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

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).

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:

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

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:

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

es insesgado ya que:

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

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:

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

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.

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:

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

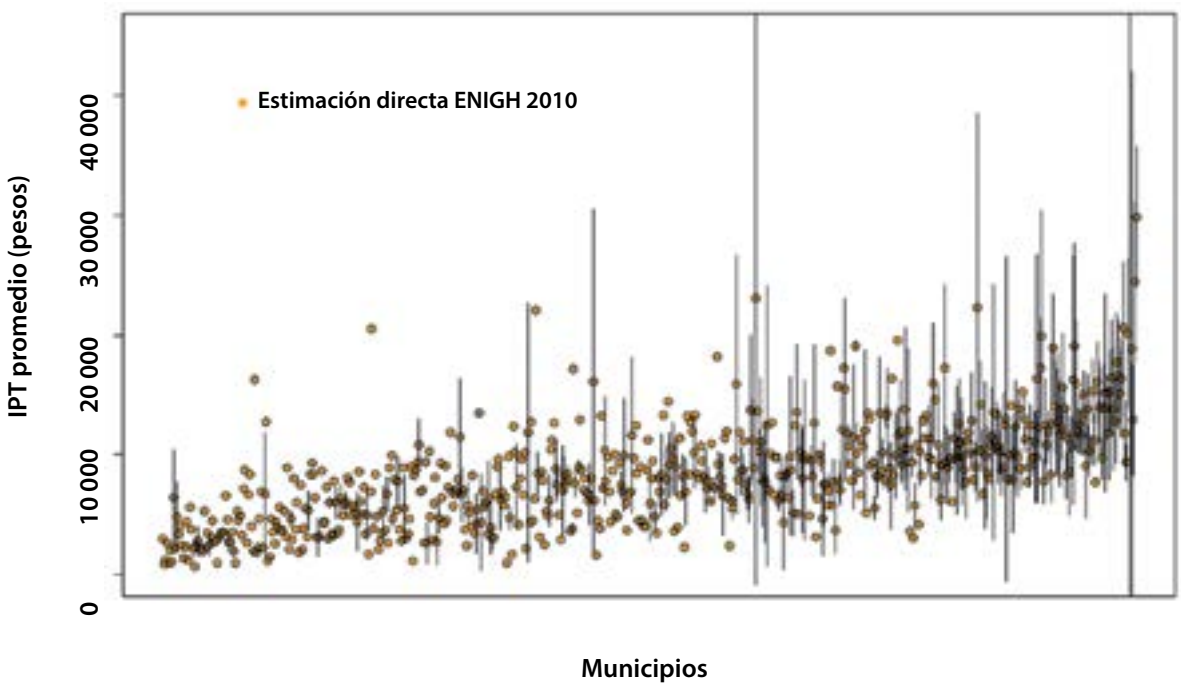
Se debe aclarar que para varios municipios no fue posible calcular la varianza debido a que sólo una UPM fue muestreada en ellos; por esta razón, en la gráfica 1 varios promedios estimados no incluyen su intervalo de confianza.

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:

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

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.

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 $\hat{\sigma}^2_{\upsilon} = 112.467$.

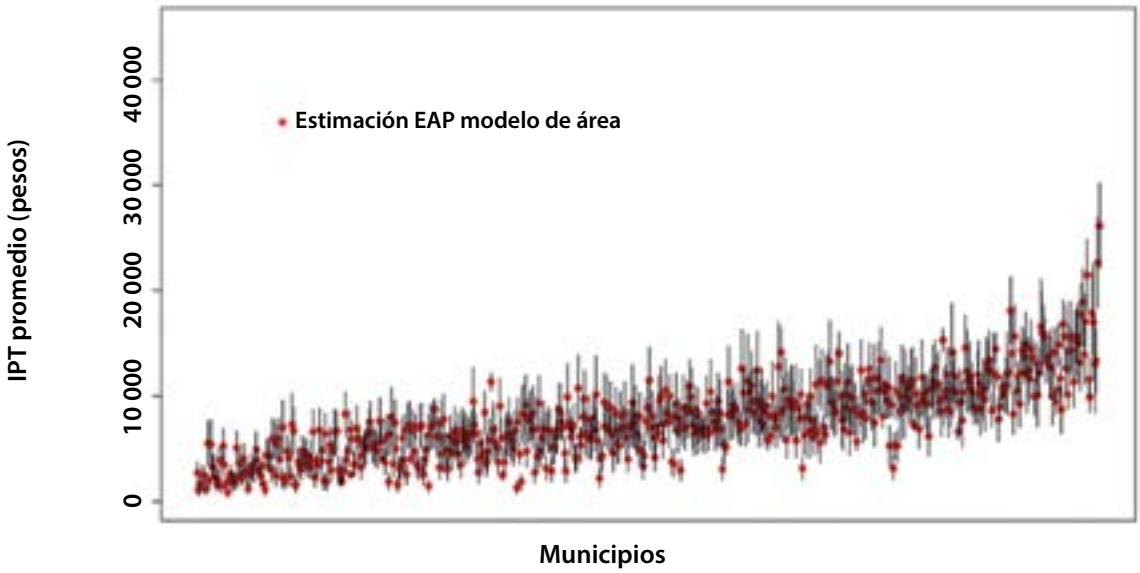
La escala de la gráfica 2 es igual a la de la 1 con el propósito de facilitar la comparación visual entre ellas; al hacerlo se aprecia, a simple vista, que las estimaciones se condensan siguiendo el modelo planteado, además de que se homogeneizan los intervalos de confianza a una longitud razonable,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



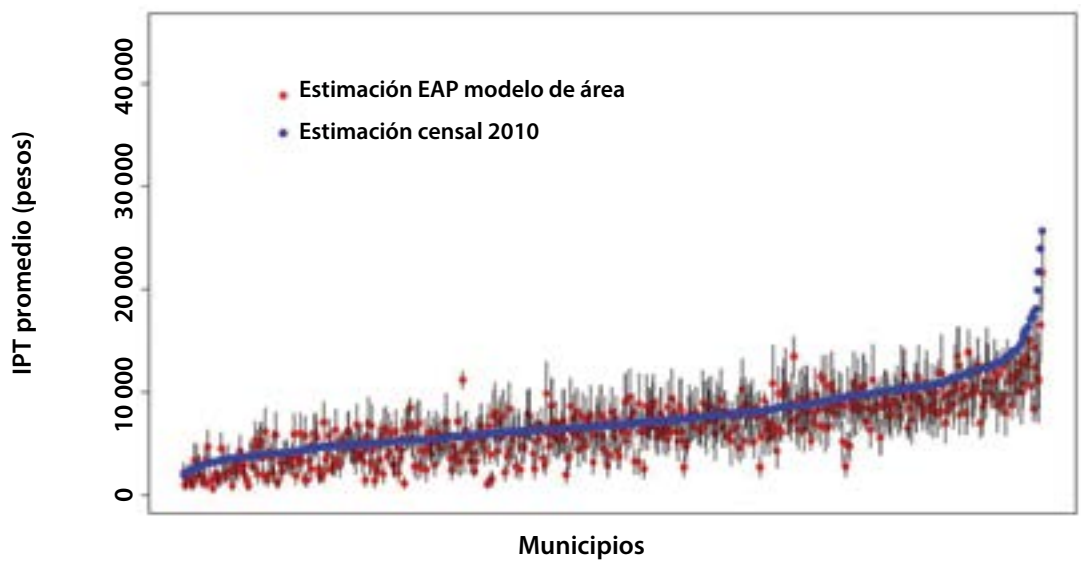
La gráfica 3 se construye tomando como base la 2, a la cual se añaden los puntos en azul que corresponden a la estimación del promedio del ingreso por trabajo en la vivienda de la muestra del CPV 2010 en los municipios con muestra de la ENIGH 2010; se observa cómo la condensación citada en la gráfica 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%



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:

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

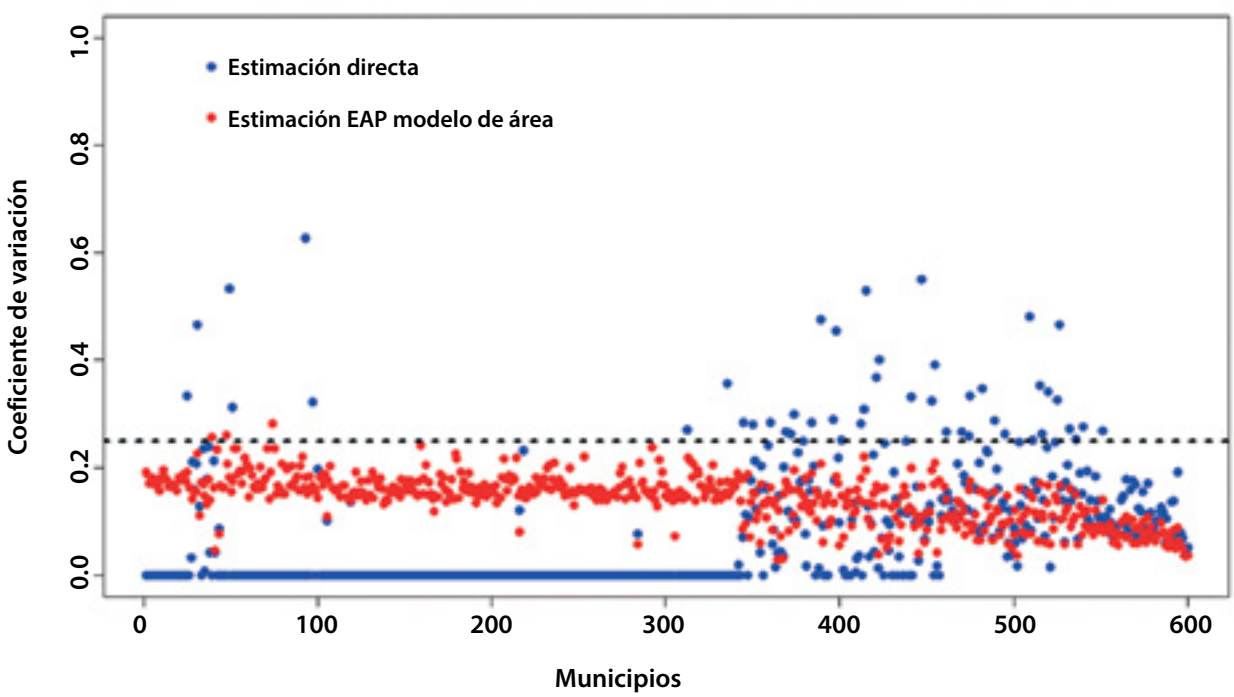
donde está considerada la transformación a las unidades originales.

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

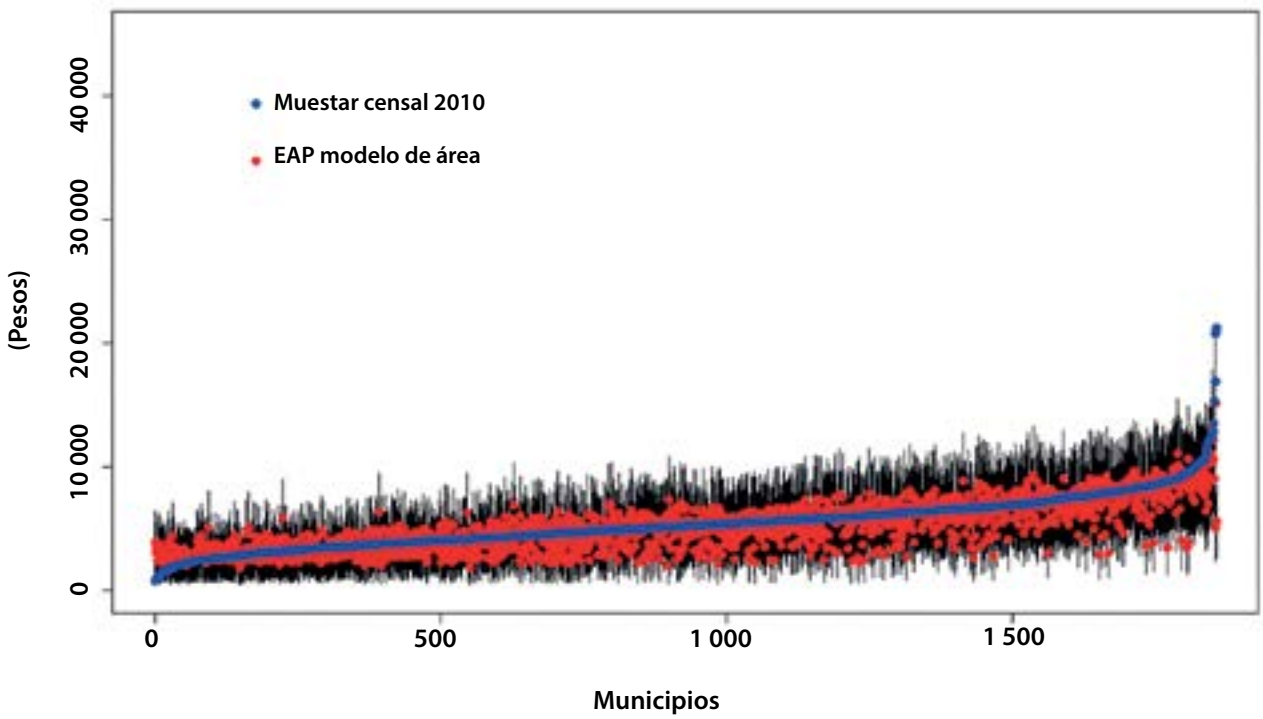
Gráfica 4

Coeficientes de variación de las estimaciones, municipios con muestra



Gráfica 5

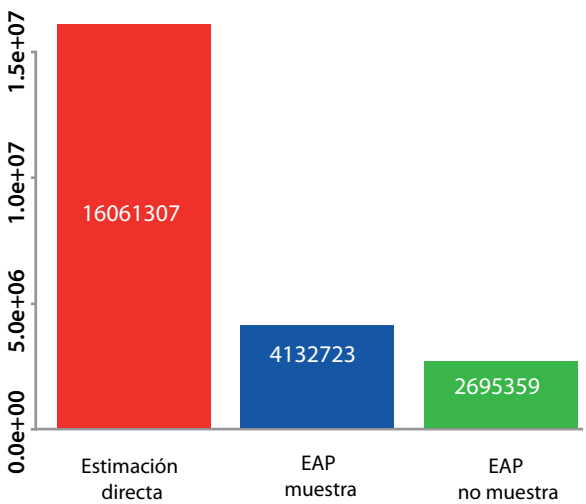
Municipios sin muestra ENIGH 2010, promedios estimados con intervalos a 95%



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

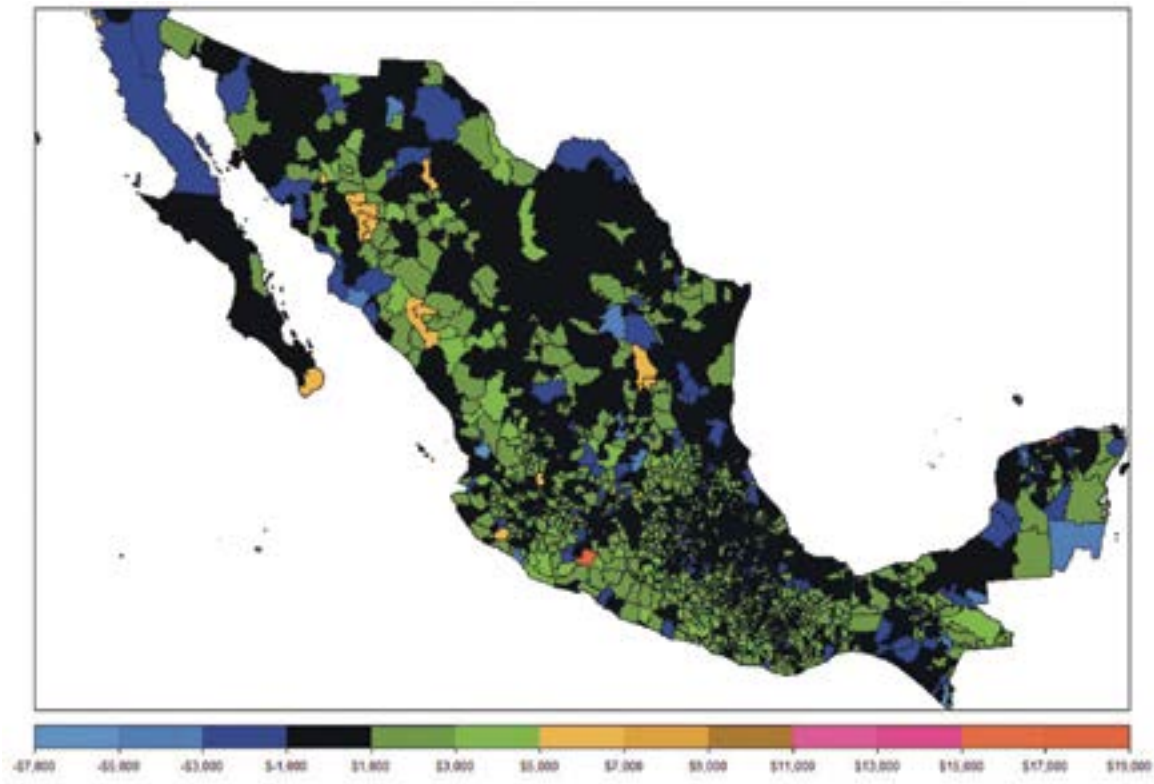
MSE empírico de las estimaciones respecto al valor de la muestra censal 2010



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.

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.