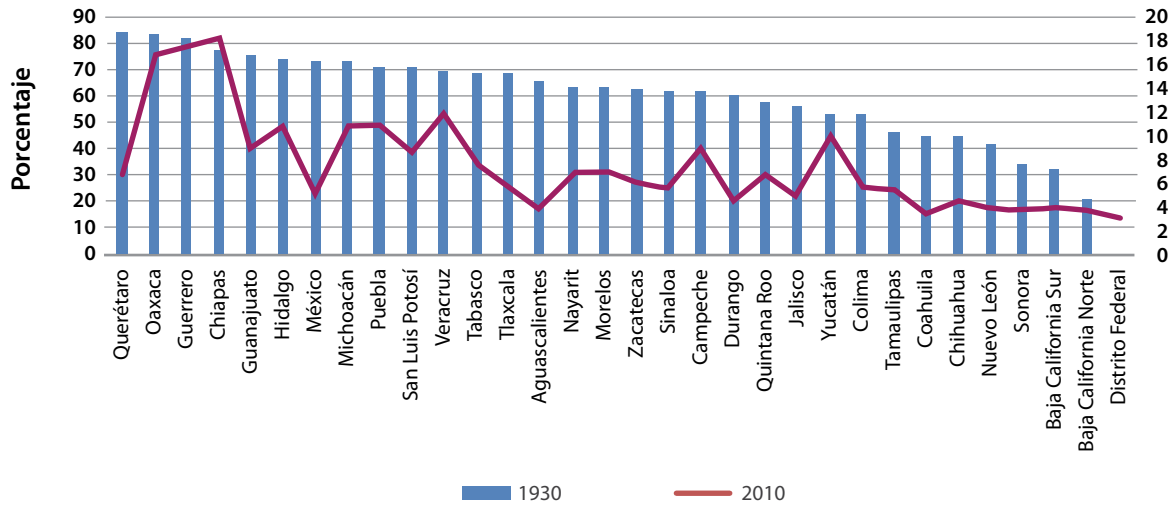
Los microdatos expandidos indican que 66.8% de la población mexicana mayor de 15 años era analfabeta (56% mujeres y 44% hombres). En un positivo contraste, los resultados del censo del 2010 indicaron que el analfabetismo en México estaba en el orden de 8%, casi 59 puntos porcentuales abajo de 1930, menor al de los países de Centroamérica, pero aún por encima de los más desarrollados de Europa, Norteamérica y Sudamérica.

En la gráfica 8 se observa que en Oaxaca, Guerrero y Chiapas, el promedio de analfabetismo en el 2010 fue de 17.5%, muy por debajo del de 1930 (81.1%); sin embargo, es notorio que el rezago respecto a las demás entidades, se mantiene hasta la fecha. No así para casos como el estado de México, Guanajuato, Tlaxcala y Durango, quienes escalaron de cinco a 15 posiciones a nivel nacional. También, ha habido retroceso de posiciones en algunos estados como Veracruz, Campeche y Yucatán; éste último presenta el mayor.

En cuanto a la escolaridad, Muñoz (1980) dice que las políticas públicas instrumentadas en la década de los 80 lograron que 35% de la pobla-

Gráfica 8

Analfabetismo por entidad federativa, 1930 y 2010



Nota: Querétaro, Quintana Roo, Colima y Aguascalientes contaban con menos de 35% de la población, por lo que algunas estimaciones como ésta deben tomarse con cautela.
Fuente: microdatos expandidos de los censos de 1930 y 2010.

ción entre 13 y 18 años de edad pudiera tener la educación posprimaria y que 11% de los jóvenes mayores de 18 años y menores de 25 alcanzaran la educación superior. En cambio, en 1930, las oportunidades de recibir estos niveles de educación sólo estaban al alcance de 1% de la población que se encontraba en estos grupos de edades.

Así, en 1930, de los niños entre 6 y 14 años sólo 36.3% asistía a la escuela, siendo Aguascalientes, BC (Distrito Norte), Yucatán y Morelos los que tenían a más niños matriculados; por el contrario, Michoacán, Veracruz y Guerrero son los que presentaron los menores índices de matriculación en este rango de edad.

En comparación, según el censo del 2010, 87.3% de la población entre 6 y 14 años de edad asistía a la escuela; los mejores indicadores de matriculación fueron en San Luis Potosí (SLP), Tlaxcala y Nuevo León. Los estados que presentaron los mayores avances con respecto a 1930 son SLP, Chiapas, Hidalgo y Guanajuato; por el contrario, los de mayor retraso, Baja California, Morelos y Sonora.

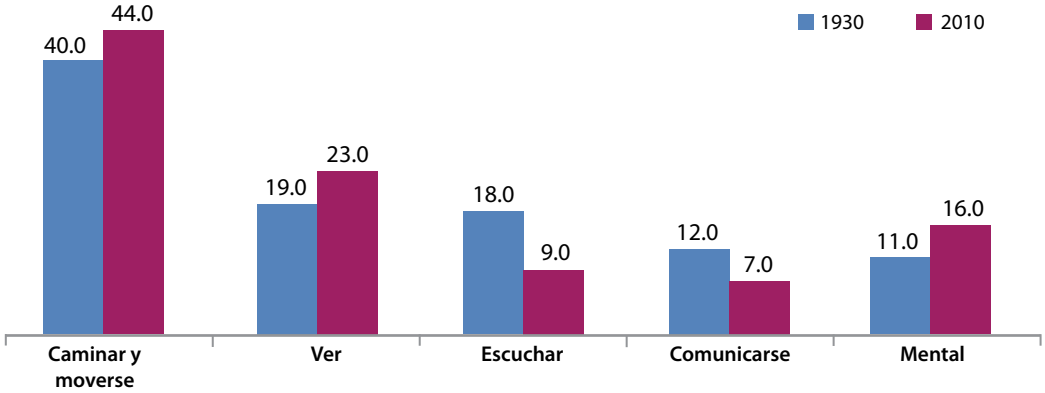
Discapacidades

En 1930, 0.72% de la población indicó que sufría algún tipo de discapacidad, prácticamente el mismo porcentaje que el reportado en 1921 (0.73%) y muy diferente al 5.1% reportado en el censo del 2010. Un comparativo con los microdatos expandidos de 1930 y los resultados del censo del 2010 se muestra en la gráfica 9. En cuanto a las causas que ocasionaron este cambio, el INEGI (2013) indica que las discapacidades para caminar y moverse se deben al proceso de envejecimiento (el cual es más notorio en las mujeres porque su esperanza de vida es mayor a la de los hombres); en los varones, un factor adicional son los accidentes que sufren debido a que sus actividades productivas implican mayor riesgo. En relación con la disminución que se presenta en las discapacidades para escuchar y comunicarse, esto ha ocurrido por la prevención de enfermedades que se hace en edades tempranas, prácticas que no se realizaban con anterioridad. Por último, las mentales en el 2010 han aumentado de forma notoria, sobre todo en la población infantil y juvenil masculina, pues están relacionadas con

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 9

Distribución porcentual de la población por tipo de discapacidad



Fuente: microdatos expandidos de 1930 y Censo de Población y Vivienda 2010.

problemas de nacimiento; esta tendencia es consistente con lo que reportó la Organización Mundial de la Salud en el 2011, al indicar que 20% de los niños y adolescentes del mundo tienen alguna discapacidad de este tipo, lo que las convierte en uno de los principales problemas de salud pública del planeta.

Lengua indígena

Con base en los microdatos, se puede apreciar que la población hablante de lengua indígena (PHLI) disminuyó 10 puntos porcentuales de 1930 al 2010, pasando de 16.2% de la población total a 6.2 por ciento. Durante este periodo se observaron oscilaciones en cuanto a los principales estados con mayor PHLI, siendo Oaxaca el estado que de manera consistente ocupó alguno de los tres primeros lugares, seguido de Puebla y Veracruz; Yucatán y Chiapas ocuparon de forma esporádica alguno de los primeros cinco lugares (Marino, 1963; Oliveira *et al*., 1982; Ortíz, 2005). El estado que presentó el mayor decremento es Oaxaca (-7.8%); por el contrario, Chiapas tuvo un incremento (12.5%).

El tzeltal y tzotzil son las únicas lenguas indígenas que se hablaban más en el 2010 que en 1930. En la gráfica 10 sólo se pueden apreciar las principales, sin embargo, en los microdatos de 1930 se tienen catalogadas 35.

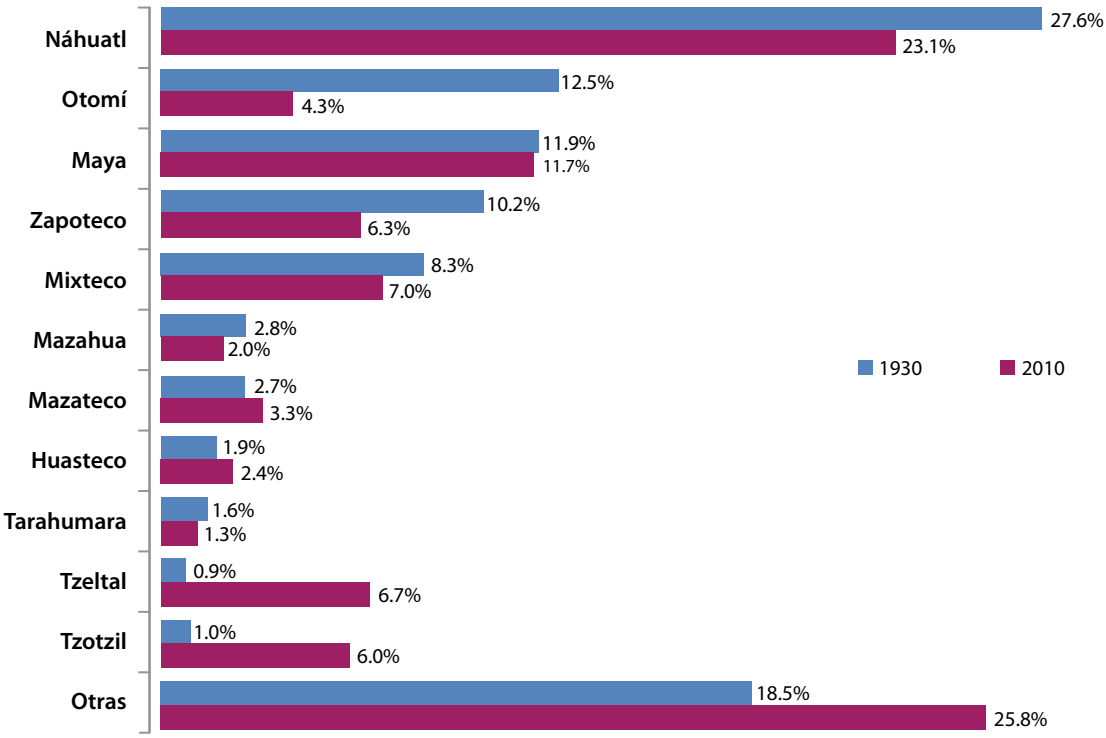
Es notorio que las lenguas indígenas están en riesgo de desaparecer; el Instituto Nacional de Lenguas Indígenas (INALI) (2015) indica que uno de los principales motivos es la estigmatización de su uso, tanto por el hablante como por el resto de la sociedad dominante.

Religión

La católica ha predominado entre la población desde la época de la Colonia y, aunque las *Leyes de Reforma* (1859-1860) y *la Ley Calles* (1926) le quitaron poder a la Iglesia y propiciaron la introducción de nuevas doctrinas, la gente continuó siendo en su mayoría católica (García, 2004). Prueba de ello es que en 1930 se estimó que 97.80% de la población practicaba esta religión, 1.07% no profesaba ninguna y el restante 1.13% pertenecía a otra (en este rubro, 0.58% se declaró protestante). En cuanto al 2010, el INEGI reportó que 82.72% de la población era católica; 4.68% no pertenecía a religión alguna y 12.60% declaró pertenecer a otra doctrina, entre las más representativas: protestantes (7.47%) y bíblicas diferentes a las evangélicas (2.26%).

Gráfica 10

Población de 5 años de edad y más hablante de lengua indígena



Fuente: microdatos expandidos de los censos de 1930 y 2010.

Ocupación

En cuanto a la población mayor de 12 años de edad de 1930, 46.2% estaba ocupada en alguna actividad económica, cifra menor a la reportada en el 2010, donde el indicador se situó en 56.2 por ciento. Un hecho importante fue la incorporación de las mujeres a las actividades productivas, ya que este indicador en el sector femenino se desplazó de 7.2% en 1930 a 33.3% en el 2010.

La gráfica 11 se realizó con las ocupaciones y profesiones que los pobladores declararon realizar al momento de cada censo[9]. Como se puede observar, 71.7% de la población económicamente activa (PEA) desarrollaba su actividad en el sector 1 —compuesto sobre todo por trabajadores del campo (88.9%) y ganaderos (6.2%)—; en el 2, 10.6% —en el cual se encontraban diversas profesiones, de las que destacaron, por su participación en el total, molendero (11.68%), obrero (10.3%), carpintero (9.7%), costurero (7.7%) y albañil (7.15%), entre otras—; y, por último, en el sector 3 —cuyas ocupaciones más frecuentes eran comerciante (29.2%), empleado (24.0%), mecánico (4.2%) y profesor (4.1%)— participaba 13.6% de la población. Éstas son sólo las actividades con mayores frecuencias, sin embargo, el catálogo obtenido de los microdatos se compone de 1 073 ocupaciones y profesiones.

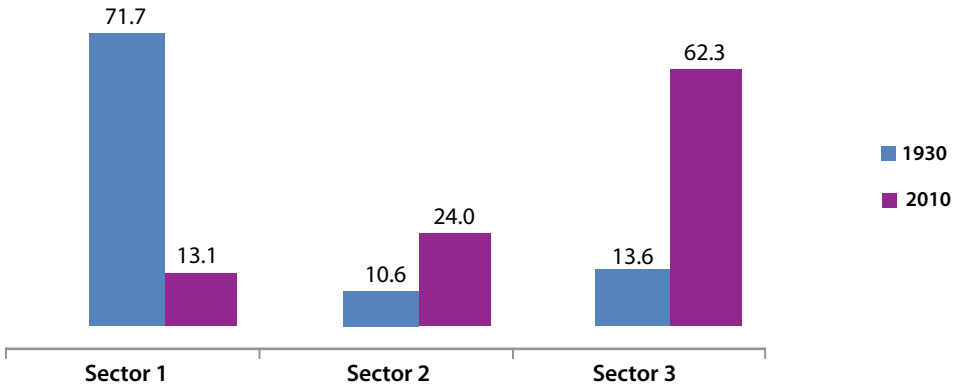
Por otro lado, es evidente que las actividades económicas han cambiado mucho en los últimos 80 años. En el 2010, el sector que tenía más importancia es el terciario, que aportó 62.3% al producto interno bruto (PIB) nacional, y los principales subsectores, la manufactura (23%), el comercio (19%) y actividades relacionadas con el subsector inmobiliario (13%). En cuanto al secundario, contribuyó con 24% al PIB en subsectores como: minería

9 La clasificación fue hecha por el equipo de investigación, basada en el catálogo del Sistema Nacional de Clasificación de Ocupación (SINCO) 2011 del INEGI.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 11

Porcentaje de la población ocupada por sector de actividad económica



Fuentes: microdatos expandidos del censo de 1930. // INEGI. *Encuesta Nacional de Ocupación y Empleo (ENOE). 2010.*

(47%), construcción (44%) y electricidad (9%). Por último, el sector primario sólo participó con 13.1% (INEGI, 2010d). Alanis (2008) considera que el resultado de una economía en crecimiento es dejar de basar primordialmente su producción en actividades agrícolas y de extracción, pasar a otras industriales y, por último, tener un sector de servicios más importante.

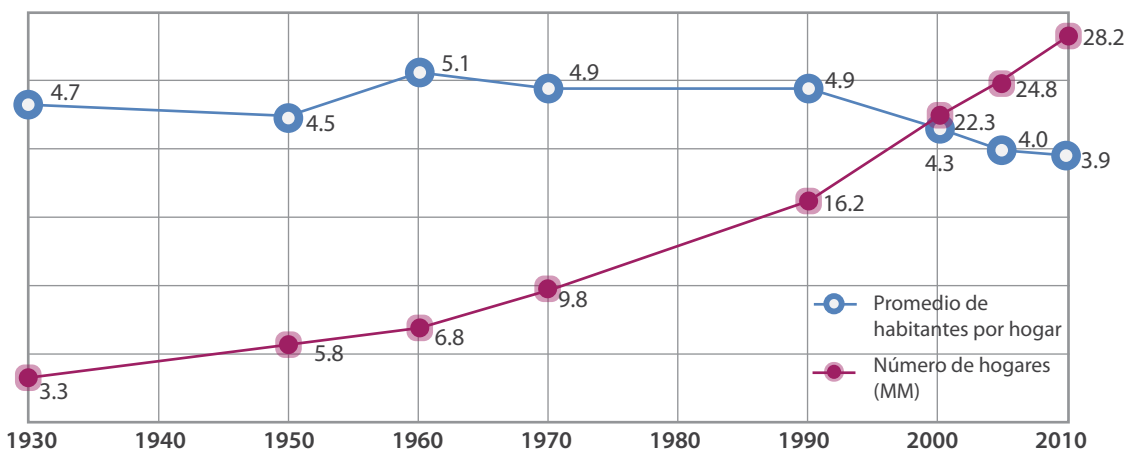
Hogares

Expertos en el tema coinciden en reconocer la centralidad de la familia en ámbitos como el control social, el funcionamiento de los sistemas de herencia y trasmisión de la propiedad, la reproducción demográfica, la socialización de los individuos, las relaciones de género y la solidaridad entre generaciones (Tuirán, 1993).

Como se menciona en INEGI (2001), en un comparativo de 1930 al 2000, el tamaño promedio del hogar permaneció hasta cierto punto estable durante las primeras décadas, pero en años posteriores existió una tendencia a la baja, la cual puede relacionarse con el descenso de la fecundidad y el incremento de la migración.

Gráfica 12

Número de hogares y promedio de integrantes por hogar



Fuente: censos de 1930 y 2010.

Los microdatos de 1930 indican que el número promedio de integrantes en la familia fue de 4.7 (en 3.3 millones de hogares registrados), el cual disminuyó a 3.9 miembros en el 2010 (en 28.2 millones de hogares), lo que, en número de hogares, representa un incremento de 749% (ver gráfica 12). Finalmente, en 1930, 18.4% de los hogares reconocía como jefa de familia a una mujer y, para el 2010, esta cifra aumentó a 24.5 por ciento.

Conclusiones y recomendaciones

En cuanto a la base de datos

La recuperación de los microdatos del Quinto Censo de Población, 15 de mayo de 1930, fue una acertada decisión por parte del INEGI. La muestra de 10% de ellos indica que este censo cumple con los estándares de calidad necesarios para hacer análisis estadísticos. Esto se confirma con las estimaciones realizadas con los microdatos presentados a lo largo del presente documento, los cuales reflejan resultados muy similares a los mostrados en los tabulados básicos de 1932.

La información recuperada será un valioso insumo para investigadores en Ciencias Sociales, pues podrán realizar estimaciones precisas por municipio; aunque puedan presentar errores, estarán en niveles aceptables.

Ya que se ha desarrollado una metodología y una plataforma para la recuperación de estos datos, se recomienda rescatar el otro 90%, lo que permitirá conocer características para cohortes poblacionales más específicas.

En cuanto a los resultados

La población en México se ha septuplicado en los últimos 80 años y, de ser una población muy joven, se ha convertido en una donde alrededor de 50% de sus habitantes está en edad laboral, lo cual representa un bono demográfico que terminará en la tercera década de este siglo.

Las políticas públicas ejecutadas en el país durante las últimas ocho décadas han abordado varios aspectos demográficos, como: la densidad de población, el aumento de la esperanza de vida, la disminución de la mortalidad y el analfabetismo, el control de la natalidad y la dependencia económica; sin embargo, no se han diseñado políticas públicas eficaces para resolver aspectos relevantes como la concentración de 50% de la población en la cuarta parte de los estados del país, el bajo índice de masculinidad y la pérdida de la población hablante de lengua indígena.

La estructura de los hogares ha cambiado, en esto ha influido la política de fecundidad, la cancelación de algunos prejuicios sociales y la inserción de las mujeres a la actividad productiva. Ahora, se encuentran familias más pequeñas, se reconoce más a una mujer como jefa de familia, el número de parejas casadas ha disminuido y ha aumentado el número de personas divorciadas; no obstante, deben observarse las implicaciones de estos cambios, ya que se ha reconocido a la familia como la unidad básica de la sociedad y como la promotora de valores en los individuos, por lo que el Estado debería indagar acerca de los efectos a mediano y largo plazo de esta reestructuración social y encauzar este proceso.

La pérdida importante en la riqueza de lenguas indígenas señala un desarrollo nacional no incluyente y unilateral, una inercia hacia la monocultura occidentalista. El estado de Chiapas se presenta hoy en día como el único que ha conservado su diversidad lingüística por el aumento del número de hablantes de lenguas indígenas, principalmente tzeltal y tzotzil. Sería recomendable analizar lo que ocasionó este fenómeno, en aras de cuidar la riqueza de lenguas en nuestro país.

Los resultados en cuanto al índice de masculinidad son de llamar mucho la atención, es necesario que se analicen y corrijan las causas de que se encuentre en un nivel tan bajo, lo que significa una situación crítica para los hombres, tanto como en la época de la Revolución Mexicana. El que ahora haya menos hombres que antes indica

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

que sus condiciones de vida son más precarias y su esperanza de vida permanece a la zaga de la de las mujeres. Esto tiene implicaciones graves, debido a que el IM está deteriorado sobre todo en los grupos de mayor actividad económica y densidad poblacional. Por lo anterior, es necesario diseñar políticas públicas de salud para aumentar la esperanza de vida de los hombres; también, se debe atender el problema de migración que, por décadas, ha mermado a la población masculina del país y, para finalizar, abordar el creciente problema de violencia que empieza a incidir en las estadísticas nacionales.

El aspecto de la educación es muy importante; aunque es notorio un cambio muy favorable en el indicador de alfabetismo, es necesario investigar las causas por las que entidades como Oaxaca, Guerrero y Chiapas no han podido recuperarse de situaciones tan precarias como otras entidades que sí lograron alcanzar mejores niveles de instrucción. Por otro lado, resulta primordial relacionar la educación con las actividades productivas que la población realiza. En la actualidad, aunque la población dejó de dedicarse a actividades del sector primario y se ha movido al de servicios, se nota que los puestos que ahora ocupan son de bajo perfil, pues la PEA se está dedicando a actividades como el comercio y la manufactura. Esta circunstancia podría ser crítica, ya que el país atraviesa por una situación demográfica favorable, la cual se está desaprovechando. Debido a esto, es necesario impulsar políticas públicas que fortalezcan la instrucción de alto nivel en la población joven, para que ésta se inserte en la dinámica de los cambios tecnológicos e informáticos que están modificando la economía mundial y que, de manera simultánea, mejore sus condiciones de vida.

Fuentes

Alanis, G. "Cambio estructural en México: crecimiento económico, apertura comercial, el TLCAN y el medio ambiente", en: *Conferencia del Cuarto Simposio de América del Norte sobre Evaluación de los Efectos Ambientales del Comercio*. Phoenix, EE.UU., 2008. Consultado en http://www3.cec.org/islandora/en/item/2338-structural-changes-in-mexico-economic-growth-trade-liberalization-nafta-and-es.pdf el 10 de abril de 2015.

Alba, F. *La población de México*. México, DF, Centro de Estudios Económicos y Demográficos. El Colegio de México, 1976, p. 14.

Asuad, N. "La ciudad de México, su región y la construcción y operación del nuevo aeropuerto de la ciudad de México (*sic*) (NAICM)", en: *XXIV Seminario de Economía Urbana y Regional*. IIEc, Facultad de Economía de la UNAM, 2014. Consultado en http://www.economia.unam.mx/cedrus/descargas/Ponencia%20Asuad.pdf el 11 de marzo de 2015.

Arrom, S. *The Women of Mexico City 1790-1857*. California, EE.UU., Stanford University Press, 1985, pp. 127-129.

Bonilla, I. "Cronología de la educación y campañas de alfabetización en México", en: *Mirada Ferroviaria* (revista digital). Núm. 15, 2011, pp. 43-52.

Chackiel, J. y G. Macció. *Evaluación y corrección de datos demográficos*. Serie B. Núm. 39. Santiago de Chile, CELADE-CEPAL, 1978, p. 4.

Cochran, W. G. *Sampling Techniques*. 3.ª edición. EE.UU., John Wiley & Sons, Inc., 1977.

CONAPO. *Envejecimiento de la población de México: reto del siglo XXI*. México, DF, CONAPO, 2004, p. 22.

_____ *República Mexicana: indicadores demográficos*. Consultado en http://www.conapo.gob.mx/es/CONAPO/Proyecciones_Datos el 10 de abril de 2015.

Cordero, E. "La subestimación de la mortalidad infantil en México", en: *Demografía y Economía*. El Colegio de México, 1968. Consultado en http://codex.colmex.mx:8991/exlibris/aleph/a18_1/apache_media/KKA9YM8XK23G1B6Y16GKGIXIIKRDE9.pdf el 10 de abril de 2015.

Secretaría de la Economía Nacional. Dirección General de Estadística. *Quinto Censo de Población. 15 de mayo de 1930. Resumen general*. México, Secretaría de la Economía Nacional, 1932, p. 3.

_____ *Quinto Censo de Población. 15 de mayo de 1930. Tabulados básicos*. México, Secretaría de la Economía Nacional, 1932.

García, M. *Credibilidad y opinión pública entre estudiantes de Ciencias de la Comunicación y Derecho: caso la Iglesia católica*. Tesis de licenciatura. Universidad de las Américas Puebla, 2004.

Garduño, M. A. "Determinación genérica de la mortalidad masculina", en: *Salud Problema*. 6(10-11), 2001.

González, G., J. Schiavon, D. Crow y G. Maldonado. *México, las américas y el mundo 2010. Política exterior: opinión pública y líderes*. Informe. México, CIDE, 2011, p. 50.

Gutiérrez, R. "Conformación del proceso migratorio al norte de México", en: *Estudios Demográficos y Urbanos*. El Colegio de México, 1995, pp. 569-605.

Gutiérrez, M. "Desarrollo y distribución de la población urbana en México", en: *Investigaciones geográficas. Boletín del Instituto de Geografía*. Núm. 50. UNAM, 2003.

INALI. *México, lenguas indígenas nacionales en riesgo de desaparición: variantes lingüísticas por grado de riesgo*. 2000. México, DF, Instituto Nacional de Lenguas Indígenas, 2015. p. 3.

INE. *Población y sociedad: aspectos demográficos*. Santiago, Chile, Instituto Nacional de Estadística, 2008, p. 31.

_____ *Indicadores demográficos básicos: metodología*. Madrid, España, Instituto Nacional de Estadística, 2014, p. 69.

INEGI. *Indicadores sociodemográficos de México (1930-2000)*. Aguascalientes, México, INEGI, 2001, pp. 101-102.

_____ *125 años de la Dirección General de Estadística 1882-2007*. Aguascalientes, México, INEGI, 2010a, pp. 27-68.

_____ *Estadísticas vitales 2010.* Consultado en http://www.inegi.org.mx/est/contenidos/proyectos/registros/vitales/consulta.asp?c=11800# el 2 de junio de 2015. 2010c.

_____ *Tabulados básicos Censo 2010*. Aguascalientes, México, INEGI, 2010b.

_____ *Sistema de Cuentas Nacionales de México. Producto interno bruto por entidad federativa 2005-2009: año base 2003*. Aguascalientes, México, INEGI, 2010d.

_____ *Diseño de la muestra censal* 2010. Aguascalientes, México, INEGI, 2011, pp. 6-13.

_____ *Las personas con discapacidad en México: una visión al 2010*. Aguascalientes, México, INEGI, 2013, pp.116 y 228.

_____ *Censo de Población y Vivienda 2010*. Consultado en http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2010/Default.aspx el 1 de abril de 2015.

Márquez y Molina. "El otoño de 1918: las repercusiones de la pandemia de gripe en la ciudad de México", en: *Desacatos*. Núm. 32, 2010, pp. 121-144.

McCaa, R. *La viuda viva del México borbónico: sus voces, variedades, y vejaciones*. El Colegio de México, 1991, pp. 299-324.

_____ *Missing Millions: The Demographic Costs of the Mexican Revolution*. Vol. 19, núm. 2, 2003, pp. 367-400.

McCaa, R.; A. Esteve; S. Ruggles y M. Sobek. "La integración de los microdatos censales de América Latina: el proyecto *ipums*-América Latina", en: *Estudios Demográficos y Urbanos*. Vol. 20, núm. 1, 2005.

McCaa, R. y A. Gomez-Galvarriato. "1930 Population census of Mexico: a 1% Pilot Microdata Sample", en: *Population Association of American Annual Meeting 2011*. (Cartel) Washington, DC, 2013.

Muñoz, C. "Educación, Estado y sociedad en México 1930-1976". Ponencia presentada en el I Foro Latinoamericano de Educación Comparada. 19-22 de marzo de 1980, Colima, Col.

Naciones Unidas, CEPAL, UNFPA. "Los datos demográficos: Alcances, limitaciones y métodos de evaluación", en: *Serie Manuales*. Santiago de Chile, 2014, p. 49.

Ordorica, M. y J. L. Lezama. "Consecuencias demográficas de la Revolución Mexicana", en: *El Poblamiento de México*. Núm. 4, 1993, pp. 32-53.

Organización Internacional para las Migraciones (OIM). "Hechos y cifras", en: *Organización Internacional para las Migraciones*. 2014. Consultado en http://oim.org.mx/hechos-y-cifras-2 el 2 de junio de 2015.

Ortíz, M. *La población hablante de lenguas indígenas en México*. México, DF, Dirección General de Publicaciones y Fomento Editorial, 2005, p. 17.

Kish, L. *Survey Sampling*. EE.UU., John Wiley & Sons, Inc., 1965.

Lohr, S. *Sampling: Design and Analysis*. EE.UU., Duxbury Press, 1999.

Riba, C. y A. Cuxart. "Aspectos metodológicos de la encuesta social europea". *VIII Congreso Español de Sociología*. Alicante, España, 2004, p. 18.

Romo, R., Y. Téllez y J. López. "Tendencias de la migración interna en México en el periodo reciente", en: *La situación demográfica de México*. México, CONAPO, 2012.

Salazar Anaya, D. "Imágenes de la presencia extranjera en México: una aproximación cuantitativa 1894-1950", en: *Dimensión Antropológica*. Vol. 3, enero-abril de 1996. México, INAH. Consultado en http://www.dimensionantropologica.inah.gob.mx/?p=1473 el 2 abril de 2015.

Scheaffer, R., W. Mendenhall y L. Ott. *Elementos de muestreo*. EE.UU., Grupo Editorial Iberoamérica, 1987, pp. 78.

Sobrino, J. *Migración interna en México durante el siglo XX*. México, CONAPO, 2010.

Tuirán, R. "Vivir en familia: hogares y estructura familiar en México, 1976-1987", en: *Comercio Exterior*. 1993. Consultado en http://revistas.bancomext.gob.mx/rce/magazines/248/8/RCE8.pdf el 12 de abril de 2015.

UNFPA, UNISDR y ONU-HABITAT. *Vínculos entre las dinámicas demográficas, los procesos de urbanización y los riesgos de desastres. Una visión regional de América Latina*. 2012.

Velázquez, C. *Diferencias políticas entre los inmigrantes chinos del noroeste de México (1920-1930). El caso de Francisco L. Yuen. Historia mexicana*. Vol. LV, núm. 2. El Colegio de México, 2005, pp. 461-512.

Verduzco, G. "La migración mexicana a Estados Unidos: estructuración de una selectividad histórica", en: *Migración México-Estados Unidos: continuidad y cambio*. México, DF, CONAPO, 1997, pp. 11-32.

Weeks, J. *Sociología de la población*. EE.UU., Alianza Universidad Textos, 1999, p. 98.

Zamudio, F.; R. Arana; A. Corona; R. Bautista; M. Andrade; J. Santibañez y M. Jiménez. *Informe estadístico sobre desarrollo humano en México: 1995-2010. Resumen*. México, Universidad Autónoma Chapingo, 2012.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Estimación del ingreso por trabajo **en los municipios y las delegaciones de México** utilizando técnicas de estimación para áreas pequeñas

Miguel Ángel Suárez Campos, Gustavo Aguilar Mata y Raúl Mejía González

La información y el conocimiento son factores esenciales para la toma de decisiones; sin embargo, para una mayor efectividad, se requiere de información más desagregada que la disponible en la actualidad. Esta necesidad de mayor detalle ha sido motivo de diferentes estudios que buscan técnicas estadísticas que satisfagan las expectativas de un importante número de usuarios. Una de ellas es la estimación para áreas pequeñas, que utiliza modelos lineales mixtos. En este trabajo se presenta un ejercicio que compara los resultados de la estimación directa con los obtenidos mediante esta técnica, referente al ingreso promedio mensual por trabajo en la vivienda para todos los municipios y delegaciones de México, a partir de los datos recabados por la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2010. Además, a manera de validación, se presenta un comparativo entre los resultados obtenidos con dicha técnica y la estimación proporcionada por la muestra del Censo de Población y Vivienda (CPV) 2010.

Palabras clave: modelos mixtos, efectos aleatorios, muestras complejas, modelo a nivel de área.

Recibido: 28 de abril de 2014.
Aceptado: 14 de mayo de 2015.

Even though Information and knowledge are essential factors for decision-making, a higher effectiveness in results would require more disaggregated information than the one available. This need for more detailed information has contributed to the search for statistical techniques that satisfy the expectations of a substantial number of users. One of these techniques is the Small Area Estimation (SAE), where mixed linear models are used. This paper presents an exercise comparing the results of direct estimation with those gathered using the SAE technique. Such results are in reference to the average monthly income in relation to work at household for all municipalities and delegations of Mexico. The data for this exercise was collected by the Income and Expenditure National Household Survey 2010. In addition and for validation purposes, this paper also introduces a comparison between the results obtained with SAE and the estimation provided by the complementary survey conducted with the Population and Housing Census in 2010.

Key words: mixed models, random effects, complex sampling, area level model.

Plaza-constitucion-veracruz-port/abalcazar/Stock photo ©

Introducción

En el muestreo de poblaciones finitas —y en particular en el sistema estadístico nacional— existe una demanda creciente de estimaciones precisas sobre promedios o indicadores de interés en áreas cada vez más pequeñas (entidades federativas, municipios o localidades, o bien, en subgrupos o pequeños dominios de la población total, como las subclases de alguna actividad económica). Estos cálculos se consideran un subproducto de los trabajos de muestreo en áreas grandes, donde se diseña una muestra para obtener valores del total o de la media de una característica de interés con una precisión prefijada. Esto implica que el número de observaciones muestrales en un área pequeña es reducido o, incluso, nulo, por lo que utilizar estimadores basados en un diseño muestral conduce a grandes errores en los resultados y, de hecho, es imposible obtenerlos en áreas no muestreadas. Para resolver esta problemática, es necesario, por un lado, aumentar el tamaño de muestra (lo cual implica elevar los costos) y, por el otro, aplicar la técnica estadística estimación para áreas pequeñas (EAP).

Por tradición, los procedimientos de inferencia sobre características de poblaciones finitas se han basado en el diseño de la muestra; sin embargo, en la actualidad son cada vez más numerosos los de aproximación basada en modelos (Cassel *et al*., 1977), en la cual se le da más peso al modelo de la distribución de la característica en estudio en la población que al proceso de diseño de la muestra y se supone la aceptación de algunos riesgos (Hansen y Madow, 1983), como el no estar seguros de que el modelo que se adopta es el correcto; aun así, deben ser aceptados, ya

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

que sólo es posible considerar la inferencia basada en el diseño libre de supuestos cuando la muestra es grande, lo cual no ocurre con la estimación para áreas pequeñas. Al recurrir a este tipo de inferencia se pueden solventar problemas no abordables por la vía del diseño, como la no respuesta o la estimación en áreas pequeñas (Särndal, 1984).

Harter (1993) recopiló diferentes métodos para la EAP para obtener valoraciones de la población en ciudades y municipios en periodos intercensales; en fecha más reciente, Rao (2003) hizo una compilación de aquéllos y otros métodos en su libro *Small Area Estimation*.

La EAP relaciona mediante un modelo a la variable de interés obtenida de una encuesta con la información emanada tanto de eventos censales como de registros administrativos y geográficos, contenida en las denominadas *variables auxiliares*, de las cuales se conoce, en unos casos, el valor para cada elemento de la población y en otros, sólo la información agregada (promedios, totales o proporciones) de cada dominio o área pequeña, de donde se tienen dos grandes tipos de modelos de EAP: los de nivel de unidad y los de nivel de área. Para los primeros, es necesario asociar la información auxiliar contenida en los elementos de la población (censos o registros administrativos) y su unidad muestral correspondiente, lo cual no siempre es posible, mientras que los segundos nada más necesitan información agregada de cada dominio o área pequeña como regresores del modelo y los estimadores directos obtenidos del muestreo como variables de respuesta.

En este trabajo se optó por utilizar la EAP basada en modelos de área con el propósito de obtener una estimación del promedio por municipio[1] de la variable *ingreso por trabajo en la vivienda*, la cual se eligió por dos razones:

- Aprovechar que se obtuvo de manera independiente en el mismo año por dos medios distintos: uno por la ENIGH 2010 y el otro por la muestra del CPV 2010; de esa manera se pueden comparar las estimaciones que logran obtenerse por distintos métodos a partir de la Encuesta con las de la muestra del Censo que, para algunos municipios, no es un resultado muestral sino censal.
- La variable seleccionada forma parte sustancial en la construcción de la variable *ingreso corriente en la vivienda*, la cual es muy importante para el cálculo de indicadores económicos estratégicos, como los referidos a la pobreza, el producto interno bruto (PIB) y la distribución del ingreso; así, la obtención futura de estimaciones de esta variable podría ser más sencilla con base en los resultados de este trabajo.

Para conseguir ese propósito, este documento se organiza de la siguiente manera: primero se muestra brevemente la ENIGH 2010, centrándose en su diseño de muestreo, los estimadores que utiliza y la variable *ingreso por trabajo;* se continúa con la presentación de la muestra del CPV 2010, haciendo hincapié en su diseño de muestreo y la misma variable; después, se comenta de forma concisa la estimación basada en el modelo y sus ventajas cuando hay escasez de muestra; se expone, además, un resumen de la base teórica del modelo a nivel de área, que es el que aquí se utiliza para realizar las estimaciones en las áreas pequeñas; se continúa con una explicación sobre la construcción del modelo, la selección de las variables que mejor se ajustan al modelo planteado (el ajuste se basa en el cumplimiento de pruebas estadísticas al modelo, mismas que se muestran en su formulación matemática); se prosigue con la obtención del promedio municipal del ingreso mensual por trabajo en la vivienda, obteniéndolo primero con la formulación de la estimación directa y, después, aplicando el modelo a nivel área, para hacer una comparación gráfica de los resultados; a manera de validación, se cotejan también gráficamente las EAP con las de la muestra del CPV 2010; por último, se exponen las conclusiones logradas con los resultados y el método de estimación aplicado.

1 Para efectos de exposición de este documento, al utilizar las palabras municipio o municipios se referirá también a la(s) delegación(es) del Distrito Federal.

ENIGH 2010

Tiene sus antecedentes en varios sondeos realizados por diferentes dependencias públicas, como la Secretaría de Industria y Comercio (SIC), el Banco de México, la Secretaría del Trabajo y Previsión Social (STPS) o la Secretaría de Programación y Presupuesto (SPP), pero fue a partir de 1984 que se integró como tal y ha sido levantada de manera formal por el Instituto Nacional de Estadística y Geografía (INEGI). Desde 1992 se ha realizado con una periodicidad bienal, con excepción del 2005, ya que fue un levantamiento ex profeso.

La ENIGH 2010 se levantó del 21 de agosto al 28 de noviembre con el objetivo de proporcionar información sobre la distribución, monto y estructura del ingreso y gasto de los hogares, así como de ofrecer datos tanto de las características sociodemográficas y ocupacionales de los integrantes del hogar como de la infraestructura de la vivienda y el equipamiento del hogar.

Los resultados de este levantamiento se encuentran a nivel nacional y para los ámbitos rural y urbano. Además, se tiene información para las entidades que, en su momento, convinieron con el INEGI una ampliación de la muestra para este operativo (Chiapas, Guanajuato, Distrito Federal, estado de México y Yucatán).

El marco de muestreo utilizado fue el de propósitos múltiples del INEGI, constituido con la información demográfica y cartográfica obtenida a partir del levantamiento del XII Censo General de Población y Vivienda 2000. El marco es probabilístico, estratificado,[2] unietápico y por conglomerados; estos últimos se consideran unidades primarias de muestreo (UPM),[3] pues es en ellos donde, en una segunda etapa, se seleccionan las viviendas que forman parte de la muestra (INEGI, c2011), por lo que la elección de la muestra para la ENIGH 2010 se realizó en dos etapas: en la primera se escogieron las UPM y en la segunda, las viviendas objeto de entrevista de la Encuesta.

Por lo tanto, el esquema de muestreo de la ENIGH 2010 se considera como complejo, siendo éste probabilístico, estratificado, bietápico y por conglomerados, donde la unidad es la vivienda.

El estimador directo que se utiliza para la media en un dominio S (entidad, ciudad, municipio, etc.) es un derivado del estimador de Horvitz-Thompson, que tiene la forma:

$$\widehat{\bar{Y}}_S^{DIR} = \sum_{i=1}^{n_S} \frac{y_i}{\pi_i} \Bigg/ \sum_{i=1}^{n_S} \pi_i \ , \qquad (1)$$

donde y_i es el valor i-ésimo de la variable de interés; n_s, el número de elementos en el dominio S (por simplicidad, las ecuaciones de esta sección se refieren al dominio S) y π_i la probabilidad de selección de y_i, dado por:

$$\pi_i = \frac{n_{Mh}^I n_{Mhj}^{II}}{N_h^I N_{Mhj}^I} \ ,$$

donde:

M = marco maestro del INEGI.

I = primera etapa de muestreo.

II = segunda etapa de muestreo.

n_{Mh}^I = número de UPM seleccionadas del marco M en el estrato h.

n_{Mhj}^{II} = número de viviendas seleccionadas en la UPM j del estrato h en el marco M; para la zona urbana es de cinco viviendas mientras que para la zona rural y el complemento urbano, 20.

N_h^I = número total de UPM en el estrato h.

2 La estratificación se realiza tomando en cuenta la división política del país, el ámbito urbano y rural, el tamaño de las localidades y las características sociodemográficas de los habitantes y de equipamiento de las viviendas (ENIGH 2010. Diseño muestral).

3 Son agrupaciones de 80 a 160 viviendas con características diferenciadas de acuerdo con el ámbito al que pertenecen (ENIGH 2010. Diseño muestral).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

N^{I}_{Mhj} = número de viviendas de la UPM j del marco M en el estrato h.

Al inverso de la probabilidad de selección $w_i = 1/\pi_i$ se le conoce como factor de expansión; para la ENIGH 2010, este factor tiene ajustes por no respuesta y por encuadre con proyecciones de población, siendo ésta la razón por la que no es un estimador de Horvitz-Thompson puro.

La varianza estimada directa del estimador de la media $\hat{\bar{Y}}_S^{DIR}$ se obtiene con la siguiente expresión:

$$\hat{V}\left(\hat{\bar{Y}}_S^{DIR}\right) = \frac{1}{N_S^2}\left[\left(N_h^I\right)^2\left(\frac{1}{n_{Mh}^I}-\frac{1}{N_h^I}\right)\frac{\sum_{j=1}^{n_{Mh}^I}\left(\hat{t}_{Mhj}^I-\hat{\bar{t}}_{Mh}^I\right)^2}{\left(n_{Mh}^I-1\right)}\right.$$

$$\left. + \frac{N_h^I}{n_{Mh}^I}\sum_{j=1}^{n_{Mh}^I}\left(N_{Mhj}^I\right)^2\left(\frac{1}{n_{Mhj}^{II}}-\frac{1}{N_{Mhj}^I}\right)\frac{\sum_{i=1}^{n_{Mhj}^{II}}\left(y_{Mhji}^{II}-\hat{\bar{y}}_j^{II}\right)}{\left(n_{Mhj}^{II}-1\right)}\right], \quad (2)$$

donde:

$\hat{t}^I_{Mhj}$ = suma total estimada de la variable en estudio de la UPM j seleccionada del marco M en el estrato h.

$\hat{\bar{t}}^I_{Mh}$ = promedio estimado de los totales de la variable en estudio de las UPM seleccionadas del marco M en el estrato h.

y^{II}_{Mhji} = valor de la variable en estudio de la vivienda i de la UPM j seleccionada del marco M en el estrato h.

$\hat{\bar{y}}^{II}_{Mhj}$ = promedio estimado del valor de la variable en estudio de la UPM j seleccionada del marco M en el estrato h.

La variable a estimar es la de *ingreso por trabajo*;[4] en la ENIGH 2010 se consideró que un integrante del hogar percibe *ingreso por trabajo* sólo si tiene o ha tenido participación directa en actividades reconocidas como económicas. Por sus fuentes,[5] los ingresos del trabajo pueden provenir de: las remuneraciones por el trabajo subordinado, los ingresos por el trabajo independiente y otros ingresos provenientes del trabajo. La referencia temporal de todos ellos es del mes inmediato anterior al levantamiento y de los cinco meses anteriores (INEGI, 2011c).

De los 2 456 municipios existentes al 2010, únicamente en 600 hubo muestra de la ENIGH 2010 con viviendas que reportaron *ingresos por trabajo*, con una o más UPM por municipio.

Muestra del CPV 2010

Se elaboró un cuestionario ampliado para el levantamiento de la muestra censal[6] 2010, que fue aplicado del 31 de mayo al 25 de junio de ese año en alrededor de 2.9 millones de viviendas en el país. Dados los requerimientos de información, se decidió incluir con certeza en la muestra las viviendas habitadas de los municipios con menos de 1 100 de éstas, así como las de los 125 con menor índice de desarrollo humano (IDH), sin importar su tamaño. Bajo este criterio, en 766 municipios se censaron la totalidad de viviendas con el cuestionario ampliado; para los 794 que tenían entre 1 101 y 4 mil viviendas habitadas el tamaño de la muestra fue de 800; en los municipios sin localidades de 50 mil y más habitantes, la muestra fue de 1 100 viviendas y para los 193 restantes el tamaño fue variable, pero mayor a 2 mil. Los tamaños de muestra fijados garantizan estimaciones municipales aceptables para proporciones cercanas a 0.01 o mayores.

La variable estimada a nivel municipal es la de *ingresos por trabajo,* definida para la muestra del CPV 2010 como la percepción monetaria que la población ocupada obtiene o recibe del(los) trabajo(s) que desempeñó en la semana de referencia. Se consideran los ingresos por concepto de ganancia, comisión, sueldo,

4 En el diseño conceptual de la ENIGH 2010 se denomina a la variable como ingreso del trabajo y el marco conceptual del CPV 2010 la nombra ingreso por trabajo; en este artículo se homologa con este último.

5 El detalle de estas fuentes de ingreso se puede ver en la nueva construcción de la ENIGH 2010.

6 Su diseño es estratificado por conglomerados.

salario, jornal, propina o cualquier otro devengado de su participación en alguna actividad económica. Los ingresos están calculados de forma mensual (INEGI, 2011c).

Estimación basada en el modelo

Permite relacionar, mediante el uso de información auxiliar, a las áreas con escasez o inexistencia de muestra con aquéllas próximas de mayor información muestral para, de esta forma, incrementar la precisión del estimador; además, el uso de la estimación basada en modelos tiene las siguientes ventajas:

- El diagnóstico del modelo puede usarse para encontrar el adecuado que mejor ajuste a los datos. Incluye análisis de residuales cuyo resultado verificaría la validez del modelo supuesto, la selección de variables auxiliares para éste, la detección y, si es necesario, la supresión de observaciones influyentes.
- Estos métodos pueden utilizarse en casos complejos tanto en casos longitudinales como transversales.
- Pueden emplearse las metodologías desarrolladas en fechas recientes para modelos de efectos aleatorios con el fin de lograr inferencias precisas en el área pequeña.

Para la estimación basada en modelos en el caso de las áreas pequeñas se tienen, como ya se dijo antes, dos grandes tipos de modelos según la disponibilidad de la información auxiliar: a nivel de área y de unidad. En este trabajo se utilizó sólo el primero para una variable cuantitativa, razón por la que este modelo se presenta más adelante con cierto nivel de detalle.

Una premisa importante en estos modelos es que el comportamiento de la variable bajo estudio en las áreas pequeñas presenta ligeras diferencias entre sí y respecto al comportamiento de la misma en el área mayor que las contiene; esto se refleja considerando efectos aleatorios de área en el modelo lineal que se plantea, con lo que se introduce a los modelos lineales mixtos, llamados así porque abarcan en su expresión efectos fijos y aleatorios.

Resumen teórico del modelo a nivel de área

Se basa en el modelo de Fay-Herriot (1979), donde se tienen p variables como información auxiliar (promedios poblacionales por área) en el vector $\mathrm{x}_a = \left(\bar{X}_1, \ldots, \bar{X}_p\right)$ y se supone que están relacionadas con la media en la subpoblación de la variable de interés $\bar{Y}_a = Y_a / N_a$, o una cierta función de ésta, a través del modelo lineal con efectos aleatorios:

$$\theta_a := g\left(\bar{Y}_a\right) = \mathbf{x}_a^T \beta + \upsilon_a, \quad a = 1, \ldots, m, \qquad (3)$$

donde β es el vector de parámetros de regresión y υ_a, el efecto aleatorio que se supone independiente e idénticamente distribuido (iid) con media 0 y varianza σ_υ^2 por lo general, se supone la normalidad de υ_a. En caso de que no todas las áreas sean seleccionadas en la muestra, se continúa bajo el supuesto de que las muestreadas (m) obedecen al modelo de población.

Por otro lado, sea $\hat{\bar{Y}}_a^{DIR}$ el estimador directo de la media de la variable $\bar{Y}$ en el área pequeña a-ésima, esto supone que el tamaño de la muestra en el área (n_a) es mayor o igual a 1.

Ahora, se tiene que:

$$\hat{\theta}_a^{DIR} = g\left(\hat{\bar{Y}}_a^{DIR}\right) = \theta_a + e_a, \qquad (4)$$

donde e_a son los errores muestrales que son independientes con media 0 y la varianza ψ_a se supone conocida; se puede relajar este supuesto reemplazando ψ_a por estimadores suavizados $\hat{\psi}_a$ basados en las varianzas calculadas de los datos a nivel unidad (Rao, 2003).

Sustituyendo (3) en (4) se obtiene el modelo lineal mixto:

$$\hat{\theta}_a^{DIR} = \mathbf{x}_a^T \beta + \upsilon_a + e_a. \qquad (5)$$

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Se observa que este modelo involucra tanto errores del diseño muestral e_a como los del modelo υ_a y se supone que e_a y υ_a son independientes.

Bajo el modelo (5), el mejor estimador lineal insesgado (BLUP, por sus siglas en inglés) $\tilde{\theta}_a^{BLUP}$ de θ_a, el cual minimiza el error cuadrático medio (MSE, por sus siglas en inglés) $MSE\left(\tilde{\theta}_a\right)=E\left(\tilde{\theta}_a-\theta_a\right)^2$ es (Rao & Molina, 2012):

$$\tilde{\theta}_a^{BLUP}=\mathbf{x}_a^T\tilde{\beta}+\tilde{\upsilon}_a,$$

donde:

$$\tilde{\beta}=\beta\left(\sigma_\upsilon^2\right)=\left(\sum_{a=1}^{m}\gamma_a\mathbf{x}_a\mathbf{x}_a^T\right)^{-1}\sum_{a=1}^{m}\gamma_a\mathbf{X}_a\hat{\theta}_a^{DIR},$$

$$\tilde{\upsilon}_a=\gamma_a\left(\hat{\theta}_a^{DIR}-\mathbf{x}_a^T\tilde{\beta}\right),\quad \gamma_a=\sigma_\upsilon^2\left(\sigma_\upsilon^2-\psi_a\right)^{-1}.$$

El estimador $\tilde{\theta}_a^{BLUP}$ puede expresarse como:

$$\tilde{\theta}_a^{BLUP}=\left(1-\gamma_a\right)\mathbf{x}_a^T\tilde{\beta}+\gamma_a\hat{\theta}_a^{DIR}. \qquad (6)$$

Se observa que el estimador BLUP de θ_a se expresa como un promedio ponderado del estimador directo $\hat{\theta}_a^{DIR}$ y el estimador de regresión sintética $\mathbf{x}_a^T\tilde{\beta}$, donde el ponderador γ_a $(0\le\gamma_a\le1)$ mide la incertidumbre en el modelizado del predictor para cada área pequeña *a* (Azula *et al*., 2004).

Como $\tilde{\theta}_a^{BLUP}$ depende de σ_υ^2 a través de $\tilde{\beta}$ y γ_a se puede utilizar el BLUP empírico EBLUP, reemplazando σ_υ^2 por su estimador $\hat{\sigma}_\upsilon^2$:

$$\hat{\theta}_a^{EBLUP}=\tilde{\theta}_a^{BLUP}\left(\hat{\sigma}_\upsilon^2\right).$$

En lo referente a la estimación de θ para las k áreas no muestreadas sólo se aplica el estimador de regresión sintética:

$$\hat{\theta}_k^{SYN}=\mathbf{x}_k^T\hat{\beta}\ \text{ donde }\ \hat{\beta}=\tilde{\beta}\left(\hat{\sigma}_\upsilon^2\right). \qquad (7)$$

Una versión más general de (6) que no se limita a la linealidad de θ_a es el mejor estimador (*B*) o también llamado estimador de Bayes, el que bajo normalidad se expresa como:

$$\tilde{\theta}_a^B\left(\beta,\sigma_\upsilon^2\right)=E\left(\theta_a\middle|\hat{\theta}_a^{DIR}\right)=\left(1-\gamma_a\right)\mathbf{x}_a^T\beta+\gamma_a\hat{\theta}_a^{DIR},$$

es insesgado ya que:

$$E_{\hat{\theta}_a^{dir}}\left(\tilde{\theta}_a^B\right)=E_{\hat{\theta}_a^{dir}}E_{\theta_a|\hat{\theta}_a^{dir}}\left(\theta_a\right)=E\left(\theta_a\right).$$

También, el mejor estimador empírico bajo normalidad de θ_a coincide con el EBLUP, esto es:

$$\hat{\theta}_a^{EB}=\tilde{\theta}_a^B\left(\hat{\beta},\hat{\sigma}_\upsilon^2\right)=\hat{\theta}_a^{EBLUP}.$$

Bajo el supuesto de normalidad de los efectos aleatorios se pueden estimar los componentes de la varianza por máxima verosimilitud —*Maximum Likelihood* (ML)— o máxima verosimilitud restringida —*Restricted Maximum Likelihood* (REML)—,[7] esta última reduce el sesgo de la estimación ML, ya que no depende de β. La función log-verosímil del modelo mixto en cuestión es:

$$l(\beta,\psi,\sigma_\upsilon^2\,|\,y)=-\frac{1}{2}\Big[n\ln\left(2\pi\psi\right)+\ln|\mathbf{V}|+\psi^{-1}\left(y+\mathbf{X}\beta\right)^T\mathbf{V}^{-1}\left(y+\mathbf{X}\beta\right)\Big],$$

donde $\mathbf{V}=diag_{1\le a\le m}(\sigma_\upsilon^2+\psi_a)$. Derivando parcialmente esta función con respecto a β y σ_υ^2 tomando a ψ como constante conocida e igualando a cero se tienen las ecuaciones para la estimación de β y σ_υ^2 (Rao, 2003):

$$\frac{\partial l}{\partial\beta}=\mathbf{X}\mathbf{V}^{-1}(y-\mathbf{X}\beta)=0$$

$$\frac{\partial l}{\partial\sigma_\upsilon^2}=-(1/2)\Big[tr\left(\mathbf{V}^{-1}\sigma_\upsilon^2\right)-(y-\mathbf{X}\beta)^T\mathbf{V}^{-1}\sigma_\upsilon^2(y-\mathbf{X}\beta)\Big]=0.$$

Es claro que la estimación ML de β es justamente su estimación por mínimos cuadrados generalizados con el parámetro $\hat{\mathbf{V}}$, la ecuación de

7 Surge al plantear que el estimador ML suele producir estimaciones sesgadas de la varianza porque no tiene en cuenta los grados de libertad que se pierden al estimar la media. Para evitar este problema, se factoriza la verosimilitud completa en dos partes independientes, una de las cuales no contiene la media, asumiendo que por usar esta parte de la verosimilitud no se pierde información con respecto a la verosimilitud completa, así, la restringida corresponde en realidad con la verosimilitud asociada a una combinación lineal de las observaciones, cuya media es nula (León E., 2004).

la varianza σ_υ^2 no tiene una solución analítica, por lo que se debe resolver de forma numérica.

La ecuación de estimación de σ_υ^2 por REML es más simple, ya que no depende de β (Rao, 2003):

$$\frac{\partial l_R}{\partial \sigma_\upsilon^2} = -(1/2)\left[tr\left(\mathbf{P}\sigma_\upsilon^2\right) - y^T \mathbf{P}\sigma_\upsilon^2 \mathbf{P} y \right],$$

con $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}$, sin embargo, tampoco tiene una solución analítica, sólo por métodos numéricos. Para esto, se utiliza el algoritmo de Fisher scoring (Rao, 2003), mismo que se describe con los siguientes pasos:

1. Inicializa las variables de paro, $k = 0$, $tol = 10^{-5}$, $\sigma_\upsilon^{2(0)} = 10$.
2. Arma matriz de varianza de rango m con diagonal $= \psi + \sigma_\upsilon^{2(0)}$.
3. Estima $\beta^{(k)}$ por mínimos cuadrados generalizados.
4. Actualiza $\sigma_\upsilon^{2(k+1)} = \sigma_\upsilon^{2(k)} + \left[I\left(\sigma_\upsilon^{2(k)}\right)\right]^{-1} s\left(\tilde{\beta}^{(k)}, \sigma_\upsilon^{2(k)}\right)$.
5. Compara si $\sigma_\upsilon^{2(k+1)} - \sigma_\upsilon^{2(k)} > tol$ sigue al paso 2.
6. Se obtienen las estimaciones definitivas $\beta = \beta^{(k+1)}$ y $\sigma_\upsilon^2 = \sigma_\upsilon^{2(k+1)}$.

Las funciones del paso 5 son la primera y segunda derivadas de la función de ML o de REML, según sea el caso. Para ML se tienen las siguientes expresiones:

$$s\left(\tilde{\beta}^{(k)}, \sigma_\upsilon^{2(k)}\right) = -\frac{1}{2}\sum_{a=1}^{m}\frac{1}{\sigma_\upsilon^{2(k)} + \psi_a} + \frac{1}{2}\sum_{a=1}^{m}\frac{(\hat{y} - X\tilde{\beta}^{(k)})^2}{(\sigma_\upsilon^{2(k)} + \psi_a)^2};$$

$$\mathrm{I}\left(\sigma_\upsilon^{2(k)}\right) = \frac{1}{2}\sum_{a=1}^{m}\frac{1}{(\sigma_\upsilon^{2(k)} + \psi_a)^2},$$

para REML se tiene:

$$s_R\left(\sigma_\upsilon^2\right) = -\frac{1}{2}tr\left(\mathbf{P}\right) + \frac{1}{2}\hat{y}^T\mathbf{P}\mathbf{P}\hat{y};$$

$$\mathrm{I}_R\left(\sigma_\upsilon^2\right) = \frac{1}{2}tr\left[\mathbf{P}\mathbf{P}\right],$$

donde $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\bar{X}\left(\bar{X}^T\mathbf{V}^{-1}\bar{X}\right)^{-1}\bar{X}^T\mathbf{V}^{-1}$. Ahora, bajo supuestos de normalidad de e_a y υ_a, el error cuadrático medio de $\hat{\theta}_a^{EB}$ con respecto al modelo (Rao, 2003) es:

$$MSE\left(\hat{\theta}_a^{EB}\right) = E\left(\hat{\theta}_a^{EB} - \theta_a\right)^2 = g_{1a}\left(\sigma_\upsilon^2\right) + g_{2a}\left(\sigma_\upsilon^2\right) + g_{3a}\left(\sigma_\upsilon^2\right), \quad (8)$$

donde:

$$g_{1a}\left(\sigma_\upsilon^2\right) := \gamma_a\psi_a$$

$$g_{2a}\left(\sigma_\upsilon^2\right) := \sigma_\upsilon^2\left(1-\gamma_a\right)^2\mathbf{x}_a^T\left(\sum_{a=1}^{m}\gamma_a\mathbf{x}_a\mathbf{x}_a^T\right)^{-1}\mathbf{x}_a$$

$$g_{3a}\left(\sigma_\upsilon^2\right) := \left(1-\gamma_a\right)^2\gamma_a\sigma_\upsilon^{-2}V\left(\hat{\sigma}_\upsilon^2\right),$$

$V\left(\hat{\sigma}_\upsilon^2\right)$ es la varianza asintótica de σ_υ^2, su forma depende del método de estimación usado para σ_υ^2.

En los sumandos de (8), g_{1a} representa el error debido al efecto aleatorio, g_{2a} representa el error por la estimación de β y g_{3a}, el que ocurre debido a la estimación de la varianza σ_υ^2.

Cuando $\hat{\sigma}_\upsilon^2$ se obtiene por REML, el estimador insesgado del MSE es:

$$MSE\left(\hat{\theta}_a^{EB}\right) = g_{1a}\left(\hat{\sigma}_\upsilon^2\right) + g_{2a}\left(\hat{\sigma}_\upsilon^2\right) + 2g_{3a}\left(\hat{\sigma}_\upsilon^2\right). \quad (9)$$

Si $\hat{\sigma}_\upsilon^2$ se obtiene por ML se debe añadir un término extra de sesgo (Rao, 2003).

El error cuadrático medio para las áreas no muestreadas es:

$$MSE\left(\hat{\theta}_k^{syn}\right) = E\left(\mathbf{x}_k^T\hat{\beta} - \theta_k\right) \approx x_k^T\hat{V}_{\hat{\beta}}x_k + \sigma_\upsilon^2 \quad (10)$$

donde $\hat{V}_{\hat{\beta}}$ es la matriz de varianzas y covarianzas de los coeficientes β estimados.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Construcción del modelo

Si se tienen p variables explicativas, se puede conformar un total de 2^p-1 modelos a elegir. A medida que el número de variables aumenta, la cantidad de cálculos necesarios se incrementa muy rápido. Cuando el número de variables es grande (p>40), el trabajo de cómputo es enorme; una forma de aligerar la carga se obtiene al aplicar el método de selección hacia adelante (*forward*). El procedimiento inicia eligiendo de las p variables aquella que tiene la mayor correlación con la que se va a explicar para después ir añadiendo una de las restantes —la que mejor contribuya al modelo— y, así, hasta llegar a un máximo determinado o una condición a cumplir o bien, hasta que nuevas variables no aporten estadísticamente más al modelo. Es pertinente repetirlo iniciado con otra variable también representativa y comparar los modelos resultantes.

Las condiciones a cumplir en la elección del modelo se dividen en tres apartados: los criterios estadísticos utilizados primordialmente para comparar dos modelos y seleccionar el mejor, la significancia estadística de los efectos fijos y el cumplimiento de los supuestos estadísticos del modelo.

Primer apartado

- Prueba de razón de verosimilitud citada por Pinheiro (2002) para determinar si dos modelos, uno con n variables regresoras y otro con las mismas n variables más una variable adicional, representan estadísticamente el mismo modelo, dada por el siguiente estadístico:

$$2\log(L_2/L_1) = 2[\log(L_2) - \mathrm{Log}(L_1)],$$

donde L_1 es la verosimilitud del modelo con n variables regresoras y L_2, la verosimilitud del modelo con n + 1 variables regresoras. Se tiene que la distribución de este estadístico, bajo la hipótesis nula de que el modelo con n variables es el adecuado, se distribuye como una X^2 con un grado de libertad.

- Además, para auxiliar la comparación de los modelos, se calculan los criterios Akaike (AIC) y de Schwarz —*Bayesian Information Criterion* (BIC)—; el primero considera el número de parámetros para comparar los modelos. Su idea es imponer una penalización por la complejidad del modelo, se define como:

$$AIC = -2\log(L) + 2(p+1),$$

donde:

$$\log(L) = -\frac{1}{2}\sum_{a=1}^{m}\left[\log\left(2\pi\left(\sigma_\upsilon^2+\psi_a\right)r_a^2 / \left(\sigma_\upsilon^2+\psi_a\right)\right)\right]$$

con $r_a = y_a - X_a\hat{\beta}$.

El criterio BIC (donde la penalización de complejidad es un poco mayor) está dado por:

$$BIC = -2\log(L) + (p+1)\log(m).$$

Segundo apartado

- Prueba t para cada una de las estimaciones de los coeficientes β del modelo.

Tercer apartado

- El estadístico de la prueba de Shapiro-Wilks para normalidad es:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\right)^2},$$

donde $x_{(i)}$ es el número que ocupa la i-ésima posición en la muestra y $\bar{x}$, la media muestral; las constantes a_i se calculan:

$$\left(a_1,\ldots,a_n\right) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}},$$

donde $m = (m_1,\ldots,m_n)^T$ siendo $m_1,\ldots,m_n$ los valores medios del estadístico ordenado, de variables aleatorias independientes e idénticamente distribuidas, muestreadas de distri-

buciones normales; V es la matriz de covarianzas de ese estadístico de orden.

El valor máximo de W es 1, lo que indica una coincidencia completa con una distribución normal; en consecuencia, el alejamiento de este valor indica menor normalidad.

- Prueba de Breusch-Pagan (1979) para la homocedasticidad del modelo, es decir, que la varianza de los errores es constante, ideada de un multiplicador de Lagrange; el estadístico de prueba es:

$$LM = \frac{1}{2}\left[\mathbf{h}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{h} - m\right] \sim \chi_p^2 ,$$

donde $\mathbf{h}$ es el vector $h_a = e_a^2 / \left(\mathbf{e}^T\mathbf{e}/m\right)$ y $\mathbf{Z}$, la matriz de valores ajustados del modelo y $\mathbf{e}$ es el vector de errores del modelo lineal de los residuales estandarizados respecto a los valores ajustados del modelo.

- Para establecer la presencia de multicolinealidad, se aplican dos mediciones. La primera es el *factor de inflación de varianza* (VIF) (Neter *et al*., 1990); el VIF para mínimos cuadrados ordinarios se obtiene como los elementos de la diagonal de la inversa de la matriz de correlación; el criterio adoptado es, si VIF > 10 en alguna variable indica con alta certeza que existe colinealidad (Chatterjee *et al*., 2000). La segunda es el *número de condición* Rachudel (1971), que se obtiene como:

$$\kappa\left(X\right) = \sqrt{\frac{\lambda_n}{\lambda_1}} \quad ,$$

donde λ_i es la raíz característica de la matriz X^TX. Según Belsley (1991), el problema de multicolinealidad es grave cuando $\kappa(X) > 30$.

Aplicación de la EAP con modelo a nivel de área para estimar el promedio municipal del ingreso por trabajo

El objetivo de la aplicación de la EAP es obtener las estimaciones del promedio del ingreso por trabajo de todos los municipios del país utilizando como insumos:

- La estimación directa del promedio del ingreso por producto del trabajo obtenida de la ENIGH 2010 para cada municipio de la muestra, que servirá como la variable a explicar en el modelo a nivel de área. Es necesario resaltar que se considera que la información de la ENIGH 2010 proviene de una encuesta diseñada bajo un muestreo complejo por lo que, al momento de estimar los promedios y las varianzas de los mismos, es necesario aplicar las ecuaciones (1) y (2) que son muy diferentes a las empleadas en un muestreo aleatorio simple. También, es importante aclarar que las UPM del diseño de la ENIGH 2010 están contenidas en su totalidad en su municipio correspondiente verificando, así, que no hay contraposición entre las áreas pequeñas y el diseño de muestreo.
- Las variables que se construyeron con información de la muestra del CPV 2010, que conceptualmente están relacionadas con el ingreso (Tarozzi & Deaton, 2009; CONEVAL, 2011; Suárez, 2011; Székely *et al*., 2006), constituyen insumo tanto para la estimación como para la estratificación y serán tomadas como regresoras en el modelo a nivel de área.

Una vez obtenidas las estimaciones del modelo a nivel de área y las directas, se realizará un comparativo de ambas con las cifras obtenidas de la muestra del CPV 2010.

Previo al cálculo de las estimaciones se realizaron dos importantes ajustes: el primero es para hacer comparable la información de los ingresos respecto a la diferencia temporal entre la ENIGH 2010 y la muestra del CPV 2010, el método utilizado es el que sugieren Cortés y Rubalcava (1994), donde se toma cada rubro en la ENIGH 2010 que forma parte del *ingreso por trabajo* para cada uno de los seis meses captados y se deflacta a la mitad del periodo de referencia del CPV 2010, es decir, a junio del 2010; el segundo es respecto a la muestra a nivel municipal para asegurar que la expansión de la muestra del número de viviendas sea el contabilizado por el

CPV 2010 para cada municipio; se hace utilizando la siguiente expresión:

$$w'_{ai} = w_{ai}\frac{N_{vacpv2010}}{\hat{N}_{va}} \text{ con } \hat{N}_{va} = \sum_{i=1}^{n_a} w_{ai},$$

donde:

w'_{ai} = factor de expansión de la vivienda i ajustado al CPV 2010 del municipio a.
w_{ai} = factor de expansión original de la ENIGH 2010 para la vivienda i en el municipio a.
$\hat{N}_{vacpv2010}$ = número de viviendas contadas por el CPV 2010 en el municipio a.
$\hat{N}_{va}$ = número de viviendas estimadas por la ENIGH 2010 en el municipio a.

La estimación directa del promedio municipal del *ingreso por trabajo* por vivienda de los municipios en los que hay muestra en la ENIGH 2010 se calcula con la ecuación (1) tomando a $w'_{ai} = 1/\pi_i$ y al dominio S como el municipio (área pequeña) a, la varianza respectiva se calcula con la ecuación (2). Estos cálculos se realizan utilizando el paquete estadístico R con la librería *survey*; en los resultados se aprecia que los 600 municipios con muestra están ordenados de menor a mayor según su número de viviendas censales (ver gráfica 1); su correspondiente intervalo de confianza se obtiene con la siguiente expresión:

$$I_a^{DIR} = \hat{\bar{Y}}_a^{DIR} \pm 1.96\sqrt{\psi_a^{DIR}}.$$

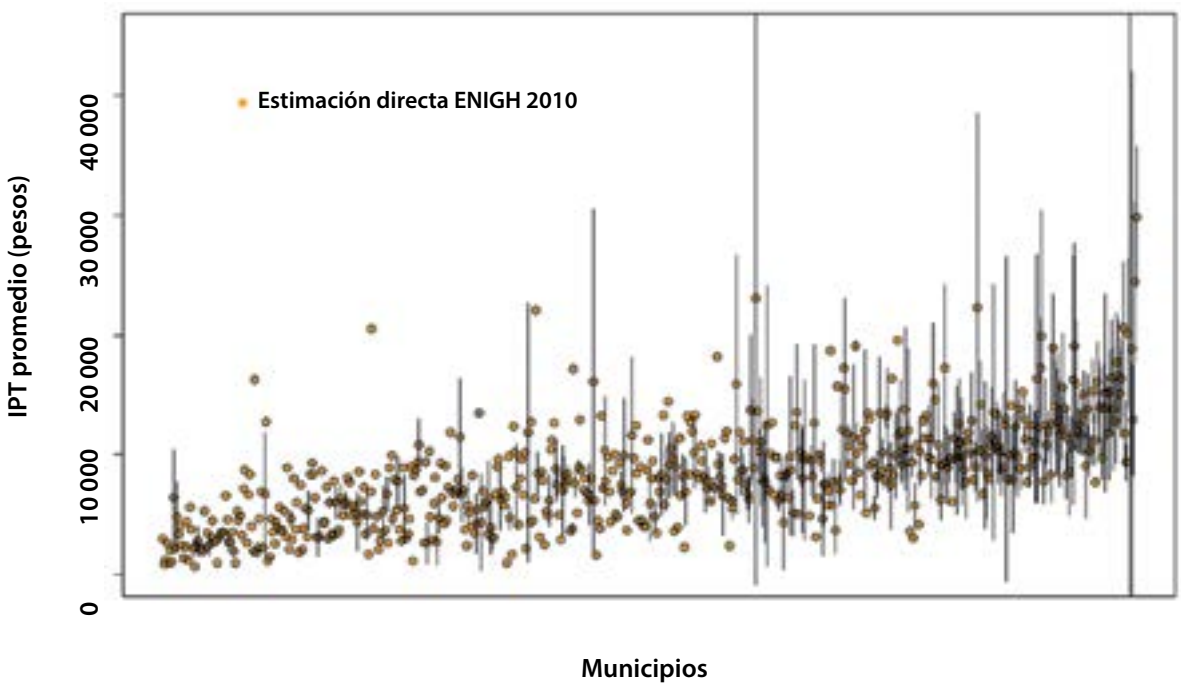
Se debe aclarar que para varios municipios no fue posible calcular la varianza debido a que sólo una UPM fue muestreada en ellos; por esta razón, en la gráfica 1 varios promedios estimados no incluyen su intervalo de confianza.

En el análisis exploratorio de la variable a estimar y de su varianza se determinó que existe una relación entre éstas debido a que su correlación es cercana a 0.5 y a que, para cumplir con el supuesto de homogeneidad de la varianza y la normalidad de la variable dependiente, era necesario realizar una transformación por lo que, partiendo de la expresión (3), el modelo propuesto ahora es:

$$\hat{\theta}_a^{DIR} = \sqrt{\hat{\bar{Y}}} = \mathbf{x}_a^T\beta + \upsilon_a + e_a. \qquad (11)$$

Gráfica 1

Municipios con muestra ENIGH 2010, promedios estimados con intervalos a 95%



Uno de los supuestos fundamentales del modelo a nivel de área es que se conoce la varianza ψ_a para cada área pequeña, pero por lo general no es así, entonces se sustituye por su estimador suavizado $\hat{\psi}_a$; éste se obtiene partiendo de la estimación directa de la varianza de los datos a nivel unidad $\hat{\psi}_a^{DIR}$, pero como premisa es sabido que los tamaños de la muestra en los municipios son pequeños, por lo que $\hat{\psi}_a^{DIR}$ no es un buen estimador, ya que puede tomar valores en extremo grandes o increíblemente pequeños. Para solucionar esta situación, las estimaciones de $\hat{\psi}_a^{DIR}$ se *suavizan* sustituyendo los valores extremos por los obtenidos del siguiente modelo lineal para la varianza:

$$\hat{\psi}_r^{DIR} = \beta_{\psi_r} \frac{\hat{\bar{Y}}_r^{DIR}}{\sqrt{n_r}} + e_{\psi_r} \quad \text{con } r \leq m. \qquad (12)$$

Para el ajuste de los parámetros del modelo (12) se toma la varianza de los municipios en muestra, excluyendo los 333 que sólo cuentan con una UPM y cuya varianza no es posible calcular, además de excluir también 13 municipios cuyo coeficiente de variación es mayor a 0.35.[8]

Así, utilizando las predicciones de este modelo lineal se pretende obtener valores más realistas de $\hat{\psi}_a^{DIR}$ para estos 13 municipios y para los 333 con una sola UPM.

Es importante comentar que los municipios fueron excluidos sólo para modelar la varianza; no lo son para las demás actividades del ajuste del modelo a nivel área.

Ahora, para la estimación en áreas pequeñas, es necesario establecer el modelo a nivel de área, que obedece a la ecuación (6) con la transformación del modelo (11) y que cumpla las condiciones señaladas en los tres apartados descritos en la sección *Construcción del modelo*; estas condiciones se verifican utilizando funciones de las librerías *stats, nnet, MASS, survival, car* y *nlme* del paquete estadístico R, así como a través de la elaboración de programas de cómputo propios.

Las variables regresoras contenidas en X_a se seleccionan bajo los criterios del primer apartado, resultando que las que mejor explican el promedio municipal del *ingreso por trabajo* en la vivienda para la ENIGH 2010 son cinco; sus mnemónicos y descripciones son los siguientes:

- (p_v_cel). Porcentaje de viviendas que disponen de teléfono celular.
- (p_v_inter). Porcentaje de viviendas que cuentan con servicio de internet en el municipio.
- (viv_cocgas). Porcentaje de viviendas particulares habitadas que usan gas como principal combustible para cocinar.
- (p_ocupados). Porcentaje de personas en el municipio de 12 a 130 años de edad que trabajaron o que no trabajaron pero sí tenían trabajo en la semana de referencia.
- (ocupnoagri). Porcentaje de personas ocupadas que en la semana de referencia no se desempeñaron en su trabajo en actividades agrícolas, ganaderas, pesqueras, forestales y de caza, de acuerdo con la Clasificación Única de Ocupaciones (todas las claves, excepto las comprendidas entre 6101 a 6999).

Los resultados de la validación de las condiciones del segundo apartado se presentan en el cuadro 1.

El cumplimiento de las condiciones del tercer apartado se aplica para verificar la normalidad de los residuales del modelo, la ausencia de multicolinealidad en los valores de $\boldsymbol{X}$, la homocedasticidad de los residuales y la normalidad de los efectos aleatorios, siendo este último el supuesto subyacente para la estimación de los componentes de la varianza.

Para verificar la normalidad de los residuales del modelo, se utiliza la prueba de normalidad de Shapiro-Wilks, aplicando la función *resid* (*modelo, type = normalized*) de R, la cual usa la descompo-

8 Escárcega, Campeche; Namiquipa, Chihuahua; Nuevo Casas Grandes, Chihuahua; Cuajimalpa de Morelos, Distrito Federal; Lerdo, Durango; Chilapa de Álvarez, Guerrero; Huixquilucan, estado de México; Santa Lucía del Camino, Oaxaca; Matlapa, San Luís Potosí; Angostura, Sinaloa; Nuevo Laredo, Tamaulipas; Nativitas, Tlaxcala y Fresnillo, Zacatecas.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Cuadro 1

Estimación del parámetro β del modelo

Variable	$\hat{\beta}$	Error std. $\hat{\beta}$	Valor de t	Valor p
Intercepto	23.2	5.23	4.4	9.3e-06
p_v_cel	0.25	0.06	4.0	5.1e-05
p_v_inter	0.50	0.07	6.8	1.3e-11
viv_cocgas	0.10	0.03	3.0	2.5e-03
p_ocupados	0.39	0.12	3.1	1.6e-03
ocupnoagri	0.13	0.05	2.7	6.2e-03

Nota: en la quinta columna, el valor para todas las variables es menor a 0.05, lo cual indica que en el modelo tanto el intercepto como las demás variables se asocian de forma satisfactoria con el ingreso.

sición de Cholesky, que elimina la dependencia entre los residuales dada la estructura de correlación (Valencia, M., 2010); el resultado es:

Shapiro-Wilk normality test
data: resnorm
$W = 0.9948$, *p-value* $= 0.04058$.

Por ello, no se rechaza la hipótesis de normalidad a un nivel de significancia de 4 por ciento.

Los resultados de la aplicación de las pruebas de multicolinealidad VIF y número de condición se observan en el cuadro 2, ambas indican la ausencia de problemas severos de multicolinealidad, ya que los VIF obtenidos son siempre menores a 10 y el número de condición es menor que 30.

Cuadro 2

Valores de las pruebas de multicolinealidad

Variable	VIF	Variable	VIF
p_v_cel	7.49	p_ocupados	2.25
p_v_inter	2.74	ocupnoagri	3.63
viv_cocgas	3.96		
Número de condición		20.36	

Para verificar la homocedasticidad en el modelo se aplica la prueba de Breush-Pagan, utilizando la función *ncvtest* de R, con el resultado siguiente:

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare $= 0.7538$, *Df* $= 1$, $p = 0.3852$.

Esto indica que la varianza del error puede considerarse como constante.

La prueba de Shapiro-Wilks es la empleada para comprobar la normalidad de los efectos aleatorios, dando el resultado siguiente:

Shapiro-Wilk normality test
data: eblup1a$randeff
$W = 0.998$, *p-value* $= 0.7216$.

El valor p observado indica que hay evidencia de normalidad en los efectos aleatorios υ.

Una vez que el modelo (6) con las variables seleccionadas cumple con las condiciones establecidas, se obtienen las estimaciones del promedio del ingreso por trabajo en las viviendas para los municipios con muestra en la ENIGH 2010, mediante la aplicación de la librería *sae* del paquete estadístico R. Como resultado se obtienen las estimaciones de los promedios que se presentan en la gráfica 2, así como la varianza de los efectos aleatorios, con un valor estimado de $\hat{\sigma}^2_{\upsilon} = 112.467$.

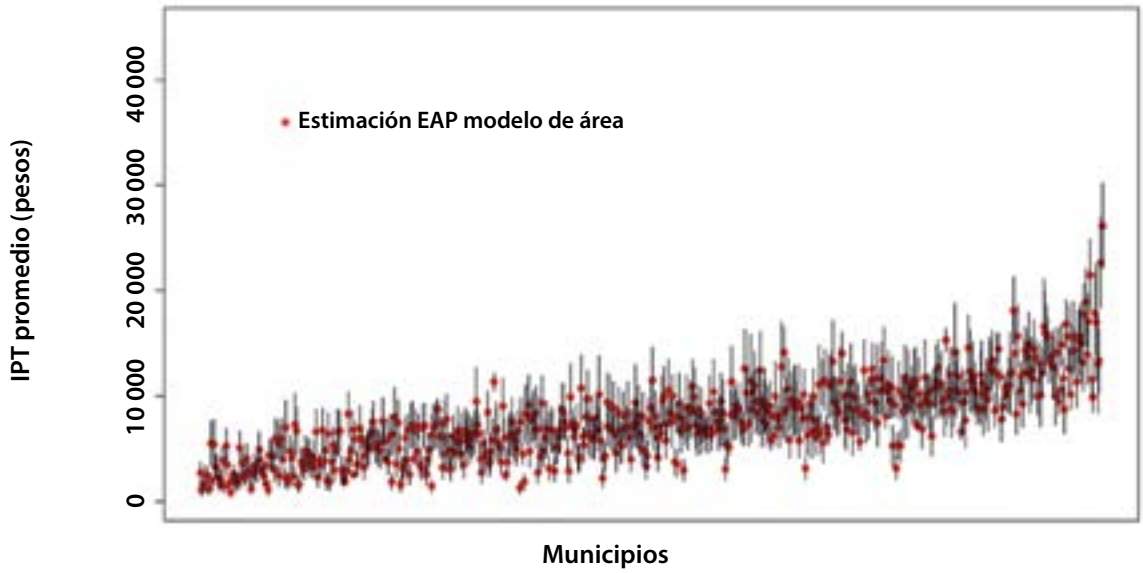
La escala de la gráfica 2 es igual a la de la 1 con el propósito de facilitar la comparación visual entre ellas; al hacerlo se aprecia, a simple vista, que las estimaciones se condensan siguiendo el modelo planteado, además de que se homogeneizan los intervalos de confianza a una longitud razonable, mismos que se obtienen aplicando la ecuación:

$$I_a^{EB} = \left[\hat{\bar{Y}}_a^{EB}\right]^2 \pm 1.96\sqrt{4MSE\left(\hat{\bar{Y}}_a^{EB}\right)\left[\hat{\bar{Y}}_a^{EB}\right]},$$

donde ya está considerada la transformación a las unidades originales (pesos mexicanos). Para transformar el MSE se aplicó el método Delta que relaciona la varianza de Y con la de $\theta(Y) = \sqrt{Y}$ (para más detalles, ver Velasco, C., 2007).

Gráfica 2

Municipios con muestra ENIGH 2010, promedios estimados con intervalos a 95%



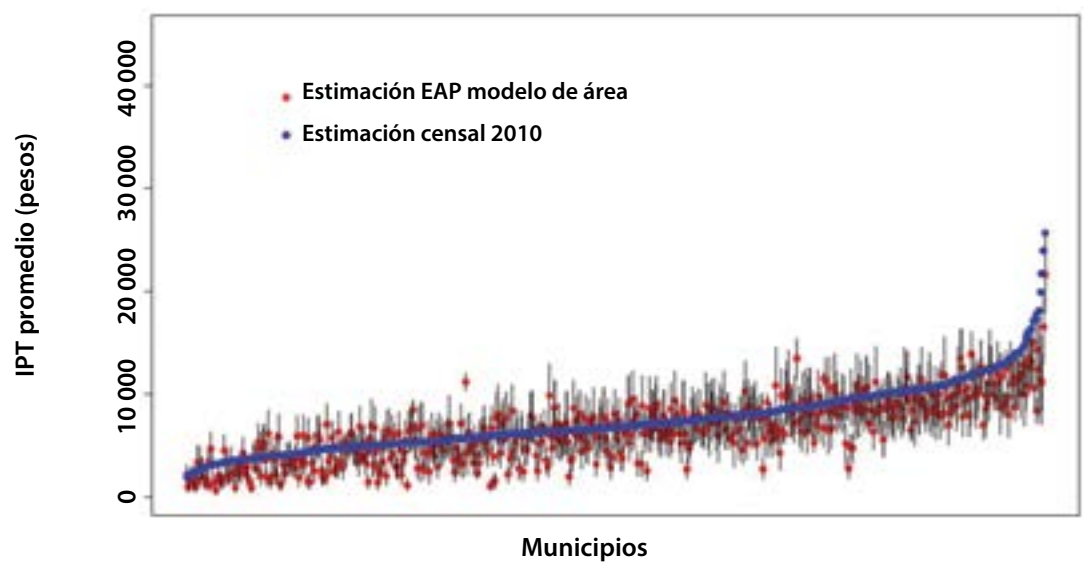
La gráfica 3 se construye tomando como base la 2, a la cual se añaden los puntos en azul que corresponden a la estimación del promedio del ingreso por trabajo en la vivienda de la muestra del CPV 2010 en los municipios con muestra de la ENIGH 2010; se observa cómo la condensación citada en la gráfica 2 corresponde con las estimaciones de la muestra censal, aun cuando la ENIGH 2010 y la muestra del CPV 2010 fueron levantamientos y diseños muestrales por completo independientes.

El coeficiente de variación (CV) es una medida que puede proporcionar, en casos específicos,[9] una

9 Medias positivas derivadas de valores positivos con distribución normal o aproximadamente normal.

Gráfica 3

Municipios con muestra ENIGH 2010, promedios estimados con intervalos a 95%



INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

buena idea de la precisión de las estimaciones. En la gráfica 4 se observan los CV de las estimaciones directas y del modelo de área de los municipios con muestra en la ENIGH 2010, mismos que están ordenados de menor a mayor tamaño de muestra. En la línea punteada se establece el valor del CV = 0.25 como umbral para indicar que los CV menores representan estimaciones razonablemente confiables; la gran mayoría de las estimaciones con el modelo de área (puntos rojos) están por debajo de este umbral. Los puntos azules con CV igual a cero representan a los municipios con sólo una UPM en muestra, para los que en realidad su CV está indeterminado.

Ahora, para la estimación del promedio municipal del *ingreso por trabajo* en la vivienda para los municipios que no tienen muestra en la ENIGH 2010, se utilizó la ecuación (7); la varianza respectiva se calcula mediante la ecuación (10) y los intervalos de confianza con la siguiente expresión:

$$I_k^{SYN} = \left[\hat{\bar{Y}}_k^{SYN}\right]^2 \pm 1.96\sqrt{4\left(\hat{\sigma}_\upsilon^2 + \hat{V}_{\hat{\beta}}\right)\left[\hat{\bar{Y}}_k^{SYN}\right]},$$

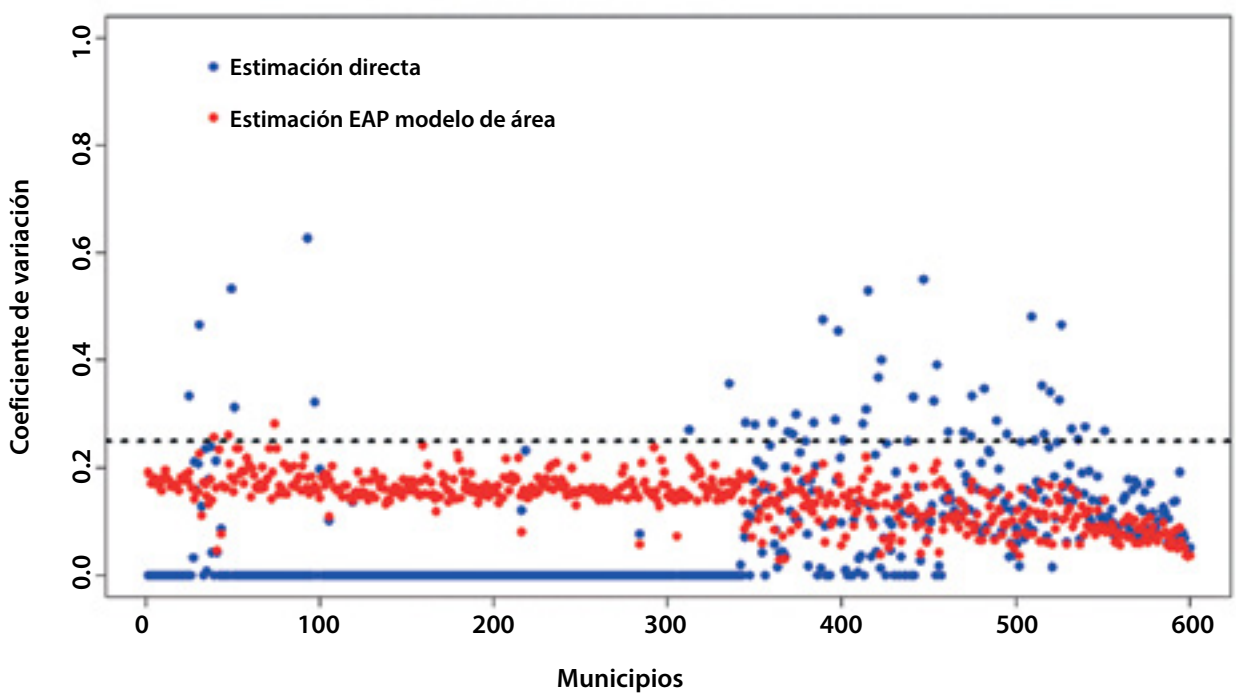
donde está considerada la transformación a las unidades originales.

Los resultados se muestran en la gráfica 5, la cual presenta, también, la estimación de la misma variable, pero de la muestra del CPV 2010. Se aprecia que, en general, las estimaciones con el modelo son menores a las del Censo. Se conserva la misma escala de la gráfica 1 para facilitar su comparación visual.

Enseguida, se realiza el cálculo del MSE empírico de las estimaciones con el método directo, y de las del modelo a nivel de área, ambas respecto a la estimación de la muestra del CPV 2010; para el segundo caso se diferencia a los municipios con muestra de los que no la tuvieron. Dichos cálculos se confrontan en la gráfica 6, en la cual se puede apreciar que el MSE empírico entre el método directo y el del modelo a nivel de área se reduce aproximadamente 3.9 veces, lo cual es razonable en virtud de que el dominio de la muestra de la ENIGH 2010 no abarca un nivel municipal. El MSE empírico en las estimaciones de los municipios no muestreados es

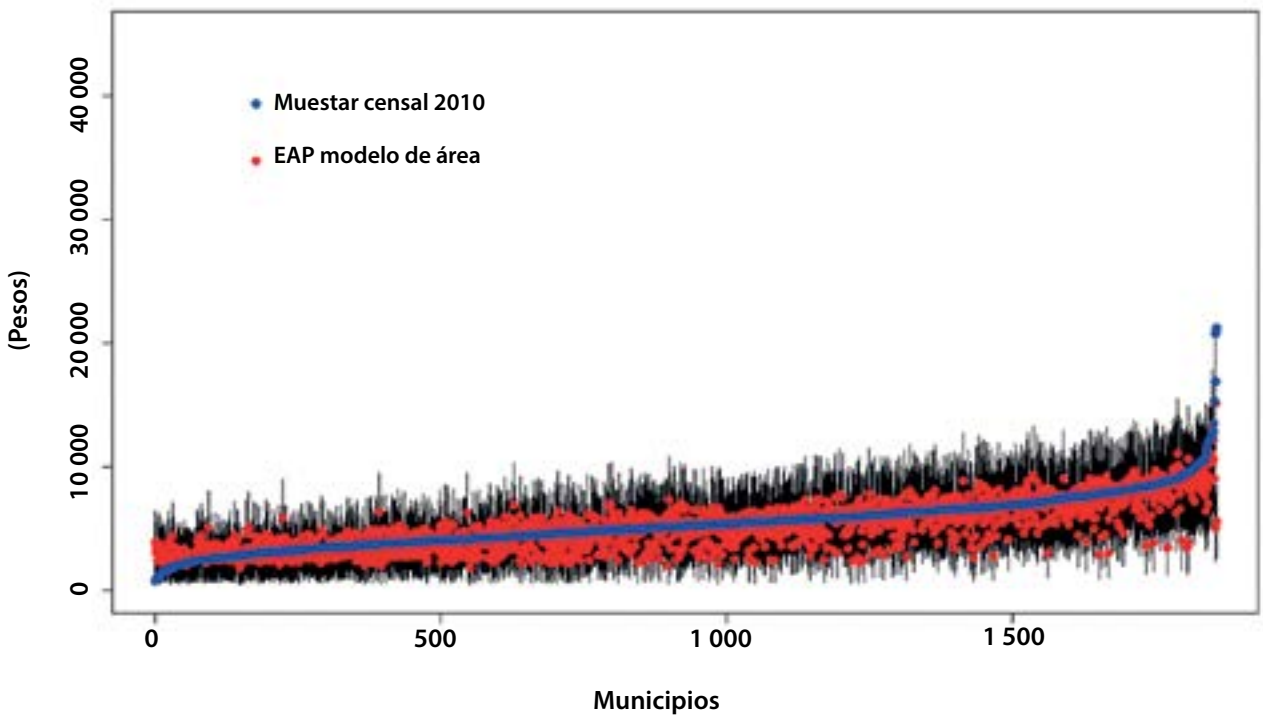
Gráfica 4

Coeficientes de variación de las estimaciones, municipios con muestra



Gráfica 5

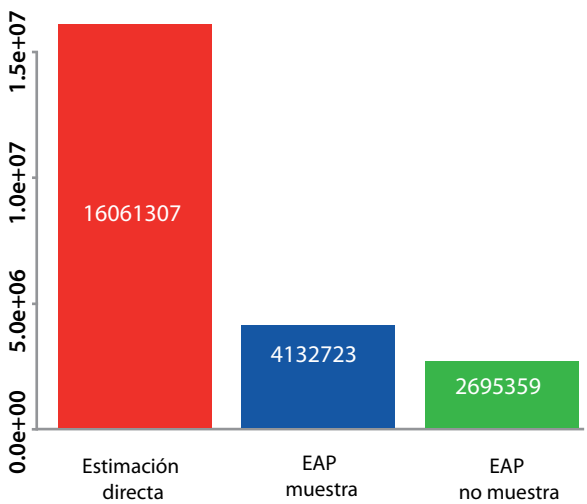
Municipios sin muestra ENIGH 2010, promedios estimados con intervalos a 95%



muy favorable, debido a que la mayoría de ellos son de ingresos bajos, por lo tanto, sus diferencias cuadráticas con los ingresos de la muestra censal son menores que las de los municipios con muestra; sin embargo, el modelo para estos municipios es conservador, ya que proporciona intervalos de confianza amplios (ver gráfica 5).

Gráfica 6

MSE empírico de las estimaciones respecto al valor de la muestra censal 2010

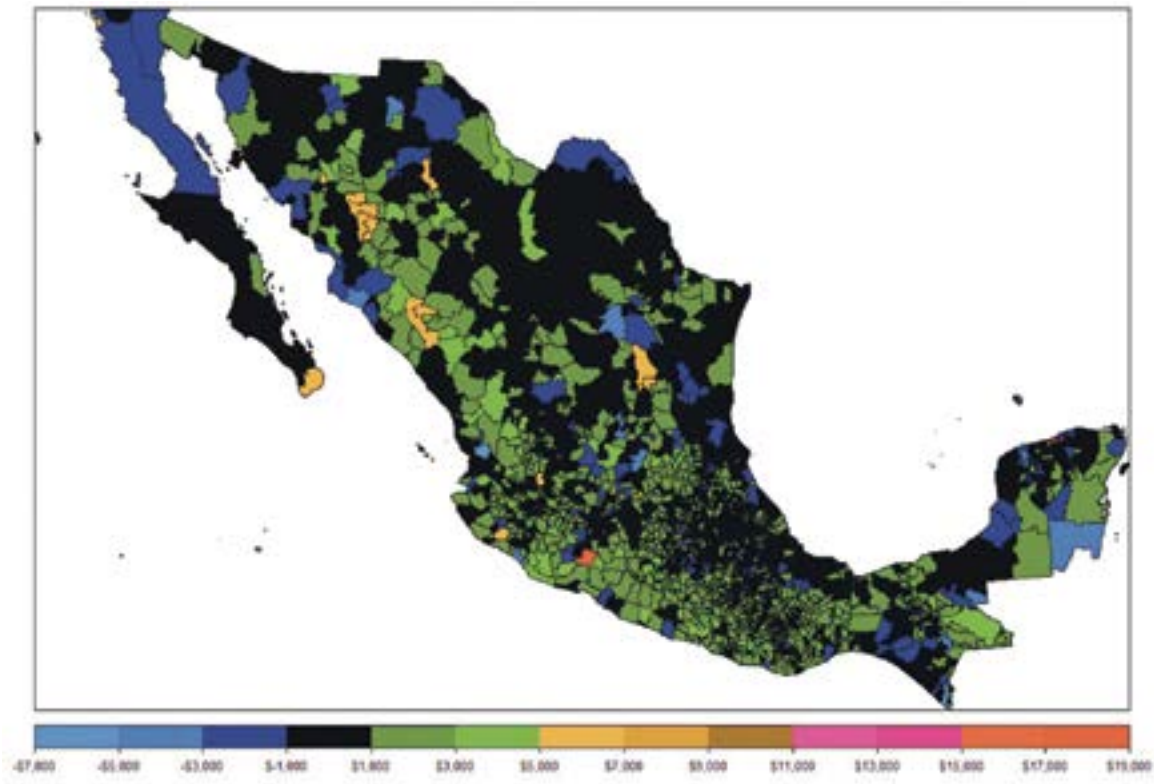


Otra forma de comparar las estimaciones se puede ver en el mapa con división municipal, donde se colorean las diferencias por rangos al restar de la estimación censal 2010 la del modelo EAP. En color negro se resaltan los municipios en los que esta diferencia absoluta es menor a mil pesos; con azul obscuro, las que están por abajo de la muestra censal entre mil y 3 mil pesos y con verde obscuro, entre mil y 3 mil pesos por arriba de la misma; en los demás colores (que son más claros) se aprecian los municipios con mayor diferencia. Es importante anotar que estos colores aparecen dispersos en todo el mapa, es decir, no hay una concentración que pueda indicar que el modelo sólo explica el comportamiento de alguna región en particular. Sin tomar en cuenta el negro, el predominio del verde obscuro indica que la estimación del modelo EAP es ligeramente menor a la de la muestra del CPV 2010.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Figura 1

Diferencia entre las estimaciones de la muestra del CPV 2010 y EAP



Conclusiones

En el ejercicio realizado se ha mostrado que la técnica de EAP con modelo a nivel de área mejora en todos los casos la eficiencia de la estimación directa y, en muchos, ésta es radical. Es claro que la disponibilidad de datos auxiliares y la selección de un buen modelo es fundamental para lograr el éxito deseado. La evidente tendencia de las estimaciones con el modelo a nivel de área hacia las estimaciones de la muestra del CPV 2010 indica una buena congruencia de los datos con la ENIGH 2010 del mismo año, sabiendo que ambos levantamientos fueron totalmente independientes.

Aún hay muchas situaciones teóricas por resolver en la gran gama de aplicaciones para la EAP, sin embargo, existen técnicas probadas para aplicar con modelos espaciales, temporales, espacio-temporales tanto para modelos a nivel de área como de unidad.

La aplicación de modelos para la estimación en áreas pequeñas en el ámbito de las estadísticas oficiales es cada vez más utilizada en los institutos nacionales de estadística debido al gran ahorro de recursos que pueden representar. En México estas técnicas aún son poco conocidas y mucho menos aplicadas, sin embargo, al emplearlas es importante asegurarse de que los métodos utilizados, las hipótesis que subyacen a los modelos, así como la calidad de los resultados sean descritos de forma clara a los usuarios.

Fuentes

Belsley, D. A. *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. 1st ed. New York, NY; John Wiley & Sons, Inc.; 1991.

Breusch, T. S. y A. R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation", en: *Econometrica*. 47. 1979, pp. 1287-1294.

Cassel, C. M., C. E. Särndal y J. H. Wretman. *Foundations of Inference in Survey Sampling*. John Wiley & Sons, Inc., 1977.

Chatterjee, S., A. S. Hadi y B. Price. *Regression Analysis by Example*. 3rd ed. New York, NY; John Wiley & Sons, Inc.; 2000.

Cortés, F. y R. Rubalcava. *El ingreso de los hogares*. Serie *Monografías Censales*. Vol. VII. Aguascalientes, México; INEGI-El Colegio de México-Instituto de Investigaciones Sociales, UNAM; 1995.

Eurarea Consortium. *Enhancing Small Area Estimation Techniques to meet European Needs. Project Reference.* Vol. 2. UK, Explanatory Appendices, 2004.

Fay, R. E. y R. A. Herriot. "Estimates of Income for Small Places: an Application of James-Stein procedures to census data", en: *Journal of the American Statistical Association*. Vol. 74. 1979, pp. 269-277.

Fox J. y S. Weisberg. *An {R} Companion to Applied Regression*. Second edition. Thousand Oaks, CA, Sage, 2011. Consultado en http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Harter, R. M. *Small Area Estimation using nested-error and for sugar models and auxiliary data*. Ph. D. Theses. Iowa State University, Estados Unidos de América (EE.UU.), 1983.

Hasen, M. H., W. G. Madow y B. J. Tepping. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys", en: *Journal of the American Statistical Association*. Vol. 78. 1983, pp. 776-793.

Instituto Nacional de Estadística y Geografía (INEGI). *Diseño de la muestra censal 2010*. México, INEGI, 2011c.

_____ *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2010. Diseño muestral*. México, INEGI, 2011c.

_____ Marco conceptual del Censo de Población y Vivienda 2010. México, INEGI, 2011c.

_____ Nueva construcción de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2010. Nueva construcción de ingresos y gastos. México, INEGI, 2011c.

Kackar, R. N. y D. A. Harville. "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models", en: *Journal of the American Statistical Association*. 79. 1984, pp. 853-862.

León, E. "Métodos de estimación de componentes de varianza en poblaciones. Una reseña histórica", en: *Revista Computarizada de Producción Porcina*. 11, 2004, pp. 28-29.

Lohr, S. L. *Muestreo: diseño y análisis*. Thomson International, 2000.

Lumley, T. *Survey: analysis of complex survey samples.* R package version 3.24. 2011.

Molina, I. y Y. Marhuenda. *Sae: Small Area Estimation*. R package version 1.0. 2012.

Molina I. y J. N. K. Rao. *Taller de Aplicación de Técnicas de Estimación para Áreas Pequeñas a las Ciencias Sociales*. México, DF; Universidad Iberoamericana; 2012.

Nel Pacheco, P. *Verificación de supuestos*. Consultado en http://www.virtual.unal.edu.co/cursos/ciencias/dis_exp/und_3/pdf/validaciondesupuestosunidad%203b.pdf el 25 de noviembre de 2013.

Neter, J., W. Wasserman y M. H. Kutner. *Applied Linear Statistical Models.* 3.ª ed. Irwin, MA, 1990.

Pinheiro J. y D. Bates. *Mixed Effects Models in S and S-PLUS*. Corrected third printing. New York, Springer-Verlag, 2002.

Pinheiro, J.; D. Bates; S. DebRoy; D. Sarkar y the R Development Core Team. *Nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.pp. 1-100. 2011.

Prasad, N. G. N. y J. N. K. Rao. "The Estimation of Mean Squared Errors of Small Area Estimators", en: *Journal of American Statistical Association*. 85. 1990, pp. 163-171.

R Development Core Team. *R: A language and environment for statistical computing*. Viena, Austria, R Foundation for Statistical Computing, 2011. Consultado en http://www.R-project.org/

Rachudel, W. J. *Multicollinearity once again*. Cambridge, Harvard Institute of Economic Research, 1971.

Rao, J. N. K. *Small Area Estimation*. New Jersey, EE.UU.; Wiley Interscience; 2003.

Roso, V. M.; F. S. Schenkel; S. P. Miller y L. R. Schaeffer. *Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation*. Consultado en http://www.journalofanimalscience.org/content/83/8/1788 el 25 de abril de 2014.

Särndal, C. E. "Design-Consistent versus Model-Dependent Estimators for Small Domains", en: *Journal of the American Statistical Association*. Vol. 79. 1984, pp. 624-631.

Särndal, C. E. y M. A. Hidiroglou. "Small Domain Estimation: A Conditional Analysis", en: *Journal of the American Statistical Association*. Vol. 84. 1989, pp. 266-275.

Suárez M. *Estimación del ingreso promedio por vivienda en los municipios del estado de Sonora*. Tesis. CIMAT, 2010.

Székely M.; L. López-Calva; A. Meléndez; E. Rascón y L. Rodríguez. *Poniendo a la pobreza de ingresos y a la desigualdad en el mapa de México*. 2006. Consultado en http://www.economiamexicana.cide.edu/num_anteriores/XVI-2/03_SZEKELY.pdf

Tarozzi A. y A. Deaton. "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas", en: *The Review of Economics and Statistics*. Vol. 91, núm. 4. Noviembre del 2009, pp 773-792.

Terry Therneau y original Splus->R port by Thomas Lumley. *Survival: Survival analysis, including penalised likelihood.* R package version 2.36-5. Consultado en http://CRAN.R-project.org/package=survival. 2011.

Uriel E., *Multicolinealidad*. Universidad de Valencia. Consultado en http://www.uv.es/uriel/material/multicolinealidad3.pdf el 26 de abril de 2014.

Velasco, C. *Curso de Estadística*. Capítulo 5. Madrid, Universidad Carlos III. Consultado en http://www.eco.uc3m.es/~cavelas/EstMEI/tema5.pdf

Valencia, M. *Estimación en modelos lineales mixtos con datos continuos usando transformaciones y distribuciones no normales*. Tesis. Universidad Nacional de Colombia, 2010.

Venables, W. N. y B. D. Ripley. *Modern Applied Statistics with S.* Fourth edition. New York, Springer, 2002.

Clasificación de cultivos agrícolas

utilizando técnicas clásicas de procesamiento de imágenes y redes neuronales artificiales

Roberto Antonio Vázquez Espinoza de los Monteros, José Ambrosio Bastián y Guillermo Alberto Sandoval Sánchez[1]

Flower crops in foothills in Michoacan state/Agostini/Getty Images

1 Los autores agradecen tanto al Fondo Sectorial CONACYT-INEGI como a la Universidad La Salle por los apoyos otorgados a través de los proyectos con claves 187637 el-061/12 el-065/12, respectivamente. Asimismo, Guillermo Sandoval agradece al Fondo Sectorial CONACYT-INEGI por la beca obtenida mediante el proyecto con clave 187637.

La percepción remota, llevada a cabo con el apoyo de satélites artificiales, ha contribuido con el paso de los años a la clasificación adecuada de los cultivos agrícolas. Así, se han identificado varios avances en este campo usando sus imágenes con diferentes resoluciones espaciales y espectrales. Su clasificación puede ser abordada como un problema de reconocimiento de patrones que utiliza procesamiento de imágenes y que está formado por varias etapas: la primera se relaciona con la selección del método o medio a través del cual se obtiene la imagen, la segunda está inmersa tanto en el procesamiento de ésta como en la extracción de rasgos y la tercera hace referencia a la selección y aplicación de algoritmos de clasificación. En esta investigación se presenta una metodología basada en técnicas de procesamiento de imágenes satelitales de baja resolución y reconocimiento de patrones que permite realizar una adecuada clasificación de cultivos agrícolas. Para validar su desempeño, se definió una región de prueba en el estado de Sinaloa, México, con una imagen de prueba de *GoogleEarth* con tres canales de información en el espectro visible; se etiquetaron manualmente cinco tipos de cultivo de donde se derivaron 24 bases de datos compuestas por 2 752 muestras asociadas a los diferentes cultivos.

Palabras clave: percepción remota, clasificación de cultivos agrícolas, reconocimiento de patrones, procesamiento de imágenes.

Recibido: 30 de junio de 2014.
Aceptado: 21 de mayo de 2015.

Remote sensing, held with artificial satellites, has contributed over the years to the proper classification of agricultural crops. We have identified several advances in crop classification using satellite images with different spatial and spectral resolutions. Crops classification can be seen as a problem of pattern recognition and it consists of several stages: the first related to the methodology by which the image is obtained; the second involves image processing and feature extraction; and the third refers to the selection and application of classification algorithms. In this study, we present a methodology to classify agricultural crops from low-resolution satellite images based on techniques of image-processing and pattern recognition. To validate the proposed methodology, we defined a region from the state of Sinaloa, Mexico. The test image was obtained through *Google Earth* with three channels of information in the visible spectrum, and it was manually labeled with 5 types of crops. Hence, there were derived a total of 24 databases made by 2 752 samples associated with different crops.

Key words: remote sensing, crop classification, pattern recognition, image processing.

Introducción

Se conoce como percepción remota a la acción de llevar a cabo mediciones de la energía electromagnética emitida por la superficie de la Tierra utilizando instrumentos como cámaras en aeronaves o sensores ubicados en satélites. Esto ha mostrado ser útil en trabajos relacionados con la identificación de recursos naturales, tal es el caso del agua (Zhong *et al*., 2009) y de los reservorios de carbono constituidos por la biomasa vegetal (Czerepowicz *et al*., 2012).

En el área de la agricultura, las aplicaciones de la percepción remota las encontramos en torno a la administración del suelo, estimación de cosechas, proyección de daños causados por fenómenos naturales y clasificación de cultivos, entre otras.

Cuando hablamos de esta última se está abordando un problema donde se pueden identificar varias etapas:

- Selección del método o medio a través del cual se obtendrá la imagen (fuentes de información).
- Procesamiento de ésta, así como selección y extracción de rasgos (extracción de características).
- Aplicación de algoritmos de clasificación.

En los últimos años, esta área de investigación se ha desarrollado y ha logrado avances significativos en cada una de las etapas previamente descritas.

Respecto al medio por el cual se obtiene la imagen, se puede mencionar la información de radar (Skriver *et al*., 1999; Schotten *et al*., 1995), la multiespectral (de Castro *et al*., 2012; Smith y Fuller, 2001) y la hiperespectral (Senthilnath *et al*., 2011; Gomez-Chova *et al*., 2003).

En cuanto a la extracción de características, se encuentran las transformaciones pixel a pixel (Shao y Lunetta, 2009; Chakraborty y Panigrahy, 1997) y las basadas en la transformada de Fourier (Nejati *et al*., 2008; Kiani *et al*., 2010), así como las técnicas sustentadas en texturas (Smith y Fuller, 2001; Dean y Smith, 2003).

Por último, en lo que se refiere a los algoritmos de clasificación, es posible mencionar aquellos trabajos basados en máxima verosimilitud (Gomez-Chova *et al*., 2003; Dean y Smith, 2003), árboles de decisión (Chakraborty y Panigrahy, 1997; Fries *et al*., 1998) y técnicas apoyadas en el cómputo inteligente, como redes neuronales artificiales (de Castro *et al*., 2012; Kiani *et al*., 2010).

Cada una de las técnicas mencionadas presentan ventajas y desventajas con respecto a otras (ver tabla 1), lo cual se ve reflejado en la eficiencia de clasificación, la erogación monetaria por la adquisición de las imágenes y el costo computacional derivado del procesamiento y análisis de éstas. En ese sentido, resulta interesante aplicar aquellas que permitan reducir los costos originados por la compra de las imágenes y disminuir el costo computacional sin afectar la eficiencia y confiabilidad de estos sistemas durante la clasificación de cultivos agrícolas.

Hoy en día, las imágenes satelitales están al alcance de cualquier usuario y se pueden obtener a un bajo precio mediante plataformas como *Google Earth*; sin embargo, su calidad no necesariamente será la mejor y sólo se tendrá acceso a imágenes con tres bandas de información (espectro visible RGB). Si se utilizara este tipo, provocaría que muchos sistemas de clasificación de cultivos disminuyeran su confiabilidad debido a que fueron diseñados para trabajar con más bandas de información.

En ese sentido, el desarrollo de una metodología que permita clasificar cultivos agrícolas en imágenes satelitales de baja resolución y bandas de información del espectro visible conlleva a un gran reto: ¿cómo construir una que sea capaz de hacerlo con un alto grado de confiabilidad?

Para abordar este reto, se plantea utilizar un enfoque basado en técnicas del área de reconocimiento de patrones, procesamiento de imágenes y cómputo inteligente, elementos indispensables para la construcción de un sistema automático de reconocimiento de objetos, donde éstos pueden representar al cultivo que se desea clasificar.

Conceptos básicos

En esta sección se revisarán aquellos que permitan al lector entender la metodología que se propone en este artículo. Principalmente, nos centraremos en los relacionados con la etapa de extracción de rasgos y técnicas de clasificación.

Es importante recordar que en cualquier sistema de reconocimiento automático de objetos se pueden identificar cinco etapas:

- Adquisición de la información: medio por el cual se obtendrán los datos que se emplearán para realizar la clasificación.
- Preprocesamiento: permite adecuar la imagen a analizar y eliminar cualquier tipo de ruido asociado a cambios de iluminación y deformaciones.
- Segmentación: primer paso para clasificar el contenido de una imagen.
- Extracción de rasgos: sobre cada región segmentada se hace una selección de características.
- Clasificación: con las características obtenidas en el paso anterior, se procede a elegir

Tabla 1

Técnicas utilizadas en la clasificación de cultivos distribuidas en función de las etapas que se involucran en los sistemas automáticos para la clasificación de cultivos

Etapas	Técnicas	Docs. consultados	%	Referencias (sección de fuentes)
Fuente de información	Radar	11/32	34.3	[3], [4], [10], [15], [16], [17], [18], [19], [20], [21] y [22]
	Multiespectral	14/32	43.7	[5], [6], [9], [12], [13], [14], [23], [24], [25], [26], [27], [28], [29] y [30]
	Hiperespectral	7/32	21.8	[5], [7], [8], [31], [32], [33] y [34]
Extracción de características	Texturas	4/32	12.5	[6], [13], [23] y [21]
	Otras transformaciones	3/32	9.3	[11], [12] y [24]
	Pixel a pixel	5/32	15.6	[9], [10], [14], [26] y [30]
Algoritmos de clasificación	Máxima verosimilitud	9/32	28.1	[8], [10], [13], [22], [25], [27], [28], [34] y [35]
	Árbol de decisión	4/32	12.5	[10], [14], [26] y [35]
	Cómputo inteligente	11/32	34.3	[5], [9], [10], [12], [27], [28], [29], [31], [32], [33] y [34]

el tipo de algoritmo que se va a utilizar para el desarrollo de esta etapa. Entre la gran variedad que existe, uno de los más populares son las redes neuronales artificiales. Este tipo de técnicas usualmente requieren de un periodo de aprendizaje donde se ajustan los modelos neuronales de tal forma que aprenden de un conjunto de muestras representativas de los objetos a clasificar. Después, en una fase de validación, dichos modelos son estimulados con información desconocida, correspondiente a alguno de los objetos a clasificar.

En las siguientes secciones se describirán tres de las técnicas utilizadas en cada etapa de la metodología propuesta en este trabajo.

Espacios de color

El color de un objeto se puede observar gracias al rango de longitudes de onda de la luz que éste refleja. Al capturar una imagen, el dispositivo que lo hace guarda valores con respecto a un espacio o modelo de color, que es una especificación estándar de los colores de una imagen por medio de coordenadas, donde un color se puede representar por medio de varios valores y que está contenido en un punto de ese plano cartesiano. Hay diversos tipos de modelos que toman en cuenta diferentes factores para especificar los colores.

El espacio estándar es el RGB, el cual define al rojo, verde y azul como colores primarios. Cada

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

pixel de una imagen está compuesto de información asociada a estos canales de color. El RGB es muy sensible a los cambios de iluminación. En ese sentido, existen otros espacios que permiten separar la componente de saturación y luminosidad de cada píxel. Entre los modelos más populares con estas características, podemos mencionar el HSV, HSI, HSL, LAB y LUV (ver figura 1).

Figura 1

Ejemplo de tres espacios de color: HSI, HSL y LAB

Por ser un tema muy extenso, en esta sección sólo se presentarán las ecuaciones con las que se puede transformar el espacio RGB a HSI. El HSL—donde cada componente representa el matiz, saturación y luminosidad— es muy similar al HSB y HSI. Para profundizar en este tema, recomendamos al lector consultar a Gonzalez y Woods (2002).

Las ecuaciones (1), (2) y (3) permiten transformar una imagen en un espacio de color RGB a uno HSI:

$$H = \begin{cases} 0, & \text{si } MAX = MIN \\ \left(60 \times \dfrac{B-R}{MAX-MIN} + 360\right) \bmod 360, & \text{si } MAX = R \\ 60 \times \dfrac{G-B}{MAX-MIN} + 120, & \text{si } MAX = G \\ 60 \times \dfrac{R-G}{MAX-MIN} + 240, & \text{si } MAX = B \end{cases} \quad (1)$$

$$L = \frac{(MAX + MIN)}{2} \quad (2)$$

$$S = \begin{cases} 0, & \text{si } MAX = MIN \\ \dfrac{MAX-MIN}{2L}, & \text{si } L \leq \dfrac{1}{2} \\ \dfrac{MAX-MIN}{2-2L}, & \text{si } L \leq \dfrac{1}{2} \end{cases} \quad (3)$$

donde R indica el nivel de gris que tiene el pixel en la componente de color R; G, el que tiene el pixel en la componente G y B, el que tiene el pixel en la componente B; MAX equivale al valor máximo de los valores (R,G,B) y MIN, al valor mínimo de los valores (R,G,B).

Matriz de co-ocurrencia

La textura es una de las características más importantes usadas en la identificación de objetos o regiones de interés en una imagen. Su cálculo se puede realizar empleando estadísticas de 1.er orden (media, desviación estándar, varianza) y de 2.º basadas en la matriz de co-ocurrencia. En Haralick *et al*. (1973), los autores desarrollaron un conjunto de 14 medidas de textura, sustentadas en la dependencia espacial de los tonos de gris, las cuales se calculan a partir de esta matriz.

Ésta describe la frecuencia de un nivel de gris que aparece en una relación espacial específica con otro valor de gris dentro del área de una ventana determinada. La matriz de co-ocurrencia considera la relación espacial entre dos pixeles, llamados de referencia y vecino. Más detalles sobre su cálculo los puede encontrar en Haralick *et al*. (1973).

Una vez que se obtiene la matriz, es necesario aplicar un proceso de normalización —donde la resultante indica la probabilidad de que una relación exista— mediante la ecuación (4):

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} V_{i,j}} \quad (4)$$

donde cada elemento de la matriz $V_{i,j}$ es el número de veces que la relación i,j se presenta en la imagen o región de interés.

Sobre esta matriz de probabilidades se pueden calcular 14 medidas de textura, dentro de las cuales podemos mencionar las siguientes: homogeneidad, contraste, disimilaridad, media, desviación estándar, entropía, correlación, segundo momento angular, energía, etc.; ver ecuaciones (5)-(12).

$$c1 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2} \quad (5)$$

$$c2 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{i,j}\,(i-j)^2 \quad (6)$$

$$c3 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{i,j}\,|i-j| \quad (7)$$

$$c4 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} iP_{i,j}\,,\; c5 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} jP_{i,j} \quad (8)$$

$$c6 = \sqrt[2]{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{i,j}\,(i-c4)^2},\; c7$$

$$= \sqrt[2]{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{i,j}\,(i-c5)^2} \quad (9)$$

$$c8 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} -P_{i,j}In\left(P_{i,j}\right) \quad (10)$$

$$c9 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{i,j}\left[\frac{(i-c4)(i-c5)}{\sqrt[2]{(\sigma_i^2)(\sigma_{ij}^2)}}\right] \quad (11)$$

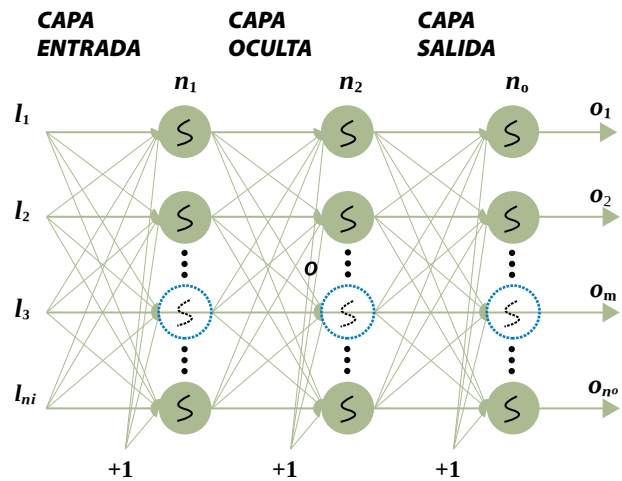
$$c10 = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{i,j}^2 \quad (12)$$

Redes neuronales de perceptrones

En la figura 2 se muestra una red perceptrón multicapa (MLP), la cual es una extensión de un perceptrón simple y uno de los modelos más utilizados debido a su facilidad de adaptación a varias aplicaciones prácticas (Hu e Hirasawa, 2002; Baomin y Bingjing, 1993; Vicen-Bueno *et al*., 2005; Hontoria, 1994). Aunado a esto, existen estudios que demuestran que una MLP constituye un aproximador universal de funciones (Funahashi, 1989; Hornik *et al*., 1989).

Figura 2

Arquitectura de una red de perceptrones



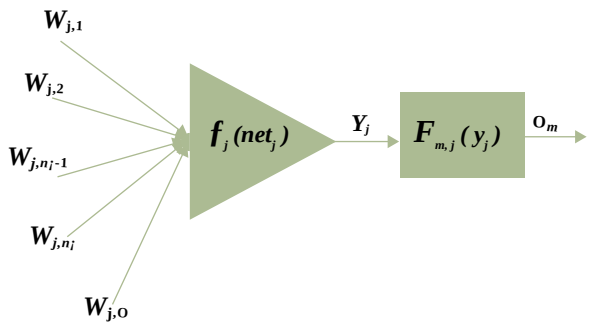
De manera general, una MLP está definida por: a) la estructura de la red, b) las funciones de activación y c) el algoritmo de aprendizaje. De estos elementos, destacaremos el último, ya que el aprendizaje es un proceso fundamental de todas las redes neuronales artificiales y el ajuste de todos sus parámetros en busca de un objetivo común (Haykin, 1999).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

El elemento esencial de la MLP está dado por el perceptrón que se muestra en la figura 3.

Figura 3

Arquitectura de un perceptrón



Así, ***la salida de la neurona j se obtiene mediante la ecuación*** (13):

$$y_j = f_j(net_j) \tag{13}$$

donde f_j es la función de activación de la neurona j y el valor de net_j es la suma de las entradas debidamente ponderadas a la neurona j definida por la ecuación (14):

$$net_j = \sum_{i=1}^{ni} w_{j,i} y_{j,i} + w_{j,0} \tag{14}$$

donde $y_{j,i}$ es la i-ésima entrada al nodo de la neurona j debidamente ponderada por el peso $w_{j,i}$ y $w_{j,0}$ es el peso del *bias* de la neurona j.

Si consideramos que los pesos se inician de manera aleatoria y que durante el aprendizaje la MLP buscará los pesos ideales que le permitan realizar la tarea especificada, entonces el proceso se convierte en iterativo, donde el desempeño de la red está en función del error, obtenido de la señal deseada y la salida actual; así, la búsqueda tiene como fin encontrar los pesos que hagan menor a la función de error (Munakata, 2008).

Para llevar a cabo esta tarea de minimización, se utiliza ampliamente el algoritmo supervisado de gradiente descendiente g, donde el aprendizaje tiene como meta llegar a un mínimo global de la función de error; esto se encuentra definido en la ecuación (15):

$$g = \frac{\partial E(x,w)}{\partial w} = \left[\frac{\partial E}{\partial w_1} \; \frac{\partial E}{\partial w_2} \cdots \frac{\partial E}{\partial w_N}\right]^T \tag{15}$$

donde x es el vector de entrada a la red; w, el vector de pesos y E, la función de error total.

La regla de actualización de pesos queda como se indica en la ecuación (16):

$$w_{k+1} = w_k - \alpha g_k \tag{16}$$

donde a es la constante de aprendizaje.

Este algoritmo, también llamado de propagación hacia atrás de errores o *retropropagación* (EBP), tiene fallos bien conocidos ya que converge de forma muy lenta y puede llegar a detenerse en un mínimo local de la función de error aun cuando éste es muy grande. Como solución a la convergencia lenta del EBP, se emplea el algoritmo de Gauss-Newton, que usa las derivadas de 2.° orden de la función de error para inicialmente modificar la regla de actualización de pesos, como se ve en la ecuación (17):

$$w_{k+1} = w_k - H_k^{-1} g_k \tag{17}$$

donde H es la matriz hessiana.

Al introducir la matriz jacobiana J, para reducir la complejidad del cálculo y considerando a e como el vector de error, tenemos finalmente la ecuación (18):

$$w_{k+1} = w_k - \left(J_k^T J_k\right)^{-1} (J_k e_k) \tag{18}$$

La ventaja de este algoritmo es su rápida convergencia, pero es inestable, es decir, puede no converger y alejarse de la solución deseada. Como alternativa a esta nueva problemática, surge el algoritmo basado en el método Levenberg-Marquardt, que aproxima a la matriz hessiana como se indica en la ecuación (19):

$$\boldsymbol{H \approx J^T J + \mu I} \quad (19)$$

donde $\boldsymbol{\mu}$ es una constante siempre positiva, llamada coeficiente de combinación, e $\boldsymbol{I}$ es la matriz identidad.

Esta aproximación nos provoca que la regla de actualización de pesos quede expresada como se indica en la ecuación (20):

$$\boldsymbol{w_{k+1} = w_k - \left(J_k^T J_k + \mu I\right)^{-1}(J_k e_k)} \quad (20)$$

Esta regla es una combinación de los dos algoritmos anteriores, es decir, cuando la constante $\boldsymbol{\mu}$ es muy pequeña se aprovecha la velocidad de convergencia del algoritmo de Gauss-Newton, y cuando $\boldsymbol{\mu}$ es muy grande el EBP asegura la estabilidad (Haykin, 1999); por ello, se decidió emplear este algoritmo de adaptación de pesos en el estudio.

Metodología propuesta

En esta investigación se realizó un análisis sobre diferentes algoritmos que existen en la literatura para realizar la extracción de rasgos sobre regiones de interés en una imagen y distintos algoritmos inteligentes para llevar a cabo la clasificación de patrones, esto con el fin de proponer una metodología que permita clasificar cultivos agrícolas temporales en imágenes de baja resolución con bandas de información en el espectro visible y estimar la superficie sembrada.

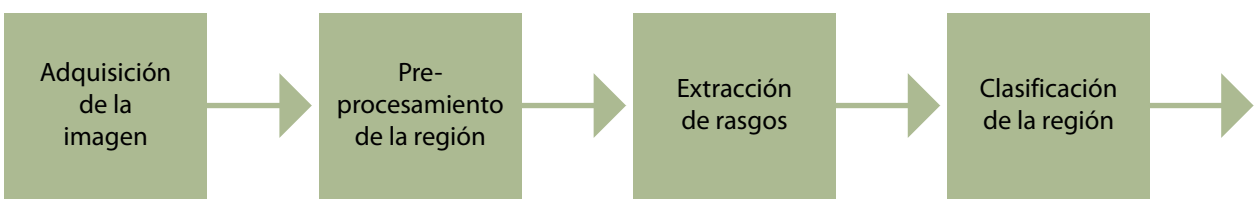
Este trabajo se dividió en cuatro etapas: la primera estuvo dedicada a la adquisición de la imagen, la segunda se enfocó a su preprocesamiento, la tercera se dedicó a aplicar los algoritmos de extracción de rasgos y la última estuvo orientada al entrenamiento de diferentes algoritmos inteligentes que permitieran clasificar los rasgos descriptivos obtenidos en la primera etapa (ver figura 4).

Las fuentes más comunes para obtener la información son imágenes provenientes de tres tipos de sensores ubicados en satélites: el primero es el Radar de Apertura Sintética (SAR, por sus siglas en inglés), el segundo se trata de multiespectrales que se caracterizan por tener bandas con rangos de 100 nanómetros y el tercero son imágenes hiperespectrales, las cuales pueden contener información separada en bandas con rangos de 10 nanómetros, por poner un ejemplo.

En el caso particular de esta investigación, se utilizó una imagen de una región de prueba del estado de Sinaloa, México, obtenida de *GoogleEarth* del tipo multiespectral tomando en cuenta sólo el espectro visible.

Figura 4

Etapas que componen a un sistema automático para la clasificación de cultivos agrícolas



INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Para poder llevar a cabo la extracción de características, fue necesario aplicar una transformación del espacio de color para posteriormente usar un proceso de segmentación de la información.

Durante la etapa de segmentación se etiquetaron de forma manual diferentes regiones de cultivos que estaban presentes dentro de la imagen generando áreas de interés asociadas con los distintos cultivos que se desea clasificar. Una de las técnicas más recurrentes para la segmentación de la información es definir polígonos o zonas en las imágenes indicando el cultivo al cual pertenece cada uno de ellos.

Una vez que las regiones se tienen identificadas, procedemos a calcular un conjunto de rasgos descriptivos con el fin de tener un vector con características que describan de forma numérica el tipo de cultivo que aparece en la región (ver figura 5). El objetivo de esta etapa es encontrar un método de extracción de rasgos que ayude a representar de mejor manera a los patrones pertenecientes a un cultivo, de tal forma que los que pertenecen al cultivo *A* sean muy similares con otros miembros de la misma clase y muy diferentes con aquellos que pertenecen a un cultivo *B*. Dentro de esta investigación se evaluó el desempeño de tres técnicas de extracción de rasgos: matriz de co-ocurrencia, transformada de Fourier y transformada wavelet.

Por último, para llevar a cabo el proceso de clasificación, un porcentaje de los vectores de rasgos son utilizados para entrenar tanto diferentes arquitecturas de redes neuronales de perceptrones como clasificadores basados en el cálculo de la distancia; el porcentaje restante es utilizado para validar el desempeño del clasificador.

Figura 5

Extracción de un vector de rasgos sobre una región de la imagen definida por un polígono

$$x = \begin{bmatrix} 0.21 \\ 1.23 \\ 0.67 \\ \vdots \\ 2.12 \end{bmatrix}$$

Resultados experimentales

Para validar la eficiencia de la metodología propuesta, de la región de prueba en Sinaloa se obtuvo la imagen satelital y, posteriormente, se aplicó un etiquetado manual para detectar cinco cultivos diferentes y se generaron 24 bases de datos compuestas por 2 752 patrones correspondientes a cinco clases de éstos. Cada base se construyó a partir de la combinación de las distintas técnicas de extracción de rasgos (matriz de co-ocurrencia, transformada de Fourier y transformada wavelet) y espacios de color utilizados (CMYK, RGB, HSV, HSI, HSL, LAB, LUV y XYZ).

Por otro lado, con el propósito de evaluar la conveniencia de utilizar una red neuronal artificial en este problema, evaluamos el desempeño de cinco técnicas de clasificación: clasificador por distancia mínima (*CityBlock*, Euclidiana, Minkowski) y redes neuronales artificiales o redes neuronales multicapa (FNN) 1 y 2.

Para poder entrenar los clasificadores y las redes neuronales, particionamos las bases de datos en dos conjuntos: uno compuesto por 50% de la información para entrenar las diferentes técnicas de clasificación y el restante para validar el desempeño.

Además, para poder comparar los resultados generados con las distintas configuraciones, utilizamos una métrica ponderada que integra el porcentaje de clasificación obtenido durante la etapa de entrenamiento (pC) y el de clasificación que se consiguió en la fase de prueba (pV) definida por la ecuación (21):

$$\boldsymbol{pT = 0.4pC + 0.6pV} \qquad (21)$$

donde el porcentaje de clasificación se obtiene a partir de la relación: patrones bien clasificados entre patrones analizados.

Para el caso de las redes neuronales artificiales multicapa, se diseñaron dos tipos de arquitecturas:

FNN1, compuesta por una capa oculta y FNN2, por dos capas ocultas; el algoritmo de entrenamiento utilizado fue el de Levenberg-Marquardt con un factor de aprendizaje de 0.1 y 2 mil épocas de entrenamiento.

En la tabla 2 se muestra el desempeño promedio de la metodología propuesta con las diferentes combinaciones que se evaluaron. De esta información, podemos hacer tres tipos de análisis: sobre los espacios de color, acerca de la técnica de extracción y del algoritmo de clasificación.

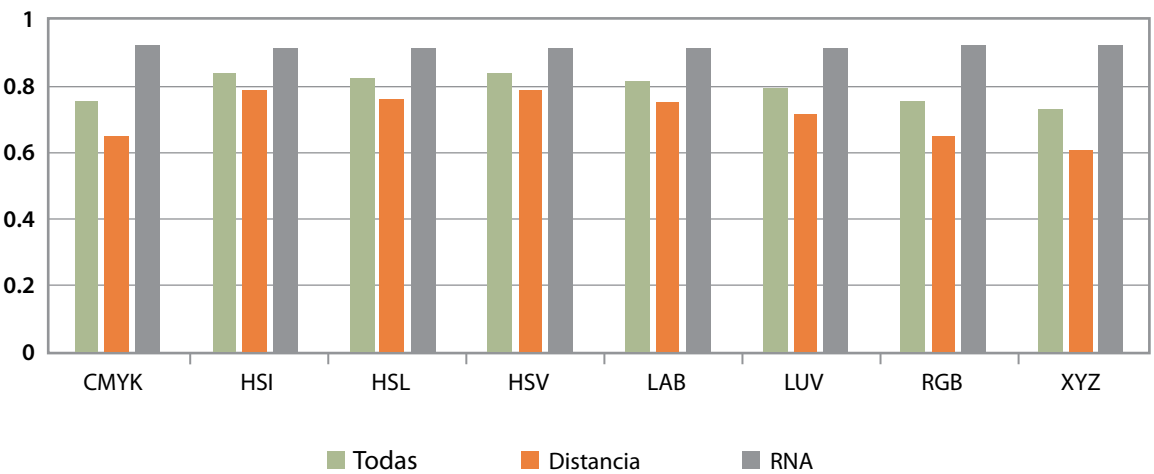
En cuanto al primero, mencionamos que, como media, el HSI fue el que mejores resultados otorgó, alcanzando una eficiencia de clasificación promedio de todos los clasificadores de 83.8%; si separamos la información y analizamos la influencia del espacio de color en los clasificadores por distancia mínima, observamos que con el HSI se obtiene el mejor porcentaje promedio de clasificación, logrando 78.93; en cuanto a la influencia del espacio de color usando redes neuronales, observamos que el mejor desempeño se obtiene con el CMYK, con un porcentaje de eficiencia promedio de 91.98. Para mayor referencia, ver la gráfica 1.

En cuanto a las técnicas de extracción de rasgos, podemos observar que la matriz de co-ocurrencia otorgó una eficiencia de 87.1% en promedio para todos los clasificadores. Si nos enfocamos sólo a los basados en la distancia, la matriz de co-ocurrencia dio una eficiencia media de 81.9%; por último, para el caso de las redes neuronales artificiales, la matriz de co-ocurrencia proporcionó una de 97.7%; en este caso, la matriz de co-ocurrencia es la que mejores resultados otorgó para los diferentes clasificadores. Para mayor referencia, ver la gráfica 2.

Por otro lado, en cuanto a las técnicas de clasificación, también observamos que las redes neuronales artificiales son las que mejores resultados presentaron. En particular, las que cuentan con una capa lograron alcanzar una eficiencia promedio de 92%; es importante mencionar que una red neuronal combinada con la matriz de co-ocurrencia en el espacio de color RGB logró el mejor porcentaje promedio, alcanzando un desempeño de 98.2; en la gráfica 3 se muestra el desempeño medio de todos los clasificadores usados, combinados con las diferentes técnicas de extracción de rasgo y espacios de color.

Gráfica 1

Desempeño promedio obtenido con diferentes espacios de color utilizando diferentes técnicas de clasificación

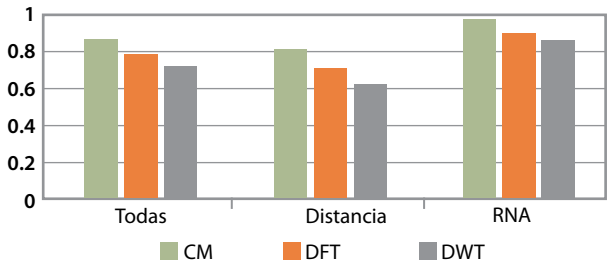


Nota: el comportamiento promedio de todos los clasificadores se puede observar en las barras etiquetadas como *Todas*; el de sólo los clasificadores por distancia, en las de *Distancia* y, finalmente, el de las redes neuronales, en las etiquetadas como *RNA*.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 2

Desempeño promedio obtenido con las diferentes técnicas de extracción de rasgos utilizando diferentes técnicas de clasificación



Nota: el comportamiento promedio de todos los clasificadores se puede observar en las barras etiquetadas como *Todas*; el de sólo los clasificadores por distancia, en las de *Distancia* y, finalmente, el de las redes neuronales, en las etiquetadas como *RNA*.

Gráfica 3

Desempeño promedio obtenido con las diferentes técnicas de clasificación

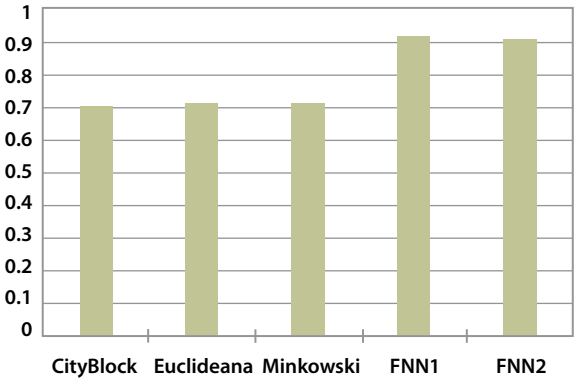


Tabla 2

Continúa

Eficiencia en términos de porcentaje en la clasificación de cultivos agrícolas utilizando diferentes técnicas de clasificación, espacios de color y técnicas de extracción de rasgos

		Clasificadores por distancia mínima			Redes neuronales artificiales	
		CityBlock	Euclideana	Minkowski	FNN1	FNN2
CMYK	**CM**	0.716±0.008	0.723±0.004	0.721±0.006	0.980±0.001	0.981±0.004
	DFT	0.632±0.006	0.655±0.007	0.654±0.007	0.917±0.007	0.899±0.020
	DWT	0.563±0.006	0.579±0.006	0.579±0.008	0.877±0.009	0.865±0.015
HSI	**CM**	0.867±0.003	0.872±0.004	0.873±0.004	0.975±0.004	0.977±0.005
	DFT	0.783±0.003	0.782±0.004	0.780±0.004	0.909±0.008	0.892±0.017
	DWT	0.722±0.005	0.713±0.005	0.712±0.006	0.867±0.005	0.855±0.023
HSL	**CM**	0.838±0.008	0.846±0.006	0.846±0.004	0.978±0.005	0.977±0.003
	DFT	0.741±0.003	0.750±0.005	0.751±0.004	0.904±0.005	0.888±0.021

Tabla 2

Concluye

Eficiencia en términos de porcentaje en la clasificación de cultivos agrícolas utilizando diferentes técnicas de clasificación, espacios de color y técnicas de extracción de rasgos

		Clasificadores por distancia mínima			Redes neuronales artificiales	
		CityBlock	Euclideana	Minkowski	FNN1	FNN2
	DWT	0.696±0.005	0.679±0.005	0.679±0.003	0.873±0.010	0.864±0.020
HSV	**CM**	0.863±0.005	0.873±0.006	0.872±0.007	0.978±0.003	0.976±0.005
	DFT	0.787±0.004	0.781±0.003	0.782±0.004	0.911±0.005	0.882±0.009
	DWT	0.717±0.004	0.713±0.004	0.715±0.003	0.871±0.007	0.847±0.022
LAB	**CM**	0.822±0.007	0.827±0.005	0.829±0.008	0.975±0.004	0.974±0.003
	DFT	0.765±0.004	0.766±0.003	0.768±0.004	0.911±0.005	0.896±0.015
	DWT	0.666±0.006	0.652±0.004	0.648±0.005	0.870±0.010	0.852±0.023
LUV	**CM**	0.766±0.006	0.800±0.007	0.800±0.008	0.976±0.004	0.974±0.004
	DFT	0.725±0.005	0.737±0.002	0.737±0.003	0.906±0.007	0.885±0.017
	DWT	0.618±0.004	0.607±0.005	0.613±0.006	0.867±0.009	0.857±0.017
RGB	**CM**	0.736±0.005	0.750±0.004	0.750±0.006	0.982±0.004	0.979±0.006
	DFT	0.619±0.004	0.642±0.009	0.641±0.009	0.923±0.005	0.907±0.021
	DWT	0.535±0.008	0.572±0.016	0.564±0.015	0.870±0.007	0.863±0.016
XYZ	**CM**	0.730±0.008	0.738±0.005	0.740±0.004	0.982±0.003	0.981±0.003
	DFT	0.579±0.008	0.585±0.006	0.590±0.006	0.918±0.006	0.903±0.018
	DWT	0.481±0.008	0.495±0.012	0.503±0.008	0.875±0.006	0.862±0.017

Tabla 3

Desempeños de clasificación obtenidos en trabajos de la literatura en función del tipo de imágenes que utilizan para realizar la detección de los cultivos

Referencia del trabajo	Tipo de información	Desempeño de clasificación	Número de cultivos	Características generales
McNairn *et al.*, 2009	Información de radar	70%	4	Información polarimétrica
		80%	4	Datos temporales
		88.7%	4	Información multifrecuencias
Del Frate *et al.*, 2003		85%	3	Información polarimétrica
		80%-88%	3	Información polarimétrica y multifrecuencia
Senthilnath *et al.*, 2011	Información multiespectral	78%	-	Uso de PCA y HAIS
Gomez-Chova *et al.*, 2003		95.21%	-	*Sequential float feature selection*
de Castro *et al.*, 2012		98.1%-100%	2	Imágenes multiespectrales e hiperespectrales

Tabla 4

Desempeños de clasificación obtenidos en trabajos de la literatura en función de la técnica de extracción de rasgos utilizada para realizar la detección de los cultivos

Referencia del trabajo	Estrategia de extracción de rasgos	Desempeño de clasificación	Número de cultivos	Características generales
Yi *et al.*, 2008	Textura	60.72%	4	Matrices de covarianza
		88.26%	4	Información a nivel de pixel
		88.94%	4	Combinación de matriz de covarianza y nivel de pixel
Doraiswamy *et al.*, 2007	Otras transformaciones	75%-82%	2	Transformación NDVI

Tabla 5

Desempeños de clasificación obtenidos en trabajos de la literatura en función de la técnica de clasificación utilizada para realizar la detección de los cultivos

Referencia del trabajo	Estrategia de clasificación	Desempeño de clasificación	Número de cultivos	Características generales
Camps-Valls *et al.*, 2003	Cómputo inteligente	94.1%-95-53%	-	Máquinas de soporte vectorial
Omkar *et al.*, 2009		98%-99.56%	4	Técnicas de inteligencia de enjambre
Eddy *et al.*, 2006		92.97%	3	Redes neuronales multicapa
Bairagi y Hassan, 2002	Máxima verosimilitud	75%-82%	3	-

Finalmente, en las tablas 3, 4 y 5 se presenta una comparativa contra los resultados obtenidos en diferentes estudios reportados en la literatura. Es importante mencionar que al no existir una base de datos común entre todos los trabajos, se intentó seleccionar un conjunto de éstos en los cuales se reportan resultados donde se clasifican de tres a cinco cultivos diferentes, buscando equiparar los alcances de esta investigación en cuanto el número de cultivos clasificados.

De igual forma, hacemos notar que el siguiente grupo de resultados fueron realizados sobre diferentes tipos de imágenes SAR, multiespectral e hiperespectral, lo cual es de suma importancia para esta comparativa, ya que en este trabajo sólo se utilizó información de las bandas de espectro visible.

Como podemos ver, si comparamos lo obtenido con la metodología propuesta en esta investigación contra los resultados logrados con otro tipo de imágenes, se puede observar que son superiores, aun cuando sólo utilizamos imágenes con bandas en el espectro visible. En cuanto a la técnica de extracción de rasgos, lo que se alcanzó al aplicar la matriz de co-ocurrencia fue superior comparado contra los resultados que se muestran en otros trabajos de la literatura.

Respecto a los algoritmos de clasificación, se puede observar que todos los trabajos que aplican técnicas de cómputo inteligente generan resultados comparables a aquellos obtenidos con la metodología propuesta.

Conclusiones

En esta investigación se pudo mostrar cómo es posible realizar la clasificación de cultivos agrícolas a partir de imágenes de baja resolución y con bandas de información en el espectro visible.

Para lograrlo, se propuso utilizar un enfoque sustentado en técnicas del área de reconocimiento de patrones, procesamiento de imágenes y cómputo inteligente empleando diferentes técnicas de extracción de rasgos sustentadas en el análisis de texturas y técnicas de cómputo inteligente basadas en redes neuronales artificiales.

Debido a que sólo se usaron imágenes con bandas de información en el espectro visible, la transformación a otros espacios de color cobró un papel muy relevante: el HSI es el que mejores resultados otorgó para los clasificadores basados en distancias y el RGB, para los que se sustentan en redes neuronales artificiales.

En cuanto a la técnica de extracción de rasgos, se observó que el cálculo de texturas en función de la matriz de co-ocurrencia proporcionó los mejores resultados para todos los clasificadores.

En particular, las redes neuronales artificiales con una capa lograron alcanzar una eficiencia promedio de 92%; es importante mencionar que una red neuronal combinada con la matriz de co-ocurrencia en el espacio de color RGB logró el mejor porcentaje promedio, alcanzando un desempeño de 98.2.

Fuentes

[1] Zhong, L., T. Hawkins, K. Holland, P. Gong y G. Biging. "Satellite imagery can support water planning in the Central Valley", en: *Calif. Agric*. Vol. 63, núm. 4, octubre de 2009, pp. 220-224.

[2] Czerepowicz, L., B. S. Case y C. Doscher. "Using satellite image data to estimate aboveground shelterbelt carbon stocks across an agricultural landscape", en: *Agric. Ecosyst. Environ*. Vol. 156, agosto de 2012, pp. 142-150.

[3] Skriver, H., M. T. Svendsen y A. G. Thomsen. "Multitemporal C- and L-band polarimetric signatures of crops", en: *IEEE Trans. Geosci. Remote Sens*. Vol. 37, núm. 5, septiembre de 1999, pp. 2413-2429,

[4] Schotten, C. G. J., W. W. L. V. Rooy y L. L. F. Janssen. "Assessment of the capabilities of multi-temporal ERS-1 SAR data to discriminate between agricultural crops", en: *Int. J. Remote Sens*. Vol. 16, núm. 14, sepiembre de 1995, pp. 2619-2637.

[5] de Castro, A. I., M. Jurado-Expósito, M. T. Gómez-Casero y F. López-Granados. "Applying Neural Networks to Hyperspectral and Multispectral Field Data for Discrimination of Cruciferous Weeds in Winter Crops", en: *Sci. World J*. Vol. 2012, mayo de 2012, p. e630390.

[6] Smith, G. M. y R. M. Fuller. "An integrated approach to land cover classification: An example in the Island of Jersey", en: *Int. J. Remote Sens*. Vol. 22, núm. 16, enero de 2001, pp. 3123-3142.

[7] Senthilnath, J., S. N. Omkar, V. Mani y N. Karnwal. "Hierarchical artificial immune system for crop stage classification", en: *2011 Annual IEEE India Conference (INDICON)*, 2011, pp. 1-4.

[8] Gomez-Chova, L., J. Calpe, G. Camps-Valls, J. D. Martin, E. Soria, J. Vila, L. Alonso-Chorda y J. Moreno. "Feature selection of hyperspectral data through local correlation and SFFS for crop classification", en: *Geoscience and Remote Sensing Symposium, 2003. IGARSS '03. Proceedings. 2003 IEEE International*. Vol. 1, 2003, pp. 555-557.

[9] Shao, Y. y R. S. Lunetta. "Comparison of sub-pixel classification approaches for crop-specific mapping", en: *2009 17th International Conference on Geoinformatics*, 2009, pp. 1-4.

[10] Chakraborty, M. y S. Panigrahy. "Comparative performance of per-pixel classifiers using ers-1 sar data for classification of rice crop", en: *J. Indian Soc. Remote Sens*. Vol. 25, núm. 3, septiembre de 1997, pp. 155-161,

[11] Nejati, H., Z. Azimifar y M. Zamani. "Using fast fourier transform for weed detection in corn fields", en: *IEEE International Conference on Systems, Man and Cybernetics, 2008. SMC 2008*. 2008, pp. 1215-1219.

[12] Kiani, S., Z. Azimifar y S. Kamgar. "Wavelet-based crop detection and classification", en: *2010 18th Iranian Conference on Electrical Engineering (ICEE)*. 2010, pp. 587-591.

[13] Dean, A. M. y G. M. Smith. "An evaluation of per-parcel land cover mapping using maximum likelihood class probabilities", en: *Int. J. Remote Sens*. Vol. 24, núm. 14, enero de 2003, pp. 2905-2920.

[14] Fries, R. S. D., M. Hansen, J. R. G. Townshend y R. Sohlberg. "Global land cover classifications at 8 km spatial resolution: The use of training data derived from Landsat imagery in decision tree classifiers", en: *Int. J. Remote Sens*. Vol. 19, núm. 16, enero de 1998, pp. 3141-3168.

[15] Skriver, H., F. Mattia, G. Satalino, A. Balenzano, V. R. N. Pauwels, N. E. C. Verhoest y M. Davidson. "Crop Classification Using Short-Revisit Multitemporal SAR Data", en: *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. Vol. 4, núm. 2, junio de 2011, pp. 423-431.

[16] Skriver, H. "Crop Classification by Multitemporal C- and L-Band Single- and Dual-Polarization and Fully Polarimetric SAR", en: *IEEE Trans. Geosci. Remote Sens*. Vol. 50, núm. 6, junio de 2012, pp. 2138-2149.

[17] McNairn, H., J. Shang, C. Champagne y X. Jiao. "TerraSAR-X and RADARSAT-2 for crop classification and acreage estimation", en: *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*. Vol. 2, 2009, pp. II-898-II-901.

[18] McNairn, H., J. Shang, X. Jiao y C. Champagne. "The Contribution of ALOS PALSAR Multipolarization and Polarimetric Data to Crop Classification", en: *IEEE Trans. Geosci. Remote Sens*. Vol. 47, núm. 12, diciembre de 2009, pp. 3981-3992.

[19] Del Frate, F., G. Schiavon, D. Solimini, M. Borgeaud, D. H. Hoekman y M. A. M. Vissers. "Crop classification using multiconfiguration C-band SAR data", en: *IEEE Trans. Geosci. Remote Sens*. Vol. 41, núm. 7, julio de 2003, pp. 1611-1619.

[20] Skriver, H., M. T. Svendsen, F. Nielsen y A. Thomsen. "Crop classification by polarimetric SAR", en: *Geoscience and Remote Sensing Symposium, 1999. IGARSS '99 Proceedings. IEEE 1999 International.* Vol. 4, 1999, pp. 2333-2335.

[21] Pingxiang, L. y F. Shenghui. "SAR image classification based on its texture features", en: *Geo-Spat. Inf. Sci.* Vol. 6, núm. 3, septiembre de 2003, pp. 16-19.

[22] Bairagi, G. D. y Z.-U. Hassan. "Wheat crop production estimation using satellite data", en: *J. Indian Soc. Remote Sens.* Vol. 30, núm. 4, diciembre de 2002, pp. 213-219.

[23] Yi, C., Y. Pan y J. Zhang. "An Integrated Approach to Agricultural Crop Classification Using SPOT5 HRV Images", en: *Computer And Computing Technologies In Agriculture*: Vol. I, US, Springer, 2008, pp. 677-684.

[24] An, Q., W. Gao, B. Yang, J. Wu, L. Yu y Z. Liu. "Research on Feature Selection Method Oriented to Crop Identification Using Remote Sensing Image Classification", en: *Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09*. Vol. 5, 2009, pp. 426-432.

[25] Sheikho, K. M., A. M. Abu Mouti, T. Yoshie y F. Al-Qurnas. "Crops classification using multiple Landsat data: a case study in arid lands", en: *Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS '98. 1998 IEEE International*. Vol. 2, 1998, pp. 794-797.

[26] Doraiswamy, P. C., A. J. Stern y B. Akhmedov. "Crop classification in the U.S. Corn Belt using MODIS imagery", en: *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*. 2007, pp. 809-812.

[27] Omkar, S. N., J. Senthilnath, D. Mudigere y M. M. Kumar. "Crop classification using biologically-inspired techniques with high resolution satellite image", en: *J. Indian Soc. Remote Sens*. Vol. 36, núm. 2, marzo de 2009, pp. 175-182.

[28] Vatsavai, R. R., E. Bright, C. Varun, B. Budhendra, A. Cheriyadat y J. Grasser. "Machine Learning Approaches for High-resolution Urban Land Cover Classification: A Comparative Study", en: *Proceedings of the 2Nd International Conference on Computing for Geospatial Research & Applications*. New York, NY, USA, 2011, pp. 11:1-11:10.

[29] Kuncheva, L. I., J. C. Bezdek y R. P. W. Duin. "Decision templates for multiple classifier fusion: an experimental comparison", en: *Pattern Recognit*. Vol. 34, núm. 2, febrero de 2001, pp. 299-314.

[30] Doraiswamy, P. C., B. Akhmedov y A. J. Stern. "Improved Techniques for Crop Classification using MODIS Imagery", en: *IEEE International Conference on Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006*. 2006, pp. 2084-2087.

[31] Zhang, C. y F. Qiu. "Hyperspectral image classification using an unsupervised neuro-fuzzy system", en: *J. Appl. Remote Sens*. Vol. 6, núm. 1, 2012, pp. 063515.

[32] Senthilnath, J., S. N. Omkar, V. Mani, N. Karnwal y P. B. Shreyas. "Crop Stage Classification of Hyperspectral Data Using Unsupervised Techniques", en: *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. Vol. 6, núm. 2, abril de 2013, pp. 861-866.

[33] Camps-Valls, G., L. Gómez-Chova, J. Calpe-Maravilla, E. Soria-Olivas, J. D. Martín-Guerrero y J. Moreno. "Support Vector Machines for Crop Classification Using Hyperspectral Data", en: Perales, F. J., A. J. C. Campilho, N. P. de la Blanca y A. Sanfeliu (Eds.). *Pattern Recognition and Image Analysis*. Springer Berlin Heidelberg, 2003, pp. 134-141.

[34] Eddy, P. R., A. M. Smith, B. D. Hill, D. R. Peddle, C. A. Coburn y R. E. Blackshaw. "Comparison of Neural Network and Maximum Likelihood High Resolution Image Classification for Weed Detection in Crops: Applications in Precision Agriculture", en: *IEEE International Conference on Geoscience and Remote Sensing Symposium*, *2006. IGARSS 2006.* 2006, pp. 116-119.

[35] Friedl, M. A. y C. E. Brodley. "Decision tree classification of land cover from remotely sensed data", en: *Remote Sens*. *Environ*. Vol. 61, núm. 3, septiembre de 1997, pp. 399-409.

[36] Gonzalez, R. C. y R. E. Woods. *Digital Image Processing*. 2nd edition. Upper Saddle River, N.J., Prentice Hall, 2002.

[37] Haralick, R. M., K. Shanmugam e I. Dinstein. "Textural Features for Image Classification", en: *IEEE Trans. Syst. Man Cybern*. Vol. SMC-3, núm. 6, noviembre de 1973, pp. 610-621.

[38] Hu, J. y K. Hirasawa. "A method for applying multilayer perceptrons to control of nonlinear systems", en: *Proceedings of the 9th International Conference on Neural Information Processing*, *2002. ICONIP '02*. Vol. 3, 2002, pp. 1267-1271.

[39] Baomin, T. y D. Bingjing. "Multilayer perceptron structures applied to adaptive echo canceller for ISDN subscriber loops", en: *1993 IEEE Region 10 Conference on TENCON '93*. *Proceedings. Computer, Communication, Control and Power Engineering*. Vol. 3, 1993, pp. 613-616.

[40] Vicen-Bueno, R., R. Gil-Pita, M. Rosa-Zurera, M. Utrilla-Manso y F. López-Ferreras. "Multilayer Perceptrons Applied to Traffic Sign Recognition Tasks", en: Cabestany, J., A. Prieto y F. Sandoval (Eds.). *Computational Intelligence and Bioinspired Systems*. Springer Berlin Heidelberg, 2005, pp. 865-872.

[41] Hontoria, J. A. L. "An application of the multilayer perceptron: solar radiation maps in Spain", en: *Sol. Energy-Sol*. *ENERG*. Vol. 79, núm. 5, 1994.

[42] Funahashi, K. I. "On the approximate realization of continuous mappings by neural networks", en: *Neural Netw*. Vol. 2, núm. 3, 1989, pp. 183-192.

[43] Hornik, K., M. Stinchcombe y H. White. "Multilayer feedforward networks are universal approximators", en: *Neural Netw*. Vol. 2, núm. 5, 1989, pp. 359-366.

[44] Haykin, S. *Neural Networks*: A Comprehensive Foundation. Prentice Hall, 1999.

[45] Munakata, T. *Fundamentals of the New Artificial Intelligence: Neural, Evolutionary*, *Fuzzy and More*. Springer Science & Business Media, 2008.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Cambios recientes en la esperanza de vida en México, **análisis por medio de su descomposición**

César Bistrain Coronado

IAAF World Youth Championships Cali 2015 - Day 1/Patrick Smith/Getty Images

Nota: el autor agradece los comentarios de Virgilio Partida Bush.

La esperanza de vida está ligada a las condiciones físicas, sociales y culturales que intervienen en el desarrollo, por ende, la relevancia de analizar sus avances o retrocesos; sin embargo, para caracterizar las transformaciones en el patrón de la mortalidad, es necesario analizar el efecto diferencial de factores como los grupos etarios o las causas de muerte. Por ello, el presente artículo analiza las modificaciones en la esperanza de vida nacional por medio de la descomposición de éstos, lo que proporciona información para apoyar la producción de medidas que aumenten el bienestar de la población. La conclusión principal que se obtiene es que la esperanza de vida nacional ha mermado su incremento debido a una sobremortalidad por causa de agresiones en hombres jóvenes durante el periodo 2006-2011.

Palabras clave: esperanza de vida, descomposición, sobremortalidad, agresiones.

Recibido: 6 de mayo de 2014.
Aceptado: 5 de mayo de 2015.

Life expectancy is linked to the physical, social and cultural conditions involved in the development, hence the importance of analyzing their progress or setbacks. However, to characterize the changes in the pattern of mortality, it is necessary to analyze the differential effect of factors such as age groups or causes of death. Therefore, this article analyzes changes in the national life expectancy by decomposing them, providing information to support the production of measures to improve the welfare of the population. The main conclusion to be drawn is that the national life expectancy reduced its increase by an excess mortality due to aggressions in young men during the period 2006-2011.

Key words: life expectancy, decomposition, excess mortality, aggressions.

Cambios en la esperanza de vida

La mortalidad es un fenómeno aleatorio, y cualquier individuo está expuesto al riesgo de muerte desde su nacimiento. Es un suceso que puede ocurrir por un amplio número de causas, desde aquellas que posiblemente son postergadas o evitadas con ciertos cuidados durante la vida (enfermedades crónico-degenerativas), hasta las que surgen sin previsión (accidentes); sin embargo, también existen las provocadas por agresiones, que son acciones violentas relacionadas de manera estrecha con las condiciones de la sociedad que los experimenta, por lo que se pueden prevenir o fomentar desde su interior.

La aleatoriedad de la mortalidad implica una enorme dificultad para su análisis individual, por lo que hay técnicas aplicables a un conjunto de individuos con el fin de identificar patrones e indicadores que nos llevan a su mejor caracterización. Entre ellos tenemos la esperanza de vida, la cual constituye una de las herramientas demográficas que dan indicios claros acerca de la situación de una población en su territorio al estar ligada a sus condiciones físicas, sociales y culturales; para un momento, señala el número promedio de años que cualquier miembro del grupo vivió a partir de cierta edad, usualmente al nacimiento, pero es hipotético al ser imposible conocer con precisión el tiempo vivido por cada individuo; por ello, la esperanza de vida se construye a partir de una cohorte ficticia, la cual intenta recoger la experiencia de muerte de las cohortes que conviven en un momento específico.

Es un indicador que puede compararse entre distintas unidades geográficas; para todas las edades, resume la experiencia de muerte sin importar la causa que la provocó, por lo cual es útil para medir el nivel de la mortalidad de cualquier población sin importar sus condiciones; no obstante, a pesar de que su análisis temporal cuantifica la modificación de los años promedio de vida, no es aceptable para caracterizar el cambio de la mortalidad, pues dos poblaciones con diferente nivel que experimenten el mismo cambio relativo de mortalidad en cada edad sufren modificaciones absolutas y relativas distintas en la esperanza de vida (Arriaga, 1996, p. 10).

Por otro lado, mientras más cerca se esté de los límites biológicos conocidos, el aumento factible para la esperanza de vida es cada vez menor, ya que el carácter crónico y degenerativo de la morbilidad en esta etapa es un obstáculo para lograr mayores ganancias.

Ante esta situación, es indispensable profundizar en la descripción de los cambios en la esperanza de vida y, con ello, tener más posibilidades para aumentar sus niveles, por lo que es preciso aplicar técnicas de descomposición, las cuales parten del principio de separar los indicadores en componentes que contribuyan a la mejor comprensión del fenómeno. Éstas se utilizan para comparar variables demográficas que pertenecen a diferentes poblaciones o cuando se comparan variables de la misma población en distintos momentos (Canudas, 2003, p. 4).

Por lo anterior, el objetivo del presente documento es profundizar en el análisis de la esperanza de vida nacional por medio de su descomposición y así identificar la situación de cada grupo etario, además de posibles cambios en el perfil reciente de mortalidad, aspectos importantes que proporcionarán información para generar medidas útiles que mejoren las condiciones de la población.

Contribución al cambio en la esperanza de vida por grupos de edad y causa de muerte según sexo

Un aspecto fundamental al analizar la esperanza de vida es identificar el comportamiento diferencial que puede tener, o no, cada grupo de edad para obtener la ganancia neta. Esta situación depende de la estructura etaria de la población, así como de las condiciones de salud y de contexto, ya que, por ejemplo, acciones enfocadas a disminuir la mortalidad infantil en regiones donde sea alta necesariamente implicarán un importante aumento en la acumulación de los años-persona vividos y, en consecuencia, de la esperanza de vida. Por otro lado, el incremento de ésta será de menor monto cuando la mortalidad prepondere en edades adultas.

Por ello, en el periodo 2001-2011 para México, esta sección contestará la pregunta: ¿cuánto contribuyó cada grupo etario al cambio en la esperanza de vida de cada sexo?, cuestionamiento crucial para identificar y resolver problemáticas intrínsecas a cada grupo y que contribuye en la planeación de políticas enfocadas a atender las condiciones de vida de los residentes habituales del país. Para responder, se aplicará la técnica de Arriaga relativa a la descomposición de los cambios en la esperanza de vida (ver anexo 2).

Las tablas de mortalidad construidas para el presente ejercicio señalan que en el 2001 la esperanza de vida en México para su población fue de 73.55 años, cifra que aumentó a 74.09 para el 2011, lo cual representó una ganancia de poco más de medio año de vida en el periodo.[1] En ese decenio, la ganancia en la esperanza de vida masculina fue de 0.30 años al pasar de 71.28 en el 2001 a 71.58 en el 2011, mientras que el aumento para las mujeres fue de 0.82, colocándose en el 2011 en 76.59 años (ver cuadros 1 y 2).

1 La forma en que se trataron las fuentes de información se mencionan en el anexo 2.

Cuadro 1

Esperanza de vida al nacimiento según sexo, 2001, 2006 y 2011

	2001	2006	2011
Total	**73.55**	**73.97**	**74.09**
Hombres	71.28	71.71	71.58
Mujeres	75.77	76.18	76.59

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Cuadro 2

Cambios en la esperanza de vida al nacimiento según sexo, 2001-2006 y 2006-2011

	2001-2011	2001-2006	2006-2011
Total	**0.54**	**0.42**	**0.12**
Hombres	0.30	0.43	-0.13
Mujeres	0.82	0.41	0.41

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

La gráfica 1 muestra para el periodo 2001-2011 que entre las edades 15 y 44 se contribuyó negativamente en el crecimiento de la esperanza de vida nacional, es decir, se tuvo una pérdida que representó -31.1% (-0.17 años). Al mismo tiempo, ocurrió una diferencia importante entre sexos, ya que para las mujeres sólo se perdió 2.4% en el grupo de 15 a 29, mientras que para los hombres la pérdida en el mismo grupo fue de 90.9% y entre los de 30 a 44 años, de 36.8%, lo cual significa que en las edades productivas y de baja mortalidad (entre los 15 y 44 años), el sexo masculino perdió más de su ganancia neta en el decenio, es decir, 127.7% que representa 0.38 años.[2]

La principal ganancia a nivel nacional estuvo en el grupo de 60 a 74 años (49.4%), seguido por el de menores de 1 año (43.5%); sin embargo, llama la atención que, para hombres, la ganancia del de 0 a 1 año prácticamente es la que se obtuvo al final (92.7%, 0.28 años). Esto no significa que otros grupos no hayan aportado, sino que, debido a la pérdida en grupos centrales, se anuló el efecto positivo de ellos, por ejemplo, la ganancia entre los 60 y 74 años de edad representó 74.1 por ciento.

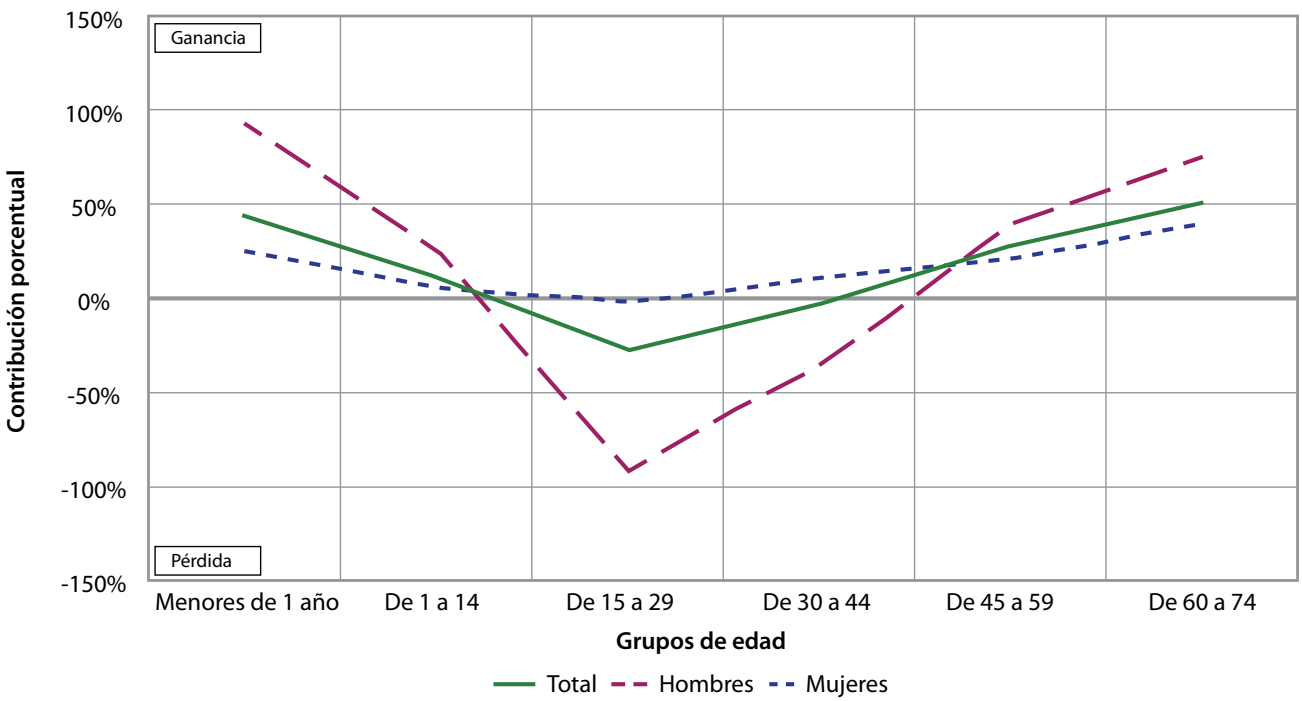
Al desagregar el decenio, tenemos que, durante el 2001 y 2006, en ningún sexo —menos para el total de población— ocurrió una pérdida porcentual en la esperanza de vida. El comportamiento entre sexos fue similar, la mayor ganancia provino del grupo de 60 a 74 años de edad y, enseguida, de la de menores de 1 año. La contribución más baja estuvo en los grupos centrales y, a partir de ellos, tuvo una tendencia creciente (ver gráfica 2).

El cambio en el indicador femenino fue de 0.41 años, la mitad de la ganancia del decenio. Para el caso masculino, fue prácticamente la misma que la femenina (0.43 años), sin embargo, para el decenio, ésta constituyó 143.3%, es decir, la ganancia decenal fue menor a la obtenida en ese quinquenio.

2 Se solicita al lector observar las escalas de las gráficas al no ser posible homologarlas, ya que pasarían desapercibidos algunos cambios.

Gráfica 1

Contribución porcentual al cambio de la esperanza de vida por sexo y grupos de edad, 2001-2011

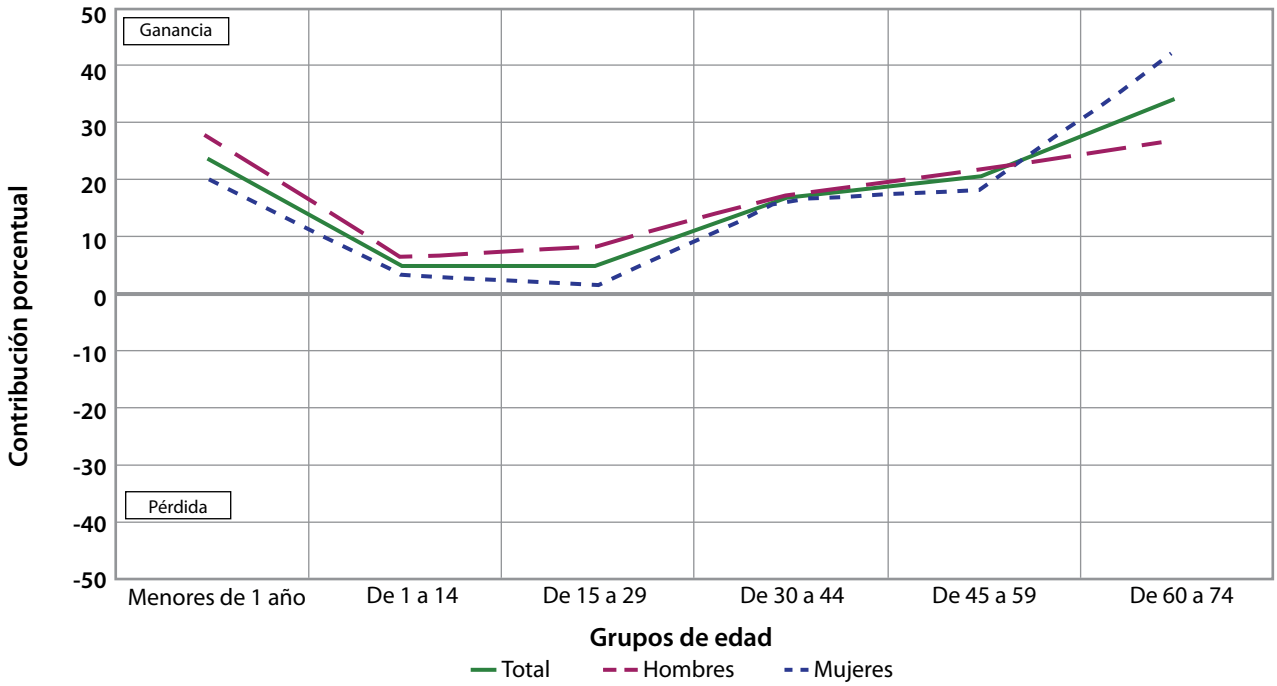


Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 2

Contribución porcentual al cambio de la esperanza de vida por sexo y grupos de edad, 2001-2006



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

La ganancia neta en la esperanza de vida nacional para el quinquenio 2006-2011 fue de 0.12 años, menos de la tercera parte de la que ocurrió en el anterior. Además, es relevante que entre las edades 15 a 44 sucedió una pérdida que significó más del doble del aumento (-210.4%, -0.26 años).

Analizando por sexo, tenemos que el aumento en el promedio de años vividos por las mujeres desde su nacimiento fue el mismo que ocurrió en el quinquenio anterior (0.41 años) y el comportamiento para los grupos etarios no muestra un comportamiento tan distinto al esperado, es decir, que las mayores ganancias se deban a los grupos etarios de los extremos y a una mínima contribución de las edades centrales; sin embargo, sucedió una pérdida de 6.2% entre las edades 15 a 29 años.

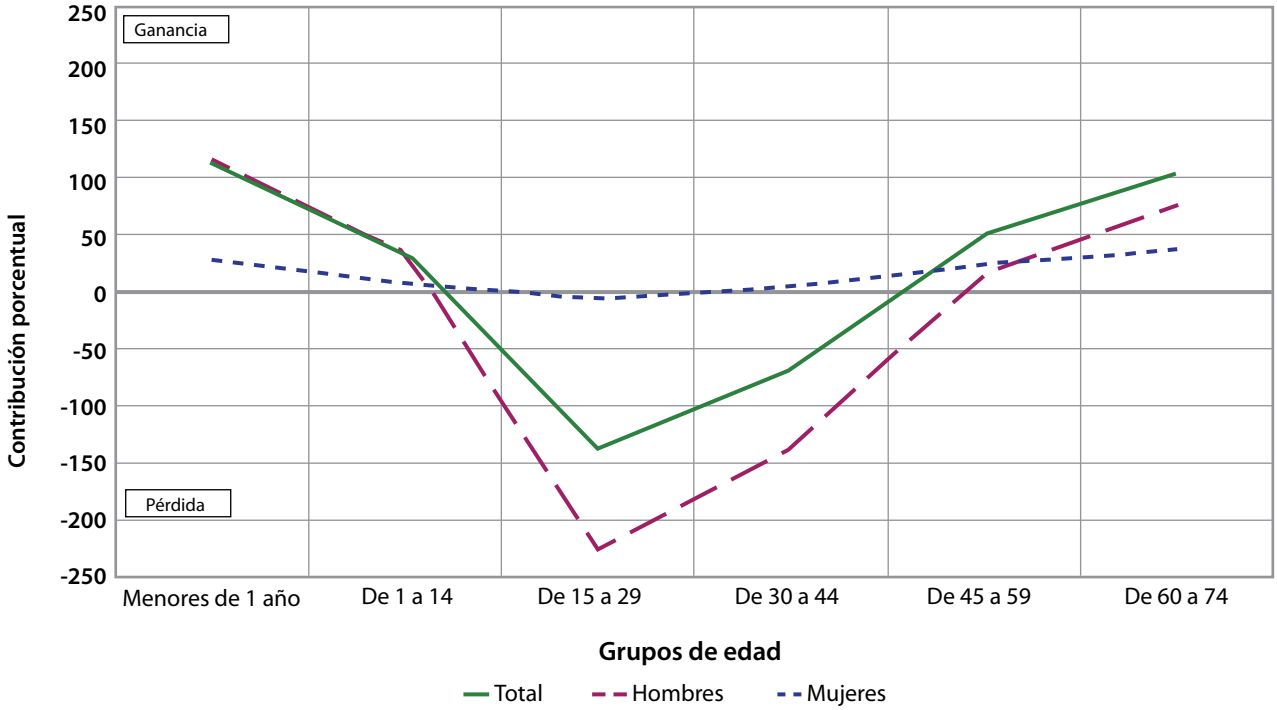
En el caso de los hombres esto no ocurre, su esperanza de vida neta decreció 0.13 años, y es notoria la contribución negativa entre los 15 y 44 años. Entre las edades 15 y 29, la disminución fue de más de dos veces la pérdida neta (-226.3%, -0.30 años) y la que ocurrió entre los 30 y 44 años fue de -137.2% (-0.18 años). Estos datos son significativos, ya que señalan un incremento grave de la mortalidad en edades jóvenes. Por otro lado, se debe mencionar que la ganancia para los menores de 1 año representó 115.9% (0.15 años), como se muestra en la gráfica 3.

Los datos hasta aquí presentados ponen en relieve la ventaja de analizar cuánto contribuyó cada grupo de edad al cambio de la esperanza de vida nacional para el periodo 2001-2011 pues, a pesar de que el indicador da indicios de una problemática al mostrar un desaceleramiento en su crecimiento hacia el periodo 2006-2011, por sí sólo no permite conocer dónde surge el problema; no obstante, con este análisis se concluye que la esperanza de vida mermó su incremento debido a que los hombres con edades entre 15 y 44 años contribuyeron negativamente entre el 2006 y 2011, lo que se interpreta como un excedente de mortalidad de población en edad productiva. En la siguiente parte se ampliará el análisis según algunas causas de muerte.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 3

Contribución porcentual al cambio de la esperanza de vida por sexo y grupos de edad, 2006-2011



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

El cambio de la mortalidad en cada causa de muerte puede ocurrir en todas o en algunas edades y, además, es posible que sea negativo o positivo en ellas (Arriaga, 1996: 9), por lo que incluir en el análisis a todas las causas por edad desplegada sobrepasa la extensión del presente trabajo, por ello se agrupó la información por edad, además de seleccionar para análisis las cinco causas con mayor frecuencia durante el 2011 para ambos sexos: diabetes mellitus, enfermedades isquémicas del corazón, enfermedades del hígado, enfermedades cerebrovasculares y agresiones (ver anexo 1).[3]

Al observar la contribución porcentual en el cambio en la esperanza de vida durante el periodo 2001-2011 de las causas de muerte seleccionadas para el total de la población, se observa que dos de ellas contribuyeron de manera negativa: diabetes mellitus (-11.4%) y agresiones (-44.7%). Para el caso femenino, la principal disminución se produjo por diabetes (-46.4%), seguida por las enfermedades isquémicas del corazón (-42.7%) y las agresiones (-6.5%).

Para los hombres, es relevante que por causa de agresiones ocurrió una contribución negativa en la esperanza de vida de casi el doble de la ganancia (-176.1%), ya que al ser de 0.30 años el aumento neto para el periodo 2001-2011, la contribución negativa por esta causa fue de -0.53 años; enseguida está la diabetes con -57.8%; además, se debe notar que la aportación de las enfermedades del hígado significaron 81.6% de la ganancia total (ver gráfica 4).

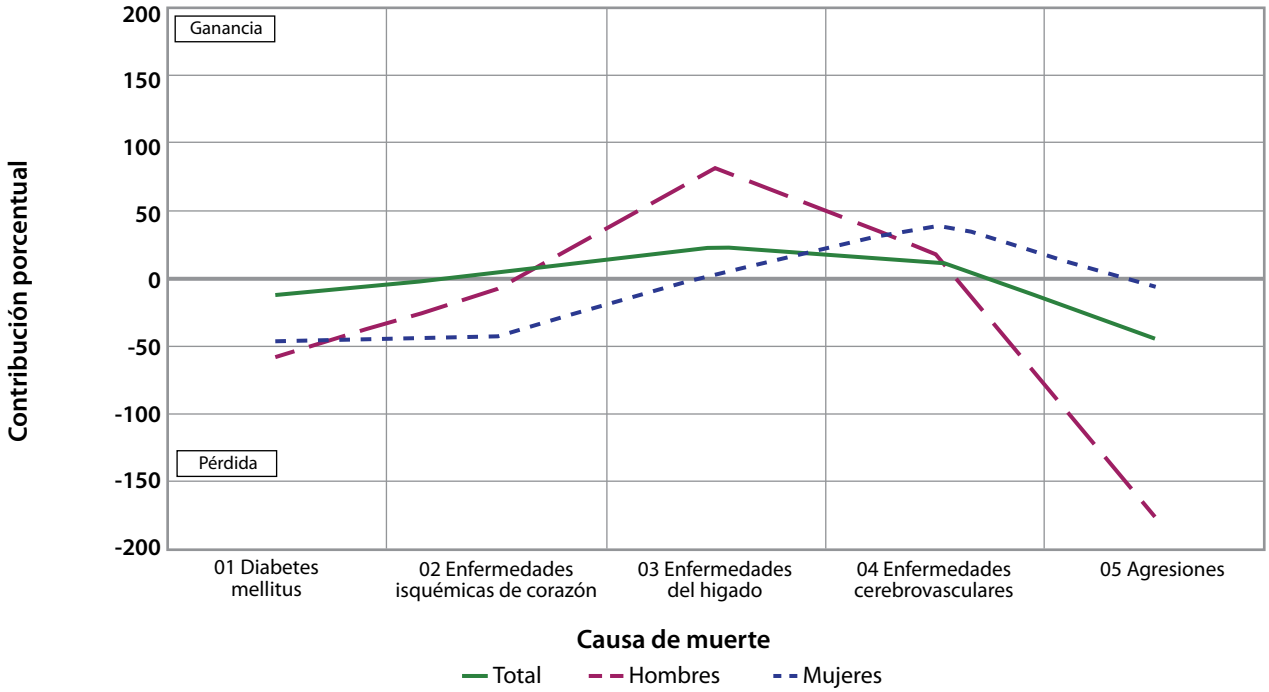
Dado que el perfil epidemiológico del país se dirigía a resolver las problemáticas de la mortalidad en menores de edad y hacia la prevalencia de

3 La agrupación por edad se vuelve necesaria para identificar la información relevante; para este ejercicio, se generó una matriz para el total de población y una para cada sexo, las cuales fueron de 7*6, resultado de siete grupos etarios y seis causas de muerte (el sexto grupo corresponde al resto de las causas).

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 4

Contribución porcentual al cambio de la esperanza de vida por sexo y causa de muerte, 2001-2011



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

las enfermedades crónico-degenerativas en edades adultas, es relevante encontrar para jóvenes un incremento muy importante en las agresiones como causa de muerte, fenómeno que representó que en la esperanza de vida nacional se dejaran de ganar 0.24 años.

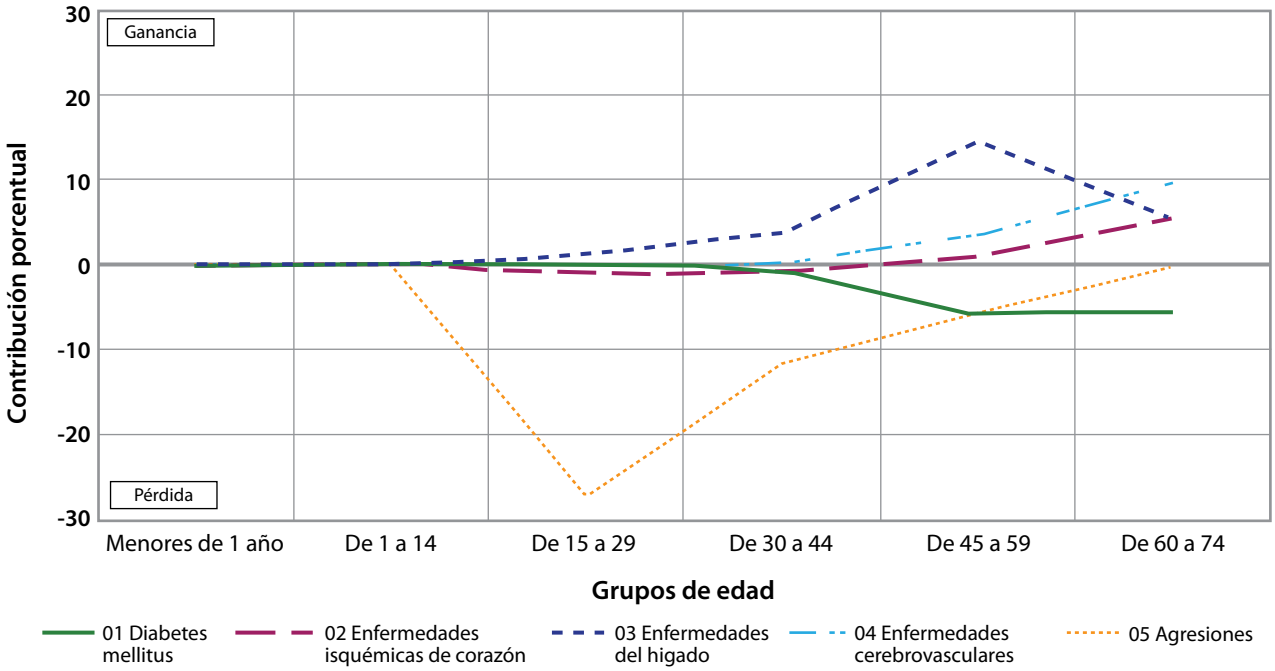
Entre los 15 y 59 años de edad, el aporte negativo de esta causa fue de -44.5%; para dimensionar esta cifra basta decir que en las mismas edades la contribución positiva de enfermedades isquémicas del corazón, del hígado y cerebrovasculares fue de 42.8%, es decir, la ganancia por estas tres causas se anuló en términos del indicador por motivo de las agresiones (ver gráfica 5).

El comportamiento para la mortalidad masculina en el periodo 2001-2011 señala que la mayor pérdida en la esperanza de vida ocurrió entre las edades 15 a 29 por muertes provocadas por agresiones (-88.8%, 0.27 años), seguida por la del grupo de 30 a 44 años con -67.9% por el mismo motivo. Estos datos muestran con claridad una sobremortalidad de la población masculina en edades productivas generada por una causa extrínseca, es decir, prevenible mediante acciones ajenas a las encaminadas a mejorar la salud de la población. La diabetes también aportó negativamente al indicador con -52.1% entre los 45 y 74 años de edad, sin embargo, las enfermedades del hígado entre los de 30 y 74 años contribuyeron con 77.7% de la ganancia (ver gráfica 6).

La mortalidad femenina entre el 2001 y 2011 muestra un comportamiento más cercano al deseable al tener que ocurrió una contribución porcentual positiva en la mayoría de los grupos de edad para las enfermedades analizadas, consecuencia de las acciones de salud que conducen a la postergación del evento muerte, o lo que es lo mismo, al incremento de la esperanza de vida. Además, causas como la diabetes mellitus con 1.3% de la ganancia entre los 45 y 59 años de edad y 2.9% entre los 60 y 74, así como las enfermedades isquémicas del corazón con 1.9 y 6.1%, respectivamente, permiten señalar que progresivamente las mayores ganancias provendrán de

Gráfica 5

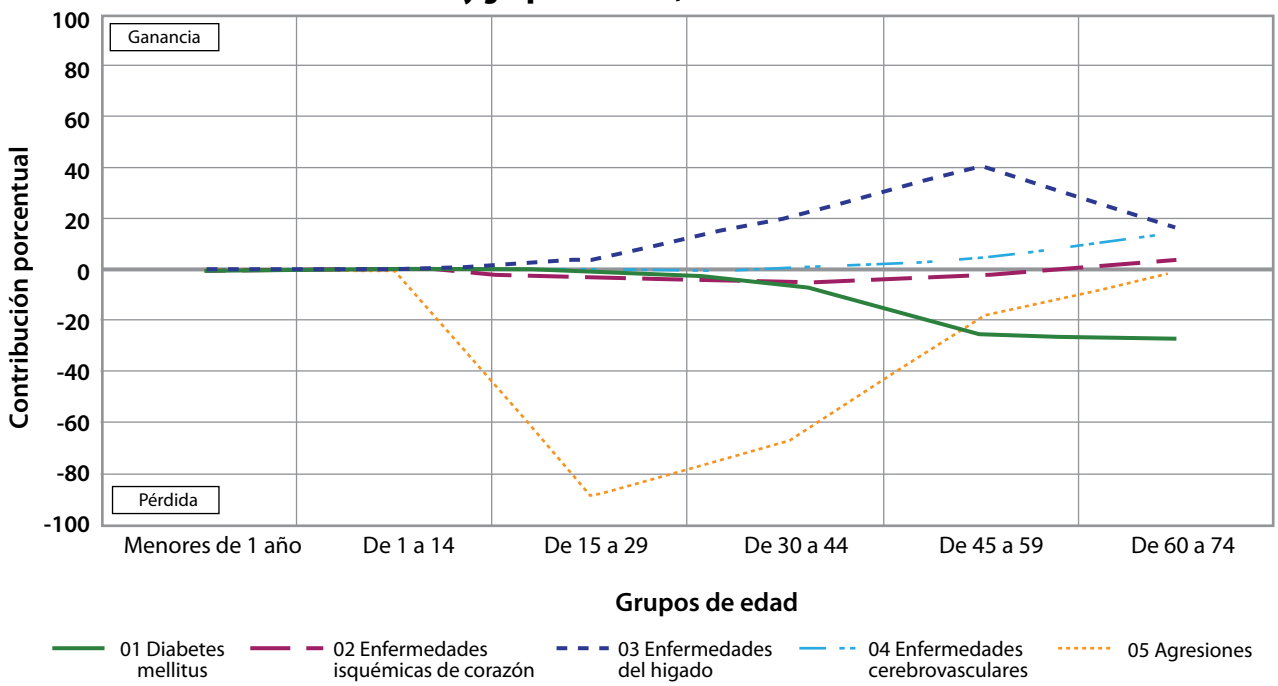
Ambos sexos: contribución porcentual al cambio de la esperanza de vida por causas de muerte y grupos de edad, 2001-2011



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Gráfica 6

Hombres: contribución porcentual al cambio de la esperanza de vida por causas de muerte y grupos de edad, 2001-2011



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

los últimos grupos etarios, es decir, de la prevención y atención de las enfermedades crónico y degenerativas (ver gráfica 7).

En el caso de las defunciones por agresiones, de igual forma ocurrió una pérdida, sin embargo, de mucho menor monto. La más grande estuvo en el grupo de mujeres jóvenes de 15 a 29 años con -3.3% seguido por la del grupo de 30 a 44 con -2.7%, es decir, si bien se perdió en la esperanza de vida femenina por causa de agresiones, no es comparable con el efecto que se presentó para la masculina.

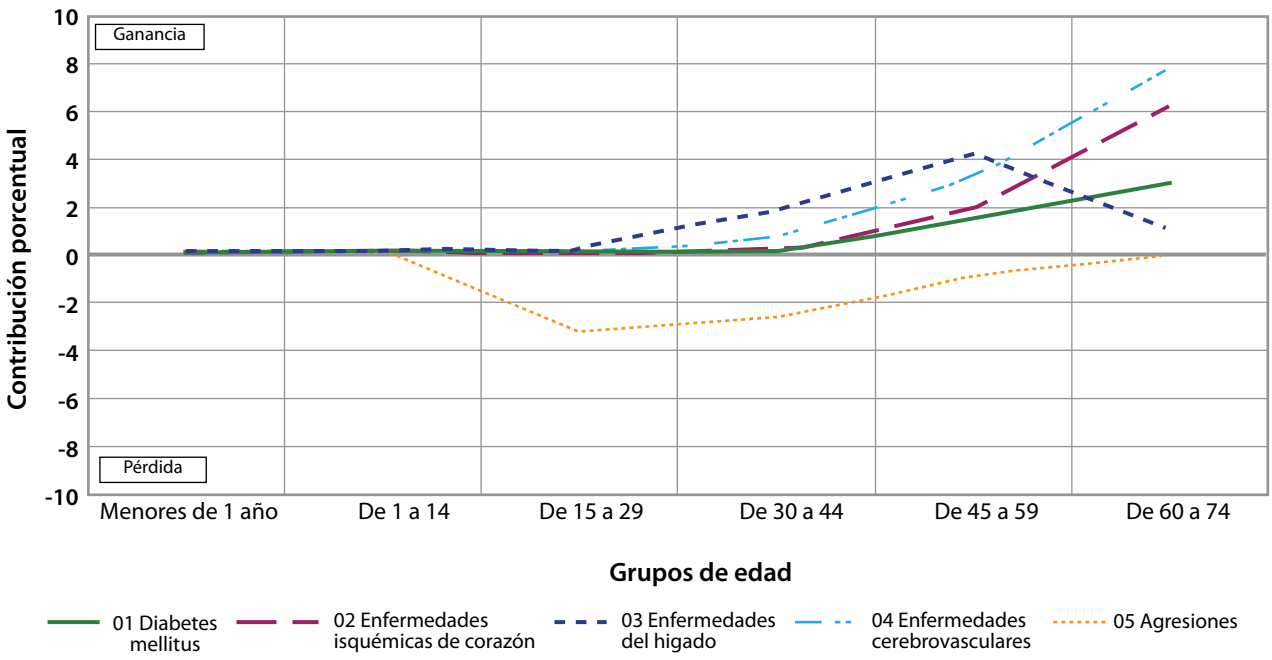
Dado el comportamiento significativo de las contribuciones negativas a la esperanza de vida masculina, se desagregó el decenio para identificar mejor el periodo de ocurrencia. De esta forma, tenemos que, durante 2001-2006, las agresiones como causa de muerte provocaron un decremento sólo en el grupo de 30 a 44 años de edad, que representó -0.9% de la ganancia total. Además, es significativo que la diabetes aportó una pérdida de -44.4% (-0.19) de la ganancia quinquenal entre los 30 y 74 años, y que la contribución positiva de las enfermedades del hígado para el mismo grupo etario representó 33.3% (ver gráfica 8).

Entre el 2006 y 2011, los resultados fueron desalentadores, el comportamiento por causa de muerte según su contribución porcentual al cambio en la esperanza de vida masculina señala una enorme sobremortalidad en edades centrales. El acumulado entre los 1 y 74 años de edad representó 450% de la pérdida quinquenal, es decir, 0.60 años de vida. El grupo que aportó la mayor pérdida es, de nuevo, el de 15 a 29 años con -216.8%, seguido por el de 30 a 44 con -162.1%; pero también es preocupante notar que en el de menores de edad (1 a 14 años) ocurrió una contribución negativa de -2.9% (ver gráfica 9).

A partir de los resultados aquí descritos, se resume que, para el periodo 2001-2011, las muertes por agresiones en el sexo masculino modificaron negativamente el comportamiento de la esperanza de vida nacional.

Gráfica 7

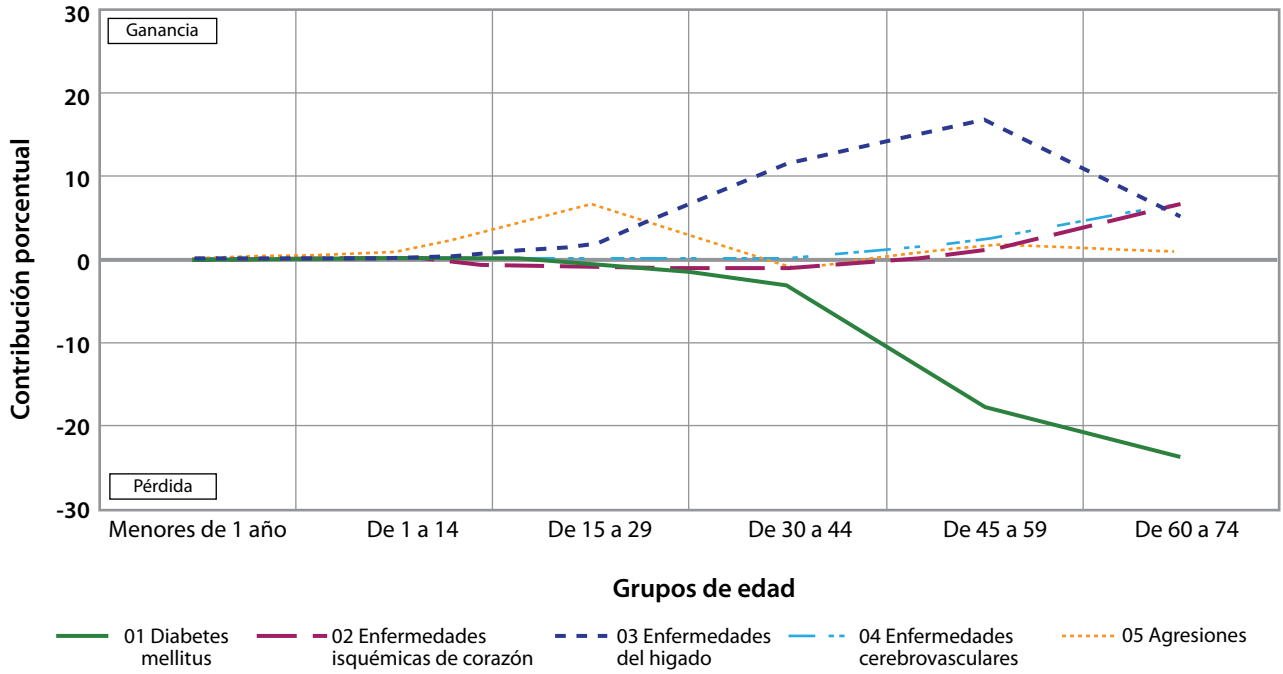
Mujeres: contribución porcentual al cambio de la esperanza de vida por causas de muerte y grupos de edad, 2001-2011



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Gráfica 8

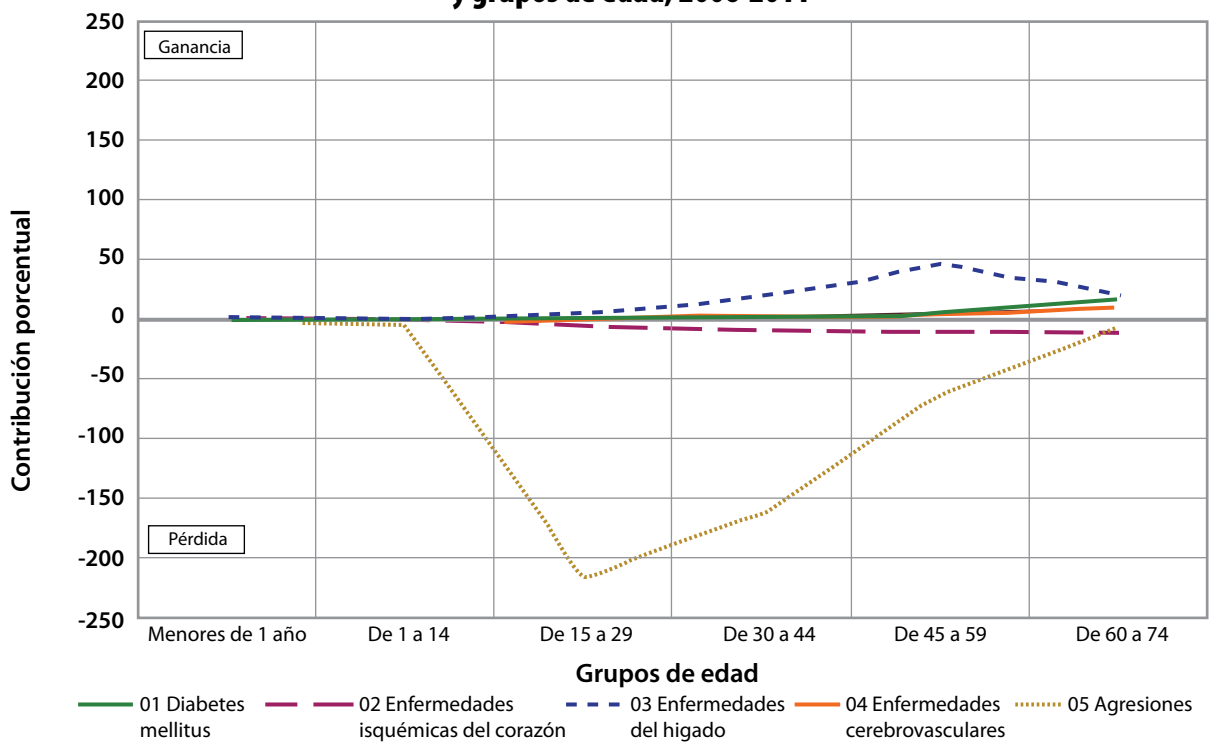
Hombres: contribución porcentual al cambio de la esperanza de vida por causas de muerte y grupos de edad, 2001-2006



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Gráfica 9

Hombres: contribución porcentual al cambio de la esperanza de vida por causas de muerte y grupos de edad, 2006-2011



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Conclusiones

La esperanza de vida es un indicador toral en Demografía al resumir información inherente a las condiciones actuales de desarrollo y de los patrones de morbilidad y mortalidad; sin embargo, su interpretación no debe ser superficial basada en su aparente sencillez, ya que cualquier modificación puede ocultar cambios drásticos en algún grupo —ya sea por edad, sexo o incluso rasgos socioculturales— y puede atenuarse por los comportamientos en otros. Por este motivo, las técnicas de descomposición apoyan para identificar situaciones específicas que deben ser atendidas, siendo deseable ampliar la desagregación geográfica del presente estudio para mejorar la identificación y cuantificación del problema.

Como conclusión principal se obtiene que en México, para el periodo 2001-2011, la esperanza de vida frenó su aumento debido a una sobremortalidad por causa de agresiones en hombres con edades entre 15 y 49 años que ocurrió, principalmente, durante el segundo quinquenio analizado. En términos demográficos se explica en que al tener un mayor número de muertes jóvenes disminuyó de manera considerable el número de años-persona vividos.

Se debe mencionar que a pesar de que las técnicas demográficas nos ayudan a cuantificar el peso de cada grupo etario y cada causa de muerte en la esperanza de vida, éstas no sirven para medir el impacto social a mediano y largo plazos. Las muertes por agresiones generan el rompimiento de las estructuras familiar, social y económica; por ello, los cambios en la esperanza de vida mostrados en este documento sólo son un destello de lo que se debe enfrentar para solucionar una enorme problemática.

Es necesario que se conozcan estos cambios negativos en la esperanza de vida nacional en distintos ámbitos y niveles, ya que proveen elementos claros que fomentan la reflexión de la situación actual del país. Además, se ilustra que en esta etapa social y demográfica no bastan los trabajos que de manera tradicional se realizan sobre el tema, por lo que es deseable se fomente un monitoreo continuo de los motivos de las modificaciones en la esperanza de vida con la finalidad de mitigar con oportunidad y prontitud el efecto negativo de eventos actuales y, con ello, tener condiciones de ofrecer mejor calidad de vida para los próximos años.

Anexo 1. Datos de mortalidad y sus causas

Las estadísticas vitales de defunciones permiten analizar con amplia desagregación temporal y causal el fenómeno de mortalidad. Para fines de seleccionar las causas de muerte analizadas en este documento, se empleó la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud. Décima Revisión (CIE-10); por esto, en las bases de registros administrativos se identificaron las variables correspondientes a los capítulos y grupos de esta clasificación (ver cuadro 3).

El comportamiento de las cuatro primeras causas señala, en general, un aumento constante del número de eventos: las mujeres tienen mayor incidencia de muerte por la diabetes mellitus (ver gráfica 10) y por enfermedades cerebrovasculares (ver gráfica 13); en los hombres predominan las muertes provocadas por enfermedades isquémicas del corazón (ver gráfica 11) y por las del hígado (ver gráfica 12).

Sin embargo, llama de manera excesiva la atención el comportamiento de las muertes por agre-

Cuadro 3

Causas de muerte seleccionadas para el estudio

Capítulo	Grupo	Descripción
4	2	Diabetes mellitus (E10-E14)
9	4	Enfermedades isquémicas del corazón (I20-I25)
11	8	Enfermedades del hígado (K70-K77)
9	7	Enfermedades cerebrovasculares (I60-I69)
20	27	Agresiones (X85-Y09)

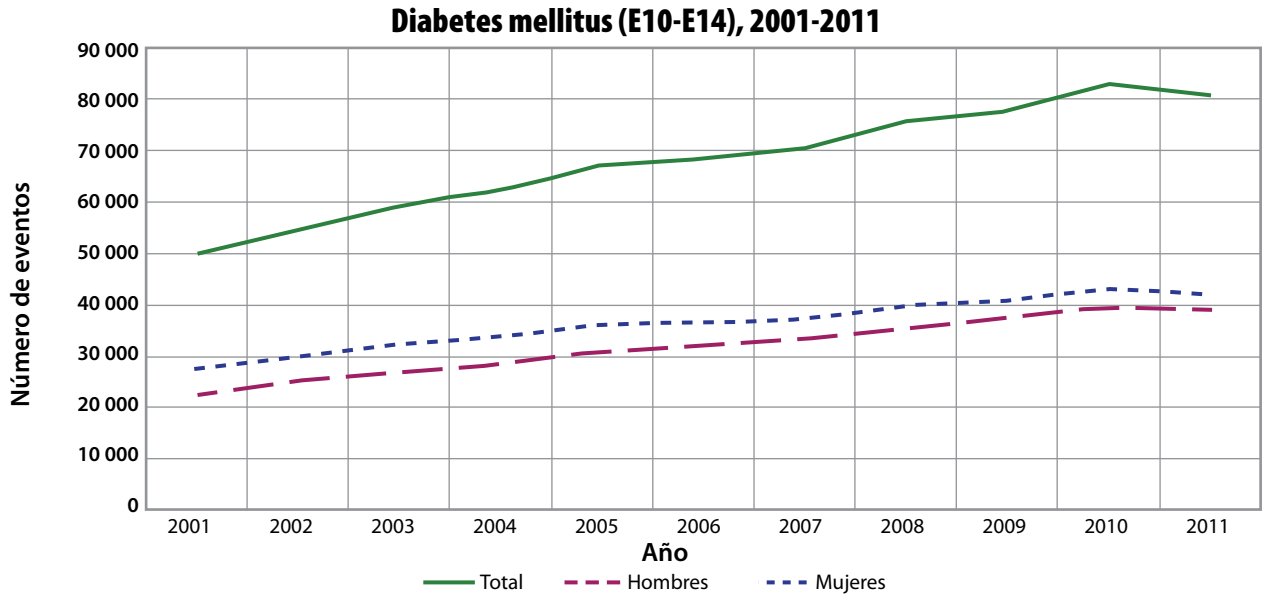
Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

siones (ver gráfica 14),[4] ya que a partir del 2007 ocurrió un incremento abrupto de casos, principalmente de hombres. Basta decir que entre el 2007 y 2011 se triplicó el número de defunciones por esta causa; de otra forma, si consideramos que entre el 2001 y 2007 sucedieron en promedio 8 500 muertes anuales, se obtiene que ocurrió una sobremortalidad de 44 mil casos para el periodo 2008-2011.

Estos datos deben motivar acciones que conduzcan a corregir las causas de un problema que aquí se describió en términos de un indicador, pero es necesario recalcar que son inmensos los costos que este fenómeno ha significado para las familias y sociedad.

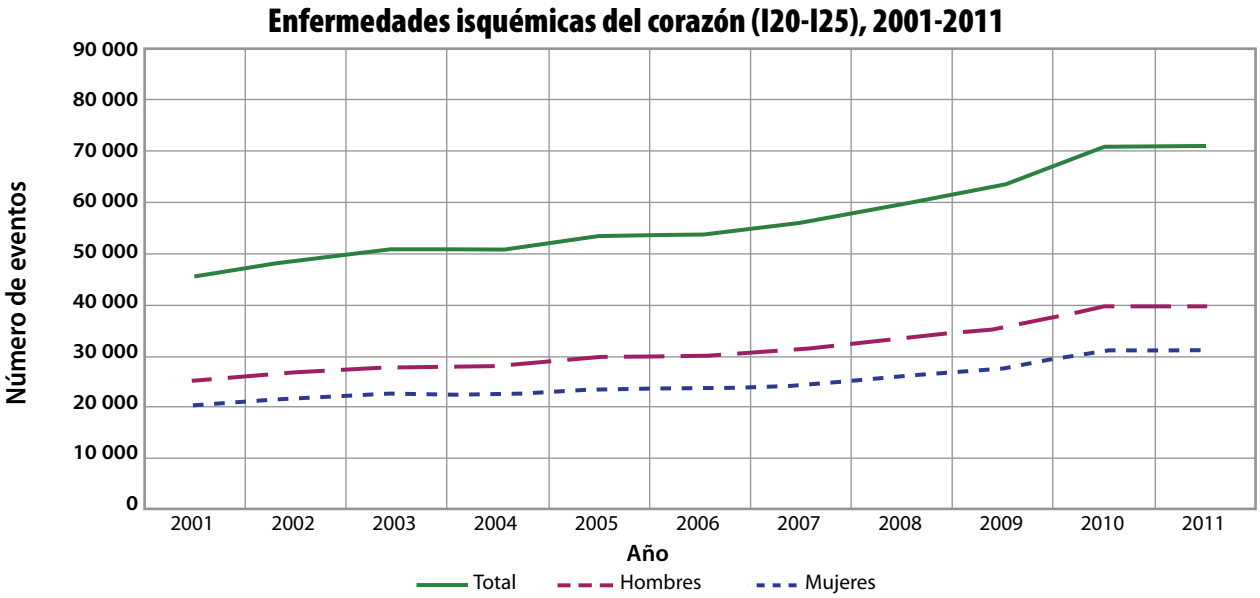
4 La causa de muerte *Agresiones* (X85-Y09) incluye el homicidio, lesiones ocasionadas por otra persona con intento de lesionar o matar, por cualquier medio, y excluye las lesiones debidas a *Intervención legal* (Y35) y *Operaciones de guerra* (Y36), ver OPS, 1995, pp. 1011-1014.

Gráfica 10



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

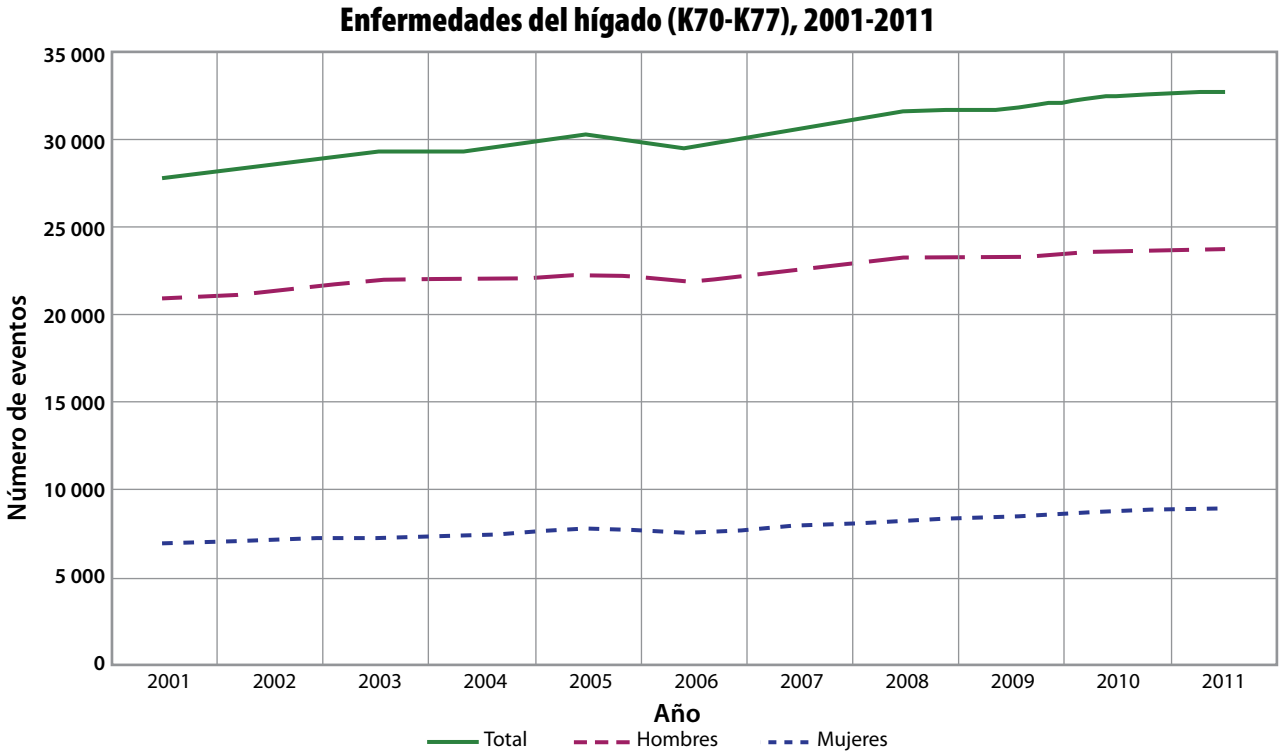
Gráfica 11



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

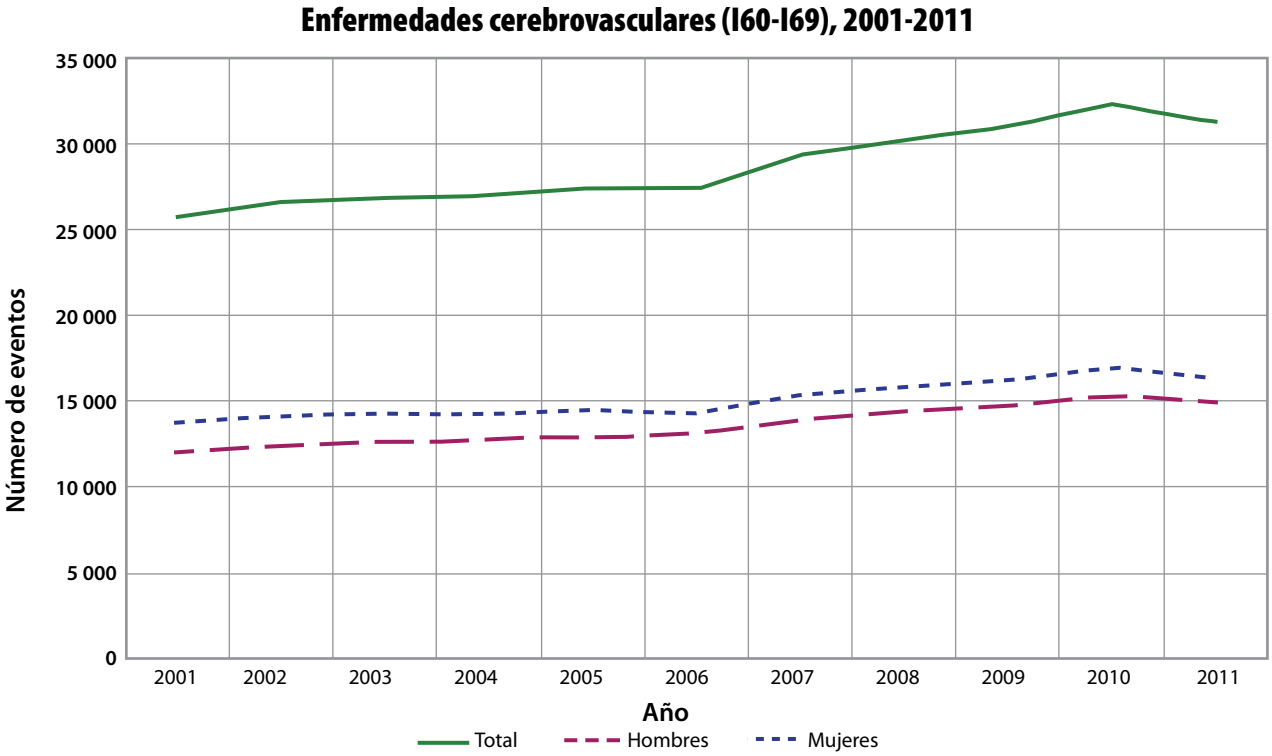
INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Gráfica 12



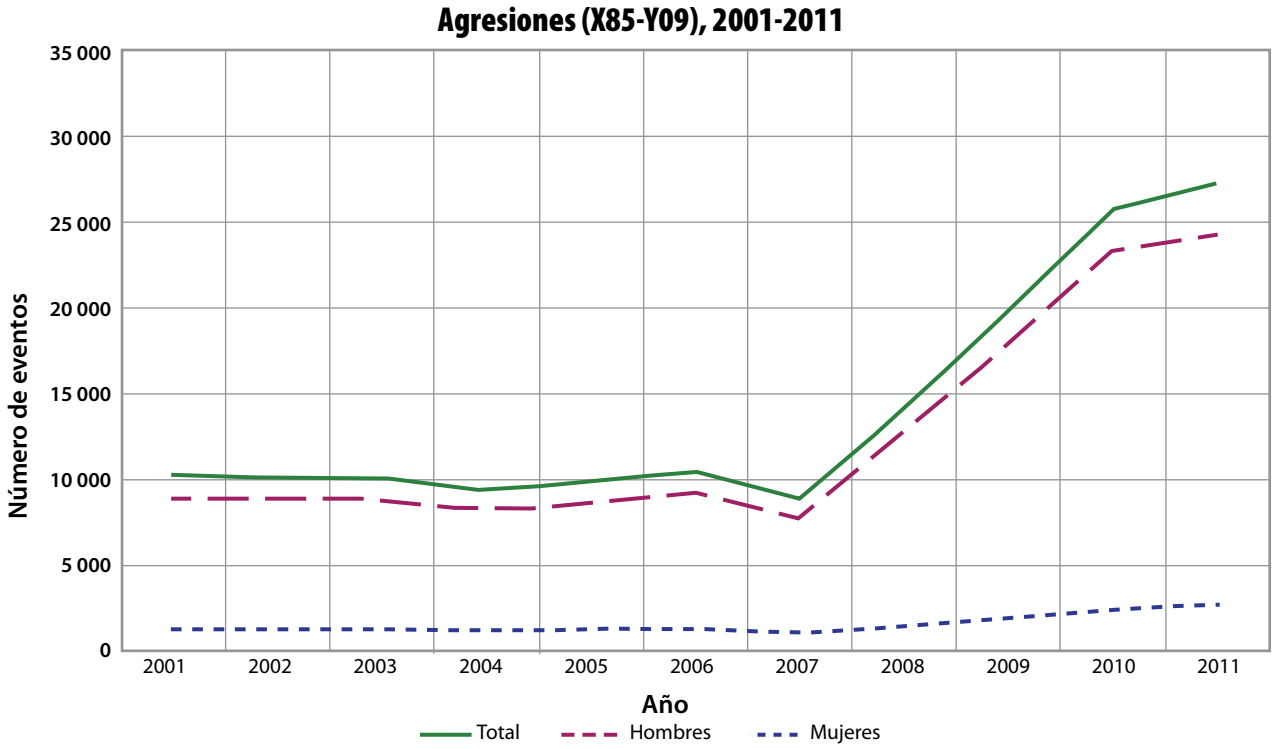
Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

Gráfica 13



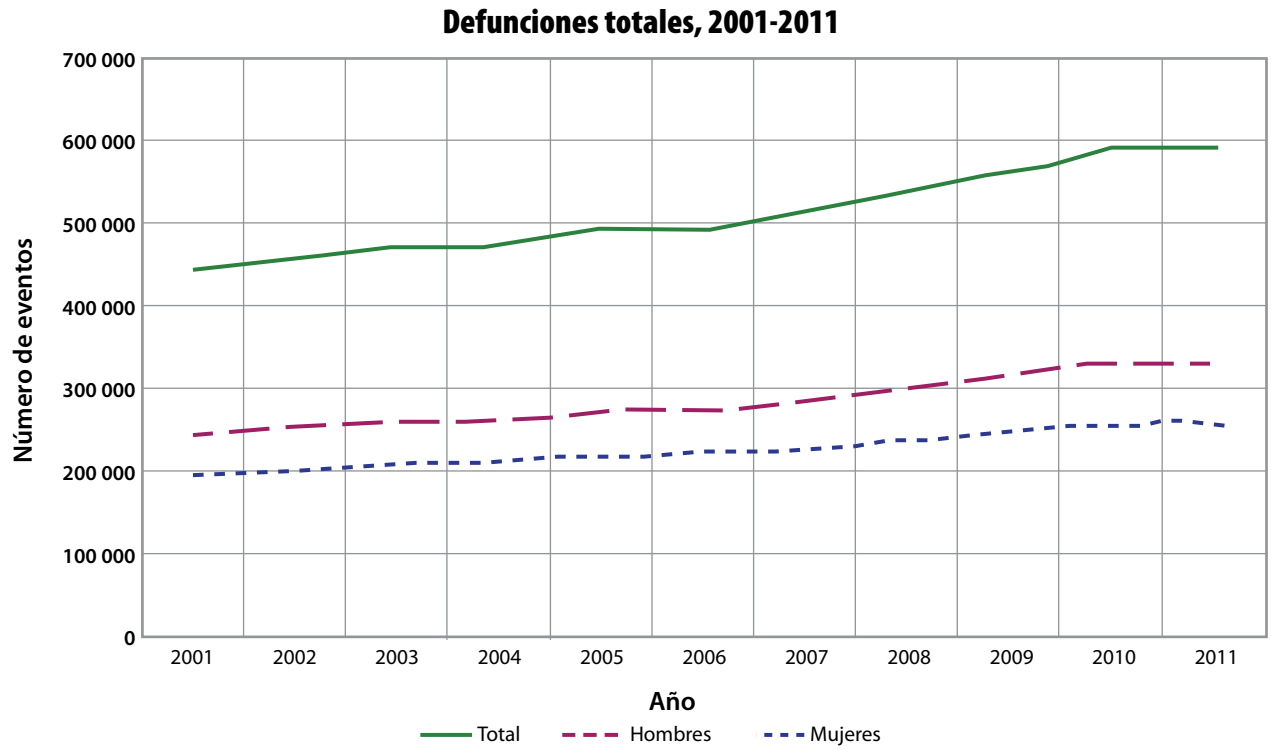
Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

Gráfica 14



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

Gráfica 15



Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011.

Anexo 2. Metodología

Construcción de las tablas de mortalidad y tratamiento de los datos

Con el fin de obtener las tablas de mortalidad utilizadas, en el presente trabajo se recurrió a los censos de población del 2000 y 2010 y a los registros de defunciones de estadísticas vitales del periodo 2001-2011. Al no tener una serie anual, fue necesario aplicar herramientas para obtener poblaciones demográficamente consistentes con esa periodicidad. En primer lugar se suavizaron las poblaciones censales por efecto de la mala declaración de la edad; después, se interpolaron las anuales por medio de diferencias divididas aplicada a cada una de las edades y extrapolando por tendencia al 2011. Con la serie anual se procedió a corregir y verificar la coherencia por cohorte y edad hasta donde fue posible su reconstrucción, lo cual se hizo aplicando de manera iterativa la regresión local (LOESS, por sus siglas en inglés) a cada tramo; de la misma forma, se corrigió la relación hombres/mujeres para cada año. Las defunciones fueron suavizadas sólo por mala declaración de edad. Con la serie de población anual y las muertes resultantes se construyeron las tablas de mortalidad, verificando que los niveles y tendencias fueran consistentes con la información de otras fuentes.

Esperanza de vida y su descomposición

La esperanza de vida surge del enfoque probabilístico que considera como una variable aleatoria al tiempo que le resta vivir a cualquier individuo a partir de la edad, definida usualmente como $T(x)$. Por ello, se define como X la variable aleatoria continua que representa la edad al fallecimiento y $F(X)$ es su función de distribución, es decir, $F(X) = Pr(X \leq x)\ x \geq 0$. A partir de ella se deduce que la probabilidad de que un recién nacido alcance la edad x es $s(x) = 1 - F(x) = Pr(X > x)\ x \geq 0$, por lo que basta obtener la probabilidad condicional de morir entre las edades x y z, dado que se llegó con vida a la edad x, que es $Pr\,(x < X \leq z \mid X > x) = \frac{F(z)-F(x)}{1-F(x)} = \frac{s(x)-s(z)}{s(x)} = T(x)\ x \geq 0$.

En términos de una población, la función $T(x)$ representa el total de años vividos por el conjunto a partir de la edad x, es decir, es la suma de los años que cada individuo vivió a partir de determinada edad.

Debido a que la esperanza de vida es el promedio de años vividos por una población, para obtenerla basta realizar el cociente entre los años vividos y el total de sus miembros, por ello, es necesario definir a l_x como los sobrevivientes de edad x; entonces, la esperanza de vida a esta edad es $e^x = \frac{T(x)}{l(x)}$. En consecuencia, la esperanza de vida al nacimiento es $e^0 = \frac{T(0)}{l(0)}$, que significa el promedio de años vividos desde el nacimiento hasta la muerte de todo miembro de una población.[5]

Adecuando al caso discreto, el cambio en la esperanza de vida al nacimiento se resume con la siguiente expresión:

$$e^0(t+h) - e^0(t) = \Delta e^0(t+h) = \sum_{i=0}^{k-1} \Delta e^i(t+h)$$

Para conocer cuánto aportó cada edad al cambio, se aplicará la técnica de Arriaga relativa a la descomposición de los cambios en la esperanza de vida ya que, como menciona Preston *et al.* (2000: 64), su aplicación para el caso discreto es fácil a partir de una tabla de mortalidad tradicional. Según esta técnica, la modificación en la esperanza de vida se explica por efectos debidos a los cambios de la mortalidad en cada grupo de edad, llamados directos e indirectos pero, además, interviene el efecto de la interacción entre ellos. Por esto, el cambio neto en la esperanza de vida al nacimiento queda resumido en la siguiente expresión:

$$\sum_{i=0}^{k-1} \Delta e^i(t+h)$$
$$= \sum_{i=0}^{k-1} [\Delta_D(x_i, x_{i+1}) + \Delta_I(x_i, x_{i+1}) + \Delta_{In}(x_i, x_{i+1})] \cdots \quad (1)$$

con $\boldsymbol{k}$ = 0,1 – 14, 15 – 29, 30 – 44, ⋯, 60 – 74 y 75 y más

5 Con el fin de tener una explicación completa de las funciones necesarias para obtener la esperanza de vida, resumidas en una tabla de mortalidad, se recomienda consultar Preston *et al.* y Bowers *et al.*

donde el cambio por el efecto directo es:

$$\Delta_D(x_i, x_{i+1}) = \frac{l(x_i;t)}{l(x_0;t)}\left(\frac{L(x_i, x_{i+1};t+h)}{l(x_i;t+h)} - \frac{L(x_i, x_{i+1};t)}{l(x_i;t)}\right) = \frac{l(x_i;t)}{l(x_0;t)}\left(e(x_i, x_{i+1};t+h) - e(x_i, x_{i+1};t)\right)$$

y el que se explica por el indirecto es:

$$\Delta_I(x_i, x_{i+1}) = \frac{T(x_{i+1};t)}{l(x_0;t)}\left(\frac{l(x_i;t) * l(x_{i+1};t+h)}{l(x_{i+1};t) * l(x_i;t+h)} - 1\right)$$

El cambio por el efecto de interacción, que involucra al efecto indirecto, queda como sigue:

$$\Delta_{In}(x_i, x_{i+1}) = \frac{T(x_{i+1};t+h)}{l(x_0;t)}\left(\frac{l(x_i;t)}{l(x_i;t+h)} - \frac{l(x_{i+1};t)}{l(x_{i+1};t+h)}\right) - \frac{T(x_{i+1};t)}{l(x_0;t)}\left(\frac{l(x_i;t) * l(x_{i+1};t+h)}{l(x_{i+1};t) * l(x_i;t+h)} - 1\right)$$

por lo cual, el cambio en la esperanza de vida al nacimiento entre dos momentos es:

$$\Delta e^0(t+h) = \sum_{i=1}^{k-1}\left[\frac{l(x_i;t)}{l(x_0;t)}\left(e(x_i, x_{i+1};t+h) - e(x_i, x_{i+1};t)\right) + \frac{T(x_{i+1};t+h)}{l(x_0;t)}\left(\frac{l(x_i;t)}{l(x_i;t+h)} - \frac{l(x_{i+1};t)}{l(x_{i+1};t+h)}\right)\right] \cdots \quad (2)$$

El primer sumando cuantifica el efecto directo del grupo de edad (x_i, x_{i+1}) y representa el cambio en el número de años vividos por el efecto de la mortalidad en el mismo grupo. El segundo describe la modificación por la exposición de los sobrevivientes a las nuevas condiciones de mortalidad, es decir, la diferencia en años vividos a partir de la edad x_{i+1} que se explica por el cambio de la mortalidad en el grupo de edad precedente.

Finalmente, como nos interesa conocer en términos porcentuales cuánto contribuyó cada grupo etario, se realiza el cociente entre el aporte de cada grupo y la ganancia total neta.

Además, es de interés identificar y cuantificar el impacto de las causas de muerte en la esperanza de vida, por lo cual se debe extender el método de forma que permita conocer el porcentaje que implicó cada causa en cualquier grupo etario.

Ya se conoce el cambio en la esperanza de vida (expresado en 1 y 2), por ello, basta aplicar la proporción de cada causa de muerte (j) para cada grupo etario (i) en cada año de referencia (t o $t + h$) de la siguiente forma:

$$\Delta e^0(t+h) = \sum_{i=1}^{k-1}\Delta e^i(i;t+h) = \sum_{j=1}^{n}\sum_{i=1}^{k-1}\Delta e^{ij}(i;t+h) \cdots \quad (3)$$

donde:

$$\Delta e^{ij}(i;t+h) = \Delta e^i(i;t+h) * \frac{R_i^j(t+h) * m_i(t+h) - R_i^j(t) * m_i(t)}{m_i(t+h) - m_i(t)} \cdots \quad (4)$$

$R^i_j(t)$ es la proporción de muertes de la causa de muerte j del grupo etario i del año t.

$m_i(t)$ es la tasa de mortalidad del grupo etario i del año t.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Anexo 3. Cuadros de resultados

Cuadro 4

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, población total, 2001-2011

Esperanza de vida							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.00	0.00	0.00	0.00	0.00	0.24	0.24
1-14	0.00	0.00	0.00	0.00	0.00	0.06	0.06
15-29	0.00	-0.01	0.01	0.00	-0.15	-0.01	-0.15
30-44	0.00	0.00	0.02	0.00	-0.06	0.03	-0.02
45-59	-0.03	0.01	0.08	0.02	-0.03	0.11	0.15
60-74	-0.03	0.03	0.03	0.05	0.00	0.19	0.27
75-+	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	**-0.06**	**0.03**	**0.13**	**0.07**	**-0.24**	**0.62**	**0.55**
Contribución porcentual							
0-1	-0.01	0.00	-0.10	0.09	0.09	43.45	43.52
1-14	0.11	0.01	-0.02	-0.12	-0.01	10.99	10.96
15-29	0.23	-1.04	1.28	-0.12	-27.28	-1.24	-28.17
30-44	-0.69	-0.81	3.67	0.25	-11.61	6.21	-2.98
45-59	-5.80	1.02	14.50	3.51	-5.63	19.44	27.04
60-74	-5.45	5.56	5.33	9.63	-0.27	34.62	49.42
75-+	0.22	0.22	0.01	-0.08	0.00	-0.16	0.21
Total	**-11.38**	**4.96**	**24.66**	**13.16**	**-44.72**	**113.32**	**100.00**

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Cuadro 5

Continúa

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, sexo maculino, 2001-2011

Esperanza de vida							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.00	0.00	0.00	0.00	0.00	0.27	0.27
1-14	0.00	0.00	0.00	0.00	0.00	0.08	0.07
15-29	0.00	-0.01	0.01	0.00	-0.26	-0.01	-0.27
30-44	-0.01	-0.01	0.06	0.00	-0.20	0.06	-0.11
45-59	-0.07	-0.01	0.12	0.01	-0.05	0.11	0.11
60-74	-0.08	0.01	0.05	0.04	0.00	0.20	0.22
75-+	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.01
Total	**-0.17**	**-0.02**	**0.24**	**0.06**	**-0.52**	**0.71**	**0.29**

Cuadro 5

Concluye

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, sexo maculino, 2001-2011

Contribución porcentual							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.05	0.00	-0.08	0.03	0.20	92.48	92.69
1-14	0.10	0.00	-0.05	-0.35	-0.25	25.87	25.34
15-29	0.32	-3.35	4.11	-0.11	-88.84	-3.07	-90.94
30-44	-4.47	-4.89	20.26	0.32	-67.86	19.82	-36.81
45-59	-25.25	-1.94	41.03	4.36	-18.59	39.23	38.84
60-74	-26.81	4.06	16.40	14.52	-0.78	66.76	74.14
75-+	-1.75	-1.91	-0.05	0.41	-0.01	0.07	-3.26
Total	**-57.82**	**-8.05**	**81.63**	**19.19**	**-176.12**	**241.16**	**100.00**

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Cuadro 6

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, sexo femenino, 2001-2011

Esperanza de vida							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.00	0.00	0.00	0.00	0.00	0.20	0.20
1-14	0.00	0.00	0.00	0.00	0.00	0.04	0.04
15-29	0.00	0.00	0.00	0.00	-0.03	0.01	-0.02
30-44	0.00	0.00	0.01	0.01	-0.02	0.09	0.09
45-59	0.01	0.02	0.03	0.03	-0.01	0.09	0.17
60-74	0.02	0.05	0.01	0.06	0.00	0.18	0.32
75-+	-0.42	-0.41	-0.03	0.23	0.00	0.63	0.01
Total	**-0.38**	**-0.35**	**0.03**	**0.33**	**-0.05**	**1.24**	**0.82**
Contribución porcentual							
0-1	-0.03	0.00	-0.12	0.11	0.04	24.39	24.40
1-14	0.12	0.02	-0.01	-0.03	0.09	5.26	5.44
15-29	0.20	-0.13	0.25	-0.11	-3.27	0.69	-2.37
30-44	0.08	-0.04	1.80	0.66	-2.66	10.68	10.52
45-59	1.34	1.88	4.17	3.16	-0.95	11.57	21.17
60-74	2.91	6.12	0.97	7.74	-0.08	22.07	39.74
75-+	-51.04	-50.60	-3.75	28.40	0.36	77.74	1.10
Total	**-46.42**	**-42.75**	**3.31**	**39.92**	**-6.46**	**152.40**	**100.00**

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Cuadro 7

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, sexo masculino, 2001-2006

Esperanza de vida							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.00	0.00	0.00	0.00	0.00	0.12	0.12
1-14	0.00	0.00	0.00	0.00	0.00	0.03	0.03
15-29	0.00	0.00	0.01	0.00	0.03	0.00	0.04
30-44	-0.01	0.00	0.05	0.00	0.00	0.05	0.07
45-59	-0.08	0.01	0.07	0.01	0.01	0.07	0.09
60-74	-0.10	0.03	0.02	0.03	0.00	0.13	0.12
75-+	-0.03	-0.01	0.00	0.00	0.00	0.00	-0.04
Total	**-0.21**	**0.01**	**0.15**	**0.04**	**0.04**	**0.40**	**0.43**
Contribución porcentual							
0-1	0.01	0.00	-0.09	0.11	0.02	27.39	27.45
1-14	0.02	0.00	-0.15	-0.13	0.68	5.87	6.30
15-29	0.33	-0.98	1.75	-0.04	6.71	0.53	8.31
30-44	-3.09	-1.07	11.50	0.08	-0.95	11.03	17.51
45-59	-17.60	1.21	16.79	2.37	1.80	17.17	21.74
60-74	-23.67	6.54	5.06	6.81	1.06	31.23	27.04
75-+	-6.01	-2.79	-0.55	0.64	0.06	0.31	-8.35
Total	**-50.00**	**2.92**	**34.32**	**9.83**	**9.39**	**93.53**	**100.00**

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Cuadro 8

Continúa

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, sexo masculino, 2006-2011

Esperanza de vida							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.00	0.00	0.00	0.00	0.00	0.15	0.15
1-14	0.00	0.00	0.00	0.00	0.00	0.05	0.05
15-29	0.00	-0.01	0.01	0.00	-0.29	-0.01	-0.30
30-44	0.00	-0.01	0.02	0.00	-0.22	0.02	-0.18
45-59	0.01	-0.02	0.06	0.00	-0.08	0.05	0.02
60-74	0.02	-0.02	0.03	0.01	-0.01	0.06	0.10
75-+	0.01	0.02	0.00	0.00	0.00	0.00	0.03
Total	**0.03**	**-0.03**	**0.11**	**0.01**	**-0.60**	**0.33**	**-0.13**

Cuadro 8 Concluye

Esperanza de vida y su contribución porcentual por grupos de edad y sexo, sexo masculino, 2006-2011

Contribución porcentual							
Grupo de edad	Diabetes mellitus	Enfermedades isquémicas del corazón	Enfermedades del hígado	Enfermedades cerebrovasculares	Agresiones	Resto de las causas	Total
0-1	0.06	0.00	0.11	-0.29	0.37	115.63	115.89
1-14	0.16	0.00	0.41	-0.32	-2.94	38.29	35.59
15-29	-0.29	-4.48	3.84	-0.13	-216.84	-8.39	-226.28
30-44	-2.42	-8.96	17.75	0.57	-162.11	17.97	-137.19
45-59	5.77	-11.26	44.45	2.07	-63.02	37.39	15.40
60-74	16.20	-11.82	19.88	10.22	-5.07	47.27	76.68
75-+	6.51	17.60	-0.87	-3.63	0.35	-0.04	19.92
Total	**25.99**	**-18.93**	**85.58**	**8.50**	**-449.26**	**248.12**	**-100.00**

Fuente: elaboración propia con base en estadísticas vitales de mortalidad 2001-2011 y censos de población 2000 y 2010.

Fuentes

Arriaga, Eduardo. "Los años de vida perdidos: su utilización para medir el nivel y cambio de la mortalidad", en: *Notas de Población*. Núm. 63. Santiago de Chile, 1996.

Bowers, Newton, Hans Gerber, James Hickman, Donald Jones y Cecil Nesbitt. *Actuarial Mathematics*. Segunda edición. Society of Actuaries. Schaumburg, 1997.

Canudas, Vladimir. *Decomposition Methods in Demography*. Amsterdam, The Netherlands, Rozenberg Publishers, 2003.

Organización Panamericana de la Salud (OPS). *Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud*. Décima Revisión. Volumen 1. Washington, DC, 1995.

Preston, Samuel, Patrick Heuveline y Michel Guillot. *Demography: Measuring and Modeling Population Processe*. Oxford, Wiley-Blackwell Publishers, 2000.

Notes on the Relationship between Trade and Employment in U.S. Manufacturing Sector, 1998-2008

Pedro I. Hancevic[1]

In this paper I study the impact of trade on employment in the U.S. manufacturing sector between 1998 and 2008. By using a production function model, I derive the labor-demand function, which incorporates the most relevant elements to analyze the relationship between trade (export and imports) and employment. The empirical evidence downplays the dominant belief that greater trade opening improves work efficiency and reduces labor demand by the firms.

JEL Classification: J21, F16

Key words: manufacturing, employment, imports, exports.

Este trabajo revisa la relación entre el comercio y el empleo en el sector manufacturero de los Estados Unidos de América entre 1998 y el 2008. Para ello, se utiliza un modelo empírico basado en una función de producción de la cual se deriva la demanda de trabajo. Se incorporan así los elementos más relevantes para analizar la relación entre el comercio (exportaciones e importaciones) y el empleo. La evidencia empírica obtenida relativiza la creencia dominante de que una mayor apertura comercial mejora la eficiencia del trabajo y reduce la demanda de dicho factor en las firmas.

Palabras clave: manufacturas, empleo, importaciones, exportaciones.

Recibido: 14 de octubre de 2014.
Aceptado: 29 de mayo de 2015.

1 I would like to thank seminar participants at the University of Wisconsin-Madison and CIDE for very helpful comments. All remaining errors are my own.

Fast Food Workers Gather To Watch State's Wage Board Decision On.../ Spencer Platt/Getty Images

1. Introduction

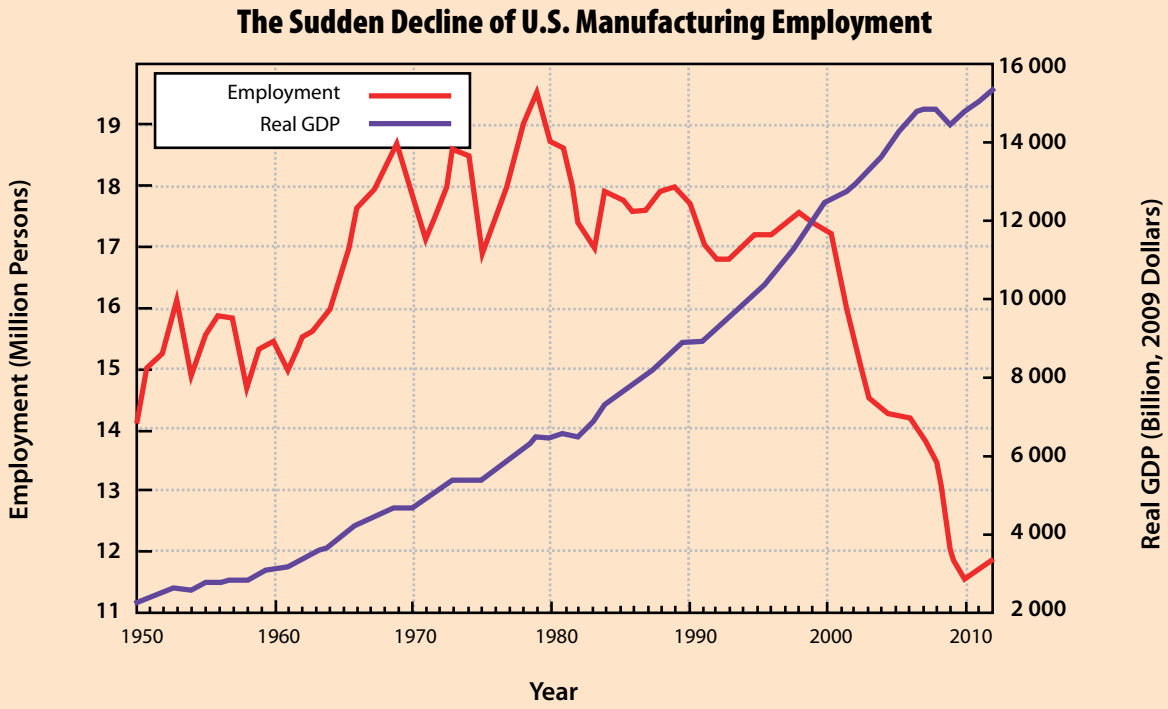
United States Manufacturing Sector has witnessed a fall in employment over the last decades. However, the output level and the (average) wage-rate have increased in this sector. Figure 1 shows the evolution of employment and output levels for the 1950–2010 period.

Several studies link this phenomenon with an increasing participation of imports coming from South-East Asia, China, Mexico, and other developing countries. Similarly, outsourcing and off-shoring activities carried out by U.S. firms are sometimes identified as "unavoidable facts due to Globalization". In addition, technological improvements and the ubiquitous aggregate consumption shift are pointed out to play a very important role by giving more participation to service industries at the expense of lower employment levels in manufacturing industries.

Another front in the debate is the evolution of (total factor) productivity in manufacturing. Some economists argue that U.S. Economy has experienced a sharp increase in productivity. Among this group of researchers, Spence and Hlatshwayo (2012) suggest that the value added per job in tradables has risen by more than 40% between 1990 and 2008. In contrast, over-

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Figure 1



Source: own elaboration based on data from the U.S. Bureau of Labor Statistics for Employment and the U.S. Bureau of Economic Analysis for Real GDP, respectively.

all productivity (including both tradables and nontradables) increased only 20% during that period. Their evidence goes in line with the classical statement about the existence of a strong relationship between tradables and productivity evolution. The almost canonical explanation is that tradables are more heavily exposed to competition than nontradables. As a result, the policy implication derived from classic models is that open economies grow faster than closed ones.

On the other hand, some studies consider that labor productivity has been overestimated as a consequence of a miscomputation of temporary labor, which was not included as part of the corresponding sector's total employment (see for example Susan Houseman *et al*. [2011]). Similarly, Luria and Rogers (2007) cast doubts on labor productivity increments that were reported in some recent estimations. They suggest a lower manufacturing productivity growth would be found if some industries like computing and high technology were excluded. Those sectors have had the greatest imputation but the least direct output measurement.

Some celebrated studies (see for example Bernard *et al*. [2006]); Harrison and McMillan (2011); and Author *et al*. [2013]) found some evidence that the import penetration and import competition deteriorate U.S. manufacturing employment. In this paper, I use more recent data and explore empirically different channels not contemplated by those previous studies, which could help to explain a negative (or positive) relationship between employment and trade in the 1998–2008 period. I build on a model that takes into account different responses coming up from different manufacturing sectors. The model goes in line with Greenaway, Hine, and Wright (1999) and Fu and Balasubramanyam (2005). It is important to distinguish which industries are more affected by the presence of foreign competitors. Moreover, it would be desirable to quantify the corre-

sponding impact on employment. Different reactions to imports and exports need to be fully understood if policymakers seek to encourage (discourage) strategic industries.

The organization of the paper is as follows: in section 2 the model is presented and the estimating equation is derived. Section 3 describes the dataset used later to quantify the impact of trade on manufacturing employment. Section 4 explains the estimation method and the main assumptions made in the empirical model. It also presents the estimation results. Finally, section 5 concludes this article.

2. Model

Based on Greenaway, Hine, and Wright (1999) and Fu and Balasubramanyam (2005), we assume a standard Cobb-Douglas production function[2] for the representative firm in industry *i* in a given year *t*

$$Q_{it} = A_{it}^{\gamma} K_{it}^{\alpha} L_{it}^{\beta} \tag{1}$$

In the above expression, Q is real output, K is capital stock, and L represents the units of work used. Variable A accounts for the total factor productivity, whereas the coefficient γ allows for improvements in firm *i*´s efficiency. The parameters α and β represent the shares of capital and labor in industry output, respectively. Assuming profit maximization, from the first order conditions we get

$$Q_{it} = A_{it}^{\gamma} \left(\frac{w}{r} \cdot \frac{\alpha}{\beta} \right)^{\alpha} L_{it}^{\alpha+\beta} \tag{2}$$

where *w* and *r* are the wage rate and the capital cost, respectively. One could normally expect that the average efficiency level of a given production process evolves over time. We hypothesize that variable A varies with time (T), export penetration (X), and import penetration (M) as follows

$$A_{i,t} = e^{\delta_0 T} X_{i,t}^{\ \delta_1} M_{i,t}^{\ \delta_2} \tag{3}$$

The variable T_i is a time trend. The parameter δ_0 is assumed to be positive but no restrictions are imposed to δ_1 and δ_2 (i.e. they could be positive or negative). Normalizing the returns to capital, assuming perfect competition, applying logarithms, and substituting (3) into (2), we finally obtain the following linear equation

$$\ln(L_{it}) = \phi_0 + \phi_1 T_i + \phi_2 \ln(X_{it}) + \phi_3 \ln(M_{it}) + \phi_4 \ln(w_{it}) + \phi_5 \ln(Q_{it}) \tag{4}$$

With the aid of expression (4) we are able to investigate the relationship between trade and employment.

3. Data

We use the OECD-STAN dataset for U.S. industry-specific data. The sample period runs from 1998 to 2008. Table 2 lists the manufacturing industries used in this study, which approximately correspond to a four-digit ISIC level of aggregation. In Table 1, the total output rose 11.2% during the period analyzed. Among the 27 industries considered here, only 13 increased their output (9 out of 17 Low-tech industries, 2 out of 4 Medium-tech industries, and only 2 out of 6 High-tech industries). There was a widespread fall in employment, totalizing a 23% fall in the manufacturing sector jointly considered. Only two marginal sectors (Pharmaceuticals and Shipbuilding) were able to increase the quantity of labor used. However, wage evolution was favorable during the period. There was a 9.5% increment in the average wage rate paid.

2 The Cobb Douglas production function has an intrinsic limitation, namely, elasticity of substitution is equal to one for all industries. The difficulty of relaxing this assumption relies on the impossibility to properly identify wages when using a general Constant Elasticity of Substitution function (CES). The use of other unconventional functional forms is not evaluated in this study.

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Table 1

Changes in manufacturing sectors between 1998 and 2008

Industry	Production (a)			Employment (b)			Wages (c)		
	1998	2008	% var.	1998	2008	% var.	1998	2008	% var.
1500	506.4	602.8	19.0%	1 734.9	1 654	-4.7%	32.6	33.9	3.9%
1600	41.3	28.9	-30.0%	35.1	21	-40.3%	54.8	50.9	-7.1%
1700	100.5	47.4	-52.8%	712.7	342.9	-51.9%	28.3	31.2	10.2%
1800	58.4	13.7	-76.5%	573.4	157.1	-72.6%	21.2	24.7	16.9%
1900	8.9	3.5	-60.4%	81.9	36	-56.0%	27.8	31.5	13.2%
2000	94.8	72	-24.1%	613	459	-25.1%	30	30.1	0.2%
2100	157.9	143.3	-9.3%	630	442	-29.8%	44.7	46.6	4.3%
2200	320	358.9	12.1%	1 843	1 553	-15.7%	49.6	54.7	10.4%
2300	139.6	593.7	325.4%	123	116	-5.7%	63.1	75.6	19.8%
2401	315.9	419.5	32.8%	756.7	577.9	-23.6%	56.6	59.6	5.2%
2423	104.3	150.5	44.2%	236.3	271.1	14.7%	65	73.2	12.7%
2500	167.8	161.7	-3.6%	931	727	-21.9%	35.3	35.9	1.6%
2600	93.4	91.9	-1.7%	539	465	-13.7%	38.2	39.1	2.4%
2700	170.9	225.2	31.8%	640	445	-30.5%	45	48.1	6.9%
2713	95.9	135.8	41.6%	361	248.1	-31.3%	49.6	52	4.9%
2723	75	89.4	19.1%	279	196.9	-29.4%	39.1	43.2	10.5%
2800	258.2	286.2	10.9%	1 752	1 541	-12.0%	37.6	39.5	5.2%
2900	280.8	280	-0.3%	1 502	1 183	-21.2%	45	47.2	5.0%
3000	114.9	55.6	-51.6%	260.4	112.8	-56.7%	60.7	70	15.4%
3100	117.1	104	-11.2%	593	421	-29.0%	38.2	44.5	16.5%
3200	234.6	153.5	-34.6%	1 021.9	630.8	-38.3%	55.1	62.4	13.3%
3300	95.8	117.8	23.0%	545.7	504.4	-7.6%	60.7	77.2	27.2%
3400	449.1	328.8	-26.8%	1 270	883	-30.5%	48.2	45.7	-5.3%
3510	18.3	23.9	30.7%	152.1	162.6	6.9%	35.4	42.4	19.8%
3529	20.1	35.7	77.1%	86	85.8	-0.3%	43.4	46.9	8.1%
3530	147.1	146.6	-0.4%	569.9	482.6	-15.3%	58.3	66.3	13.7%
3637	180.3	188.3	4.4%	1 412	1 115	-21.0%	33.4	37.5	12.2%
Total	**161.8**	**179.9**	**11.20%**	**713.1**	**549.4**	**-23.00%**	**44.3**	**48.5**	**9.50%**

Source: OECD. Industries defined according to 4-digit ISIC.
(a) Production in billions of USD, base year = 2000.
(b) Employment measured in total number of employees.
(c) Annual wage rate in thousands of USD, base year = 2000.

Table 2

Industry codes

ISIC	Description	ISIC	Description
Low-tech industries			Medium-tech industries
1500	Food products and beverages	2401	Chemicals excluding pharmaceuticals
1600	Tobacco products	2900	Machinery and equipment, n.e.c.
1700	Textiles	3400	Motor vehicles, trailers and semi-trailers
1800	Wearing apparel, dressing and dyeing of fur	3529	Railroad equip. and Transport equip. n.e.c.
1900	Leather, leather products and footwear		High-tech industries
2000	Wood and products of wood and cork	2423	Pharmaceuticals
2100	Pulp, paper and paper products	3000	Office, accounting and computing machinery
2200	Printing and publishing	3100	Electrical machinery and apparatus, n.e.c.
2300	Coke, refined petroleum products and nuclear fuel	3200	Radio, television and communication equip.
2500	Rubber and plastics products	3300	Medical, precision and optical instruments
2600	Other non-metallic mineral products	3530	Aircraft and spacecraft
2700	Basic metals		
2713	Iron and steel		
2723	Non-ferrous metals		
2800	Metal prod. (except machinery & equip.)		
3510	Building and repairing of ships and boats		
3637	Manufacturing n.e.c. and recycling		

Table 3 presents a brief description of trade patterns during the period of analysis. Only seven sectors have remained with a positive net-export sign (i.e. a trade surplus): tobacco; printing and publishing; chemicals excluding pharmaceuticals; machinery and equipment not included in other sectors; medical, precision and optical instruments; shipbuilding; aircraft and spacecraft.

Table 3

Continues

Export and import penetration during 1998-2008

Industry	Exports (a)			Imports (b)		
4-digit ISIC	1998	2008	% change	1998	2008	% change
1500	0.058	0.077	33.5%	0.064	0.094	45.4%
1600	0.122	0.024	-79.9%	0.015	0.019	32.2%
1700	0.103	0.218	112.1%	0.223	0.492	121.0%
1800	0.118	0.18	51.8%	0.458	0.81	76.9%
1900	0.255	0.677	165.1%	0.757	0.956	26.2%
2000	0.05	0.059	16.9%	0.142	0.16	12.6%
2100	0.085	0.115	35.7%	0.101	0.126	25.4%
2200	0.027	0.023	-16.2%	0.016	0.019	21.1%

Table 3 Concludes

Export and import penetration during 1998-2008

Industry	Exports (a)			Imports (b)		
4-digit ISIC	1998	2008	% change	1998	2008	% change
2300	0.049	0.083	69.7%	0.109	0.137	24.9%
2401	0.195	0.266	35.9%	0.149	0.227	52.4%
2423	0.113	0.222	97.5%	0.139	0.311	124.3%
2500	0.101	0.148	45.8%	0.113	0.203	79.5%
2600	0.062	0.089	43.5%	0.122	0.159	30.2%
2700	0.12	0.205	71.6%	0.225	0.315	39.6%
2713	0.065	0.121	85.6%	0.194	0.254	30.8%
2723	0.189	0.333	76.2%	0.267	0.41	53.7%
2800	0.055	0.067	21.9%	0.073	0.114	55.7%
2900	0.288	0.419	45.9%	0.273	0.424	55.4%
3000	0.431	0.699	62.2%	0.553	0.837	51.3%
3100	0.256	0.38	48.3%	0.325	0.503	54.7%
3200	0.308	0.494	60.5%	0.343	0.62	80.9%
3300	0.39	0.509	30.4%	0.34	0.487	43.4%
3400	0.149	0.289	94.1%	0.264	0.415	57.4%
3510	0.1	0.111	10.2%	0.069	0.062	-9.8%
3529	0.139	0.128	-7.5%	0.223	0.178	-20.1%
3530	0.443	0.518	17.0%	0.22	0.29	31.7%
3637	0.087	0.17	95.0%	0.249	0.354	42.1%
Total	**0.161**	**0.245**	**52.0%**	**0.223**	**0.332**	**49.0%**

Source: own elaboration based on OECD data.
(a) Exports/Production
(b) Imports/(Production+Imports-Exports).

4. Estimation

4.1 Dynamics and industry fixed effects

It is natural to allow for lags on employment. First, any exogenous shock to employment entails an adjustment cost which may cause a deviation of employment from its steady state path. This fact favors the introduction of lags on employment into our equation (4). The necessary number of lags in order to cope with adjustment costs depends on the heterogeneity observed in the different industries and markets. Hence, if all the industries have a homogeneous adjustment process to a given shock in the labor market and also react swiftly and smoothly, then one lag is enough. Otherwise, the econometrician needs to provide the model with more than one lag.[3]

Other potential sources of heterogeneity appear when technology shocks are serially correlated or when workers negotiation power is strong enough to prompt longer bargaining periods (which extend beyond one calendar year). Finally, with the purpose of avoiding a judgment of causality between employment and the set of explanatory variables used here, we introduce a distributed lag

3 For a discussion on this topic see Nickell (1986).

structure for all the independent variables considered in the empirical model.

We also assume that the independent variables have common impacts across industries. Thus, differencing expression (4) we are able to eliminate the industry fixed effects, and the general dynamic equation to be estimated becomes:

$$\begin{aligned}\Delta \ln(L_{it}) = {} & \lambda_0 + \lambda_{10}\Delta \ln(X_{it}) + \lambda_{11}\Delta \ln(X_{it-1}) + \ldots \\ & + \lambda_{20}\Delta \ln(M_{it}) + \lambda_{21}\Delta \ln(M_{it-1}) + \ldots \\ & + \lambda_{30}\Delta \ln(w_{it}) + \lambda_{31}\Delta \ln(w_{it-1}) + \ldots \\ & + \lambda_{40}\Delta \ln(Q_{it}) + \lambda_{41}\Delta \ln(Q_{it-1}) + \ldots \\ & + \lambda_{51}\Delta \ln(L_{it}) + \ldots + \Delta\varepsilon_{it} \qquad (5)\end{aligned}$$

As it was stated above, the explanatory variables are assumed to be endogenous and causality runs in both directions, so the regressors may be correlated with the error term. To avoid bias in the coefficients of the lagged dependent variable, we adopt a GMM approach using the Arellano and Bond (1991) method. Provided the differenced equation is free of second and higher order serial correlation, this technique proportions unbiased and consistent estimates of the desired parameters. To justify our empirical strategy, we present the Arellano and Bond test statistics to monitor the potential serial correlation. Additionally, we validate our instruments using the traditional Sargan's test, which allows us to check for over-identifying restrictions in the regression model.[4]

4.2 Regression Analysis

4 The Sargan test is based on the observation that the residuals should be uncorrelated with the set of exogenous variables when the instruments are truly exogenous.

Table 4

Continues

Employment equations for United States manufacturing sector

	(1)	(2)	(3)
$\Delta \ln(\text{employment}_{t-1})$	0.959***	0.891***	0.902***
	(0.00851)	(0.0307)	(0.0479)
$\Delta \ln(\text{wage}_t)$	-0.207***	-0.243***	-0.0157
	(0.00883)	(0.0562)	(0.138)
$\Delta \ln(\text{wage}_{t-1})$	0.109***	0.0578	-0.00947
	(0.0132)	(0.0461)	(0.107)
$\Delta \ln(\text{output}_t)$	0.377***	0.351***	0.376***
	(0.00713)	(0.0165)	(0.0256)
$\Delta \ln(\text{output}_{t-1})$	-0.283***	-0.246***	-0.261***
	(0.00768)	(0.0160)	(0.0228)
$\Delta \ln(\text{impo}_t)$		0.0893***	-3.081
		(0.0198)	(2.093)
$\Delta \ln(\text{impo}_{t-1})$		-0.107***	2.676
		(0.0207)	(1.822)
$\Delta \ln(\text{expo}_t)$		0.0451***	1.335
		(0.00965)	(1.120)
$\Delta \ln(\text{expo}_{t-1})$		-0.0606***	-1.660
		(0.0161)	(1.462)

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Table 4 Concludes

Employment equations for United States manufacturing sector

	(1)	(2)	(3)
$\Delta \ln (impo_t) . \Delta \ln (wage_t)$			0.296
			(0.196)
$\Delta \ln (impo_{t-1}) . \Delta \ln (wage_{t-1})$			-0.260
			(0.171)
$\Delta \ln (expo_t) . \Delta \ln (wage_t)$			-0.120
			(0.105)
$\Delta \ln (expo_{t-1}) . \Delta \ln (wage_{t-1})$			0.148
			(0.137)
Observations	268	268	268
$Chi^2_{(29)}$ - Sargan Test	25.59	25.13	20.94
Z - Arellano-Bond Test	-1.92	-1.75	-1.82

Standard error in parenthesis. Significance levels: ***p<0.01, **p<0.05, *p<0.1

Table 4 presents the estimates of the coefficients for three different specifications.

Column (1) reports the base specification. As expected, output causes increases in the level of derived labor demand, whereas wages have a negative effect. The lag employment coefficient is positive, which indicates persistence in wage and output effects on the level of employment. Column (2) incorporates import and export penetration to the basic employment equation. Coefficients associated with wages and output remain unchanged and of similar magnitudes. With regards to trade shares, both export and import have a slightly positive effect in time *t*. However, the impacts are mitigated since the corresponding lag variables have opposite signs of relatively similar magnitudes. Finally, Column (3) explores the impact of trade changes on the slope of the derived labor demand function. In this specification, we evaluate the possibility that increased openness could make it easier to substitute foreign workers for domestic workers and thus, trade variables are interacted with wages. There is no clear effect on the wage elasticity and the corresponding variables are not statistically significant at conventional levels.

5. Concluding comments

After a period of lower exposure to trade during the 1930s and 1940s, the post-war period resulted in a growth of world integration. Trade was the star variable in most industrialized countries and led the economic growth. Even a large economy like United States' that has traditionally had a small share of trade in its GDP was part of the globalization process. And more recently, the impact of expanding trade on labor markets and the corresponding adjustments throughout the economy have generated a renewed interest of many analysts. One of the main concerns emerged from the effect that the rapid expansion of low-wage economies (especially in East Asia) caused on manufacturing jobs. However, a detailed quantification of this decline in employment during the last decades still needs to be studied in depth. Our paper contributes with a simple empirical model that helps to understand the relationship between trade and labor markets at the industry level.

Using a sample of 27 manufacturing industries for the period 1998-2008, we investigate the impact of trade on employment. When trade is introduced in the estimating equation, we find that increases in export and import penetration generate slight increases in the level of labor demand. However, these effects are balanced out in the long run. This evidence seems to be (partially) at odds with the view that increased openness serves to increase the efficiency with which labor is used in the firm and then reduce employment. Instead, it tends to support the idea expressed in Rodrik (1997) that substitution of foreign workers for domestic workers increases the wage elasticity of labor demand.

Further research should focus on the analysis of a richer disaggregation of import and export data in order to see whether the country of origin (of destination) affects labor demand differentially. Additionally, an analysis based on the skill gap literature that incorporates some disaggregation into different categories of labor might be useful to better understand the phenomena observed in U.S. manufacturing sector during the last decades.

References

Abowd, J. *The effects of international competition on collective bargaining agreements in the United States*. Princeton University, unpublished, 1987.

Arellano, M., and S. Bond. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", in: *Review of Economic Studies*, 1991.

Autor, D., D. Dorn, and G. Hanson. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States", in: *American Economic Review* 103(6), 2013.

Bernard, A., J. Jensen, and P. Schott. "Survival of the Best Fit: Exposure to Low-Wage Countries and the (Uneven) Growth of U.S. Manufacturing Plants", in: *Journal of International Economics*, 68(1), 2006.

Baldwin, R. "The effects of trade and foreign direct investment on employment and relative wages", in: *OECD Economic Studies,* Num. 23, 1995.

Caves, R. *Adjustment to International Competition: Short-Run Relations of Prices, Trade Flows, and Inputs in Canadian Manufacturing Industries*. Canadian Government Pub Centre, May 1990.

Greenaway, D., R. Hine, and Wright, P. "An empirical assessment of the impact of trade on employment in the United Kingdom", *European Journal of Political Economy*, 1999.

Fu, X., and V. Balasubramanyam. "Exports, Foreign Direct Investment and Employment: The Case of China", in: *World Economy*, 28(4), 2005.

Harrison, A., and M. McMillan. "Offshoring Jobs? Multinationals and U.S. Manufacturing Employment", in: *Review of Economics and Statistics*, 93(3), 2011.

Houseman, S., C. Kurz, P. Lengermann, and B. Mandel. "Offshoring Bias in U.S. Manufacturing", in: *Journal of Economic Perspectives*. Spring, 2011.

Nickell, S. "Dynamic models of labour demand", in: *Handbook of Labour Economics*. Vol. 1, 1986.

Rodrik, D. *Has Globalization Gone too Far*? Institute for International Economics, 1997.

Rogers, L. *Retooling for Growth. Building a 21st century Economy in America's older industrial areas*. Brookings Institution Press, 2007.

Spence, M., and S. Hlatshwayo. "The evolving structure of the American economy and the employment challenge", in: *Comparative Economic Studies*. Vol. 54, Num. 4, 2012.

Colaboran en este número

Carlos Alberto Francisco Cruz

Es licenciado en Economía por la Universidad Nacional Autónoma de México (UNAM) y maestro en Gobierno y Asuntos Públicos por la Facultad Latinoamericana de Ciencias Sociales (FLACSO). Actualmente, trabaja como consultor en métodos estadísticos en el Proyecto de Evaluación de Políticas de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) en México y se desempeña como profesor de asignatura en la Facultad de Economía de la UNAM.

Contacto: carlos.francisco@fao.org

Jorge Lara Álvarez

Se tituló como licenciado en Economía por el Centro de Investigación y Docencia Económicas (CIDE) y es *MSc* en *Ag & Resource Economics* por la Universidad de Arizona, así como *MPhil* en *Economics* por la Universidad de Oxford. En la actualidad, trabaja en la Subdirección de Investigación Económica de Fideicomisos Instituidos en Relación con la Agricultura (FIRA).

Contacto: jlara@fira.gob.mx

Juan Francisco Islas Aguirre

Cuenta con estudios de Licenciatura y Maestría en Economía por el CIDE. Se desempeña como profesor de asignatura en la Facultad de Economía de la UNAM. Es consultor en métodos estadísticos en el Proyecto de Evaluación de Políticas de la FAO en México.

Contacto: juanfrancisco.islas@fao.org

Ana Karen Díaz Méndez

Es licenciada en Economía por la UNAM. Trabajó en el Programa de las Naciones Unidas para el Desarrollo (PNUD) México en el área de Reducción de Pobreza. Posteriormente, colaboró en el Proyecto de Evaluación de Políticas de la FAO en México. En ambas agencias participó en evaluaciones de impacto, costo-beneficio y resultados. En la actualidad, estudia una maestría en Economía Cuantitativa en la Universitat Autónoma de Barcelona.

Contacto: akaren.diaz.mendez@gmail.com

Felipe Pérez Gachuz

Es maestro en Economía Aplicada por el Posgrado de la Facultad de Economía de la UNAM. Tiene una especialidad en Teoría Económica y es licenciado en Economía. Actualmente, es consultor del Proyecto de Evaluación de Políticas de la FAO.

Contacto: felipe.perez@fao.org

Francisco José Zamudio Sánchez	**Es** *PH. D.* (*Major in Statistics*) por la Universidad Estatal de Iowa. Se desempeña como profesor-investigador en la Universidad Autónoma Chapingo (UACh) desde 1975 y es coordinador del posgrado de la Maestría en Ciencias Forestales desde el 2010, así como director del Programa Nacional sobre Desarrollo Humano en México de la UACh desde el 2000. Ha publicado artículos sobre temas de género, desarrollo humano y recursos forestales en diversas revistas. Es árbitro de *Agrociencia* (ISI y CONACYT), *Fitotecnia Mexicana* (ISI y CONACYT), *El Colegio de Sonora* (CONACYT), *Estudios Sociales* (CONACYT, Red ALyC, LATINDEX y otros) y la *Revista Chapingo. Serie Ciencias Forestales y del Ambiente* (ISI, CONACYT, JCR). **Contacto:** zafra1949@gmail.com
Roxana Ivette Arana Ovalle	Es maestra en Ciencias (Estadística) por el Colegio de Postgraduados, México, y licenciada en Estadística por la UACh. Trabajó como investigadora asociada del 2010 al 2013 en el proyecto de investigación: Actitudes, Prácticas, Factores que Inciden y Espacios donde se Producen y Reproducen la Violencia de Género y Sexismo en la UACh, del Fondo Sectorial SEP-CONACYT, y durante 2012-2014 fue investigadora del proyecto Muestreo Probabilístico para la Recuperación de los Microdatos del Censo General de 1930, del Fondo Sectorial CONACYT-INEGI. En fechas recientes, publicó artículos sobre género, finanzas y desarrollo humano. Ha sido asesora de tesis de alumnos de la UACh en temas de violencia y género. **Contacto:** roxarana@icloud.com
Waldenia Cosmes Martínez	Es licenciada en Estadística por la UACh, becaria y coordinadora de construcción de base de datos del proyecto de investigación Muestreo Probabilístico para la Recuperación de los Microdatos del Censo General de 1930, del Fondo Sectorial CONACYT-INEGI. **Contacto:** cosmeswal@gmail.com
Javier Santibáñez Cortés	Es licenciado en Estadística por la UACh. Se desempeñó como coordinador de análisis estadístico en el proyecto de investigación Muestreo Probabilístico para la Recuperación de los Microdatos del Censo General de 1930, del Fondo Sectorial CONACYT-INEGI, y es asistente de investigador en la UACh. En fecha reciente publicó *Informe estadístico sobre desarrollo humano en México 1995-2010*, editado por la UACh en el 2012. **Contacto:** javsantibanezcor@gmail.com
Margoth Laredo Rojas	Es licenciada en Estadística por la UACh. Trabajó como coordinadora operativa en el proyecto de investigación Muestreo Probabilístico para la Recuperación de los Microdatos del Censo General de 1930, del Fondo Sectorial CONACYT-INEGI. **Contacto:** lrm.210390@gmail.com

Miguel Ángel Suárez Campos

Es licenciado en Física y Matemáticas por el Instituto Politécnico Nacional (IPN) y tiene el grado de maestro en Estadística Oficial por el Centro de Investigación en Matemáticas (CIMAT). Ha trabajado en el INEGI por más de 31 años participando en la sistematización de la Encuesta Nacional de Ocupación y Empleo, antes Encuesta Nacional de Empleo Urbano, y de los registros administrativos demográficos y sociales; los últimos 10 años los ha dedicado a la aplicación de metodología estadística a los registros administrativos y al estudio de la estimación para áreas pequeñas.

Contacto: miguel.suarez@inegi.org.mx

Gustavo Aguilar Mata

Es actuario por la UNAM y ostenta el grado de maestro en Estadística Oficial por el CIMAT. Ha realizado trabajo actuarial en el mercado de seguros por espacio de 17 años; en el INEGI tiene 20 años en los que ha realizado la sistematización de las estadísticas judiciales y laborales; los últimos 10 los ha dedicado a la aplicación de metodología estadística a los registros administrativos y al estudio de la estimación para áreas pequeñas.

Contacto: gustavo.mata@inegi.org.mx

Raúl Mejía González

Es ingeniero químico por la UNAM y concluyó el plan de estudios de la Maestría en Ingeniería Química en la misma universidad. Ha trabajado en el INEGI por siete años, durante los cuales ha colaborado en el estudio de la estimación para áreas pequeñas, en la validación y explotación de varias encuestas especiales solicitadas al INEGI y en la elaboración de análisis estadísticos aplicados a dichas encuestas.

Contacto: raul.mejia@inegi.org.mx

Roberto Antonio Vázquez Espinoza de los Monteros

Es ingeniero por la Escuela Superior de Cómputo del IPN, así como maestro y doctor en Ciencias de la Computación por el Centro de Investigación en Computación del IPN. Actualmente, es profesor-investigador en la Facultad de Ingeniería de la Universidad La Salle, campus Ciudad de México. Sus principales áreas de interés son la inteligencia artificial, redes neuronales artificiales, neurociencias computacionales, reconocimiento de patrones, análisis de imágenes y computo evolutivo.

Contacto: ravem@lasallistas.org.mx

José Ambrosio Bastián

Es ingeniero en Electrónica y Comunicaciones por la Universidad Veracruzana, campus Poza Rica, así como maestro en Ciencias de Ingeniería en Microelectrónica y doctor en Comunicaciones y Electrónica por la Sección de Estudios de Posgrado e Investigación de ESIME Culhuacán del IPN. En la actualidad, es profesor-investigador en la Facultad de Ingeniería de la Universidad La Salle. Sus principales áreas de interés son redes neuronales, lógica difusa, reconocimiento de patrones e inteligencia artificial.

Contacto: jose.ambrosio@lasallistas.org.mx

INEGI. REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA. Vol. 6, Núm. 3, septiembre-diciembre 2015. 2022

Guillermo Alberto Sandoval Sánchez

Es ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Culiacán y maestro en Ciencias, Área Cibertrónica por la Universidad La Salle. Hoy en día, es ingeniero de *software* en *Critical Mix*. Sus principales áreas de interés son procesamiento de imágenes y de voz, así como en control inteligente.

Contacto: guillermo.sandoval@lasallistas.org.mx

César Bistrain Coronado

Es actuario por la Facultad de Ciencias de la UNAM y maestro en Población y Desarrollo por la FLACSO. Actualmente, se desempeña como jefe del Departamento de Evaluación de Procesos en el Instituto Nacional de Estadística y Geografía. Cuenta con una trayectoria profesional enfocada a la demografía y construcción de indicadores estadísticos. Ha proporcionado asesorías a instituciones nacionales y extranjeras.

Contacto: cbistrain@gmail.com

Pedro I. Hancevic

Es doctor, *magister* y licenciado en Economía por la *University of Wisconsin-Madison*, la Universidad Nacional de La Plata y la Universidad Nacional de Córdoba, en ese orden. En la actualidad, se desempeña como profesor-investigador titular en el CIDE, sede Región Centro. Anteriormente, trabajó como economista júnior en la Fundación de Investigaciones Económicas Latinoamericanas (FIEL), así como de asistente de investigación en el *Center for Financial Security* (CFS) y en el *Center on Wisconsin Strategy* (COWS). Sus áreas de interés son la organización industrial y defensa de la competencia, microeconomía y microeconometría aplicada, además de la economía de la energía y el medio ambiente.

Contacto: pedro.hancevic@cide.edu

Lineamientos para publicar en
Realidad, datos y espacio.
Revista Internacional de Estadística y Geografía

Los trabajos presentados a Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía deberán tratar temas de interés relativos a la situación actual de la información estadística y geográfica.

Sólo se reciben para su posible publicación trabajos inéditos, en español o inglés. Por ello, es necesario anexar una carta dirigida al editor de Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía, en la que se proponga el artículo para su publicación y se declare que es inédito y que no se publicará en otro medio. En esta carta deben incluirse los datos completos del autor o autores, institución, domicilio completo, correo electrónico y teléfono. El envío de los artículos debe dirigirse a la atención de la M. en C. Virginia Abrín Batule, virginia.abrin@inegi.org.mx (tel. 5278 10 00, ext. 1161).

Los trabajos se tienen que presentar en versión electrónica (formato *Word* o compatible), en la cual se inclu-yan las imágenes, gráficas y cuadros (en el formato de los programas con que fueron generados y en archivos independientes, tales como Adobe Photoshop, Adobe Illustrator, TIF, EPS, PNG o JPG, con una resolución de 300 dpi y en un tamaño de 13 x 8 cm). Las expresiones y/o algoritmos, enviarlas con el formato anterior. Se sugiere una extensión de 15 cuartillas, tipo de letra Helvética, Arial o Times de 12 puntos e interlineado de 1.5 líneas.

Los artículos deben incluir: título del trabajo, nombre completo del autor o autores, institución donde trabaja y cargo que ocupa, teléfonos, correo electrónico, breve semblanza del autor o autores (que no exceda de un párrafo de cinco renglones), 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) y bibliografía u otras fuentes.

Las referencias bibliográficas deberán presentarse al final del artículo de la siguiente manera: nombre del autor comenzando por el o los apellidos; título del artículo (entrecomillado); título de la revista o libro donde apareció publicado (en cursivas); editor o editorial; lugar y año de edición. En el caso de las fuentes electrónicas (páginas *Web*) se seguirá el mismo orden que en las bibliográficas, pero al final entre paréntesis se pondrá DE (dirección electrónica), la fecha de consulta y la liga completa. Omitir las que se mencionen como notas a pie de página.

Todos los artículos recibidos serán sometidos a evaluación y el proceso de dictaminación será de acuerdo con la metodología de doble ciego (autores y dictaminadores anónimos).

GUIDELINES FOR PUBLISHING IN
Reality, Data and Space.
International Journal of Statistics and Geography

The papers submitted to Reality, Data and Space. International Journal of Statistics and Geography, must deal with issues of interest relating to state-of-the-art statistical and geographical information.

Only unpublished works, in English or Spanish will be accepted for possible publication. Therefore, it is required to attach a letter addressed to the Publisher of Reality, Data and Space. International Journal of Statistics and Geography, proposing the article for publication and stating it is unpublished material and it will not be published in any other way. The letter must include the full details of the author or authors, institution, full address, e-mail and telephone number. The dispatch of the articles should be directed to the attention of the M. C. Virginia Abrín Batule, virginia.abrin@inegi.org.mx (tel. 5278 10 00, ext. 1161).

Contributions must be submitted in electronic format (Word format or compatible), containing the images, charts and tables (in the original format of the software they were created on, and in separate files, such as Adobe Photoshop, Adobe Illustrator, TIF, EPS, PNG or JPG, with a resolution of 300 dpi and a 13 x 8 cm size of). The equations and or the algorithm send it in the same form. An extension of 15 pages, Helvetica, Arial or Times 12 points typeface, and a spacing of 1.5 lines is suggested.

The articles should include: title, full name of the author or authors, institution where he/she works and her/his position, phone, e-mail, a brief biography of the author or authors (not exceeding a 5 lines paragraph), summaries of the work, in English and Spanish (not exceeding a 10 lines paragraph), keywords, in English and Spanish (minimum 3, maximum 5) and bibliography reference list.

Bibliographical references must appear at the end of the article as follows: Author's name beginning with the surname; article's Title (in quotation marks); Title of the magazine or book where it was published (in italics); Publisher or editorial; house and year of the edition. In the case of electronic sources (Web pages) it will be used the same arrangement as for bibliographical references, but it will be followed by the mention DE (dirección electronica, in Spanish) between backets, the date of consultation and the full link.

All contributions received will be subject to evaluation and the approval process will be carried according to the methodology of double-anonymity (anonymous authors and adjudicators).

Lo que dicen los números...

Indicadores sobre **corrupción y programas anticorrupción**

Administración pública y corrupción

Experiencias de corrupción en trámites públicos

47.6% de la población cree que existen actos de corrupción en el trámite que realizó.

12.1% de la población vivió un acto de corrupción al realizar un trámite o solicitud de servicio público.

Administraciones públicas estatales

Auditorías y/o revisiones practicadas

12 037

Sanciones aplicadas

12 110

Servidores públicos sancionados

11 259

Principales trámites que incluyen los programas anticorrupción

Administraciones públicas con programas anticorrupción

Corrupción por tipo de trámite*

* Se refiere al porcentaje de contactos institucionales con servidores públicos en los que al menos en una ocasión el usuario experimentó una situación de corrupción.

Fuentes: INEGI. *Encuesta Nacional de Calidad e Impacto Gubernamental 2013.*
_____*Censo Nacional de Gobiernos Municipales y Delegacionales 2013.*
_____*Censo Nacional de Gobierno, Seguridad Pública y Sistema Penitenciario Estatales 2014.*

Conociendo México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa **@INEGI_INFORMA**

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

7 02825 07838 6

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

MÉXICO