ISSN 1680-8789

SERIE

ESTUDIOS ESTADÍSTICOS 109

Estimación en áreas pequeñas de los indicadores de pobreza en América Latina

Una aplicación basada en modelos de regresión multinivel con posestratificación

Andrés Gutiérrez
Xavier Mancero
Stalyn Guerrero

NACIONES UNIDAS

CEPAL

Gracias por su interés en esta publicación de la CEPAL

NACIONES UNIDAS

CEPAL

Si desea recibir información oportuna sobre nuestros productos editoriales y actividades, le invitamos a registrarse. Podrá definir sus áreas de interés y acceder a nuestros productos en otros formatos.

Deseo registrarme

Conozca nuestras redes sociales y otras fuentes de difusión en el siguiente link:

https://bit.ly/m/CEPAL

SERIE

ESTUDIOS ESTADÍSTICOS | 109

Estimación en áreas pequeñas de los indicadores de pobreza en América Latina

Una aplicación basada en modelos de regresión multinivel con posestratificación

Andrés Gutiérrez
Xavier Mancero
Stalyn Guerrero

NACIONES UNIDAS

CEPAL

Este documento fue preparado por Andrés Gutiérrez, Experto Regional en Estadísticas Sociales; Xavier Mancero, Jefe de Unidad, y Stalyn Guerrero, Consultor, todos de la Unidad de Estadísticas Sociales de la División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL), en el marco de las actividades del 13er tramo de la Cuenta de las Naciones Unidas para el Desarrollo. Se agradecen los aportes iniciales al documento realizados por Gabriel Nieto, Consultor de la Unidad de Estadísticas Sociales de la División de Estadísticas de la CEPAL.

Publicación de las Naciones Unidas
ISSN: 1680-8789 (versión electrónica)
ISSN: 1994-7364 (versión impresa)
LC/TS.2024/137
Distribución: L

Impreso en Naciones Unidas, Santiago
S.2401206[S]

Esta publicación debe citarse como: A. Gutiérrez, X. Mancero y S. Guerrero, "Estimación en áreas pequeñas de los indicadores de pobreza en América Latina: una aplicación basada en modelos de regresión multinivel con posestratificación", serie *Estudios Estadísticos*, N° 109 (LC/TS.2024/137), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL), 2025.

Índice

Cuadros

Gráficos

Mapas

Resumen

El diseño y formulación de políticas públicas efectivas requiere de información sobre las condiciones de vida de la población que permita identificar a las personas más necesitadas y las zonas en las que habitan. Las metodologías de estimación en áreas pequeñas han ganado creciente aceptación como un mecanismo para producir estadísticas desagregadas, con una mayor precisión de la que permiten las encuestas de hogares por sí solas. Este documento presenta una aplicación para generar cifras de pobreza, pobreza extrema e ingreso medio, desagregadas por sexo, edad, nivel educativo y grupo étnico, a nivel de la primera división administrativa de 17 países de América Latina, utilizando un modelo de regresión multinivel con postestratificación. El método combina la información proveniente de las encuestas de hogares con la que proveen los censos de población, a la que se aplica un procedimiento de actualización de estructuras demográficas, y las imágenes satelitales.

Introducción

La necesidad de contar con información oportuna y representativa de diversos grupos de la población es expresada de manera clara en la Agenda 2030 para el Desarrollo Sostenible y en su mandato de "no dejar a nadie atrás". El marco global de indicadores para el seguimiento de los Objetivos de Desarrollo Sostenible (ODS) expresa que estos deben desagregarse, cuando sea pertinente, por ingreso, sexo, edad, raza, etnia, condición migratoria, discapacidad, ubicación geográfica u otras características, de acuerdo con los Principios Fundamentales de las Estadísticas Oficiales (Naciones Unidas, 2019a).

Las encuestas de hogares son una de las fuentes de información habitualmente utilizadas para dar cuenta de las condiciones de vida de la población, y casi un tercio de los indicadores globales de los ODS pueden derivarse de esta fuente de información (Naciones Unidas, 2019b). En los países de América Latina, es habitual que los indicadores obtenidos a partir de las principales encuestas de hogares puedan desagregarse al menos por área (urbana o rural), sexo, grupos de edad y quintil de ingresos. En un conjunto más restringido de países y encuestas es posible contar con datos desagregados según grupo étnico, condición de discapacidad y situación migratoria. No obstante, las encuestas están limitadas en su capacidad para producir estimaciones representativas cuando la población de interés corresponde a una combinación de diversas categorías de desagregación (por ejemplo, la tasa de ocupación para personas inmigrantes de 30 a 40 años de edad) o cuando se quiere contar con información a menores niveles geográficos (por ejemplo, provincias o municipios).

En años recientes, las metodologías de estimación en áreas pequeñas (*Small Area Estimation* o SAE, en inglés) han cobrado relevancia como una alternativa para producir estimaciones desagregadas de buena calidad. Los métodos SAE son empleados para obtener estimaciones desagregadas de parámetros poblacionales para mejorar la calidad de la inferencia cuando la desagregación directa no cumple con los criterios requeridos para su utilización. Esta metodología funciona a partir de la integración de información de diversas fuentes para obtener estimaciones más precisas y ganar fuerza predictiva; es decir, a la información obtenida de las encuestas, se suma la información de otras fuentes con mayor capacidad de desagregación, como los censos de población, las imágenes satelitales, los registros administrativos u otro tipo de información que se tenga disponible para todas las unidades de observación.

En los últimos años se ha multiplicado la cantidad de recursos para difundir los métodos de SAE y su adopción por parte de las oficinas nacionales de estadística y otros organismos productores de estadísticas

oficiales. Algunos ejemplos son la "caja de herramientas" sobre SAE de la División de Estadísticas de Naciones Unidas (véase https://unstats.un.org/wiki/display/SAE4SDG/), algunos manuales generados por organismos internacionales (Corral y otros, 2022; Asian Development Bank, 2020), junto con cursos y seminarios realizados en América Latina por la CEPAL.

Varios países de la región han incursionado en el uso de modelos de estimación en áreas pequeñas; por ejemplo, desde el año 2011, el Ministerio de Desarrollo Social de Chile ha venido publicando de forma bianual la estimación de la incidencia de la pobreza a nivel comunal con este tipo de metodologías (véase https://observatorio.ministeriodesarrollosocial.gob.cl/pobreza-comunal); asimismo, el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) publica cada cinco años las estimaciones de la incidencia de la pobreza multidimensional para todos los municipios de México (véase https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipio-2010-2020.aspx); de la misma forma, el Instituto Nacional de Estadística e Informática publicó en 2018 el mapa de pobreza monetaria provincial y distrital (véase https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1718/Libro.pdf) y el Departamento Nacional de Estadística en Colombia ha impulsado el uso de este tipo de estadísticas dentro de sus procesos (https://www.cepal.org/sites/default/files/presentations/proyecto-sae-colombia_natalia-arteaga-dane_nov2023.pdf).

Por su parte, la División de Estadísticas de la CEPAL recopila y armoniza diversas encuestas de hogares de los países de América Latina, con las que produce anualmente más de 100 indicadores sobre la situación socioeconómica de la población, en temas como pobreza, distribución del ingreso, educación, empleo, vivienda y servicios básicos, entre otros. Para algunos de estos indicadores, la CEPAL ha implementado un proceso de estimación en áreas pequeñas que permita su desagregación a nivel municipal, utilizando como fuentes de información complementaria los censos de población y variables derivadas de datos satelitales (Gutiérrez y otros, 2022).

Uno de los desafíos que se enfrenta al utilizar los censos de población como fuente secundaria de datos para los modelos de SAE es la desactualización de la información. Estos modelos requieren generalmente que el censo y la encuesta correspondan a un período similar. Por tanto, en la medida que pasa el tiempo, la posibilidad de utilizar el censo para tomar prestada la capacidad de desagregación va disminuyendo.

Este trabajo presenta una alternativa metodológica que permite producir estadísticas desagregadas de pobreza de manera continua, sin dejar de utilizar el censo como la fuente de información complementaria a la encuesta. Para ello, se recurre a un modelo de regresión multinivel con post estratificación (*Multilevel Regression with Postestratification* o MRP en inglés), que se combina con un proceso de actualización de la estructura de la población del censo. Como se muestra a continuación, este modelo ofrece una alternativa viable para producir estadísticas de pobreza desagregadas según combinaciones de distintas características del individuo, hasta el nivel geográfico de primera división administrativa en los países de América Latina.

El documento está estructurado de la siguiente manera. Después de una breve introducción, el primer capítulo presenta las diferentes técnicas utilizadas para la estimación en áreas pequeñas, particularmente en el contexto de la medición de la pobreza. Se presenta el modelo de regresión multinivel propuesto y se describe sus principales características, así como las similitudes y diferencias con un modelo de unidad "tradicional", como el mejor predictor empírico. El segundo capítulo describe las tres fuentes de información utilizadas para la aplicación de la metodología, las encuestas de hogares, los censos de población y las imágenes satelitales, junto con su preparación para la estimación en áreas pequeñas de los indicadores de pobreza en América Latina. El tercer capítulo presenta el proceso de actualización de tablas censales con base en encuestas de hogares, como una forma de superar la pérdida de vigencia de la información censal a medida pasa el tiempo desde la realización del censo. Este proceso busca mantener la correlación natural entre las variables del censo, lo cual es esencial para obtener resultados más precisos mediante metodologías de estimación en áreas pequeñas. El cuarto capítulo describe el ajuste de los modelos para los indicadores de ingreso medio, pobreza y pobreza extrema y la predicción de estos indicadores a nivel de las áreas pequeñas de interés. Se utiliza un modelo de regresión multinivel basado

en postestratos, combinando efectos aleatorios y fijos. Para el ingreso medio, se asume normalidad en los efectos aleatorios y errores; mientras que para los indicadores de pobreza y pobreza extrema se emplea un modelo logístico con enlace Logit, adaptado al enfoque bayesiano. Se realizan chequeos predictivos posteriores y se valida la convergencia de las cadenas con la estadística R-hat. El quinto capítulo describe el proceso de *benchmarking*, clave para comparar las estimaciones de pobreza obtenidas con modelos frente a los indicadores directos derivados de encuestas. El *benchmarking* asegura que las estimaciones del modelo a nivel individual sean consistentes con las estimaciones directas y reduce el sesgo. Esto se logra mediante la calibración multivariada, ajustando los ponderadores en las simulaciones por métodos de Montecarlo basados en cadenas de Markov (*Markov Chain Montecarlo* o MCMC, en inglés). El proceso incluye validaciones adicionales para verificar la precisión de las estimaciones ajustadas, comparándolas con los datos de la encuesta y realizando análisis visuales de consistencia. Este capítulo aborda también la estimación del error cuadrático medio. El sexto capítulo expone algunas conclusiones resultantes de la aplicación de esta metodología en los países de América Latina. Finalmente, en los anexos se presentan los resultados obtenidos mediante el modelo descrito en su aplicación para 17 países de América Latina.

I. Metodologías de estimación en áreas pequeñas

A. Estimadores directos

Un estimador directo de área es aquel que toma únicamente la información de la encuesta disponible en esa área específica. Los estimadores directos son ampliamente utilizados para la obtención de estadísticas oficiales gracias a sus propiedades de insesgamiento, eficiencia y consistencia.

Como lo expresa Molina (2019), el estimador de Horvitz-Thompson (HT) es un estimador insesgado de la variable Y con respecto al diseño muestral de la media del área *d*, $\widehat{Y}_d$. Para emplear este estimador es imprescindible conocer el tamaño real del área N_d donde se desea estimar el indicador, además de los pesos muestrales $w_{di} = \pi_{di}^{-1}$ para los individuos en el área *d* que pertenecen a la muestra. De acuerdo con lo anterior, el estimador de *HT* de $\widehat{Y}_d$ se expresa de la siguiente manera:

$$\widehat{Y}_d = N_d^{-1} \sum_{i \epsilon s_d} w_{di} Y_{di} \tag{1}$$

Ahora bien, se debe tener en cuenta que para estimar el total de la variable *Y* para el área *d*, $Y_d = \sum_{i=1}^{N_d} Y_{di}$, el estimador *HT* toma la forma $\widehat{Y}_d = \sum_{i \epsilon s_d} w_{di} Y_{di}$, por lo cual no es necesario que el tamaño poblacional del área N_d sea conocido de antemano.

Otro estimador directo ampliamente utilizado es el estimador de Hájek; para la estimación de la media $\widehat{Y}_d$ no se hace necesario conocer el tamaño del área N_d, debido a que este estimador es igual a la media ponderada en las observaciones del área, en el que los pesos muestrales son empleados como ponderadores así:

$$\widehat{Y}_d^{HA} = \widehat{N}_d^{-1} \sum_{i \epsilon s_d} w_{di} Y_{di}, \qquad \text{donde} \quad \widehat{N}_d = \sum_{i \epsilon s_d} w_{di} \tag{2}$$

Un estimador "directo" siempre se rige bajo un diseño de muestreo (principio de representatividad) que utiliza los datos de la encuesta y los factores de expansión para poder realizar inferencias sobre los parámetros de interés del estudio (Gutiérrez, 2016). De este modo, gracias a algunas medidas de precisión como el error estándar y el error cuadrático medio, es posible determinar si los resultados obtenidos cumplen con los criterios necesarios para ser aceptados. En general, una estadística se hace oficial si

cumple con los criterios de confiabilidad para su publicación, lo que significa, que el error de muestreo relativo o coeficiente de variación (CV) no supere un umbral de precisión definido por la institución que la genera, entre otros criterios de calidad (Gutiérrez y otros, 2020).

Los dominios usualmente empleados en las encuestas de hogares pueden ser geográficos, sociodemográficos y mixtos. Los dominios geográficos corresponden a la estructura político-administrativa de un país (departamentos, estados, provincias, comunas, municipios, entre otros) o a distribuciones territoriales especiales (distritos escolares, zonas de análisis de transporte, zonas de servicios de salud, cuadrantes de seguridad ciudadana, entre otras) (Janicki y Vesper, 2017; Pratesi, 2016). Los dominios sociodemográficos hacen referencia a las clasificaciones de la población basadas en características propias como el sexo, la etnia, el nivel educativo, el grupo de edad, entre otras. Los grupos mixtos se forman a partir de combinaciones de los dominios geográficos y sociodemográficos, como por ejemplo la desagregación por etnia, nivel de escolaridad y municipio (Morales y otros, 2021).

La confiabilidad de la inferencia basada en los estimadores directos disminuye a medida que aumentan los niveles de desagregación y el tamaño de muestra se reduce. Según Rao y Molina (2015) un área (dominio) es "pequeña" si su respectivo tamaño de muestra resulta escaso para garantizar una adecuada precisión en la estimación directa del parámetro de interés. Por lo tanto, el concepto de "área pequeña" no necesariamente se vincula directamente con el tamaño u otras características de un dominio, sino que depende del tamaño de muestra y del coeficiente de variación.

B. El método del mejor predictor empírico (EBP)

Los métodos SAE representan un conjunto de técnicas estadísticas empleadas para obtener estimaciones desagregadas de parámetros poblacionales cuando la estimación directa no cumple con los criterios de calidad requeridos para su publicación. Esta metodología funciona a partir de la integración de información de diversas fuentes para ganar fuerza predictiva. De esta manera, a la información obtenida de las encuestas, se suma la información de los censos, imágenes satelitales, registros administrativos u otro tipo de información que se tenga disponible para todas las unidades de observación.

Los modelos SAE se clasifican a partir de las variables auxiliares que se tengan disponibles y su nivel de agregación (Rao y Molina, 2015). Las categorías más comúnmente utilizadas son los modelos de área, que se ocupan de estimar en niveles de agregación predefinidos; y los modelos de unidad, los cuales se ocupan de realizar las estimaciones al nivel de todos los individuos.

Uno de los modelos de unidad habitualmente utilizados para la desagregación de indicadores de pobreza y otros indicadores sociales es el modelo denominado como Mejor Predictor Empírico (*Empirical Best Predictor* o EBP, en inglés), propuesto por Molina y Rao (2010). Este ha sido aplicado por la CEPAL para producir desagregaciones a nivel municipal de las estimaciones de pobreza y pobreza extrema que realiza habitualmente esta institución (Gutiérrez y otros, 2022).

En el caso de los indicadores de pobreza, el indicador a estimar es la proporción de personas en situación de pobreza, indicador que forma parte de la familia de indicadores FGT (Foster, Greer y Thorbecke, 1984), que se pueden expresar de la siguiente forma:

$$FGT_{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{z - y_i}{z}\right)^{\alpha} I(y_i < z) \qquad (3)$$

Donde N corresponde al tamaño de la población, y_i se refiere al ingreso per cápita del i-ésimo individuo, z hace referencia a la línea de pobreza (o indigencia), I es una función indicadora. Por su parte, α corresponde a un parámetro con valores iguales o superiores a cero (0), que determina las propiedades que cumple el índice. La tasa de pobreza (o indigencia) corresponde al caso en que $\alpha = 0$.

En Molina y Rao (2010) se propone un modelo con errores anidados que permite estimar indicadores no lineales, en el que la variable de interés no guarda una distribución normal, como el ingreso, la incidencia y la brecha de pobreza entre algunos otros. Este modelo relaciona linealmente los valores de una variable de interés Y_{di}, con un individuo *i* en un área o dominio con los valores de covariables auxiliares para el mismo individuo. De acuerdo con lo que propone Molina (2019), este método permite estimar un indicador FGT bajo un modelo que asume una relación lineal entre la transformación Log-Shift del ingreso (Y_{di}) del i-ésimo hogar dentro del área con un conjunto de covariables, como se expresa a continuación:

$$y^{*}_{di} = \times^{T}_{di}\beta + u_d + e_{di} \;\|\| i = 1, \ldots, N_d, \| d = 1, \ldots, D, \tag{4}$$

En donde $Y^{*}_{di} = log\left(Y_{di} + c\right)$, β corresponde al vector de coeficientes de regresión asociado a las covariables auxiliares, u_d hace referencia al efecto aleatorio del área, tal que $u_d \overset{i.i.d}{\sim} N\left(0, \sigma^{2}_{u}\right)$ y $e_{di} \overset{iid}{\sim} N\left(0, \sigma^{2}_{e}\right)$ son los errores a nivel del individuo, pero que son independientes de los efectos aleatorios. Cabe destacar que, $c > o$ es una constante que debe ser estimada por algún método de optimización que garantice una distribución aproximadamente normal (Molina y otros 2014).

El estimador del indicador de pobreza FGT se obtiene al desagregarlo en dos partes, tal como lo propone Molina y Rao (2010) y Rao y Molina (2015), la primera parte va asociada a las observaciones que se encuentran incluidas dentro de la muestra y la otra parte, a las observaciones que están por fuera de la muestra como información exógena de la siguiente manera:

$$\tilde{F}^{B}_{\alpha d}\left(\theta\right) = \frac{1}{N_d}\left(\sum_{i \epsilon S_d} F_{\alpha, di} + \sum_{i \epsilon r_d} \tilde{F}^{B}_{\alpha, di}\left(\theta\right)\right) \tag{5}$$

Es primordial, en primera medida, definir la capacidad de predicción de las variables auxiliares incluidas en el modelo. Con el fin de lograr una plena identificación del poder predictivo del modelo podrían generarse diversos escenarios de prueba, a partir de la combinación de las distintas covariables disponibles y sus respectivas interacciones, para luego comparar la bondad de ajuste de cada uno de ellos. Adicionalmente, como lo proponen Jiming y Lahiri (2006), debe tenerse en cuenta el número de variables significativas arrojadas por el modelo, y las estadísticas de pronóstico en el análisis y comparabilidad de los modelos.

Gutiérrez y otros (2022) abordaron el problema de la estimación de la incidencia de la pobreza usando este tipo de modelos en algunos países de la región, encontrando que, debido a la alta desigualdad en los ingresos de los países estudiados, los supuestos de normalidad en los errores del modelo no se cumplían en todos los casos, y que el efecto de la transformación logarítmica podría impactar las predicciones. De la misma manera, dado que el interés de la inferencia en áreas pequeñas no está supeditado solamente a las desagregaciones geográficas, la inclusión de un solo efecto aleatorio a nivel de división administrativa menor puede no ser suficiente para tratar la heterogeneidad de todos los subgrupos de interés.

Por ejemplo, si el interés recae en la estimación de la pobreza para las personas en situación de discapacidad con escolaridad baja en los municipios de un país, sería cuestionable no incluir efectos aleatorios para ambas desagregaciones (discapacidad y escolaridad), dado que la encuesta de hogares difícilmente es representativa para todas las clasificaciones de ambas variables. Además de lo anterior, al ser un modelo de unidad, el EBP implica el uso de la información de cada observación (individuo) del censo de población, por lo que la capacidad computacional y las altas cargas de procesamiento siguen siendo un cuello de botella a la hora de ajustar este tipo de metodologías, sobre todo en la fase de predicción con millones de observaciones en el censo y la correspondiente estimación de los errores cuadrados medios asociados a cada predicción. En el caso de esta última medida de calidad, los tiempos de procesamiento computacional se multiplican para cada subgrupo poblacional en el que se pretenda tener una predicción.

Por último, en la medida que la fecha de la encuesta y del censo se hacen más distantes, el supuesto de que ambos instrumentos corresponden a una misma población se torna más débil. Por tanto, ante el interés de producir de manera regular desagregaciones de la información generada por las encuestas, resulta de interés explorar otras alternativas metodológicas.

C. Modelos multinivel con múltiples efectos aleatorios y postestratificación

Cuando la variable dependiente es de tipo binario, como lo es la situación de pobreza (pobre o no pobre), es posible emplear modelos de regresión logística, en los que se obtienen estimaciones de la probabilidad del evento estudiado con base en una transformación (usualmente logit) que define un vínculo entre la esperanza de esta variable aleatoria con la covariables predictoras y con los efectos aleatorios sobre las desagregaciones de interés, como se muestra en la siguiente ecuación:

$$ln\left(\frac{\rho_{id}}{1-\rho_{id}}\right) = x_{id}\beta + z_{id}\gamma \qquad (6)$$

Las variables en la matriz $\boldsymbol{X}$ corresponden a los efectos fijos, mientras que las variables en el vector $\boldsymbol{\gamma}$ dan cuenta de los efectos aleatorios. En la literatura estadística, este tipo de modelos se conoce como modelos mixtos. Para que la inferencia resultante de estos modelos sea adecuada, es importante diferenciar correctamente entre ambos efectos. Los efectos fijos se refieren a cualquier subgrupo para el cual la encuesta induce inferencias precisas (generalmente los dominios de interés de la encuesta), mientras que los efectos aleatorios se refieren a cualquier subgrupo que cumpla al menos una de las siguientes condiciones: a) que alguna de las categorías del subgrupo no esté incluida en la muestra (por ejemplo, no todos los municipios están en la muestra de la encuesta) o b) que el tamaño de muestra para todas las categorías no sea suficiente para garantizar la representatividad en cada una de las categorías (por ejemplo, edades simples).

En el modelo utilizado en este estudio, los efectos fijos incluyen variables a nivel individual, como sexo, grupo de edad y nivel educativo, así como variables agregadas al nivel de la división administrativa mayor. Estas últimas covariables añaden un enfoque multinivel (con una jerarquía natural, en donde los individuos pertenecen a las agregaciones geográficas) a la inferencia y pueden obtenerse de agregados censales (porcentaje de hogares con hacinamiento, tasa de desocupación promedio, porcentaje de viviendas con materiales precarios, etc.), registros administrativos (número de escuelas públicas, tasa de homicidios, porcentaje de hogares que reciben transferencias gubernamentales, etc.), imágenes satelitales (intensidad lumínica, fracción de suelo urbano, índice de modificación humana, etc.), entre otras.

Ahora bien, en la ecuación anterior, los coeficientes β hacen referencia a los efectos fijos de las variables x_{ji} sobre las probabilidades de que la i-ésima persona sea pobre; por otro lado, los coeficientes expresan los efectos aleatorios sobre las covariables $\boldsymbol{\gamma}$. De acuerdo con esto, la característica principal de este tipo de modelos radica en que los efectos aleatorios no son considerados como parámetros, sino que son tomados como una realización de variables aleatorias.

Una de las similitudes entre los modelos lineales y los logísticos es que los signos de la ecuación estimada pueden interpretarse de manera similar en ambos casos. En particular, el signo de la pendiente indica la relación de la variable independiente con la probabilidad de ocurrencia del evento que explica la variable dependiente. Un signo positivo asociado a una covariable indica un aumento en la probabilidad de que ocurra el evento cuando se cumplen las características de dicha covariable; mientras que un signo negativo indica una disminución en esa probabilidad bajo las mismas condiciones.

A partir del conjunto de covariables que se tengan disponibles para ser incluidas en el modelo y, luego de estimar los parámetros del modelo de regresión logística multinivel, con el vector de probabilidad $\hat{\rho} = [\hat{\rho}_{id}]$ es posible predecir si una persona es o no pobre.

Teniendo en cuenta que las variables que hacen referencia a la información personal principalmente son de tipo categórico, solo habrá un cierto número de valores posibles para $\hat{\rho}$. Por consiguiente, es necesario identificar el número de personas que componen todas y cada una de las posibles combinaciones para todos los cruces viables de las covariables de postestratificación. En particular, los cruces de las variables

de información personal se denotarán como postestratos. Como se explicará más adelante, debido a las restricciones impuestas por los procesos de estandarización entre los censos de población y las encuestas de hogares, en este estudio se utilizan las siguientes variables de postestratificación:

- DAM (división administrativa mayor del país).
- Área (ubicación de la vivienda en dos áreas: urbanas o rurales).
- Sexo (clasificación del individuo en dos grupos: hombres o mujeres).
- Edad (clasificación del individuo en cinco grupos de edad).
- Años de estudio (clasificación del individuo en seis grupos de acuerdo con los años de estudio).
- Pertenencia étnica (clasificación del individuo en tres grupos).

De esta forma, existirán tantos postestratos como combinaciones de las anteriores variables en el censo de población y vivienda. Por ejemplo, si en un país hay 20 DAM, entonces se definirán automáticamente 20 X 2 X 2 X 5 X 6 X 3 = 7.200 postestratos. De manera general, suponiendo que existan Q combinaciones posibles (postestratos), se denotarán como $N_{s1}, \ldots, N_{sj}, \ldots, N_{sQ}$ a los conteos de individuos en los cruces, los cuales provendrán directamente del procesamiento censal y se denotarán como $N_{s1}, \ldots, N_{sj}, \ldots, N_{sQ}$. Por lo anterior, un estimador para la proporción de personas en condición de pobreza dentro de las áreas (subgrupos) de interés vendrá dada por:

$$\hat{\rho}_s = \frac{\sum_{j=1}^{Q} N_{sj} \hat{\rho}_{sj}}{\sum_{j=1}^{Q} N_{sj}} \tag{7}$$

Es decir, los Q valores posibles de $\hat{\rho}_{ij}$ se ponderan por el tamaño estimado de todos los cruces posibles (postestratos) de las covariables dentro del dominio de interés. Por ende, se pueden lograr estimaciones en áreas pequeñas a partir del tratamiento de estos cruces, los cuales a su vez pueden agregarse convenientemente a un nivel superior. Una aplicación de este tipo de modelos se presenta en Zhang y otros (2013), en donde, a partir de un modelo logístico multinivel con efectos aleatorios a nivel de estado y condado anidado, se desarrolla una propuesta con la que se obtuvo una estimación cruzada de la proporción de casos de enfermedad pulmonar obstructiva crónica (EPOC) mediante los conteos de personas en bloques censales de interés (obtenidos del Censo de EE. UU. de 2010). Las estimaciones a nivel de bloque censal pueden ser agregadas de acuerdo con la necesidad a niveles geográficos superiores.

Siguiendo lo propuesto por Gelman y Hill (2006) y Park, Gelman y Bafumi (2016), el modelo presentado en este documento emplea tres fuentes de información:

i) Las encuestas de hogares contienen información sobre los ingresos y el nivel de pobreza del hogar y sus miembros e información sobre edad, etnia, área urbana/rural, nivel educativo, sexo, estado de discapacidad, etc.

ii) Información agregada a nivel departamental sobre covariables censales e imágenes satelitales (luces nocturnas, fracción de cobertura urbana y fracción de cobertura de cultivos).

iii) Información actualizada agregada a nivel departamental sobre el número total de personas en cada combinación posible de las variables de información personal sobre las que se realizará el proceso de estimación. Estos datos provienen directamente de un procedimiento estadístico llamado SPREE (*Structure Preserving Estimation*, por sus siglas en inglés).

El ajuste de este modelo se compone principalmente de dos partes: en primer lugar, se debe ajustar un modelo de regresión multinivel utilizando la primera y la segunda fuente de información; posteriormente, usando los conteos censales actualizados, es posible predecir cada una de las celdas de postestratificación usando la tercera fuente.

Además de estimar la pobreza extrema y pobreza mediante el modelo de regresión logística descrito anteriormente, se realiza una estimación del ingreso medio per cápita (escalado a las líneas de pobreza). En este caso se sigue un razonamiento similar usando un modelo de regresión mixto para variables continuas, expresado de la siguiente manera:

$$y_{id} = x_{id}\beta + z_{id}\gamma + \varepsilon_{id} \tag{8}$$

En este modelo la variable respuesta y es el ingreso medio percápita, mientras que las covariables en la matriz X representan los efectos fijos, en Z se definen los efectos aleatorios y ε_{id} representa los errores del modelo que siguen una distribución normal, de la siguiente manera:

$$\varepsilon_{id} \backsim normal\left(0, \sigma_{\varepsilon}^{2}\right) \tag{9}$$

Al igual que con el caso de la regresión logística, este modelo se adapta al paradigma bayesiano no informativo, con un enfoque multinivel de postestratificación. De esta forma, es posible predecir en cada postestrato la media del ingreso $\hat{\mu} = \left[\hat{\mu}_{id}\right]$. Tal como en el caso anterior, suponiendo que existen Q postestratos, un estimador para la media del ingreso dentro de las áreas (subgrupos) de interés vendrá dada por:

$$\hat{\mu}_{s} = \frac{\sum_{j=1}^{Q} N_{sj}\hat{\mu}_{sj}}{\sum_{j=1}^{Q} N_{sj}} \tag{10}$$

Adaptar la inferencia al enfoque bayesiano también genera ventajas computacionales, haciendo más eficientes los tiempos de procesamiento. Como se verá en las secciones posteriores, este enfoque inferencial permite obtener buenos resultados y, a diferencia de la metodología EBP, puede escalarse rápidamente para generar estimaciones desagregadas en los subgrupos de interés en todos los países de la región.

Cabe destacar que, bajo el paradigma bayesiano no informativo y usando las mismas covariables, las estimaciones resultantes de ambos modelos deberían coincidir. Esto se debe a que, en los modelos estadísticos, el mejor predictor lineal insesgado es el promedio muestral. Por lo tanto, ajustar un modelo a nivel individual con efectos fijos categóricos es equivalente a ajustar un modelo ponderado a nivel de postestrato, donde la variable respuesta es la media del postestrato. Esta coincidencia algebraica ha favorecido el uso de modelos postestratificados en lugar de los modelos tradicionales EBP, permitiendo predecir el nivel promedio de pobreza para una combinación específica de características individuales.

II. Fuentes de información

A. Encuestas de hogares

Para la aplicación de los modelos de estimación en áreas pequeñas, las encuestas de hogares proveen la información sobre la variable de interés (ingreso medio y tasas de pobreza y pobreza extrema) que se desea desagregar para diversos grupos de población.

Este trabajo utiliza las encuestas de hogares del Banco de Datos de Encuestas de Hogares (BADEHOG), repositorio de la CEPAL de microdatos provenientes de los países de América Latina. Las encuestas incluidas en el repositorio son aquellas que se utilizan como fuente para producir las estimaciones oficiales de pobreza y desigualdad del ingreso en los países, y que además utiliza la CEPAL para las estimaciones que realiza con propósitos de comparabilidad regional. Los instrumentos son de diversos tipos e incluyen encuestas de mercado laboral, condiciones de vida e ingresos y gastos. En el cuadro 1 se presentan las encuestas de hogares utilizadas para la aplicación del modelo de estimación en áreas pequeñas.

Cuadro 1
Encuestas de hogares utilizadas para la aplicación del modelo

País	Encuesta	Año
Argentina	Encuesta Permanente de Hogares	2022
Bolivia (Estado Plurinacional de)	Encuesta Nacional de Hogares	2021
Brasil	*Pesquisa Nacional por Amostra de Domicílios Contínua*	2022
Chile	Encuesta de Caracterización Socioeconómica Nacional	2022
Colombia	Gran Encuesta Integrada de Hogares	2021
Costa Rica	Encuesta Nacional de Hogares	2022
Dominica	Encuesta Nacional Continua de Fuerza de Trabajo	2022
Ecuador	Encuesta Nacional de Empleo, Desempleo y Subempleo	2022

País	Encuesta	Año
Guatemala	Encuesta Nacional de Condiciones de Vida	2014
Honduras	Encuesta Permanente de Hogares de Propósitos Múltiples	2019
México	Encuesta Nacional de Ingresos y Gastos de los Hogares	2022
Nicaragua	Encuesta Nacional de Hogares sobre Medición de Nivel de Vida	2014
Panamá	Encuesta Multipropósito	2022
Perú	Encuesta Nacional de Hogares	2022
Paraguay	Encuesta Permanente de Hogares Continua	2022
El Salvador	Encuesta de Hogares de Propósitos Múltiples	2022
Uruguay	Encuesta Continua de Hogares	2022

Fuente: Banco de Datos de Encuestas de Hogares (BADEHOG) de la CEPAL.

La División de Estadísticas de la CEPAL lleva a cabo un proceso de preparación de bases de datos y armonización de las variables más utilizadas de las encuestas de hogares. En primer lugar, el proceso implica la conformación de una base de datos única, que integra la información proveniente de las bases de datos de individuos, hogares y viviendas, así como de eventuales módulos adicionales de interés. En esa base se construyen variables sociodemográficas y de mercado laboral, manteniendo las categorías de respuesta originales de la encuesta. Uno de los aspectos principales es el cálculo de los agregados de ingresos. Para ello, se construyen variables para cada una de las fuentes de ingreso captadas en la encuesta, expresadas en términos mensuales, y se organizan según la fuente de percepción: ingresos del empleo, ingresos por la propiedad de activos, ingresos por jubilaciones y pensiones, ingresos por otras transferencias, alquiler imputado. En algunos casos, en los que la institución no realiza ya un proceso similar, se aplica un modelo de imputación por no respuesta a las preguntas de ingreso. Estas bases incluyen variables referidas al ingreso per cápita (ingcorte) y las líneas de pobreza extrema (li) y pobreza (lp) calculadas por la CEPAL. Además, en todos los casos se incluyen los factores de expansión, mientras que solo algunas bases cuentan con la información del diseño complejo de la encuesta resumido en estratos y unidades primarias de muestreo.

Posteriormente, se crean nuevas variables armonizadas que se utilizan para el cálculo de diversos indicadores. Las variables incluidas abarcan características personales (grupo étnico, estado civil, parentesco con el jefe, etc.), variables de ocupación (rama de actividad, horas trabajadas, tamaño del establecimiento, informalidad en el empleo), educación (nivel educativo completado), servicios básicos (privaciones en acceso a agua, saneamiento, energía, entre otros), protección social (afiliación y cotización a sistemas de previsión y salud), etc.

Con base en esta información, se recodifican las categorías de respuesta de las variables que definen los postestratos de interés[1]. Como se indicó anteriormente, las variables de postestratificación son DAM (división administrativa mayor del país), área (ubicación de la vivienda en dos áreas: urbanas o rurales), sexo (clasificación del individuo en dos grupos: hombres o mujeres), edad (clasificación del individuo en cinco grupos de edad), años de estudio (clasificación del individuo en seis grupos de acuerdo con los años de estudio) y pertenencia a pueblos indígenas o afrodescendientes (clasificación del individuo en tres grupos).

El cuadro 2 detalla la estructura y recategorización de la encuesta original importada de los repositorios de BADEHOG; proceso que se realiza para todos y cada uno de los países en estudio.

[1] Cabe mencionar que, dado que los datos ya presentan una estandarización previa, todas las variables se encuentran disponibles para todos los países, pero no todas ellas cuentan con información, ya que algunos países en sus encuestas no incluyen algunas variables; por lo tanto, no disponen de dicha información en sus cuestionarios.

Cuadro 2
Estructura, codificación y creación de los postestratos

Característica	Variable original	Categorías de la variable original en BADEHOG	Codificación bajo la estandarización	Valores
Postestratos	DAM	Según el país	DAM	Según el país
	Área	1 = Urbano	Área	1 = Urbano
		2 = Rural		0 = Rural
	Sexo	1 = hombre	sexo	1 = Hombre
		2 = mujer		2 = Mujer
	Edad	Desde 1 año hasta la edad máxima registrada	Edad	1 = 0-14 años
				2 = 15-29 años
				3 = 30-44 años
				4 = 45-64 años
				5 = más de 65 años
	Pertenencia étnica	-1 = NA	Etnia	1 = Indígena
		0 = No indígena ni afrodescendiente		2 = Afrodescendiente
		1 = Indígena		3 = Otro
		2 = Afrodescendiente		
		99 = Ignorado, NS/NR		
	Años de estudio	Expresado en años de educación aprobados	Anoest	1 = Sin educación
				2 = 1 a 6 años
				3 = 7 a 12 años
				4 = Más de 12 años
				5 = No aplica
				6 = Otro
Indicadores de interés	lp	Línea de pobreza	lp	Valores diferentes de acuerdo con cada país
	li	Línea de pobreza extrema (línea de indigencia)	li	
	ingcorte	Ingreso per cápita del hogar	Ingcorte	
Variables del diseño muestral	_upm	Valores de acuerdo con codificación de cada país, sujeto al diseño muestral	upm	Valores diferentes de acuerdo con cada país
	_estrato		Estrato	
	_fep		fexp	

Fuente: Elaboración propia.

Luego del proceso de estandarización, se obtiene una nueva base de datos que cuenta con el mismo número de registros de la base original, pero incluyendo únicamente las variables de la estandarización.

B. Censos de población y vivienda

Los censos de población y vivienda son una fuente de información ampliamente utilizada para la aplicación de modelos de estimación de áreas pequeñas, debido a su alta capacidad de desagregación geográfica y para grupos de población. Para este ejercicio se utiliza la información disponible en el banco de datos censales de la CEPAL, administrado por la División de Población (CELADE). El cuadro 3 consolida la información referente a los censos de población y vivienda utilizados en este ejercicio.

Cuadro 3
Censos más recientes disponibles en el banco de datos censales del CELADE

País	Censo	Año
Argentina	Censo Nacional de Población, Hogares y Viviendas 2010	2010
Bolivia (Estado Plurinacional de)	Censo de Población y Vivienda 2012	2012
Brasil	Censo Demográfico 2010	2010
Chile	Censos de Población y Vivienda	2017
Colombia	Censo Nacional de Población y Vivienda - CNPV - 2018	2018
Costa Rica	X Censo Nacional de Población y VI de Vivienda	2011
Dominica	IX Censo Nacional de Población y Vivienda 2010	2010
Ecuador	VII Censo de Población y VI de Vivienda	2011
Guatemala	XII Censo Nacional de Población y VII de Vivienda	2018
Honduras	XVII Censo de población y VI de Vivienda 2013	2013
México	Censo de Población y Vivienda 2020	2020
Nicaragua	Censo de Población y Vivienda de Nicaragua de 2005	2005
Panamá	Censo 2010	2010
Perú	XII Censo de Población, VII de Vivienda y III de Comunidades Indígenas o Censo peruano de 2017	2017
Paraguay	II Censo Nacional Indígena de Población y Vivienda 2002. Pueblos Indígenas del Paraguay	2002
El Salvador	VI Censo de Población y V de Vivienda 2007	2007
Uruguay	Censo de Población 2011	2011

Fuente: Elaboración propia.

Para la utilización de la información censal en los modelos de estimación de áreas pequeñas, es necesario adecuar los insumos provistos e integrarlos al proceso general. El *software* REDATAM, desarrollado por CELADE (2021), permite el procesamiento, tratamiento y manipulación de los datos para realizar el adecuamiento necesario para la construcción de los modelos. Para este proceso se utiliza la librería redatam disponible para el *software* R.

Al igual que en el caso de las encuestas, se realizó un proceso de armonización de las categorías de respuesta para conformar los mismos postestratos indicados anteriormente. A manera de ejemplo, el cuadro 4 muestra la recategorización de variables aplicada al censo de Uruguay. Un proceso similar se realizó con los censos de los 16 países restantes.

Cuadro 4
Uruguay: ejemplo de recategorización de variables originales del censo

Nombre de la variable en el censo	Categorías de la variable en el censo	Nombre de la variable estandarizada	Categorías de la variable estandarizada
URBRUR2	1 = Urbano	Área	1 = Urbano
	2 = Rural		0 = Rural
NA014	Desde 1 año hasta la edad máxima registrada	Edad	1 = 0 a 14 años
			2 = 15 a 29 años
			3 = 30 a 44 años
			4 = 45 a 64 años
			5 = más de 65 años

Nombre de la variable en el censo	Categorías de la variable en el censo	Nombre de la variable estandarizada	Categorías de la variable estandarizada
ER025	1 = Afro o Negra	Etnia	1 = Indígena
	2 = Asiática o Amarilla		2 = Afrodescendiente
	3 = Blanca		3 = Otro
	4 = Ignorado		
	5 = Indígena		
	6 = Ninguna (no hay una principal)		
	8 = No relevado		
	9 = otra		
AESTUDIO6	Expresado en años de educación cursados	Anoest	1 = Sin educación
			2 = 1 a 6 años
			3 = 7 a 12 años
			4 = Más de 12 años
			5 = No aplica
			6 = Otro

Fuente: Elaboración propia.

A manera de resumen, el cuadro 5 muestra la disponibilidad de las variables recodificadas en todos los países. Las únicas variables que no cuentan con información completa en todos los países son el área, puesto que en Argentina solo se dispone de información para el área urbana, y la etnia, que no está disponible en República Dominicana, Panamá y Paraguay.

Cuadro 5
Disponibilidad de postestratos en cada país de la región

País	DAM	Área	Sexo	Años de estudio	Edad	Etnia
Argentina	X		X	X	X	X
Bolivia (Estado Plurinacional de)	X	X	X	X	X	X
Brasil	X	X	X	X	X	X
Chile	X	X	X	X	X	X
Colombia	X	X	X	X	X	X
Costa Rica	X	X	X	X	X	X
Dominica	X	X	X	X	X	
Ecuador	X	X	X	X	X	X
Guatemala	X	X	X	X	X	X
Honduras	X	X	X	X	X	X
México	X	X	X	X	X	X
Nicaragua	X	X	X	X	X	X
Panamá	X	X	X	X	X	
Perú	X	X	X	X	X	X
Paraguay	X	X	X	X	X	
El Salvador	X	X	X	X	X	X
Uruguay	X	X	X	X	X	X

Fuente: Elaboración propia.

C. Información geoespacial

El uso de imágenes satelitales en modelos de estimación en áreas pequeñas ha ganado popularidad en los últimos años, en la medida en que los avances computacionales y en capacidad de almacenamiento permitieron que esta información estuviera disponible a gran escala. En comparación con otras fuentes de datos no tradicionales, como la telefonía móvil o la actividad en internet, la información satelital tiene la ventaja de contar con una amplia variedad de indicadores de acceso libre (Newhouse, 2023). La información que se obtiene mediante sensores remotos suele tener un nivel de desagregación geográfica muy alto. El uso de imágenes satelitales, particularmente las luces nocturnas, la fracción de cobertura urbana y la fracción de cobertura de cultivos, brinda a los países la opción de compensar la falta de censos de población y encuestas detalladas (Naciones Unidas, 2019).

Con respecto al uso de imágenes satelitales en modelos de estimación de áreas pequeñas, Newhouse y otros (2022) concluyeron que combinar covariables satelitales con datos de encuestas de hogares mejora sustancialmente la precisión de las estimaciones de pobreza municipal, en comparación con las estimaciones que utilizan solo los datos de las encuestas (estimaciones directas). Asimismo, para todas las estimaciones basadas en información satelital, las predicciones dentro de la muestra son sustancialmente más precisas y exactas que las predicciones fuera de la muestra.

Para el modelo de estimación presentado en este documento, se utilizan variables derivadas de imágenes satelitales obtenidas desde la plataforma Google Earth Engine[2]; cuyo procesamiento involucra diferentes lenguajes de programación como Javascript, Python y R, mediante el paquete rgee, vinculado en 2021 (Aybar C., 2021). El archivo público de datos que proporciona esta herramienta incluye imágenes históricas de la tierra que se remontan a más de cuarenta años y cuenta con una amplia variedad de bases de datos sobre información geofísica, clima, tiempo e incluso datos demográficos. Además, es una herramienta versátil que permite cargar datos ráster como datos vectoriales y visualizar datos de manera rápida (para mayor información véase el tutorial de introducción a Google Earth Engine[3]).

Esta metodología hace uso de algunas variables satelitales que tienen un vínculo empírico con la pobreza, la desigualdad y las condiciones de vida en los países estudiados, y que se detallan a continuación:

- Las luces nocturnas (*night-lights*) capturadas por el Sistema de Escaneo de Línea Operacional del Programa de Meteorología de Defensa proporcionan una representación visual de la actividad humana y el desarrollo económico durante la noche. Estas imágenes muestran la intensidad de la luz visible, permitiendo a los investigadores inferir patrones de desarrollo urbano, crecimiento económico y cambios en la infraestructura a lo largo del tiempo. Estudios recientes (Andreano y otros, 2020) muestran que esta covariable es una fuente clave de información para derivar estimaciones desagregadas de la proporción de personas en situación de pobreza en los países de América Latina y el Caribe.
- El porcentaje de cubrimiento de cultivos (*crops-cover fraction*) y el porcentaje de cobertura urbana (*urban-cover fraction*) son variables que miden la proporción del área de una celda de 100 metros dedicada a cultivos y superficies urbanas, respectivamente. Estas variables continuas se usan para el monitoreo agrícola y la planificación urbana, ofreciendo información sobre el uso del suelo y la extensión de la agricultura y la urbanización en diferentes regiones. Las áreas urbanas a menudo tienen mejor acceso a infraestructura y servicios, lo cual está asociado con menores tasas de pobreza. Las disparidades en la distribución de estos recursos pueden llevar a desigualdades significativas en los países estudiados. Andreoli y otros (2022) muestran la relación existente entre la pobreza y la distribución de las áreas urbanas. Además, Mendoza (2020) muestra la asociación empírica que existe entre este tipo de variables satelitales con las condiciones socioeconómicas.

[2] https://earthengine.google.com/.
[3] Tutorial disponible: https://www.google.com/intl/es_es/earth/outreach/learn/introduction-to-google-earth-engine/.

- El índice que modificación humana (*Global Human Modification*) mide la modificación humana de las tierras terrestres a nivel global, considerando factores como asentamientos humanos, agricultura, transporte, minería, y producción de energía e infraestructura eléctrica. Estos valores reflejan la proporción de una ubicación modificada por actividades humanas y la intensidad de dicha modificación. Masaki y otros (2020) usaron este tipo de imágenes en relación con la pobreza no monetaria encontrando que la inclusión de estas covariable en los modelos de estimación de áreas pequeñas favorecía su poder predictivo, en comparación con modelos que no las tenían en cuenta.
- Las variables tiempo de viaje al hospital o clínica más cercana (*accessibility*) y tiempo de viaje utilizando transporte no motorizado (*walking only accessibility*) miden la accesibilidad geográfica a los servicios de salud. Estas variables proporcionan información sobre el tiempo de viaje en minutos desde cualquier ubicación hasta la instalación de salud más cercana, considerando todos los medios de transporte y solo el transporte a pie, respectivamente. En algunos trabajos, se ha documentado cómo la distribución espacial de la pobreza está fuertemente correlacionada con el acceso a carreteras, escuelas y centros de salud, y otras variables de accesibilidad e infraestructura (World Bank Group, 2011).

Las imágenes satelitales se construyen a partir de polígonos, los cuales son únicos para cada país. Por esta razón, para obtener las variables deseadas es necesario contar con *shapefiles*[4] de cada país. No siempre existe correspondencia entre las áreas geográficas provenientes del censo y de las imágenes satelitales, en cuyo caso se realiza un proceso de recodificación de los códigos disponibles. La información que se obtiene debe ser analizada para corroborar que cumple con las características deseadas. A manera de ejemplo, en el caso de la información sobre luces nocturnas, luego de importar la información, se extrae la variable que corresponde al número de pixeles (y que da cuenta de la luminosidad) por unidad geográfica (división administrativa mayor). Debe considerarse entre utilizar el total de pixeles o su promedio. Esta última opción puede no traer buenos resultados cuando se trabaja con áreas muy pequeñas (que pueden tener un promedio alto) o con áreas amplias que contienen terrenos rurales, donde el promedio de pixeles puede ser bajo a pesar de tratarse de un lugar con buena infraestructura de iluminación.

[4] *Shapefiles*: Es un formato sencillo empleado para guardar o archivar la ubicación geométrica y la información referente a las características propias de las entidades geográficas. Las cuales, pueden ser representadas mediante puntos, líneas o polígonos (áreas), siendo este último el más comúnmente utilizado.

III. Actualización de conteos intercensales

Los censos de población y vivienda constituyen la principal fuente de información detallada sobre la población y sus características sociodemográficas. Sin embargo, al realizarse aproximadamente cada 10 años, su información tiende a desactualizarse de forma gradual en los períodos intermedios. Para obtener resultados precisos en la desagregación de los indicadores de interés es necesario actualizar los datos censales y así evitar sesgos en los resultados. Esto se puede lograr mediante la actualización de las tablas censales utilizando la información proporcionada por las encuestas de hogares, que se implementan con mayor frecuencia.

En esta sección se presenta el proceso de actualización de los conteos de población obtenidos del censo, utilizando los conteos marginales que se obtienen a partir de la encuesta. Las metodologías para generar tablas censales actualizadas provienen del campo de la demografía, con el fin de obtener conteos post censales. Rao (2003) provee una revisión de los métodos demográficos para actualizar estas tablas. Más recientemente, se han adoptado técnicas de actualización de tablas para generar datos poblacionales sintéticos, donde se seleccionan muestras a partir de los datos de encuesta que coincidan con las proporciones de una tabla censal actualizada (Suesse y otros, 2017).

A. Conceptos y notación básica

Bajo este procedimiento se hace uso del término tabla para hacer referencia a una tabla de contingencia de múltiples vías, y la suma de los conteos por filas o por columnas se les denomina totales por fila y columna o márgenes. Las celdas son los conteos en cada combinación fila-columna de la tabla. Por simplicidad en la notación, considérese una tabla de contingencia a dos vías, como la descrita en el cuadro 6, en donde las unidades de estudio *i* se organizan en $\alpha = 1, \ldots, A$ filas, $j = 1, \ldots, J$ columnas. Como ejemplo, si las unidades de estudio son personas, las filas de la tabla podrían corresponder a las divisiones administrativas mayores y las columnas a las categorías de edad.

Cuadro 6
Tabla de contingencia a dos vías

	1	j	J	Total fila
1	y_{11}	y_{1j}	y_{1J}	$y_{1.}$
a	$y_{.a}$	y_{aj}	y_{aJ}	$y_{a.}$
A	$y_{.A}$	y_{Aj}	y_{AJ}	$y_{A.}$
Total columna	$y_{.1}$	$y_{.j}$	$y_{.J}$	$y_{..}$

Fuente: Elaboración propia.

Las celdas de la tabla contienen conteos correspondientes a cada combinación entre áreas y grupos de edad. Además, los totales por fila (márgenes fila, $y_{a.}$) y los totales por columna (márgenes columna $y_{.j}$) contienen los totales poblacionales según áreas y grupos de edad, respectivamente. Nótese que los totales por fila y por columna deben ser consistentes. El valor observado (conteo) en la celda a, j, se denota en minúscula y_{aj}. Asumiendo que la tabla Y es la tabla censal de interés, el producto de la actualización será $\widehat{Y}$. El procedimiento básico requiere de un conjunto de márgenes nuevos o actualizados, $y_{a.}$ y $y_{.j}$, que tradicionalmente se obtienen de la encuesta más reciente.

En la literatura estadística, se conoce con las siglas SPREE (*Structure Preserving Estimation*) a los procesos de actualización de conteos censales que conservan la estructura de correlación entre las variables, mientras ajustan todas las celdas de una tabla de contingencia, respetando los conteos marginales de las variables categóricas de interés. En la metodología que se propone en este documento, la preservación de las estructuras de correlación entre las covariables de postestratificación es de suma importancia para asegurar la exactitud y precisión de la inferencia estadística.

Aunque Purcell and Kish (1980) identifican seis escenarios para aplicar los métodos SPREE, de acuerdo con la disponibilidad de datos, en general, el escenario ideal es contar al menos con una tabla censal, así como márgenes recientes y confiables. La idea de trabajar con ambos conjuntos de datos es que la encuesta es confiable únicamente en sus totales, pero no en las relaciones entre variables de interés, contrario al caso del censo. Koebe y otros (2020) resume los principales requerimientos para realizar una actualización de tablas censales mediante este tipo de métodos:

i) Los datos de la tabla censal deben estar organizados en categorías (tablas de contingencia).

ii) Los márgenes de la tabla censal y de la encuesta deben ser de la misma dimensión (mismo número de filas y de columnas).

iii) Los totales de las filas y de las columnas deben sumar la misma cantidad (consistencia).

iv) Las variables de la tabla censal que se deben actualizar deben presentes en el censo y en la encuesta al mismo tiempo (por ejemplo, no se puede actualizar el número de pobres si esa información no fue recolectada en el censo).

Suesse y otros (2017) explican que trabajar con un número grande de dimensiones incrementa el número de restricciones y el algoritmo de optimización se vuelve inestable. En estos casos, los autores sugieren, por ejemplo, aplicar el procedimiento de actualización de cuadros por medio del IPF para cada área o dominio por separado.

B. El IPFP como método para preservar la estructura de los censos

La familia de los métodos SPREE es popular en el contexto de estimaciones en áreas pequeñas. La primera versión de estos métodos fue propuesta por Purcell and Kish (1980) como una herramienta para actualizar conteos o proporciones de una o más variables categóricas de interés de acuerdo con dominios de estudio, para años postcensales. En ese procedimiento, la obtención de tablas actualizadas se lleva a cabo por

medio del ajuste proporcional iterativo (*Iterative Proportional Fitting Procedure,* IPFP). Según Suesse y otros (2017), el algoritmo IPFP es, en general, uno de los métodos más utilizados para actualizar tablas censales. Este algoritmo, propuesto por Deming y Stephan (1940), también se conoce en la literatura de estimadores de calibración como *raking ratio*, o *raking* multiplicativo, y ajusta los conteos de una tabla de contingencia con base en un conjunto de márgenes dados (Deville y Sarndal, 1992).

Continuando con la notación descrita anteriormente, si *Y* es la tabla censal objetivo en el periodo actual t_1, se espera estimar una tabla $\widehat{Y}$ partir de una tabla censal *Z* en un periodo anterior, t_0, y un conjunto de márgenes provenientes de una encuesta reciente $Y_{a.}$ y $Y_{.j}$. Utilizando el censo anterior *Z* como base o punto de partida, el primer ciclo del algoritmo se describe a continuación:

i) Las celdas del cuadro se ajustan a los primeros márgenes fila de $Y_{.j}$:

$$\widehat{Y}_{aJ}^{(1)} = Z_{aJ} \frac{Y_{.J}}{Z_{.j}}, \tag{11}$$

ii) Las celdas ajustadas en el paso anterior se vuelven a ajustar, ahora con los márgenes columna de: $\widehat{Y}^{(1)}$:

$$\widehat{Y}_{aJ}^{(2)} = \widehat{Y}_{aJ}^{(1)} \frac{Y_{a.}}{\widehat{Y}_{a}^{(1)}}, \tag{12}$$

iii) Se ajustan las celdas nuevamente, con los márgenes fila de $\widehat{Y}^{(2)}$:

$$\widehat{Y}_{aJ}^{(3)} = \widehat{Y}_{aJ}^{(2)} \frac{\widehat{Y}_{.J}^{(2)}}{Y_{.J}}, \tag{13}$$

De esta forma, el algoritmo IPFP sigue un proceso iterativo, repitiendo los últimos dos pasos finales hasta alcanzar convergencia. Este estimador minimiza la distancia X^2, dada por la siguiente expresión:

$$X^2 = \sum_{a=1}^{A} \sum_{j=1}^{J} \frac{\left(Y_{aJ} - \widehat{Y}_{aj}\right)^2}{\widehat{Y}_{aj}}, \tag{14}$$

Además, Ireland and Kullback (1968) demostraron que, para valores iniciales positivos, este procedimiento logra encontrar una solución óptima, de acuerdo con la medida divergencia de Kullback-Leibler; en otras palabras, el IPFP logra minimizar la entropía relativa:

$$KL = \sum_{a=1}^{A} \sum_{j=1}^{J} Y_{aJ} \, Log \frac{Y_{aJ}}{\widehat{Y}_{aj}}, \tag{15}$$

En Bishop y otros (2007) y Zaloznik (2011) es posible encontrar más información sobre este algoritmo. Debido a que el SPREE (por medio del uso del IPFP) hace uso de estimadores directos, que son precisos y confiables, y que usualmente se trata de totales marginales provenientes de datos de encuestas de hogares, Rao and Molina (2015b) clasifican esta técnica como parte del grupo de estimadores sintéticos. Estos autores hacen mención de la similitud de este método con el estimador generalizado de regresión (*Generalized Regression Estimator*, GREG) como método de calibración, que también buscan encontrar valores que minimicen, por ejemplo, la distancia X^2 entre pesos muestrales y pesos calibrados.

Como su nombre lo indica, este procedimiento preserva la estructura del censo *Z* correspondiente al tiempo t_0. Para entender cómo el método logra cumplir con esta característica, se representa una tabla de doble entrada correspondiente al censo *Z*, con un modelo log-lineal saturado. De igual forma se puede organizar la información proveniente de una encuesta (Y_{aj}) y representarla con un modelo log-lineal:

$$log\, Y_{aj} = \alpha_0^Y + \alpha_a^Y + \alpha_j^Y + \alpha_{aj}^Y, \tag{16}$$

Purcell and Kish (1980) dividen al conjunto de los primeros tres términos α_0^Y, α_a^Y y α_j^Y y lo denominan estructura de asignación que contiene la información de los totales filas y columnas (márgenes). El último término, α_{aj}^Z, lo llaman estructura de asociación y provee las interacciones entre filas y columnas. De la ecuación anterior, los términos α_0^Z, α_a^Z, α_j^Z y α_{aj}^Z están restringidos a las siguientes condiciones:

$$\alpha_0^Z = \frac{1}{AJ}\sum_{a=1}^{A}\sum_{j=1}^{J} log\, Z_{aj}, \tag{17}$$

$$\alpha_a^Z = \frac{1}{J}\sum_{j=1}^{J} log\, Z_{aj} - \alpha_0^Z, \tag{18}$$

$$\alpha_j^Z = \frac{1}{A}\sum_{a=1}^{A} log\, Z_{aj} - \alpha_0^Z, \tag{19}$$

$$\alpha_{aj}^Z = log\, Z_{aj} - \alpha_a^Z - \alpha_j^Z - \alpha_0^Z. \tag{20}$$

Esto debe cumplir las mismas condiciones mencionadas anteriormente. El método SPREE asume la estructura de asignación que se obtiene de una encuesta actualizada provee márgenes recientes y confiables (A filas y J columnas). Alternativamente, esta estructura se puede obtener, por ejemplo, por medio de proyecciones de población.

Por otra parte, se asume que la estructura de asociación que puede brindar el censo es más sólida; es decir, logra representar las celdas (interacciones entre filas y columnas) de una manera más adecuada que los datos de encuesta. En este sentido, el método SPREE asume que estas interacciones se mantienen constantes en los años postcensales. El supuesto "estructural" es el siguiente:

$$\alpha_{aj}^Y + \alpha_{aj}^Z. \tag{21}$$

Siguiendo la notación de Luna-Hernández (2016), el proceso para obtener una tabla actualizada mediante SPREE ($\widehat{Y}_{aJ}^S$) se puede representar como:

$$\widehat{Y}_{aJ}^S = IPF\left[\!\left[exp\left(\widehat{a}_{aj}^Y\right), Y_a, Y_J \right]\!\right] \tag{22}$$

en donde Y_a y Y_J representan los márgenes (totales) confiables que provienen de datos de la encuesta (filas y columnas) y $\widehat{a}_{aj}^Y = a_{aj}^Z$. Para ahondar en los detalles de la relación entre los modelos log-lineales y los cuadros de contingencia, se recomienda consultar a Agresti (2002).

Tratar la actualización de tablas censales como un problema de optimización que puede ser resulto con algoritmos tradicionalmente utilizados en los estimadores de calibración implica que no es necesario imponer una estructura específica entre filas y columnas desde la encuesta, así como sus respectivas anidaciones, únicamente es necesario tener los totales de cada variable de interés. Lo anterior representa una ventaja significativa en comparación con otros algoritmos más complejos: esto evita procesos engorrosos al tratar de ajustar las tablas censales con los datos de la encuesta para que tenga la misma dimensión, creando valores faltantes o eliminando combinaciones existentes.

La idea básica de los métodos de calibración es utilizar información auxiliar relevante para ajustar los pesos muestrales, de tal forma que se puedan mejorar (afinar) las estimaciones basadas en los datos de encuesta. Si bien es cierto los métodos de calibración nacen para buscar soluciones a problemas en el campo de encuestas por muestreo, se han encontrado diversas utilidades adicionales, por ejemplo, su aplicación en el contexto de estimaciones en áreas pequeñas (Pfeffermann, 2013b). Además, Deville y Sarndal (1992) señalan que los métodos de calibración comparten características con otros métodos bajo determinados escenarios. Los autores señalan, por ejemplo, la conexión entre los métodos de calibración y GREG.

Además, con base en Deville y otros (1993), se evidencia la relación directa que existe entre los métodos de calibración basado en ranking multiplicativo y este tipo de métodos SPREE, ya que ambos son en realidad aplicaciones del IPFP propuesto por Deming and Stephan (1940). Aunque existen distintos estimadores de calibración (por ejemplo, lineal y logit), en este apartado se explica únicamente el método de calibración ranking multiplicativo como una alternativa para actualizar tablas censales. Más detalles de ésta y otras técnicas de calibración se pueden encontrar, por ejemplo, en Deville y otros (1993).

Siguiendo la descripción y notación de Gutiérrez (2016), se define *U* una población finita de tamaño *N* y *s* una muestra probabilística seleccionada de dicha población. Además, y_k representa el valor de la variable de interés para cada uno de los k individuos en la población y x_k el valor de una variable auxiliar. Los valores de ambas variables son conocidos y confiables. Con base en la información auxiliar, se puede obtener un total poblacional $t_x = \sum_k x_k$. Siendo t_y el total poblacional de la variable de interés, su estimador se define como $\hat{t}_{(y)} = \sum_k w_k y_k$, el cual equivale al estimador de Horvitz-Thompson con $d_k = \frac{1}{\pi_k}$. Entonces, el objetivo de la calibración es encontrar nuevos pesos w_k que mantengan su cercanía d_k. Existen distintas alternativas que permiten alcanzar este objetivo, es decir, reducir la "pseudo" distancia $G\left(\frac{w_k}{d_k}\right)$ entre w_k y d_k.Desde un punto de vista de optimización, se puede plantear:

$$\sum_{k\in S} d_k \frac{G\left(\frac{w_k}{y_k}\right)}{q_k} \tag{23}$$

con $q_k\,(k\in S)$ ponderaciones positivas y conocidas y los nuevos pesos deben cumplir con que:

$$\sum_{k\in S} w_k x_k = \sum_U x_k = t_x \tag{24}$$

Con ayuda de los multiplicadores de Lagrange (vector λ) y una función $F(\cdot)$, se puede solucionar el problema de optimización y finalmente obtener nuevos pesos:

$$w_k = d_k F(q_k \lambda' x_k) = d_k g_k. \tag{25}$$

Los detalles de las derivaciones poder llegar a este resultado se pueden consultar en Gutiérrez-Rojas (2016). Un estimador de calibración genérico, se define entonces como:

$$\widehat{t_y}, cal = \sum_{k\in S} w_k y_k = \sum_{k\in S} d_k F(q_k \lambda' x_k) y_k = \sum_{k\in S} d_k g_k y_k \tag{26}$$

en donde se debe especificar la distancia G(.) y la función F(.) a utilizar (por ejemplo, distancia Ji-cuadrado). El método multiplicativo ranking asume el uso de la distancia de entropía que define la distancia como $G_{(k)} = x_{log}(x) - x + 1$ y, además, $F_{(u)} = exp(u)$. De esta forma, los pesos calibrados son $w_k = d_k\, g_k$.

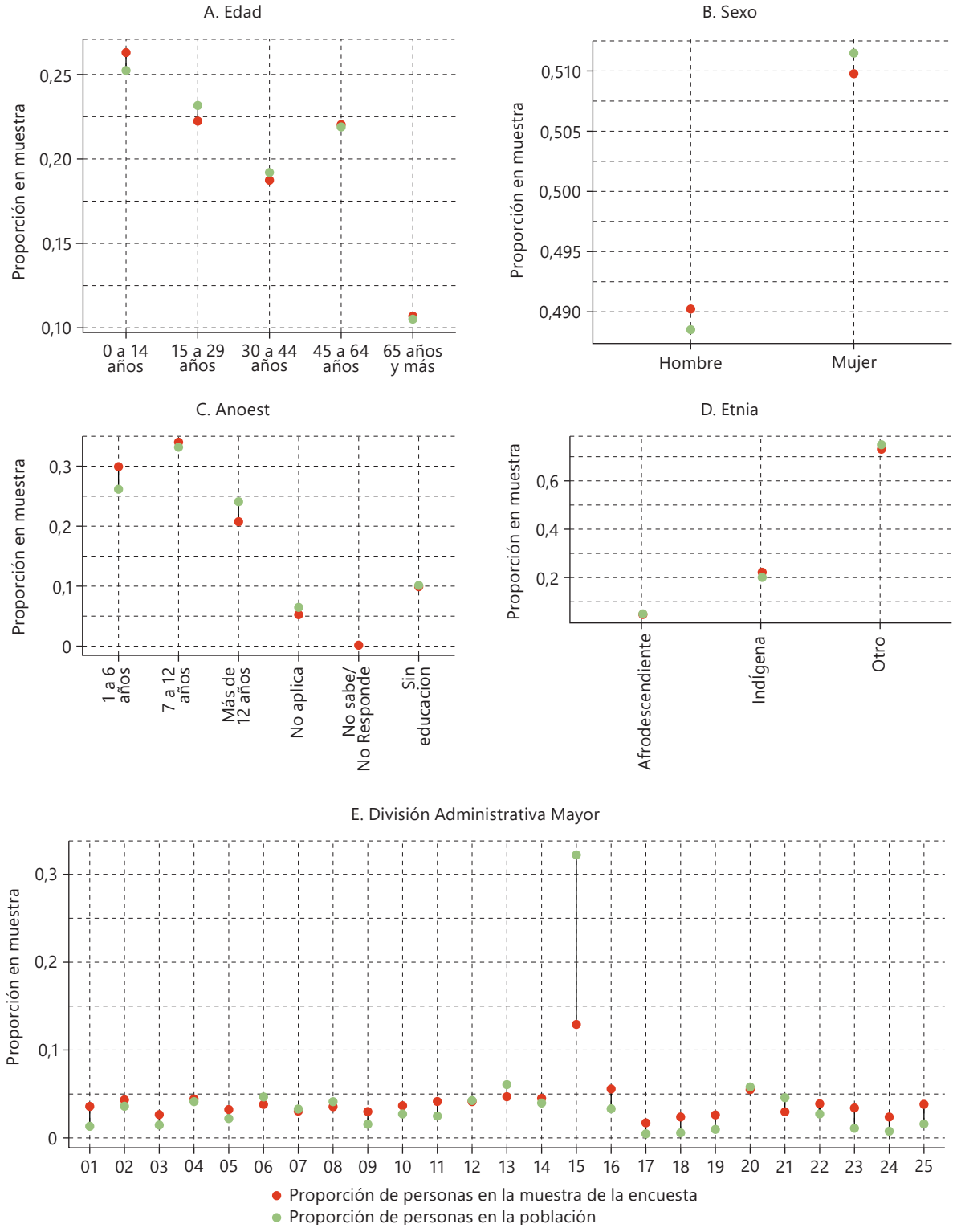
Existen algunas otras alternativas que pueden ser aplicadas para actualizar las tablas censales, como por ejemplo los basados en modelos lineales generalizados y en métodos de máxima verosimilitud. Con respecto al primero, Marker (1999) y Noble y otros (2002) explican que si una tabla de contingencia puede representarse como un modelo log-lineal, entonces el procedimiento SPREE puede incluirse dentro del marco de modelos lineales generalizados. Para poder preservar la estructura del censo, como es deseado en los métodos SPREE, la tabla censal puede ser incluida como un *offset* en un modelo Poisson (Noble y otros, 2002; Isidro y otros, 2016). Para más detalles de la relación entre un modelo log-lineal y los modelos lineales generalizados, se recomienda consultar a Agresti (2002).

C. Ejemplo de la actualización censal

Para ilustrar las metodologías de actualización censal, esta sección describe los pasos y verificaciones realizadas, utilizando como ejemplo un país de la región. Una vez definidos los postestratos de interés utilizando las mismas categorías en ambos conjuntos de datos, se evalúan la composición y proporción de cada categoría de las variables a actualizar. Este paso es crucial para verificar que no existan diferencias notables entre el censo y la encuesta o, en el caso de haberlas, identificar una explicación coherente.

El gráfico 1 muestra el ajuste de las covariables de interés (área, etnia, edad, años de estudio y departamento), donde el punto verde representa la proporción de cada categoría en el censo y el rojo en la encuesta sin expandir. Las variables sexo, años de estudio, edad y etnia muestran valores muy cercanos entre ambas fuentes . Para la variable departamento, el ajuste es aceptable, aunque se observa una diferencia significativa en el departamento 15 (Lima Metropolitana), debido al sobremuestreo en la capital del Perú en la encuesta.

Gráfico 1
Perú: exploración de ajuste entre variables del censo 2017 y la encuesta 2021
(Proporción de personas)



Fuente: Elaboración propia.

Siguiendo con el ejemplo del Perú, el cuadro 7 presenta los valores obtenidos para algunas covariables que definirán los postestratos. Por conveniencia en la representación de este ejemplo, sólo se incluyeron las covariables de área, sexo, etnia, edad y años de estudio. El cuadro presenta tres instantes en el tiempo. El primero de ellos está definido por el censo: las columnas tres y cuatro del cuadro muestran respectivamente el conteo de casos y la proporción de casos en el censo realizado en el año 2017. El segundo instante es inducido por la encuesta realizada en el año 2021: las columnas cinco, seis y siete muestran respectivamente el conteo no ponderado de casos (tamaño de muestra), la proporción no ponderada de casos en la encuesta y la estimación del número de individuos en la categoría (conteo ponderado por el factor de expansión). El tercer instante está dado por la actualización de esta tabla censal: las últimas dos columnas representan respectivamente los conteos y la proporción de casos actualizada para cada categoría.

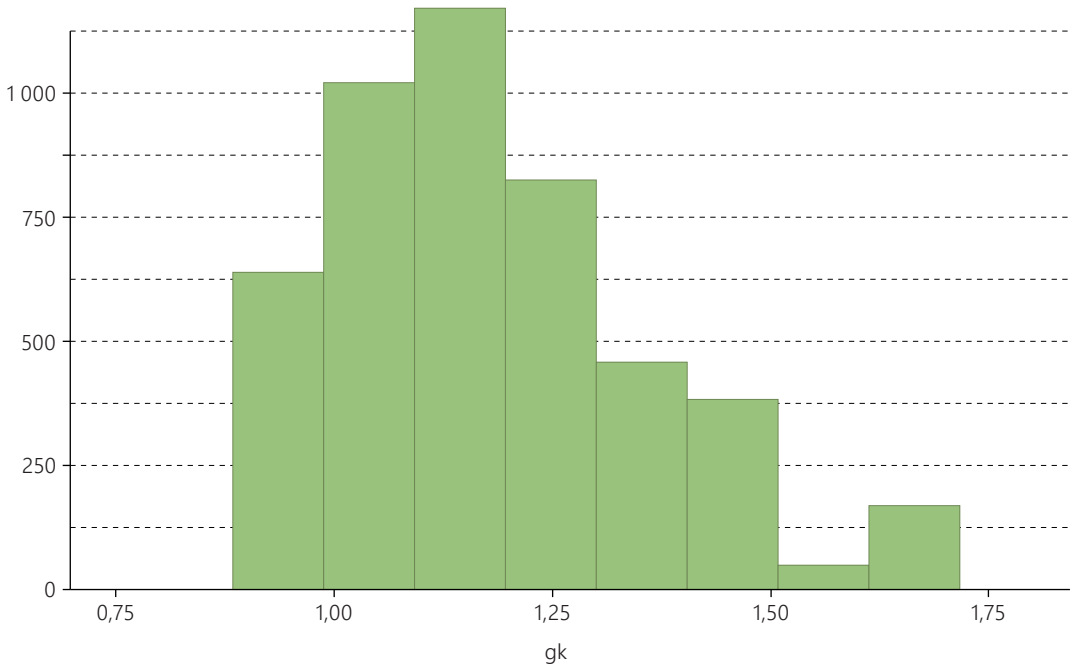
Cuadro 7
Perú: proceso de actualización de tablas censales

Variable	Categoría	Conteo de casos en el censo	Proporción de casos en el censo	Tamaño de muestra en la encuesta	Proporción de casos en la encuesta	Tamaño estimado en la encuesta	Conteo actualizado en el censo	Proporción actualizada en el censo
Área	Rural	5 658 358	0,181	44 134	0,367	6 930 312	690 6150	0,208
	Urbano	25 579 027	0,819	76 212	0,633	26 291 292	26 281 665	0,792
Sexo	Hombre	15 467 946	0,495	58 997	0,490	16 229 105	16 242 531	0,489
	Mujer	15 769 439	0,505	61 349	0,510	16 992 499	16 945 284	0,511
Etnia	Indígena	6 383 284	0,204	26 693	0,222	6 664 965	6 663 342	0,201
	Afrodescendiente	877 429	0,028	5 692	0,047	1 650 657	1 651 300	0,050
	Otro	23 976 672	0,768	87 961	0,731	24 905 981	24 873 173	0,750
Edad	0 a 5 años	3 254 581	0,104	10 205	0,085	2 750 815	2 745 014	0,083
	6 a 14 años	5 014 971	0,161	21 443	0,178	5 632 463	5 623 884	0,169
	15 a 29 años	7 869 821	0,252	26 763	0,222	7 696 362	7 686 746	0,232
	30 a 44 años	6 744 742	0,216	22 553	0,187	6 377 228	6 372 847	0,192
	45 a 64 años	5 797 210	0,186	26 516	0,220	7 273 669	7 268 960	0,219
	65 años o más	2 556 059	0,082	12 866	0,107	3 491 067	3 490 364	0,105
Años de estudio	Sin Educación	2 956 239	0,095	11 921	0,099	3 036 906	3 221 868	0,097
	De 1 a 6 años	7 784 611	0,249	36 009	0,299	9 125 079	9 038 352	0,272
	De 7 a 12 años	10 555 271	0,338	40 937	0,340	12 134 971	11 134 195	0,335
	Más de 12 años	7 816 767	0,250	24 975	0,208	7 162 527	8 001 597	0,241
	No aplica	2 124 497	0,068	6 318	0,053	1 697 171	1 791 803	0,054
	Otro	0	0,000	189	0,002	64 950	0	0,000

Fuente: Elaboración propia.

Como se detalló anteriormente, la actualización de conteos censales se realiza mediante el procedimiento IPFP. En este estudio se empleó la función *calib* del *software* R, utilizando el método *logit*. El resultado es un conjunto de ponderadores únicos para cada *postestrato*, que permite ajustar los valores censales a los de la encuesta. Para verificar que el proceso haya convergido correctamente en cada país, se utilizó la función *checkcalibration*, la cual confirma si el ajuste fue óptimo. En general, se espera que los ponderadores estén cercanos a la unidad, aunque en el caso de censos desactualizados estos valores pueden ser mayores. El histograma presentado en el gráfico 2 para Perú muestra que en este caso la mayoría de los ponderadores tiene valores mayores que la unidad.

Gráfico 2
Perú: histograma de los ponderadores para la actualización de los conteos censales
(Proporción de personas)



Fuente: Elaboración propia.

IV. Ajuste del modelo

En este capítulo se analizan los modelos de estimación para los indicadores de interés (ingreso promedio per cápita, incidencia de la pobreza e incidencia de la pobreza extrema) en áreas pequeñas. Como se mencionó en el capítulo I, para los indicadores de pobreza se emplea un modelo de regresión multinivel basado en postestratos que incluye efectos aleatorios y fijos. Para estimar el ingreso medio se emplea un modelo de regresión mixto para variables continuas, cuyos elementos teóricos siguen los mismos supuestos planteados en Gutiérrez y otros (2023), que pueden ser generalizados para modelos multinivel con postestratificación. Por lo anterior, esta sección se enfoca en los modelos logísticos para la desagregación de los indicadores de pobreza, utilizando la función de enlace logit, dada la naturaleza dicotómica de dichos indicadores.

A. Elementos de una regresión logística con efectos aleatorios

El tipo de modelo a emplear depende de la naturaleza de la variable de respuesta y de las covariables disponibles. En este caso, dado que la variable dependiente es binaria (toma el valor de uno (1) si el individuo se encuentra por debajo de la línea de pobreza (o pobreza extrema) y toma el valor de cero (0) en caso contrario), se empleará una regresión con vínculo logístico entre las covariables del modelo y la probabilidad de estar en situación de pobreza (o pobreza extrema). Siguiendo la notación del primer capítulo, la probabilidad de que el evento ocurra se expresa mediante una función logística de la siguiente manera:

$$logit(\rho_{id}) = ln\left(\frac{\rho_{id}}{1 - \rho_{id}}\right) = x_{id}\beta + z_{id}\gamma \tag{27}$$

Es posible reescribir esta expresión en términos de las covariables de regresión como se expresa a continuación:

$$\rho_{id} = \frac{e^{(x_{id}\beta + z_{id}\gamma)}}{e^{(x_{id}\beta + z_{id}\gamma)+1}} = \frac{1}{e^{-(x_{id}\beta + z_{id}\gamma)+1}} \tag{28}$$

En este estudio se emplean modelos con efectos aleatorios, que son distintos a los modelos clásicos de regresión logística. Estos modelos tienen dos componentes: efectos fijos, que suponen que hay una parte de la respuesta que es posible explicar en términos de las covariables, y efectos aleatorios, que asumen que hay una parte de la variabilidad de la respuesta que no es explicada por las covariables.

Como en cualquier modelo observacional, es común no incluir todas variables que pueden influir sobre la variable de interés, ya sea por desconocimiento o por tratarse de factores no observables o sin información disponible. Al no ser posible estimar la variabilidad asociada a estos factores, la varianza del término aleatorio se verá afectada por su propia varianza y la de las variables omitidas.

Los modelos con efectos aleatorios solucionan este inconveniente, ya que permiten separar la variabilidad en dos partes: la varianza original y la varianza de los efectos aleatorios, de la siguiente manera:

$$Respuesta = efectos\ fijos + efectos\ aleatorios + error \quad (29)$$

En este estudio, los efectos fijos corresponden a las partes del modelo que pueden ser incluidas explícitamente, mientras que los efectos aleatorios, aunque no se pueden controlar, pueden ser modelados. Una diferencia fundamental entre los modelos de efectos fijos y los modelos mixtos es que, en el primer caso, la probabilidad de éxito se calcula a partir del promedio de éxitos de la variable de interés en cada una de las categorías definidas por las covariables de los efectos fijos; mientras que en el segundo, el punto de partida es un promedio ponderado entre las proporciones de los cruces de las covariables de los efectos fijos y el promedio global. Por ejemplo, si se quiere modelar la pobreza en términos del sexo y el área (ambos dominios de interés con una buena representación en la muestra), el punto inicial de la estimación sería la proporción de personas pobres en cada una de las categorías de estas variables. Sin embargo, si además se requieren estimaciones para niveles más bajos de desagregación (por ejemplo el cruce entre DAM, etnia y escolaridad), el punto inicial será un promedio ponderado entre las proporciones inducidas por sexo y área, la proporciones inducidas por cada cruce disponible entre DAM, etnia y escolaridad, y la proporción de personas pobres a nivel nacional.

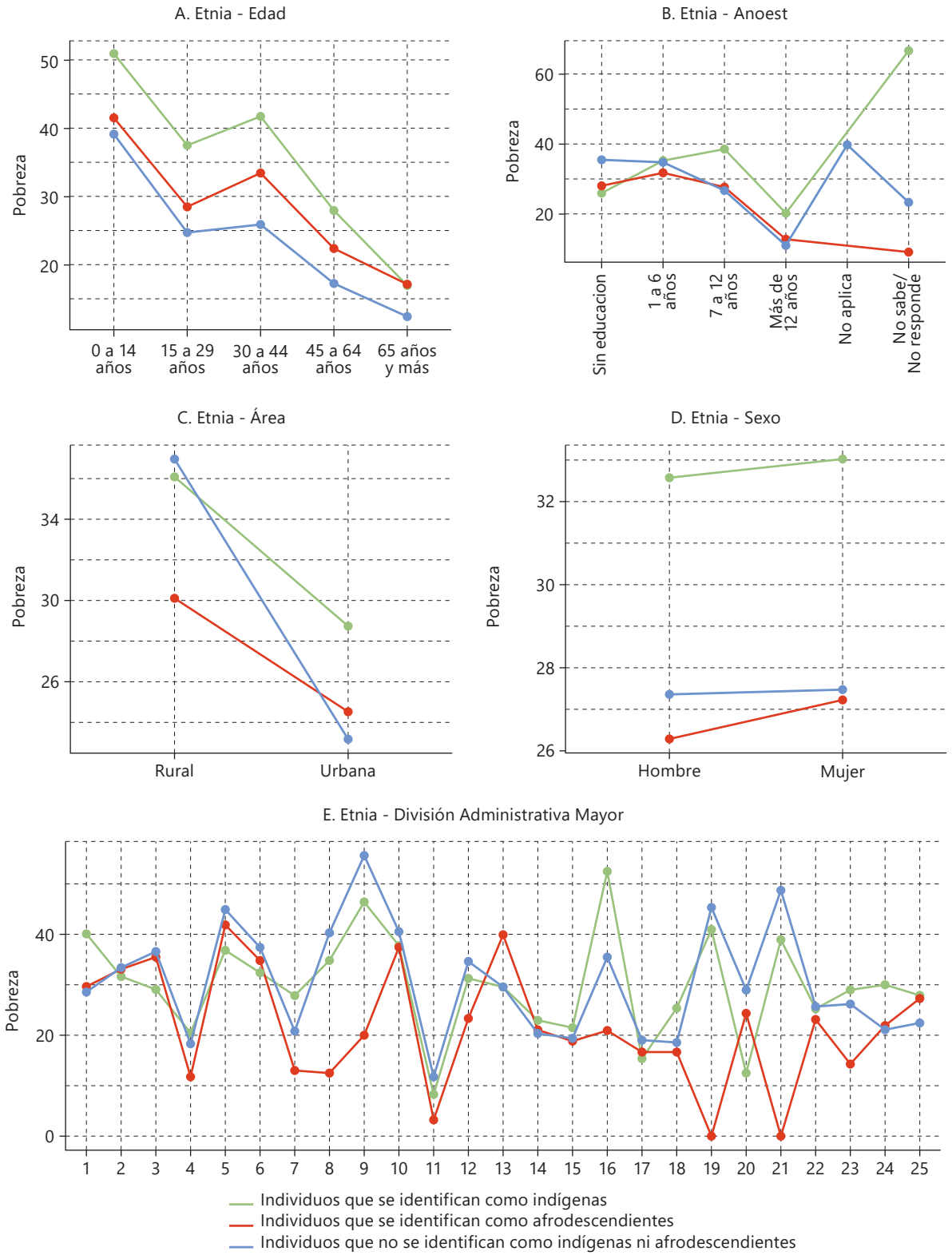
B. Interacciones entre las covariables de postestratificación

Existen fenómenos que requieren no solo la inclusión de las variables que afectan la respuesta, sino también del estudio de posibles interacciones entre ellas. Por ejemplo, la baja escolaridad suele estar asociada a la situación de pobreza. No obstante, es posible que existan personas mayores de 65 años y con una baja escolaridad que tienen un ingreso por encima de la media y no son pobres, debido a que tienen acceso a una jubilación. En este caso, la situación de pobreza no queda adecuadamente explicada en términos de las covariables por sí solas, sino que requiere considerar su la interacción.

En esta sección se presentan validaciones gráficas para analizar las interacciones entre las distintas covariables y determinar cuáles deben incluirse en el modelo. Este paso es importante, ya que las interacciones identificadas se integrarán al modelo final. El análisis consiste en elaborar un diagrama de puntos entre dos variables. En el eje de las ordenadas se muestra el porcentaje no ponderado de personas en situación de pobreza, mientras que en el eje de las abscisas se representan las categorías de una covariable de postestratificación. Cada línea del gráfico, diferenciada por color, representa una categoría de la otra covariable. Si las líneas se cruzan, esto indica la presencia de una interacción entre las covariables, que debe ser tenida en cuenta en el modelo.

Continuando con el ejemplo de Perú, el gráfico 3 muestra la interacción de la variable etnia con las demás covariables de interés. Las líneas verdes representan a los individuos que se identifican como indígenas, las rojas a quienes se identifican como afrodescendientes, y las azules a los que no se identifican como indígenas ni afros. Se observa una ligera interacción entre la etnia y los años de estudio, el área y la división administrativa mayor (DAM), pero no para edad y sexo. Por lo tanto, la interacción entre etnia y sexo no debería incluirse en los modelos a estimar.

Gráfico 3
Perú: ejemplo de las interacciones de las covariables de postestratificación
(Proporción de personas en pobreza)



Fuente: Elaboración propia.

La definición de las interacciones entre las diferentes covariables aumenta el poder predictivo de los modelos de estimación en áreas pequeñas. Sin embargo, debe decidirse, con base en la información metodológica de cada encuesta, si la interacción debe incluirse como un efecto fijo o aleatorio. Como se mencionó anteriormente, si el tamaño de muestra no es suficientemente grande en cada subgrupo definido por la interacción, esta debería considerarse como parte de los efectos aleatorios.

C. Efectos fijos y efectos aleatorios

Después de realizar las anteriores validaciones, se realiza el ajuste del modelo con efectos fijos y aleatorios para cada indicador de interés. Los efectos fijos se pueden clasificar dependiendo del nivel al que se encuentre expresada la covariable:

- A nivel individual, las covariables pueden describir características de una persona o de su hogar, como el departamento y área en que habita, el sexo o la edad. Para ser incluidas como efectos fijos, cada una de las categorías de estas covariables deberían tener asociado un tamaño de muestra lo suficientemente grande como para asegurar su representatividad.
- A nivel agregado, las covariables pueden describir características asociadas con la división administrativa mayor (que se considera representativa en la encuesta). La inclusión de estas covariables en el modelo le da el carácter multinivel al modelo. Estas covariables provienen de la agregación de datos censales (necesidades básicas insatisfechas, estadísticas del mercado de trabajo, tenencia de enseres, acceso a servicios, entre otras) o geoespaciales (intensidad lumínica, fracción de suelo urbano y fracción del suelo cultivado, índice de modificación humana, tiempo de llegada al centro de salud más cercano caminando y en vehículo motorizado).

Por otra parte, se incluyen como efectos aleatorios cada una de las covariables de postestratificación para las cuales la muestra no contenga todas sus categorías o en donde su tamaño de muestra no sea lo suficientemente grande. En general, tanto la pertenencia étnica como los años de estudio siempre serán considerados como efectos aleatorios, así como la mayoría de las interacciones incorporadas en el modelo.

Por ende, para los indicadores de pobreza extrema y pobreza, el modelo multinivel toma la siguiente forma:

$$logit(\rho_{id}) = x_{id}\beta + z_d\gamma + \varepsilon_{id} = \beta_0 + x^1_{id}\beta_1 + x^2_d\beta_2 + z_{id}\gamma \quad (30)$$

En esta notación, denota el intercepto del modelo, las covariables en la matriz x^1_{id} representan la información a nivel individual, mientras que las covariables en x^2_d representan la información agregada. A manera de ejemplo, considérese un país en el que las covariables del modelo incluyen:

- Como efectos fijos, las siguientes covariables de postestratificación en x^1_{id} el departamento (DAM), el área, la edad y el sexo.
- Como efectos aleatorios, las siguientes covariables de postestratificación en z_{id} la etnia y los años de escolaridad.
- Como efectos fijos, las siguientes covariables geoespaciales a nivel de DAM en x^2_d: luces nocturnas, cubrimiento rural, cubrimiento urbano, modificación humana, accesibilidad a hospitales y accesibilidad a hospitales caminando.
- Como efectos fijos, las siguientes covariables censales a nivel de DAM en x^2_d: porcentaje de personas por sexo, porcentaje de personas por categoría de edad, porcentaje de personas por categoría de años de escolaridad, porcentaje de personas por etnia, porcentaje de personas en cada dimensión del índice NBI, porcentaje de viviendas con pisos de tierra, porcentaje de viviendas con material de las paredes inadecuado, porcentaje de viviendas con techos inadecuados, porcentaje de hogares con acceso a electricidad e internet, porcentaje de personas desocupadas, porcentaje de personas con alguna discapacidad, entre otras.

- Como efectos aleatorios las siguientes interacciones de postestratificación en $z_{id'}$ el cruce entre departamento y sexo, el cruce entre departamento y edad, el cruce entre departamento y años de estudio, el cruce entre área y años de estudio, el cruce entre sexo y edad, el cruce entre sexo y años de estudio, el cruce entre edad y años de estudio.

Una diferencia importante entre los efectos fijos y los aleatorios es que el vector de parámetros β no asume ningún comportamiento aleatorio, mientras que γ sí lo hace. De hecho, uno de los parámetros más importantes en este tipo de modelos es la matriz de covarianzas de este vector, denotada como . Esta matriz puede asumirse con diferentes estructuras de correlación. En particular, este documento asume que la estructura de Σ es simple y que las covarianzas entre los efectos aleatorios son nulas. Suponiendo que existen *H* covariables definidas como efectos aleatorios en el vector γ, entonces la estructura de esta matriz será como sigue:

$$\sum = var(\gamma) = diag(var(\gamma_1), \ldots, var(\gamma_H)) \tag{31}$$

D. Modelamiento bayesiano

Independientemente del paradigma estadístico escogido (bayesiano o frecuentista), el modelamiento estadístico demanda la estimación del vector de parámetros de efectos fijos, denotada como $\hat{\beta}$, la estimación del vector de parámetros de efectos aleatorios, denotada como $\hat{\gamma}$, así como de su matriz de covarianzas, la cual se supone diagonal en esta investigación (proveyendo una estructura simple de autocorrelación). Para mantener la simpleza en la notación, supongamos que el vector de parámetros en los modelos se denotará como $\theta = (\beta, \gamma, \sum)'$ y su estimación como $\hat{\theta} = (\hat{\beta}, \hat{\gamma}, \hat{\sum})$.

El enfoque bayesiano empieza con la especificación de un modelo para los datos observados, *Y*, dado un vector de parámetros desconocidos, θ, en forma de distribución condicional, $p(Y \mid \theta)$. Además, supone que la incertidumbre alrededor de estos parámetros es aleatoria y que se puede modelar mediante una distribución de probabilidad previa. En términos de estimación, inferencia y predicción, el enfoque Bayesiano supone dos momentos o etapas:

i) Antes de la recolección de los datos, en donde el investigador propone, basado en su conocimiento, experiencia o fuentes externas, una distribución de probabilidad previa, denotada como $p(\theta)$, para el parámetro de interés. Con esta distribución es posible calcular estimaciones puntuales e intervalos de credibilidad con el fin de confirmar que la distribución propuesta se ajusta al problema de estudio. En esta etapa, basados en la distribución previa, también es posible hacer predicciones de cantidades observables.

ii) Después de la recolección de los datos. Siguiendo el teorema de Bayes, el investigador actualiza su conocimiento acerca del comportamiento probabilístico del parámetro de interés mediante la distribución posterior de este. Con esta distribución posterior, denotada como $p(Y \mid \theta)$, es posible calcular estimaciones puntuales y por intervalo justo como en el enfoque frecuentista. En esta etapa, basados en la distribución posterior, también es posible hacer predicciones de cantidades observables y pruebas de hipótesis acerca de la adecuación del mejor modelo a los datos observados.

Una vez recolectados los datos en la encuesta, y habiendo definido la forma funcional de la distribución previa, toda la inferencia estadística descansará en la relación condicional del vector de parámetros respecto a los datos observados. Esta relación toma la siguiente forma y se conoce en la literatura como la regla de Bayes:

$$p(\theta \mid Y) \propto p(\theta)p(Y \mid \theta) \tag{32}$$

Es decir, la distribución posterior, $p(\theta \mid Y)$, es directamente proporcional al producto entre la distribución previa, $p(\theta)$, y la distribución de los datos, $p(Y \mid \theta)$. Con base en lo anterior, es posible calcular la probabilidad posterior de que el vector de parámetros θ esté en cualquier región *G*, condicionada a los datos observados, de la siguiente manera:

$$Pr(\theta \epsilon G \mid Y) = \int p(\theta \mid Y) d\theta \tag{33}$$

Además, también es posible calcular una estimación puntual para el vector θ dados los datos observados *Y*, la cual está dada por alguna medida de tendencia central para la distribución $p(\theta \mid Y)$. En particular, si se escoge la media, entonces, la estimación puntual estará dada por la siguiente expresión:

$$\hat{\theta} = \mathrm{E}(\theta \mid Y) = \int \theta p(\theta \mid Y) d\theta \tag{34}$$

Un beneficio clave del enfoque de SAE es su capacidad para hacer predicciones en subgrupos que no estaban directamente incluidos en la muestra de la encuesta. Desde el punto de vista de la inferencia predictiva, también es posible utilizar los principios del modelamiento bayesiano para obtener no solo estimaciones sobre los conjuntos de datos observados, denotado como *Y*, sino también sobre el conjunto de datos no observado, denotado como $\tilde{Y}$. Siguiendo el enfoque de este documento, los conjuntos de datos observados estarán supeditados a los cruces de los postestratos existentes en las encuestas de hogares, mientras que los conjuntos de datos no observados estarán inducidos por los cruces de los postestratos que no fueron observados en las encuestas, pero que sí existen en el censo.

De este modo, aunque muchos cruces de los postestratos no estén representados en la muestra, es posible hacer predicciones basadas en las relaciones observadas en otros subgrupos y en la información auxiliar disponible. Por ejemplo, si una encuesta no incluye datos de todos los cruces entre sexo, etnia, área y escolaridad, el enfoque presentado en este documento permite predecir en esos cruces utilizando las relaciones provistas por los modelos, aprovechando la información en áreas similares con características comunes con los cruces que sí fueron observados.

Por ejemplo, la Gran Encuesta Integrada de Hogares de Colombia permite realizar estimaciones directas de la tasa de pobreza en solo 24 de los 33 departamentos del país. En este caso, no solo no existe muestra para nueve DAM, sino que no existe muestra para ningún cruce de los postestratos de interés dentro de estas DAM. Sin embargo, utilizando los principios predictivos de la inferencia bayesiana sí es posible obtener una predicción sobre el comportamiento promedio de las personas y hogares dentro de estos subgrupos que no tienen una representación adecuada en la encuesta.

Por ende, después de la recolección de los datos, el investigador está interesado en conocer qué distribución tiene el conjunto de datos no observados. Esta distribución se conoce con el nombre de distribución predictiva posterior, la cual se denota como $p(\tilde{Y} \mid Y)$, y está dada por la siguiente expresión:

$$p(\tilde{Y} \mid Y) = \int p(\tilde{Y} \mid \theta) p(\theta \mid Y) d\theta \tag{35}$$

En donde $p(\tilde{Y} \mid \theta)$ es la distribución de los datos evaluada en los nuevos valores $\tilde{Y}$. Con esta distribución posterior es posible definir medidas de tendencia central, como predicciones puntuales, y medidas de dispersión, para la creación de los intervalos de credibilidad. En particular, la predicción puntual estaría dada por la esperanza de la distribución predictiva, dada por $E(\tilde{Y} \mid Y)$. Así mismo, la estimación puntual para el conjunto de datos observado estaría dada por la esperanza de la distribución de los datos evaluada en las realizaciones de los valores para el vector de parámetros θ, es decir, $E(Y \mid \hat{\theta})$.

Teniendo en cuenta lo anterior, y basados en lo expuesto en la primera sección de este documento, se supone que existen *Q* combinaciones posibles (cruces entre covariables de postestratificación). Sin pérdida de generalidad, asumimos que las primeras *K* combinaciones son observadas en la encuesta y que las últimas *Q-K* combinaciones no son observadas en la encuesta, pero sí en el censo. De esta forma, el estimador para la proporción de personas en condición de pobreza dentro de las áreas (subgrupos) de interés vendrá dada por:

$$\hat{\rho}_s = \frac{\sum_{j=1}^{K} N_{sj}\hat{\rho}_{sj} + \sum_{j=K+1}^{Q} N_{sj}\hat{\rho}_{sj}}{\sum_{j=1}^{Q} N_{sj}} = \frac{\sum_{j=1}^{K} N_{sj}E_{sj}\left(Y \mid \hat{\theta}\right) + \sum_{j=K+1}^{Q} N_{sj}E_{sj}(\tilde{Y} \mid Y)}{\sum_{j=1}^{Q} N_{sj}} \quad (36)$$

Por lo tanto, $\hat{\rho}_{sj} = E\left(Y \mid \hat{\theta}\right)$ para los cruces observados, mientras que $\hat{\rho}_{sj} = E_{sj}\left(\tilde{Y} \mid Y\right)$, para las combinaciones no observadas.

Por otro lado, la escogencia de una distribución previa, $p(\theta)$, es muy importante en el análisis bayesiano puesto que afecta directamente la forma funcional de la distribución posterior, $p(\theta \mid Y)$, tal como lo ilustra la regla de Bayes. En primer lugar, la distribución previa debe describir adecuadamente los conocimientos del investigador sobre los parámetros de interés en la estimación. Por ejemplo, si se cree que un parámetro toma valores cercanos a un cierto valor, entonces la distribución previa escogida para representarla debería rondar en valores cercanos a este mismo. Por otro lado, dado que en la literatura existe un gran número de distribuciones, algunas muy similares entre ellas, a la hora de escoger una distribución previa también se debe tener en cuenta las implicaciones sobre los cálculos en la estimación puntual o del intervalo de credibilidad, procurando en la mayoría de los casos, obtener una distribución posterior tratable y fácil de manejar computacionalmente.

Cuando no existe una información previa muy precisa sobre el parámetro de interés o cuando existe total ignorancia de parte del investigador acerca del comportamiento probabilístico de los parámetros, es necesario definir distribuciones previas que sean no informativas. Es decir, que jueguen un papel mínimo en términos de influencia en la distribución posterior. Una característica de estas distribuciones es que su forma es vaga, plana o difusa. En los anteriores términos, la distribución uniforme define una distribución previa que cumple con las características de no información en la mayoría de los escenarios, incluyendo aquellos en donde el parámetro de interés está limitado a un espacio de muestreo acotado. Sin embargo, no en todos los problemas encaja la distribución uniforme. Nótese, por ejemplo, que en el caso en que los modelos de regresión, los coeficientes de regresión podrían tomar cualquier valor en los reales, por lo que la distribución uniforme sobre un rango infinito (–∞,∞) sería teóricamente problemática, ya que no es una distribución propiamente dicha (no tiene una densidad que integre a la unidad). Asimismo, la varianza de los errores está acotada al espacio (0, ∞) en cuyo caso la distribución uniforme (0, ∞) tampoco sería adecuada en el espacio de muestreo del parámetro de interés.

Para resolver estos inconvenientes, este documento plantea distribuciones previas propias y no informativas, para que sea la evidencia proporcionada por los datos observados la que predomine en la inferencia. Por ende, se propone que cada componente β_1 en el vector β sigue una distribución previa normal centrada en cero con una varianza muy alta (esto refleja una falta de conocimiento específico, permitiendo que los coeficientes tomen cualquier valor con una probabilidad casi igual):

$$\beta_1 \sim normal\left(0, 1000^2\right) \quad (37)$$

Para el caso de los efectos aleatorios, se supone que cada componente Y_r en el vector Y tiene una componente de varianza desconocida y de interés; por ende:

$$\gamma_r \sim normal(0, \sigma_r^2) \quad (38)$$

Asimismo, la distribución una distribución previa no informativa para el parámetro de varianza de los efectos aleatorios está dada por:

$$\sigma_r^2 \sim inversa - \text{gamma}(0.001, 0.001) \quad (39)$$

La anterior escogencia permite obtener una distribución posterior conjugada y tratable analíticamente. Existen otro tipo de distribuciones previas que permiten ajustar este tipo de modelos; por ejemplo, la distribución Half-Cauchy es una distribución no informativa para la desviación estándar (y por ende para la varianza) de los efectos aleatorios en modelos de regresión bayesianos. Esta distribución con un parámetro de escala grande es considerada no informativa en el sentido de que no impone una estructura previa sobre la varianza de los errores.

Como se resaltó anteriormente, el ajuste de los modelos sigue un paradigma bayesiano y la inferencia se realiza a través de métodos de Monte Carlo basados en Cadenas de Márkov (MCMC, por sus siglas en inglés). Dado que una gran parte de la inferencia bayesiana está ligada a la programación e implementación de los métodos MCMC para realizar inferencias posteriores de los parámetros de interés, en este documento se siguió el razonamiento y recomendaciones de Gelman y Hill (2020), los cuales pueden ser resumidas a continuación:

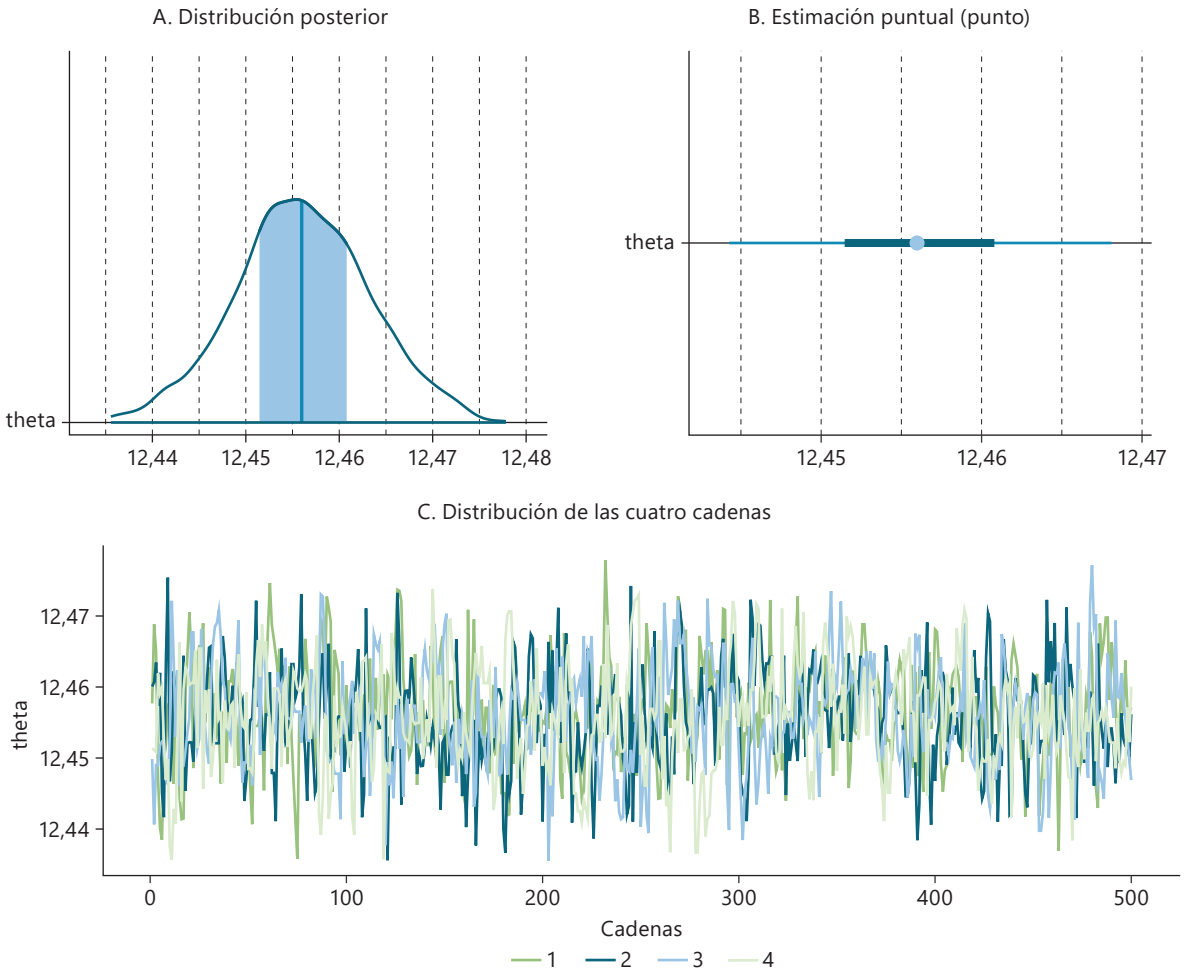
- Simulación de cuatro cadenas de forma paralela con al menos mil realizaciones. Los valores iniciales de cada cadena deben estar dispersos entre sí. Simular varias cadenas MCMC es importante para asegurar la convergencia del muestreo y obtener estimaciones posteriores confiables. Varias cadenas, inicializadas en puntos diferentes del espacio de parámetros, ayudan a evitar que el muestreo quede atrapado en una región local del espacio de parámetros. Al comparar las cadenas, se puede verificar si convergen a la misma distribución y si las muestras son representativas de la distribución posterior. Esto proporciona una verificación adicional de la calidad y estabilidad del muestreo.
- Comprobación de la convergencia de las cadenas mediante el descarte de la primera mitad de los valores generados en las cadenas. Esta etapa se conoce como *burning stage*. Deshacerse de la mitad de las realizaciones ayuda a eliminar el efecto de las condiciones iniciales y permite que las cadenas se estabilicen en la distribución posterior. Al descartar las primeras muestras, se reduce el sesgo causado por la proximidad a los valores iniciales, asegurando que las muestras restantes sean más representativas de la distribución posterior verdadera.
- Una vez que las cadenas converjan, mezclar los cuatro conjuntos de valores generados por las cadenas. Esto garantiza, en primera instancia, que las cadenas no estén auto correlacionadas. Combinar muestras de diferentes cadenas ayuda a mejorar la representación de la distribución posterior y a obtener estimaciones más robustas. Cuando se utilizan varias cadenas, cada una puede explorar diferentes áreas del espacio de parámetros y, al mezclar las muestras, se obtiene una visión más completa y precisa de la distribución posterior. Esto también ayuda a verificar la convergencia de las cadenas, ya que las muestras mezcladas deben reflejar una distribución estacionaria si todas las cadenas han convergido adecuadamente.
- Además de realizar esta mezcla, descartar valores intermedios mediante un muestreo sistemático. Esta etapa se conoce como *thinning stage*. Este submuestreo reduce la correlación entre muestras consecutivas y hace que las muestras sean más independientes entre sí. Dado que las cadenas MCMC tienden a producir muestras correlacionadas, esta etapa ayuda a disminuir esta dependencia, lo que facilita la estimación de la distribución posterior verdadera. Al final se recomienda almacenar una cantidad elevada de valores simulados.

En este documento se utiliza el software STAN (Stan Development Team, 2022) para realizar la simulación de las distribuciones posteriores en los modelos. Este es un software usado en inferencia bayesiana que utiliza un enfoque de modelación probabilística que permite la especificación de modelos complejos de manera flexible. STAN permite realizar la simulación de las cadenas de forma eficiente mediante métodos de muestreo avanzados, como el Hamiltonian Monte Carlo (HMC), junto con la técnica de "*no U-turn sampling*" (NUTS). Estos algoritmos permiten explorar eficientemente el espacio de parámetros de los modelos complejos usados en esta investigación. STAN ofrece varias interfaces a varios lenguajes de programación, incluyendo R, lo que facilitó su integración con los flujos de trabajo asociados a las anteriores etapas de la modelación.

Para ejemplificar el comportamiento de las cadenas para un coeficiente de regresión de los efectos fijos, obsérvese el gráfico 4. En la parte superior izquierda, se muestra la distribución posterior resultante de la mezcla de todas las realizaciones, luego de las etapas de *burning* y *thinning*. Es posible verificar que la distribución posterior sigue un comportamiento normal y que el rango de variación de este parámetro es positivo, lo que implica que la covariable asociada al coeficiente de regresión es significativa en el modelo.

En la parte superior derecha se observa estimación puntual (punto), junto con los intervalos de credibilidad al 50% (barra oscura) y al 95% (barra delgada). La parte inferior del gráfico muestra que la distribución de las cuatro cadenas es estacionaria y que el proceso computacional ha alcanzado convergencia.

Gráfico 4
Ejemplo del comportamiento de las cadenas para un coeficiente de regresión
(Valores del parámetro de regresión en las cadenas de Markov)



Fuente: Elaboración propia.

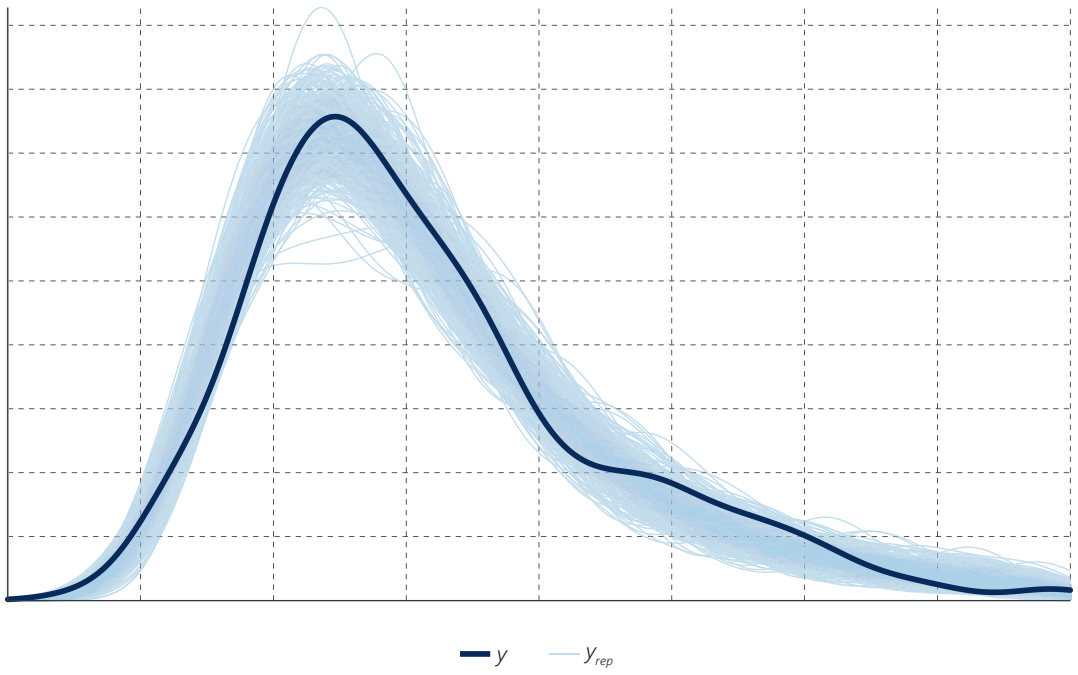
La simulación MCMC tiene un alto requerimiento computacional, ya que al ejecutar cuatro cadenas con mil realizaciones cada una, se generan cuatro mil pseudo-censos. Dado que estos modelos suelen ser complejos, con numerosos parámetros y estructuras jerárquicas, el espacio de parámetros también es considerable. Los métodos MCMC deben recorrer este espacio, lo que puede requerir un mayor número de iteraciones para lograr una adecuada mezcla y cobertura del espacio de parámetros. Además, como las muestras generadas deben ser almacenadas en memoria, un gran número de pseudo-censos implica un uso significativo de memoria. Los cálculos necesarios para generar y procesar las muestras también consumen memoria y tiempo. Por ejemplo, si cada iteración tardara diez segundos en completarse, el tiempo total para 4000 iteraciones sería de 40 000 segundos, es decir, aproximadamente 11 horas.

Además de seguir las recomendaciones anteriores, es necesario realizar chequeos predictivos posteriores (PPC, por sus siglas en inglés). Estos consisten en comparar las predicciones del modelo con

datos observados, para verificar si el modelo captura adecuadamente la variabilidad y la estructura de los datos. Estas verificaciones ayudan a identificar posibles deficiencias del modelo y determinar si este es adecuado para generar estimaciones y predicciones precisas, asegurando no solo un buen ajuste a los datos observados, sino también un buen desempeño en la predicción de datos no observados.

El gráfico 5 ejemplifica cómo los chequeos posteriores predictivos son útiles para evaluar la calidad del ajuste del modelo. En él se compara la distribución observada en los datos de la encuesta con las distribuciones generadas por el modelo. La línea solida representa los datos reales que se recolectaron y que se usaron para ajustar el modelo; específicamente denota la densidad suavizada del histograma de la variable de interés ingreso medio per cápita. Las líneas tenues son los datos generados por el modelo usando los parámetros obtenidos durante el proceso de muestreo en STAN. La comparación visual entre las distribuciones de los datos observados con las distribuciones generadas por el modelo se superpone y son similares, esto indica que el modelo está capturando bien las características de los datos.

Gráfico 5
Ejemplo de PPC para el ingreso medio per cápita



Fuente: Elaboración propia.

Por último, la convergencia de las cadenas también se valida con la prueba de Gelman-Rubin, también conocida como estadística *R-hat*, la cual se utiliza para evaluar que efectivamente todas las cadenas generadas provienen de la misma distribución posterior y por ende no dependen de los valores iniciales. Si se asume que se generaron $m = 4$ cadenas, cada una produciendo $n = 1000$ realizaciones de cada componente θ_j del vector de parámetros θ, entonces esta estadística toma la siguiente forma:

$$\widehat{R}(\theta_j) = \sqrt{\frac{V(\theta_j)}{W(\theta_j)}} \tag{40}$$

En donde

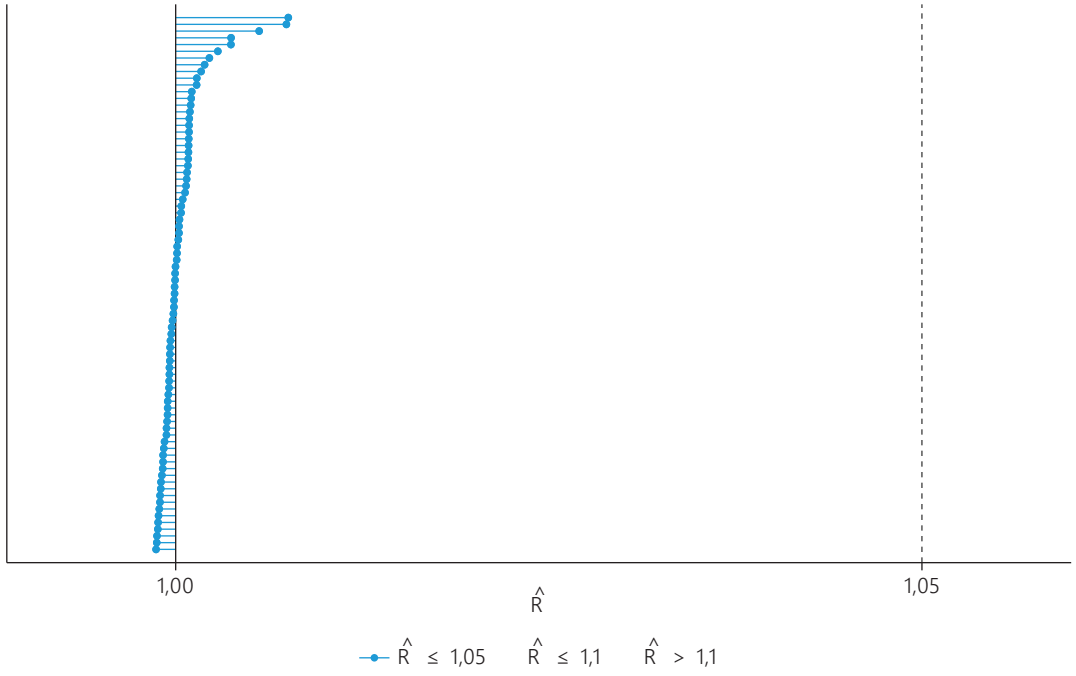
$$V(\theta_j) = (n-1/n)\,W(\theta_j) + (1/n)\,B(\theta_j) \tag{41}$$

$$W(\theta_j) = (1/m)\,Var(\theta_j) \tag{42}$$

$$B(\theta_j) = (n/m-1)\sum_{j=1}^{m}(\theta_j-\bar{\theta})^2 \quad (43)$$

Nótese que V es una combinación lineal de la varianza dentro de las cadenas, denotada como $W(\theta_j)$, y la varianza entre las cadenas, denotada como $B(\theta_j)$. Se espera que el valor de la estadística *R-hat* esté cercano a uno, indicando que las cadenas han convergido y que la simulación de los valores para los parámetros efectivamente viene de la misma distribución, sugiriendo que el modelo es adecuado. Sin embargo, si esta estadística produce valores más grandes que la unidad, es un indicio de que las simulaciones podrían provenir de diferentes partes del espacio de parámetros, y sería una evidencia de que se necesita más muestreo o que algo está mal con el modelo o su configuración. El gráfico 6 muestra los valores de la estadística R-hat para los parámetros del modelo de pobreza en Bolivia. Un resultado similar se obtuvo en todos los países analizados en este documento.

Gráfico 6
Ejemplo de la convergencia de las cadenas según la estadística *R-hat* para el modelo de pobreza



Fuente: Elaboración propia.

V. *Benchmarking* y estimación del error del modelo

Una etapa fundamental al producir desagregaciones de los indicadores de pobreza es la comparación de los resultados obtenidos con los indicadores estimados directamente mediante la encuesta, tanto a nivel nacional, como a nivel de las áreas urbana y rural y por división administrativa mayor. Estas son las estimaciones directas obtenidas por la CEPAL, con base en las encuestas del repositorio BADEHOG, y que se publican para fines de comparabilidad regional.

A. Construcción de los ponderadores

Para que las estimaciones agregadas de los indicadores de interés coincidan con las estimaciones directas es preciso realizar un ajuste en las predicciones a nivel del individuo. Este procedimiento no solo permite obtener cifras consistentes con las estimaciones directas, sino que contribuye a reducir el sesgo producido por una especificación del modelo que no sea totalmente adecuada. Para esto, se hace uso de la calibración multivariada de razones, propuesta en Gutiérrez, Zhang y Rodríguez (2016), en cada una de las simulaciones producidas por los MCMC.

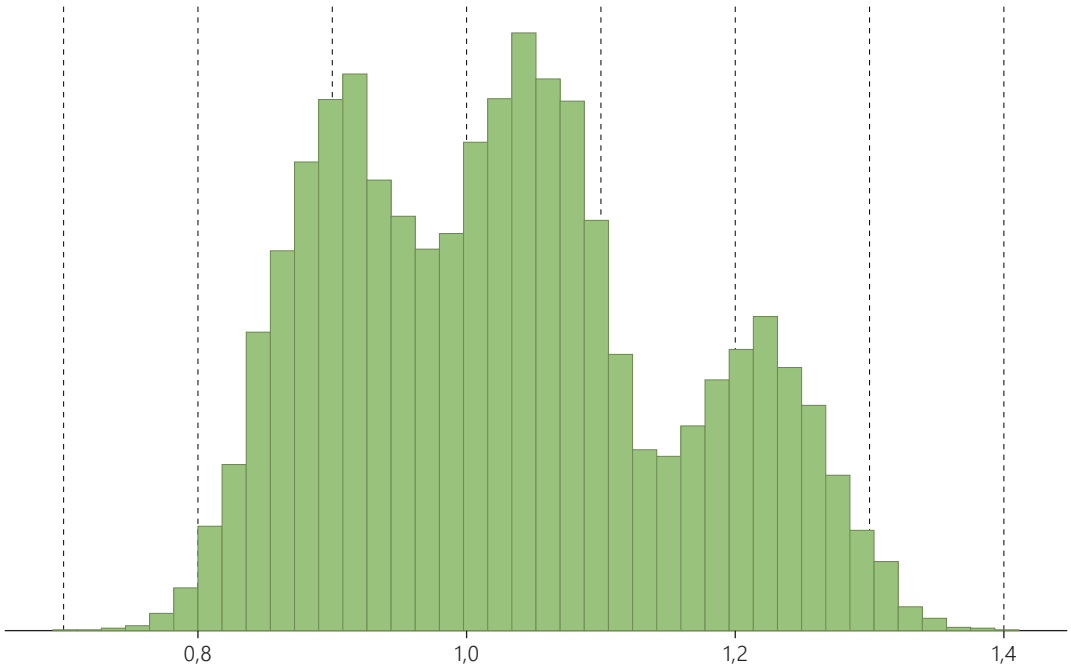
Suponiendo que la estimación directa de la proporción de personas pobres en el subgrupo *s* -ésimo está dada por $\widehat{R}_s$, entonces el algoritmo se enfoca en encontrar un conjunto de pesos *w*, que cumplan con la siguiente restricción:

$$\widehat{R}_s = \frac{\sum_{j=1}^{Q} \boldsymbol{w}_{sj\boldsymbol{x}} \widetilde{y}_{qk}}{\sum_{j=1}^{Q} \boldsymbol{w}_{sj\boldsymbol{x}} N_{sj}} \tag{44}$$

En donde $\widetilde{y}_{qk}$ es una realización de la distribución predictiva posterior $p\left(\widehat{Y} \mid Y\right)$. Por ejemplo, tomando como punto de partida el ingreso (o la tasa de pobreza o de pobreza extrema), para los niveles total rural y total urbano en cada país, el algoritmo de referencia calibra estas estimaciones ($\widehat{R}_{urbano}$ y $\widehat{R}_{rural}$) para todos y cada uno de los pseudo censos generados en el proceso MCMC, y calcula el conjunto de ponderaciones que satisfagan ambas restricciones al mismo tiempo.

Debe notarse que el *benchmarking* se realiza únicamente para las covariables de postestratificación que hayan sido parte de los efectos fijos del modelo. Esto quiere decir que se realiza un filtro en el que se incluyen únicamente las variables que se encuentran tanto en el censo como en la encuesta, validando que el número de categorías de cada una de estas variables coincida en ambos casos. El gráfico 7 muestra el histograma de los ponderadores *w*, los cuales se ajustan a lo esperado, rondando en su mayoría a valores cercanos a uno (1).

Gráfico 7
Bolivia (Estado Plurinacional de): histograma de los pesos de *benchmarking*
(Valores de los ponderadores)



Fuente: Elaboración propia.

B. Validaciones adicionales

Luego de realizar el proceso de *benchmarking* para cada cadena MCMC, se procede a verificar que la calibración realizada fue exitosa por medio de cuadros comparativos para cada una de las covariables de *benchmarking*. Estos cuadros contienen las estimaciones agregadas obtenidas con las estimaciones del modelo original, las estimaciones directas desde la encuesta y las estimaciones finales obtenidas con el proceso de *benchmarking*. El cuadro 8, muestra cómo se construyen estas comparaciones.

Para ejemplificar este proceso en el caso de la estimación de la pobreza extrema, el cuadro 9 presenta los resultados para República Dominicana de los valores obtenidos para la encuesta con los valores expandidos por el factor de expansión, las estimaciones originales producto del modelo ajustado y las obtenidas luego del proceso de *benchmarking*. Es posible observar que los indicadores obtenidos luego del proceso de ajuste de los ponderadores son cercanos a las estimaciones directas para cada División Administrativa Mayor (DAM).

Cuadro 8
Resumen de la comparación de estimaciones antes y después del *benchmarking*

Variable	Estimación directa	Estimaciones con el modelo original	Estimaciones finales después del *benchmarking*
Tamaño	Suma de los factores de expansión de la encuesta	Tamaño original de la agregación de interés	Suma de los tamaños luego de ajustar los ponderadores *w*
Totales	Suma de las variables de interés (ingreso, pobreza y pobreza extrema) ponderada por el factor de expansión de la variable de interés (ingreso (ingcorte), pobreza (lp), pobreza extrema (li)) por el factor de expansión de la encuesta	Suma de las predicciones del modelo ponderadas por el tamaño original de los postestratos	Suma de las predicciones del modelo ponderadas por el tamaño original de los postestratos y corregidas por los ponderadores *w*
Medias	Estimación directa de la media del indicador de interés (ingreso, pobreza , pobreza extrema) ponderadas por el factor de expansión de la encuesta	Estimación de la media de las predicciones del modelo original en los postestratos	Estimación de la media de las predicciones del modelo en los postestratos corregidas por los ponderadores *w*

Fuente: Elaboración propia.

Cuadro 9
República Dominicana: comparativo de indicadores agregados a nivel de DAM para la estimación de la pobreza extrema

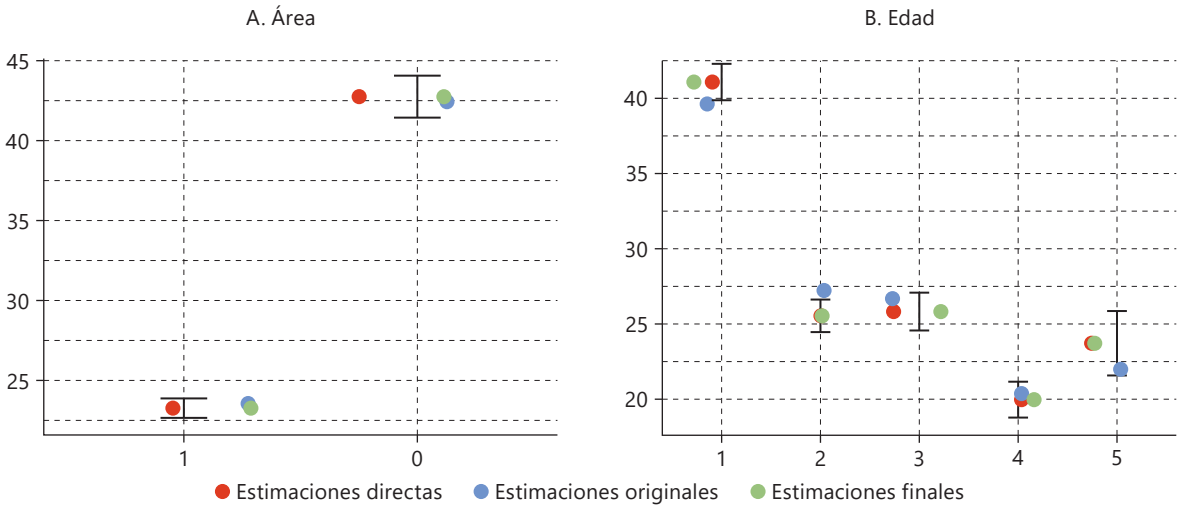
DAM	Tamaños			Totales			Medias *(En porcentajes)*		
	Directa	Modelo	*Benchmarking*	Directa	Modelo	*Benchmarking*	Directa	Modelo	*Benchmarking*
Distrito Nacional	1 199 397	1 199 372	1 199 397	50 959	57 989	50 959	4,20	4,80	4,20
Azúa	253 592	253 541	253 592	10 924	13 953	10 924	4,30	5,50	4,30
Baoruco	109 107	109 058	109 107	17 214	18 317	17 214	15,80	16,80	15,80
Barahona	181 463	181 421	181 463	21 530	23 706	21 530	11,90	13,10	11,90
Dajabón	71 735	71 688	71 735	7 390	6 890	7 390	10,30	9,60	10,30
Duarte	323 208	323 162	323 208	5 774	6 998	5 774	1,80	2,20	1,80
Elías Piña	69 872	69 822	69 872	9 422	12 343	9 422	13,50	17,70	13,50
El Seibo	90 489	90 441	90 489	4 916	5 970	4 916	5,40	6,60	5,40
Espaillat	219 263	219 212	219 263	2 315	3 664	2 315	1,10	1,70	1,10
Independencia	67 988	67 944	67 988	9 975	10 001	9 975	14,70	14,70	14,70
La Altagracia	329 501	329 451	329 501	7 808	9 821	7 808	2,40	3,00	2,40
La Romana	310 065	310 017	310 065	19 017	19 894	19 017	6,10	6,40	6,10
La Vega	417 507	417 459	417 507	15 805	15 407	15 805	3,80	3,70	3,80
María Trinidad Sánchez	138 237	138 185	138 237	1 678	2 370	1 678	1,20	1,70	1,20
Monte Cristi	116 715	116 665	116 715	1 610	931	1 610	1,40	0,80	1,40
Pedernales	26 287	26 242	26 287	3 649	4 266	3 649	13,90	16,20	13,90
Peravia	198 003	197 952	198 003	6 626	9 384	6 626	3,30	4,70	3,30
Puerto Plata	334 769	334 722	334 769	3 737	4 612	3 737	1,10	1,40	1,10
Hermanas Mirabal	73 477	73 427	73 477	266	208	266	0,40	0,30	0,40

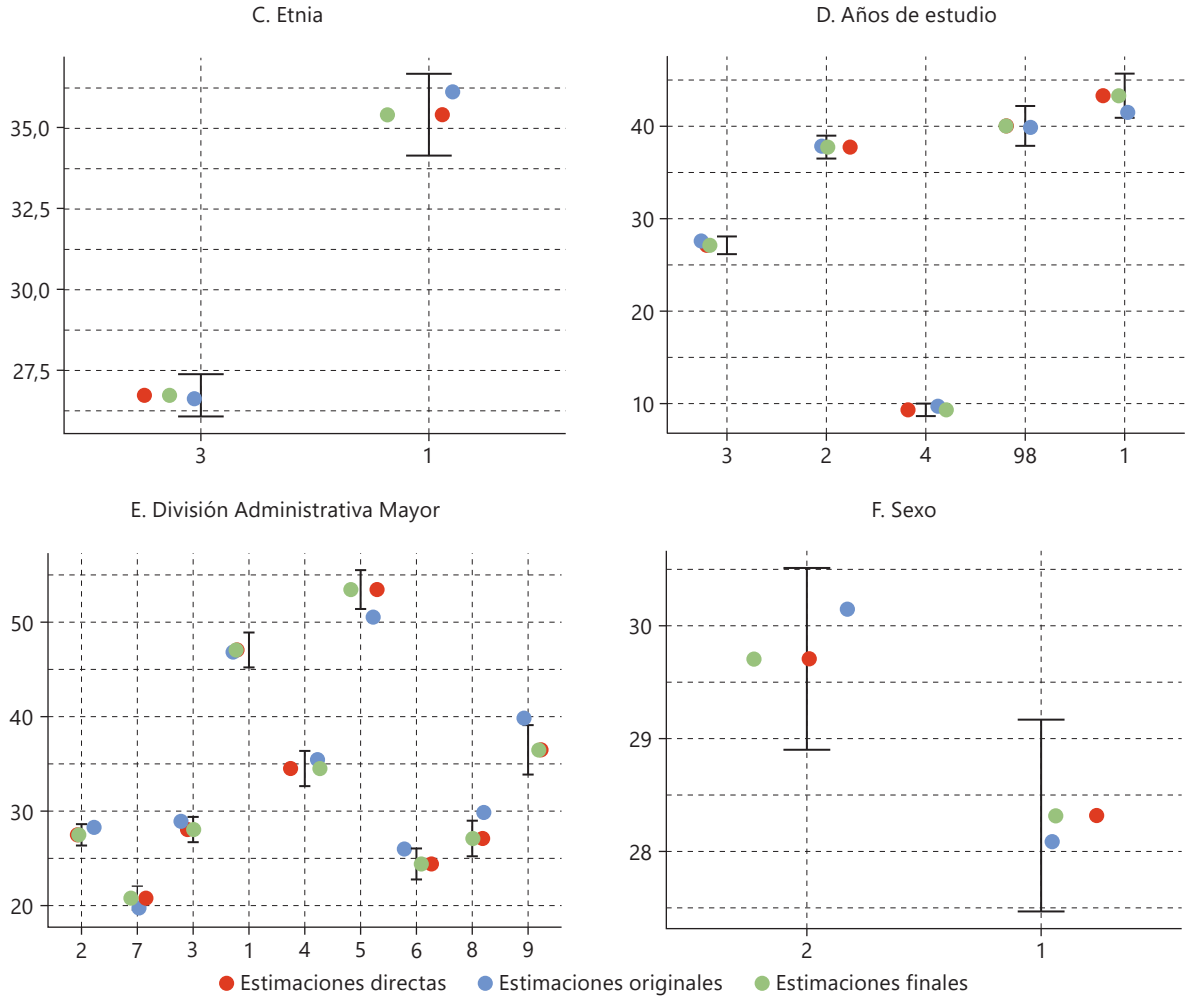
DAM	Tamaños			Totales			Medias *(En porcentajes)*		
	Directa	Modelo	*Benchmarking*	Directa	Modelo	*Benchmarking*	Directa	Modelo	*Benchmarking*
Samaná	110 554	110 511	110 554	2 197	1 763	2 197	2,00	1,60	2,00
San Cristobal	611 766	611 719	611 766	56 658	59 783	56 658	9,30	9,80	9,30
San Juan	213 682	213 629	213 682	15 277	17 809	15 277	7,20	8,30	7,10
San Pedro de Macorís	336 991	336 943	336 991	11 206	15 545	11 206	3,30	4,60	3,30
Sánchez Ramírez	165 233	165 183	165 233	2 660	5 855	2 660	1,60	3,50	1,60
Santiago	1 072 275	1 072 226	1 072 275	42 015	48 124	42 015	3,90	4,50	3,90
Santiago Rodríguez	57 695	57 649	57 695	2 154	2 163	2 154	3,70	3,70	3,70
Valverde	173 421	173 374	173 421	5 979	6 091	5 979	3,40	3,50	3,40
Monseñor Nouel	156 334	156 287	156 334	11 324	10 206	11 324	7,20	6,50	7,20
Monte Plata	177 723	177 676	177 723	31 078	28 612	31 078	17,50	16,10	17,50
Hato Mayor	68 574	68 528	68 574	2 197	1 553	2 197	3,20	2,30	3,20
San José de Ocoa	54 982	54 932	54 982	1 206	1 444	1 206	2,20	2,60	2,20
Santo Domingo	2 806 972	2 806 923	2 806 972	159 101	174 346	159 101	5,70	6,20	5,70

Fuente: Elaboración propia.

Como paso a seguir, se realiza una validación visual sobre el proceso de *benchmarking*, garantizando que el proceso se haya realizado correctamente. Este paso consiste en graficar los valores de las estimaciones derivadas del proceso de *benchmarking* para compararlas con las producidas por la encuesta junto con los intervalos de confianza de la estimación directa. El gráfico 8 ejemplifica esta comparación con datos de Bolivia, para las agregaciones de área (izquierda), edad (derecha), etnia (izquierda), escolaridad (derecha), DAM (izquierda) y sexo (derecha).

Gráfico 8
Bolivia (Estado Plurinacional de): validación visual para la estimación de la pobreza
(En porcentajes)





Fuente: Elaboración propia.

En general, se observa que las estimaciones originales (punto azul) están dentro del intervalo de confianza (líneas sólidas horizontales) de las estimaciones directas (punto rojo). Evidentemente, todas las estimaciones finales (punto verde), luego del ajuste de los ponderadores w, también están dentro del intervalo de confianza del estimador directo puesto que coinciden plenamente con la estimación resultante de la encuesta.

Es deseable que el modelo estadístico produzca estimaciones cercanas al estimador directo antes de hacer el proceso de *benchmarking*. Si el modelo ya está cerca del estimador directo, este ajuste será menos drástico y más controlado. Se considera que el modelo está bien especificado y que captura adecuadamente las características de la muestra si, a nivel agregado, produce estimaciones que estén dentro del intervalo de confianza del estimador directo.

Por último, una consecuencia de esta etapa es que el estimador postestratificado a nivel de las agregaciones definidas por los efectos fijos siempre cumplirá la restricción de igualdad con el estimador directo; es decir $\hat{\rho} = \widehat{R}_s$. Una consecuencia de lo anterior es que el estimador postestratificado a nivel nacional siempre reproducirá el estimador directo de las encuestas a nivel nacional.

En esta instancia debe recalcarse que no en todos los países la DAM puede modelarse como un efecto fijo; Por ejemplo, como se mencionó anteriormente, la Gran Encuesta Integrada de Hogares de Colombia permite realizar estimaciones directas de la tasa de pobreza en solo 24 de los 33 departamentos del país. En

este caso, los modelos para Colombia no pueden utilizar a la DAM como efecto fijo; por ende, el proceso de *benchmarking* no está contemplado para esta desagregación, aunque sí para el área (urbano/rural), el sexo y los grupos de edad. Una situación similar se presenta en Argentina y en Costa Rica en donde la definición de las DAM sigue un proceso *ad hoc*.

Nótese entonces que, las diferencias observadas en los estimadores finales a nivel subnacional podrán atribuirse más claramente a diferencias entre métodos o fuentes, y no a deficiencias en el modelo. Además, aunque no es el objetivo principal de la metodología, en muchos casos se reduce la varianza de las estimaciones finales al aprovechar la información auxiliar adicional resultante de los estimadores directos.

C. Estimación del error cuadrático medio (ECM) basado en réplicas MCMC

Para el cálculo del error ECM de las estimaciones subnacionales, se hace uso de las cadenas MCMC generadas en el proceso de simulación bayesiana. Como se mencionó en las secciones anteriores, los modelos a nivel de unidad ajustados a los datos de la encuesta se usan para predecir a la variable de interés utilizando los microdatos provenientes del censo (para cada país). El algoritmo considerado en esta etapa se describe a continuación:

- Como primer paso, ajustar el modelo SAE para obtener el vector de estimaciones $\widehat{\theta} = (\widehat{\beta}, \widehat{\gamma}, \widehat{\Sigma})$. En esta instancia en cada iteración de las cadenas se generarán diferentes valores para $\widehat{\theta}$.
- Para cada cadena, a partir de las estimaciones $\widehat{\theta}$, se obtiene un conjunto de predicciones para cada unidad en el censo, denominados pseudo censos.
- Para cada pseudo censo se generan estimaciones de los indicadores de interés (ingreso, pobreza y pobreza extrema) en cada desagregación subnacional. Además, en cada pseudo censo se realiza el proceso de *benchmarking* con las estimaciones directas. Suponiendo que se tienen D dominios de interés, y en el caso de la pobreza, entonces se obtendrán D estimaciones $\hat{\rho}_d^{(j)}$, que representa el porcentaje estimado de personas pobres en la j-ésima muestra de la cadena MCM para el d-ésimo domino de interés.
- Suponiendo que en total se obtuvieron J muestras MCMC para cada dominio d, entonces se generaría el siguiente vector de estimaciones $\hat{\rho}_d^{(1)}, \ldots, \hat{\rho}_d^{(j)}, \ldots, \hat{\rho}_d^{(J)}$.
- La estimación puntual bayesiana para $\hat{\rho}_d$ para la proporción de personas pobres en el d-ésimo dominio estará dada por la media del vector de estimaciones; es decir:

$$\hat{\rho}_d = J^{-1} \sum_{j=1}^{J} \hat{\rho}_d^{(j)} \tag{45}$$

- El estimador del error cuadrático medio para $\hat{\rho}_d$, viene dado por:

$$\widehat{MSE}(\hat{\rho}_d) = J^{-1} \sum_{j=1}^{J} \left(\hat{\rho}_d^{(j)} - \hat{\rho}_d \right)^2, d = 1, \ldots, D \tag{46}$$

En un modelo bayesiano, no se obtiene una única estimación puntual del parámetro de interés, sino una distribución posterior que refleja la incertidumbre sobre dicho parámetro. Con el error cuadrático medio estimado es posible definir otras estadísticas que describan la calidad de las estimaciones y predicciones. En particular, el coeficiente de variación y los intervalos de credibilidad. En el primer caso, el coeficiente de variación se define como una medida que permite definir la calidad de las estimaciones y predicciones. Su forma funcional está dada por la siguiente expresión:

$$\widehat{\mathrm{CV}}\left(\hat{\rho}_d\right) = \frac{\sqrt{\widehat{\mathrm{MSE}}\left(\hat{\rho}_d\right)}}{\hat{\rho}_d} * 100 \qquad (47)$$

Tanto el MSE como el CV son medidas de precisión utilizadas para evaluar la calidad del procedimiento. Desde una perspectiva bayesiana, estas medidas están supeditadas a la distribución posterior. Si el modelo es coherente deberá producir estimaciones homogéneas para cada muestra de las cadenas MCMC. Por otro lado los intervalos de credibilidad se utilizan para representar la incertidumbre en las estimaciones de los parámetros. A diferencia de los intervalos de confianza en la inferencia estadística clásica, los intervalos de credibilidad tienen una interpretación probabilística directa: si un parámetro tiene un intervalo de credibilidad del 95%, significa que, dados los valores observados en la encuesta y en las covariables y supeditado al modelo, hay un 95% de probabilidad de que el verdadero valor del parámetro se encuentre dentro de ese intervalo. Cuando se usan procedimientos basados en réplicas MCMC, los intervalos de credibilidad se definen seleccionando el percentil 2.5% y el percentil 97.5% de las muestras ordenadas.

Un valor bajo del ECM indica que las estimaciones generadas por el modelo bayesiano son cercanas al valor verdadero del parámetro, lo que refleja una baja incertidumbre. Como resultado, los intervalos de credibilidad serán más estrechos, lo que sugiere un buen ajuste del modelo. Por el contrario, un ECM elevado denota una mayor variabilidad o incluso sesgo en las estimaciones, lo cual puede indicar que el modelo no está capturando adecuadamente la realidad subyacente.

VI. Resultados

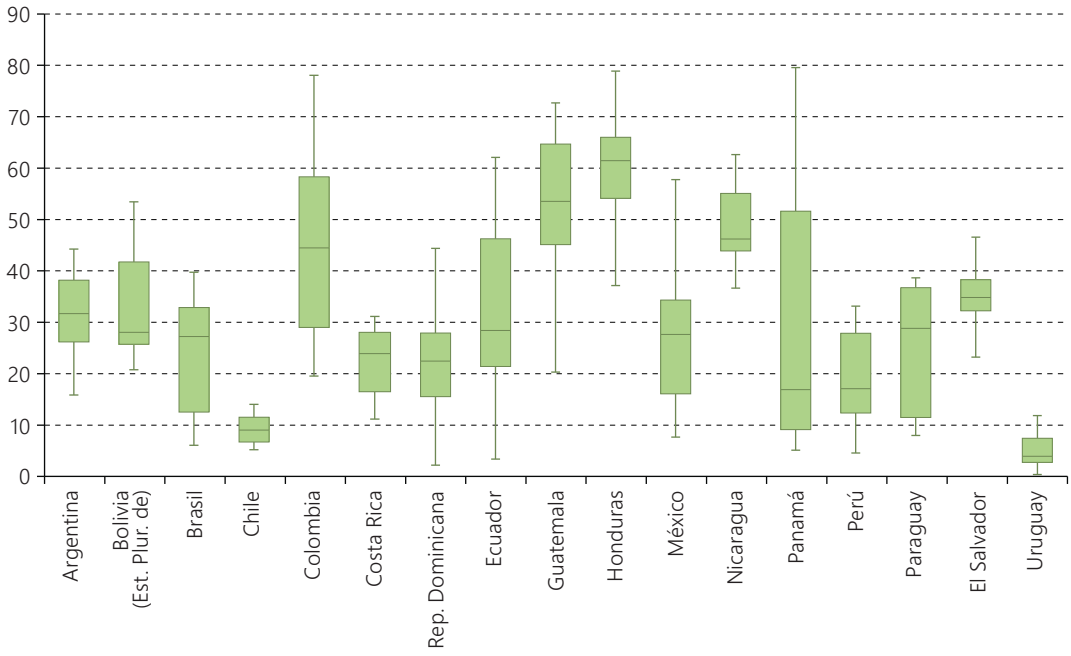
Los resultados indican que la pobreza se presenta de manera desigual en la región según la ubicación geográfica. Incluso en países con bajos niveles de pobreza en el contexto regional, pueden existir áreas geográficas con tasas de pobreza comparables o superiores a las de los países con las mayores incidencias de pobreza. Es posible encontrar divisiones administrativas mayores con tasas de pobreza superiores al 50% en Bolivia, Colombia, Ecuador, Guatemala, Honduras, México, Nicaragua y Panamá, aun cuando en solo tres de ellos la pobreza nacional supera el 33%.

Las amplias discrepancias territoriales también quedan en evidencia en la brecha en las tasas de pobreza entre las divisiones administrativas en cada país. En todos los países, con la excepción de Chile y Uruguay, la diferencia entre el valor más alto y más bajo es de al menos 20 puntos porcentuales. Dicha brecha supera los 50 puntos porcentuales en Colombia, El Salvador, Guatemala, México y Panamá (véase el gráfico 9).

Una de las ventajas de contar con información detallada a niveles geográficos menores es la posibilidad de visualizar los resultados en un mapa e identificar patrones relacionados con la geografía. En el anexo se presentan mapas por país para los indicadores de ingreso medio, pobreza y pobreza extrema, a nivel de división administrativa mayor.

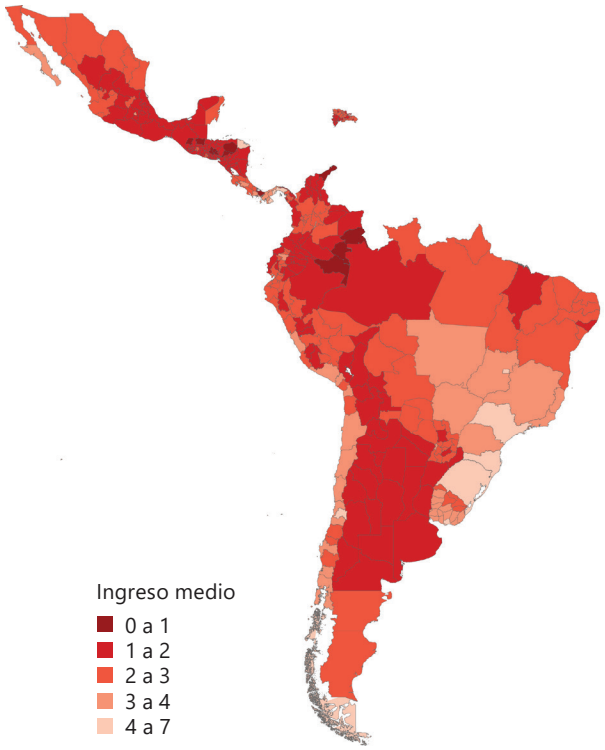
El mapa 1 ofrece una visión general de la distribución territorial del ingreso medio per cápita (expresado como múltiplo de la línea de pobreza) en América Latina. Se observa una gran cantidad de regiones en donde el ingreso medio es inferior a la línea de pobreza, es decir que en promedio las personas no cuentan con ingresos suficientes para cubrir sus necesidades básicas. Países como México y Brasil tienen una amplia heterogeneidad de sus divisiones administrativas mayores en cuanto al ingreso medio, existiendo regiones en donde esta variable es muy baja y otras donde es considerablemente más alta. Por otro lado, países como Costa Rica, Uruguay y Chile presentan una mayor homogeneidad entre territorios. Contar con este tipo de información a nivel desagregado geográficamente es un insumo útil para las políticas públicas orientadas a mejorar la distribución de los recursos en los territorios nacionales.

Gráfico 9
América Latina (17 países): población en situación de pobreza, por división administrativa mayor, 2022[a]
(En porcentajes)



Fuente: Elaboración propia.
[a] Datos de 2022, excepto en Bolivia (2021), Colombia (2021), Guatemala (2014), Honduras (2019) y Nicaragua (2014).

Mapa 1
Latinoamérica (17 países): estimación del ingreso medio per cápita (en líneas de pobreza), por División Administrativa Mayor

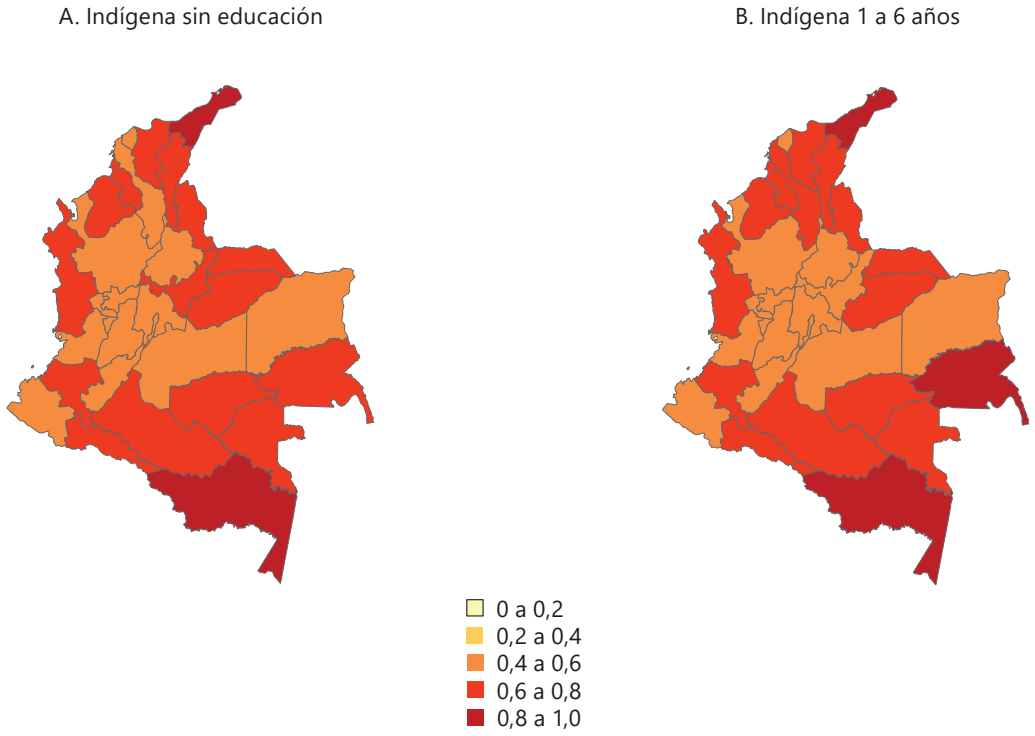


Fuente: Elaboración propia.

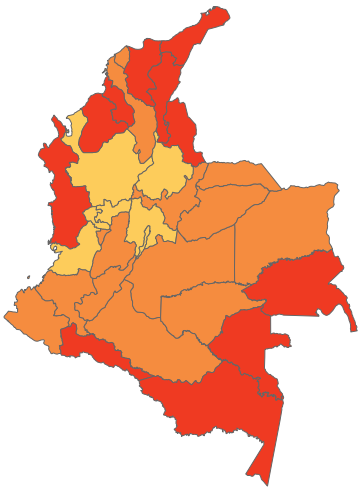
Por otro lado, la posibilidad de obtener estimaciones desagregadas no solamente por el ámbito geográfico, sino según las características individuales, permite hacer mapas comparativos entre diversas combinaciones de las mismas. Como ejemplo, el mapa 2 ilustra la relación entre etnia, nivel educativo y pobreza en Colombia. Se observa que, a medida que disminuyen los años de escolaridad, la tasa de pobreza aumenta en todas las regiones del país. En cuanto a los grupos étnicos, los datos muestran que las comunidades indígenas son las más afectadas, con tasas de pobreza cercanas o superiores al 50% en varios departamentos, mientras que la población afrodescendiente también registra índices considerablemente mayores a los de la población que no es indígena ni afrodescendiente. La visualización conjunta de ambas variables revela cómo las mayores tasas de pobreza se presentan, en todas las áreas, para la combinación de bajos niveles educativos y la pertenencia a pueblos originarios. Este ejemplo también ilustra cómo el método de estimación en áreas pequeñas aplicado permite generar resultados para áreas no cubiertas por las encuestas de hogares, ya que el mapa incluye predicciones para los departamentos no cubiertos por la encuesta de Colombia.

Los resultados para los 17 países de la región a nivel de las distintas combinaciones de sexo, edad, etnia o raza, nivel escolaridad y situación de discapacidad pueden encontrarse en CEPALSTAT, la principal base de datos estadística de la CEPAL (https://statistics.cepal.org/portal/cepalstat/).

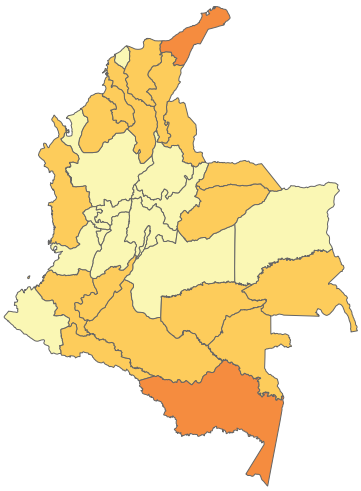
Mapa 2
Colombia: proporción estimada de la población en situación de pobreza, por División Administrativa Mayor, años de estudio y etnia, 2021



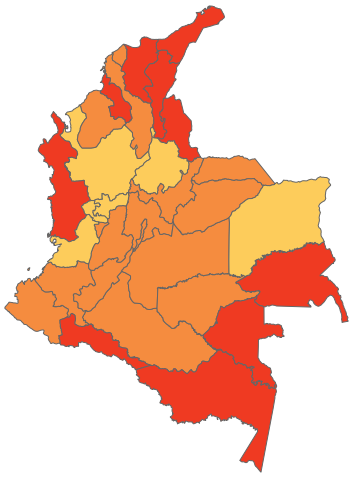
C. Indígena 7 a 12 años



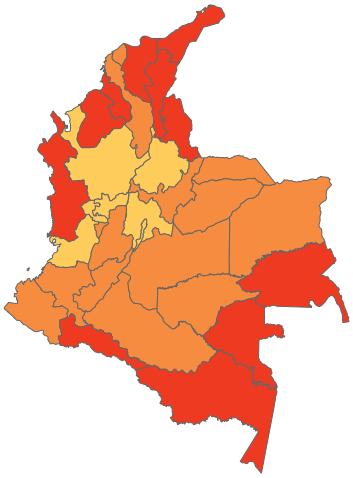
D. Indígena más de 12 años



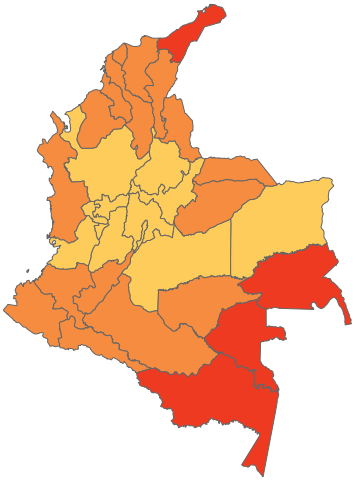
E. Afrodescendiente sin educación



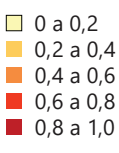
F. Afrodescendiente 1 a 6 años



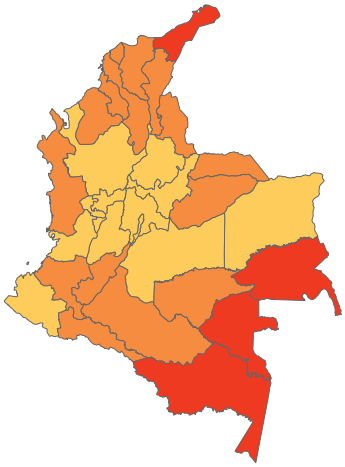
G. Afrodescendiente 7 a 12 años



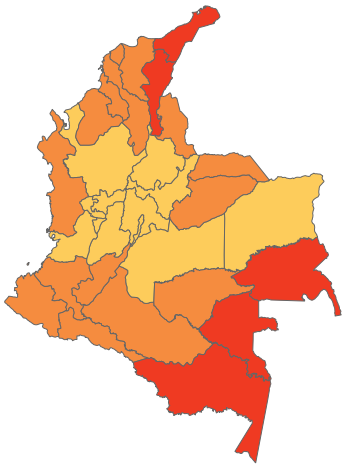
H. Afrodescendiente más de 12 años



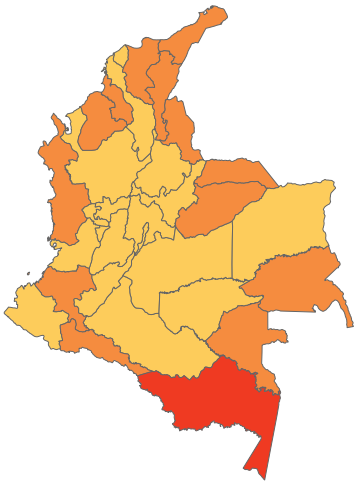
I. Otro sin educación



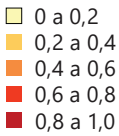
J. Otro 1 a 6 años



K. Otro 7 a 12 años



L. Otro más de 12 años



Fuente: Elaboración propia.

VII. Conclusiones

La adopción de métodos de estimación en áreas pequeñas en América Latina puede contribuir de manera significativa a la producción de estadísticas desagregadas, proporcionando estimaciones más precisas que las que se obtienen directamente de las encuestas, y que son cruciales para el diseño de políticas públicas en la región.

En esta investigación se describe la aplicación de los modelos de regresión multinivel para estimar indicadores de ingresos, pobreza extrema y pobreza a nivel subnacional en varios países de América Latina, región en la que una gran parte de la población no cuenta con los recursos suficientes para atender sus necesidades básicas, y en la que la disponibilidad de datos a escalas geográficas menores es todavía limitada. El método integra tres fuentes de información. En primer lugar, las encuestas de hogares de propósitos múltiples, que proporcionan datos sobre los ingresos, la condición de pobreza y diversas características de los individuos y sus hogares, pero que tienen limitaciones para producir estimaciones representativas en áreas geográficas menores. En segundo lugar, los datos censales agregados a nivel de categorías postestratificación, es decir, las combinaciones de variables de información sobre los individuos o sus hogares a nivel subnacional. La tercera fuente de información son los datos derivados de imágenes satelitales, procesadas al nivel de la división administrativa mayor.

El proceso comienza con el ajuste de un modelo de regresión multinivel usando las fuentes de información descritas anteriormente, seguido de la predicción de celdas de postestratificación utilizando datos censales que han sido actualizados usando técnicas estadísticas SPREE. Un aspecto crucial del proceso es el *benchmarking* de los resultados obtenidos, que permite que las estimaciones del modelo sean consistentes con las estimaciones directas obtenidas de las encuestas en los niveles en que estas son representativas. El proceso pone especial cuidado en que las estimaciones provenientes del modelo original estén dentro del intervalo de confianza de las estimaciones directas, para asegurar su fiabilidad y la ausencia de sesgos significativos.

La metodología propuesta cumple con el propósito de producir estimaciones desagregadas para grupos de población con características específicas, de manera periódica y para un conjunto amplio de países. Para ello, el modelo incluye efectos aleatorios relacionados con características personales y no solo con el ámbito geográfico, para una mayor precisión en las desagregaciones por grupos de población. Además, el modelo contempla la actualización de conteos censales, para minimizar el posible sesgo causado por la

pérdida de vigencia de las estructuras de la población reflejadas en los censos a medida que el tiempo avanza. Finalmente, el uso del paradigma bayesiano en los procesos de inferencia estadística de los modelos y en la obtención de las estimaciones en cada desagregación de interés permite obtener directamente tanto las estimaciones puntuales como las medidas de precisión, sin necesidad de recurrir a procesos computacionales complejos como el bootstrap, común en modelos SAE basados en el enfoque frecuentista.

Finalmente, es importante destacar que existen diversas maneras de aplicar los métodos de estimación en áreas pequeñas, según los objetivos, el contexto y las restricciones de información específicos. El modelo presentado destaca por su alta eficiencia computacional y su capacidad para generar resultados adecuados en la desagregación de pobreza para grupos de población específicos, al nivel de división administrativa mayor. No obstante, este modelo puede no ser idóneo para otros fines, como la desagregación de la pobreza a nivel municipal en un año en el que se dispone tanto de una encuesta de hogares como de un censo de población. Las Oficinas Nacionales de Estadística y otras entidades públicas interesadas en producir estimaciones desagregadas deben evaluar en qué medida los distintos modelos disponibles responden mejor a las necesidades a nivel nacional.

Bibliografía

Agresti, A. (2002), Categorical data analysis. John Wiley & Sons, Inc.

Andreano, M.S., Benedetti, R., Piersimoni, F. y otros (2021), Mapping Poverty of Latin American and Caribbean Countries from Heaven Through Night-Light Satellite Images. *Soc Indic Res* 156, 533–562 https://doi.org/10.1007/s11205-020-02267-1.

Andreoli, F., Mertens, A., Mussini, M. y otros (2022), Understanding trends and drivers of urban poverty in American cities. Empir Econ 63, 1663–1705 https://doi.org/10.1007/s00181-021-02174-5.

Amat, J. (2016), Regresión logística simple y múltiple.

Asian Development Bank. (2020), Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices (0 ed.). Asian Development Bank. https://doi.org/10.22617/TIM200160-2.

Aybar, C. (2020), rgee: An R package for interacting with Google Earth Engine. The Journal of Open Source Software 5(51):2272, DOI:10.21105/joss.02272.

Bell, S. & Robinson, S. (2020), Small Area Income and Poverty Estimates: 2019 Current Population Reports, Washington, D.C.: United States Census Bureau.

Bell, W. R., Chung, H. C., Datta, G. S. & Franco, C. (2019), Measurement Error in Small Area Estimation: Functional vs. Structural vs. Naïve Models. Survey Methodology, Volumen 45, pp. 61-80.

Bell, W. R. & Franco, C. (2015), Borrowing information over time in binomial/logit normal models for small area estimation. Statistics in Transition new series, 16(4), pp. 563-584.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (2007), Discrete multivariate analysis: theory and practice. Springer Science & Business Media.

Centro Latinoamericano y Caribeño de Demografía (CELADE) (2021), REDATAM [Versión 7]. Santiago de Chile: Comisión Económica para América Latina y el Caribe (CEPAL). https://www.cepal.org/es/temas/redatam.

Comisión Económica para América Latina y el Caribe (CEPAL), (2021), Estimaciones subnacionales de la pobreza para América Latina. ISSN: 2788-5828.

_____(2019), Aspectos conceptuales de los censos de población y vivienda: desafíos para la definición de contenidos incluyentes en la ronda 2020. Santiago: Comisión Económica para América Latina y el Caribe (CEPAL), Series Seminarios y Conferencias, N° 94 (LC/TS.2019/67).

_____(2018), Propuesta de indicadores y sus metadatos para el seguimiento regional del Consenso de Montevideo sobre Población y Desarrollo. Conferencia Regional Sobre Población y Desarrollo de América Latina y el Caribe.

_____(2018b), Medición de la pobreza por ingresos—Actualización metodológica y resultados. http://repositorio.cepal.org/bitstream/handle/11362/44314/1/S1800852_es.pdf.

Corral, P., Molina, I., Cojocaru, A., & Segovia, S. (2022), Guidelines to Small Area Estimation for Poverty Mapping. World Bank. https://doi.org/10.1596/37728.

Deming, E. and F. Stephan (1940), On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics 11(4), 427–444.

Deville, J.-C., C.-E. Särndal, and O. Sautory (1993), Generalized raking procedures in survey sampling. Journal of the American statistical Association 88(423), 1013–1020.

_____(1992), Calibration estimators in survey sampling. Journal of the American Statistical Association 87(418), 376–382.

Earth Engine Data Catalog, Catálogo de datos, https://developers.google.com/earth-engine/datasets.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003), Micro-level estimation of poverty and inequality. Econometrica 71(1), 355–364.

Fay, R. E. & Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366), pp. 269-277.

Feres, J. C. & Mancero, X. (2001), Enfoques para la medición de la pobreza: breve revisión de la literatura. Series de la CEPAL: Estudios Estadísticos y Prospectivos, N. 4, Volumen LC/L.1479-P, p. Comisión Económica para América Latina y el Caribe (CEPAL).

Foster, J., Greer, J. & Thorbecke, E. (1984), A class of decomposable poverty measures. Econometrica, 52(3), pp. 761-766.

Gelman, A., Hill, J. & Vehtari A. (2020), Regression and other stories, ISBN 978-1-107-02398-7.

Gelman, A., Hill, J. (2020), Data Analysis Using Regression and Multilevel/Hierarchical Models, ISBN: 9780521867061.

Google Earth Engine, https://earthengine.google.com/.

Google Earth Engine, Education, https://www.google.com/intl/es_ALL/earth/education/tools/google-earth-engine/#:~:text=Google%20Earth%20Engine%20es%20una,de%20sat%C3%A9lite%20de%20nuestro%20planeta.

Google Earth solidario, Introducción a Google Earth Engine https://www.google.com/intl/es_es/earth/outreach/learn/introduction-to-google-earth-engine/

Guadarrama, M., Molina, I. & Rao, J. (2018), Small area estimation of general parameters under complex sampling designs. Computational Statistics and Data Analysis, Volumen 121, pp. 20-40.

Gutiérrez, A., y otros (2023), Modelos de unidad para la generación de mapas de pobreza a nivel subnacional. *Serie Estudios Estadísticos*, N° 105 (LC/TS.2022/191), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).

_____(2020), Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares. Una aplicación a la migración internacional. *Serie Estudios Estadísticos*, N° 101 (LC/TS.2020/52), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).

Gutiérrez, A. (2018), Limitaciones de las encuestas de hogares en la medición de indicadores sociales. Presentación en «Taller regional sobre desagregación de estadísticas sociales mediante metodologías de estimación en áreas pequeñas». (https://www.cepal.org/sites/default/files/courses/files/limitaciones-encuestas-hogare-medicion-indicadores Referencias.

_____(2016), Estrategias de Muestreo, diseño de encuestas y estimación de parámetros. 2 ed. Bogotá: Ediciones de la U.

Gutiérrez, A., Zhang, H. & Rodriguez, N. (2016), "The Performance of Multivariate Calibration on Ratios, Means and Proportions." Revista Colombiana de Estadística 39(2): 281.

INEI (2019), Perú - Encuesta Demográfica y de Salud Familiar 2018. Instituto Nacional de Estadística e Informática (INEI).

Ireland, C. T. and S. Kullback (1968), Contingency tables and with given and marginals. Biometrika. 55(1), 179 – 188.

Isidro, M., S. Haslett, and G. Jones (2016), Extended structure preserving estimation for updating small area estimates of poverty. The Annals of Applied Statistics 10(1), 451–476. DOI: https://doi.org/10.1214/15-AOAS900.

Janicki, R. & Vesper, A. (2017), Benchmarking Techniques for Reconciling Small Area Models at Distinct Geographic Levels. Statistical Methods Applications, Volumen 26, pp. 557-581.

Jiang, J. & Rao, J. S. (2003), Consistent procedures for mixed linear model selection. Sankhyā: The Indian Journal of Statistics, 65(1), pp. 23-42.
Jiang, J., Lahiri S.M, & Chien, H. W. (2002), Jackknifing in The Fay-Herriot Model with an example.
Jiang, J., Rao, J. S., Gu, Z. & Nguyen, T. (2008), Fence methods for mixed model selection. The Annals of Statistics, 36(4), pp. 1669-1692.
Jiming, J. & Lahiri, P. (2006), Mixed model prediction and small area estimation. TEST, 15(1).
Koebe, T., A. Arias-Salazar, N. Rojas-Perilla, T. Schmid, and N. Tzavidis (2020), Intercensal updating using structure-preserving methods and satellite imagery. Journal of the Royal Statistical Society: Series A (Statistics in Society), 1–27.
Luna-Hernández, A. (2016), Multivariate structure preserving estimation for population compositions. Ph. D. thesis, University of Southampton.
Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2020), Small Area Estimation of Non-Monetary Poverty with Geospatial Data. World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-9383.
Marker, D. A. (1999), Organization of small area estimators using a generalized linear regression framework. Journal of Official Statistics 15(1), 1–24.
Mendoza, Daniel L. (2020), "The Relationship between Land Cover and Sociodemographic Factors" Urban Science 4, no. 4: 68. https://doi.org/10.3390/urbansci4040068.
Molina, I. (2019), Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas. Series Estudios Estadísticos, No 97, (LC/TS.2018/82/Rev.1) ed. Santiago de Chile: Comisión Económica para América Latina y el Caribe, (CEPAL).
Molina, A. y otros (2015), Mapa de Pobreza y Desigualdad por consumo Ecuador 2014. 1 ed. Quito-Ecuador: Instituto Nacional de Estadística y Censos y Banco Mundial (INEC-BM).
Molina, I., Nandram, B. & Rao, J. (2014), Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. The Annals of Applied Statistics, 8(2), p. 852–885.
Molina, I. & Rao, J. (2010), Small area estimation of poverty indicators. Canadian Journal of Statistics, 38(3), p. 369–385.
_____(2010), Small area estimation of poverty indicators. The Canadian Journal of Statistics, 28(3), pp. 369-385.
Molina, I., Rao, J. N. K. & Guadarrama, M. (2019), Small Area Estimation Methods for Poverty Mapping: A Selective Review. Statistics and Applications, 17(1), pp. 11-22.
Morales, D., Esteban, M. D., Pérez, A. & Hobza, T. (2021), A course on small area estimation and mixed models. 1 ed. New York, New York: Springer.
Naciones Unidas (2019), Achieving the Full Potential of Household Surveys in the SDG Era. Background document to the 50th session of the UN Statistical Commission. UN Statistical Commission.
Newhouse, D., Merfeld, J. D., Ramakrishnan, A. P., Swartz, T., & Lahiri, P. (2022), Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning. 72.
Noble, A., S. Haslett, and G. Arnold (2002), Small area estimation via generalized linear models. Journal of Official Statistics 18(1), 45–60.
Park, D., Gelman A., Bafumi, J., (2006), State Level Opinions from National Surveys: Poststratification Using Multilevel Logistic Regression.
Pfeffermann, D. (2013), New Important Developments in Small Area Estimation. Statistical Science, 28(1), pp. 40-68.
Pfeffermann, D., Eltinge, J. & Brown, L. (2015), Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. Journal of Survey Statistics and Methodology, Volumen 3, pp. 425-483.
Prasad, N. G. N. & Rao, J. N. K. (1990), The Estimation of the Mean Squared Error of Small-Area Estimators. Journal of the American Statistical Association, 85(409), pp. 163-171.
Pratesi, M. (2016), Analysis of Poverty Data by Small Area Estimation. New York, New York: Wiley.
Purcell, N. J. and L. Kish (1980), Postcensal estimates for local areas (or domains). International Statistical Review 48(1), 3–18.
Rao, J. & Molina, I. (2015), Small Area Estimation. 2 ed. Hoboken, New Jersey: Wiley.
_____(2003), Small area estimation. New York: Wiley.

Singh, M.P., Gambino, J. & Mantel, H.J. (1994), Issues and strategies for small area data. Survey Methodology, Vol 20, pp. 3–22.

Stan Development Team. (2022), Stan: A probabilistic programming language (Version 2.30) [Software]. https://mc-stan.org/.

Suesse, T. F., M.-R. Namazi-Rad, P. Mokhtarian, and J. Barthelemy (2017), Estimating cross-classified population counts of multidimensional tables: an application to regional australia to obtain pseudo-census counts. 18.

Tzavidis, N. y otros (2015), From start to finish: A framework for the production of small area official statistics.

Tzavidis, N., L.-C. Zhang, A. Luna Hernandez, T. Schmid, and N. Rojas-Perilla (2018), From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 181(4), 927–979.

Tutorial Google Cloud, Instalación google cloud, https://cloud.google.com/sdk/docs/install.

United Nations General Assembly (2015), Res 70/1. transforming our world: The 2030 agenda for sustainable development. Technical report, United Nations General Assembly.

World Bank Group (2011), Nepal small area estimation of poverty 2011; volume 1. Washington, D.C. http://documents.worldbank.org/curated/en/959781468290482736/Nepal-small-area-estimation-of-poverty-2011.

Zaloznik, M. (2011), Iterative proportional fitting - theoretical synthesis and practical limitations. Ph. D. thesis, University of Liverpool.

Zhang, L.-C. and R. L. Chambers (2004), Small area estimates for crossclassifications. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 66(2), 479–496. DOI: https://doi.org/10.1111/j.1369- 7412.2004.05266.x.

Anexo A1

Mapas de ingreso medio, pobreza y pobreza extrema, por país

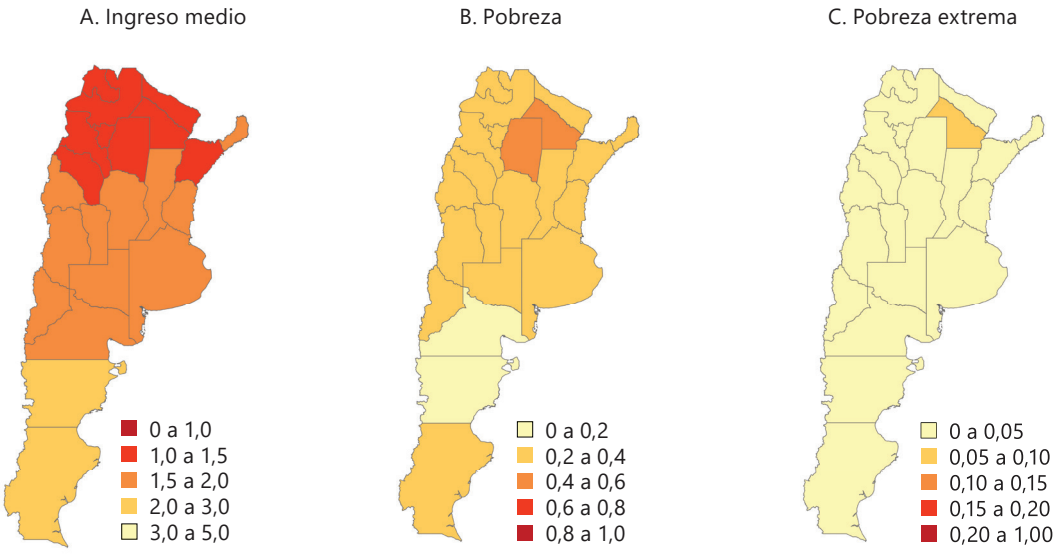
En esta sección se presentan mapas con las estimaciones del ingreso medio (expresado en líneas de pobreza) y las tasas de pobreza extrema y pobreza para diecisiete países de Latinoamérica, a nivel de la división administrativa mayor.

A. Argentina

El mapa A1.1 muestra que en Argentina, el ingreso medio de las divisiones administrativas varía considerablemente, con la Ciudad Autónoma de Buenos Aires presentando un ingreso significativamente más alto (2.73 veces la línea de pobreza), mientras que provincias como Catamarca y Corrientes muestran ingresos medios más bajos, cercanos a 1.4 veces la línea de pobreza. Este patrón refleja disparidades importantes en los niveles de ingresos a lo largo del país.

En cuanto a los niveles de pobreza extrema, la proporción de personas por debajo de la línea de indigencia es baja en todo el país, aunque se observan diferencias entre provincias. Buenos Aires y Corrientes tienen tasas de pobreza moderadas (0.32 y 0.39 bajo la línea de pobreza, respectivamente), mientras que áreas como Córdoba mantienen una proporción más baja de personas en situación de pobreza (0.28). Estos datos evidencian una polarización en la distribución de la pobreza en Argentina, con algunas regiones enfrentando mayores desafíos socioeconómicos.

Mapa A1.1
Argentina 2022: ingreso medio, pobreza y pobreza extrema, por provincia



Fuente: Elaboración propia.

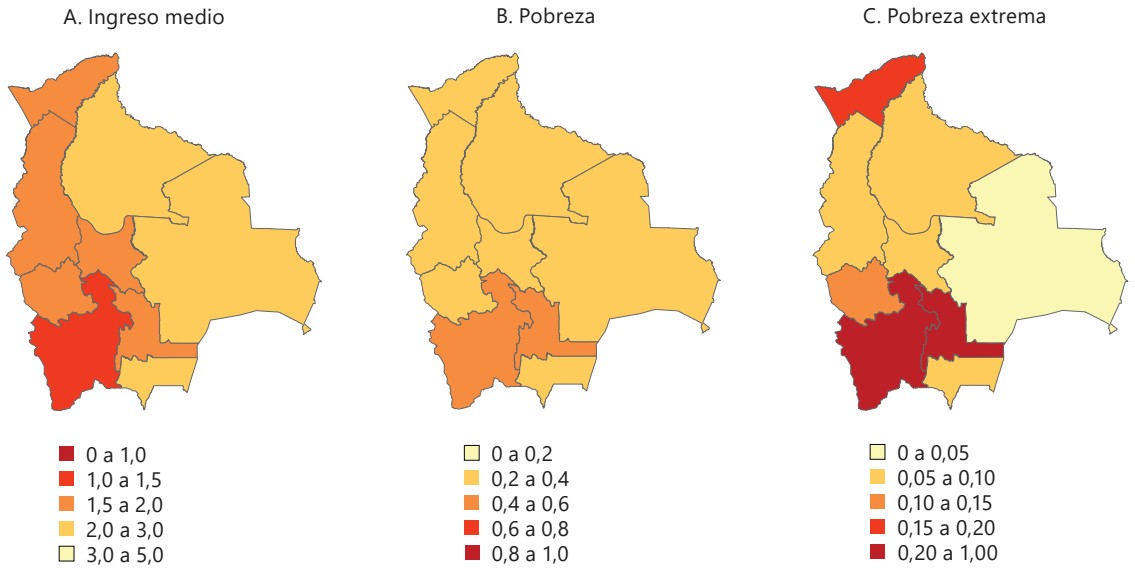
B. Bolivia (Estado Plurinacional de)

En Bolivia, el ingreso medio en las divisiones administrativas mayores (DAM) varía de manera notable. Como se observa en el mapa A1.2, La Paz y Cochabamba destacan con ingresos medios más altos, aproximadamente 1.98 y 1.91 veces la línea de pobreza, respectivamente. En contraste, regiones como

Potosí y Chuquisaca presentan ingresos medios más bajos, con un valor cercano a 1.5 veces la línea de pobreza. Estas diferencias reflejan la distribución desigual de los ingresos en las diversas regiones del país.

En cuanto a la pobreza, se observa una marcada variación entre las DAM. La Paz presenta la menor proporción de personas por debajo de la línea de indigencia (cercana al 8%), mientras que en Chuquisaca y Potosí esta cifra alcanza casi el 30%. En términos de pobreza bajo la línea de pobreza, Potosí se destaca con una proporción muy alta (53.4%), lo que indica una grave situación de vulnerabilidad en esta región. En cambio, La Paz y Cochabamba tienen proporciones significativamente menores (27-28%), reflejando mejores condiciones socioeconómicas en comparación con otras áreas del país.

Mapa A1.2
Bolivia (Estado Plurinacional de) 2021: ingreso medio, pobreza y pobreza extrema, por departamento



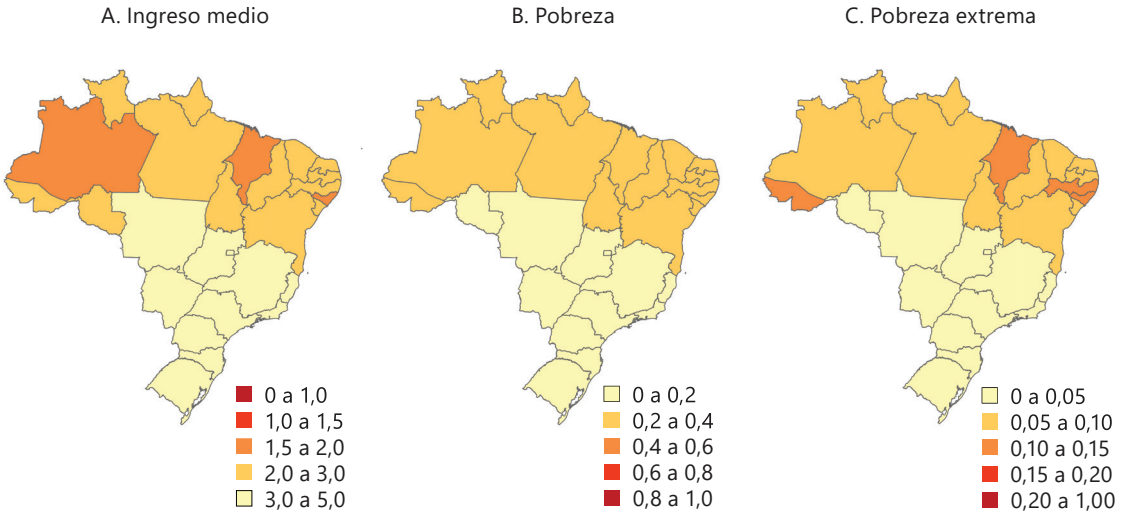
Fuente: Elaboración propia.

C. Brasil

Como se visualiza en el mapa A1.3, los resultados un panorama mixto en Brasil, con algunas regiones alcanzando ingresos más altos y enfrentando menores niveles de pobreza, mientras que otras áreas siguen siendo más vulnerables. Los niveles de ingreso medio varían entre las divisiones administrativas. Rondônia presenta uno de los ingresos más altos, equivalente a 2.86 veces la línea de pobreza, mientras que regiones como Amazonas y Acre tienen ingresos más bajos, en torno a 2 veces la línea de pobreza. Estas diferencias reflejan las disparidades en los niveles de ingresos a lo largo del país.

En cuanto a la pobreza, la proporción de personas por debajo de la línea de indigencia es baja en la mayoría de las regiones, con Rondônia y Pará mostrando tasas relativamente bajas (alrededor del 5-7%). Sin embargo, en Acre, la proporción de personas en indigencia es más alta (12.2%). En cuanto a la pobreza general, Amazonas tiene la mayor proporción (37.8%), lo que destaca las dificultades socioeconómicas de esta región en comparación con otras áreas como Rondônia, donde la pobreza es considerablemente menor (17.8%).

Mapa A1.3
Brasil 2022: ingreso medio, pobreza y pobreza extrema, por estado



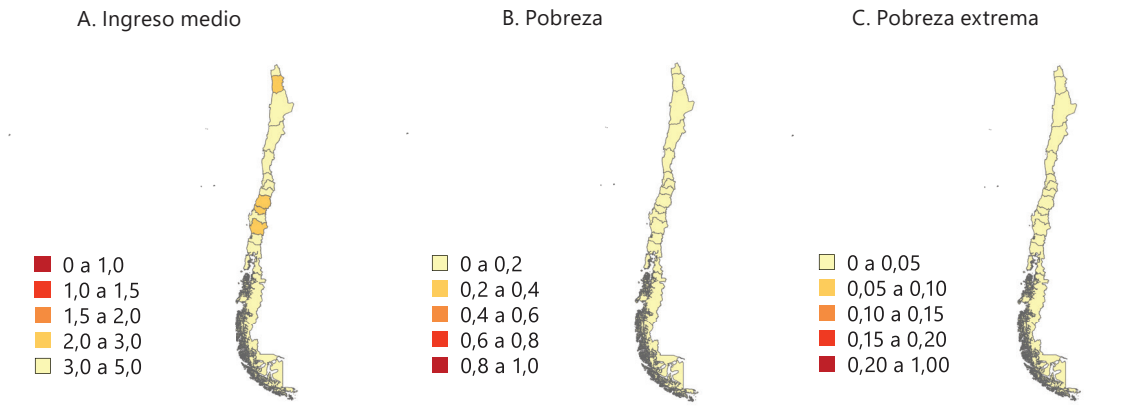
Fuente: Elaboración propia.

D. Chile

En general, como lo muestra el mapa A1.4, Chile muestra una situación socioeconómica sólida, con altos ingresos medios y bajas tasas de pobreza e indigencia en todas sus regiones. Los niveles de ingreso medio son notablemente altos en comparación con otros países de la región. La Región Metropolitana tiene el ingreso medio más alto, pero otras regiones regiones como Aysen y Magallanes destacan con ingresos medios de aproximadamente 4.2 veces la línea de pobreza, lo que indica un alto nivel económico en estas áreas. Todas las regiones presentan ingresos superiores a 2.9 veces la línea de pobreza, lo que sugiere una distribución más equitativa en términos de ingresos.

Respecto a la pobreza, tanto la proporción de personas por debajo de la línea de indigencia como por debajo de la línea de pobreza son bajas. Antofagasta tiene la menor proporción de personas en situación de pobreza (9%), mientras que otras regiones, como Tarapacá y Coquimbo, tienen una proporción algo mayor, pero aún baja en términos generales (14% y 10%, respectivamente). La tasa de indigencia es baja en todas las regiones, con valores que oscilan entre el 2% y el 3%.

Mapa A1.4
Chile 2022: ingreso medio, pobreza y pobreza extrema, por región



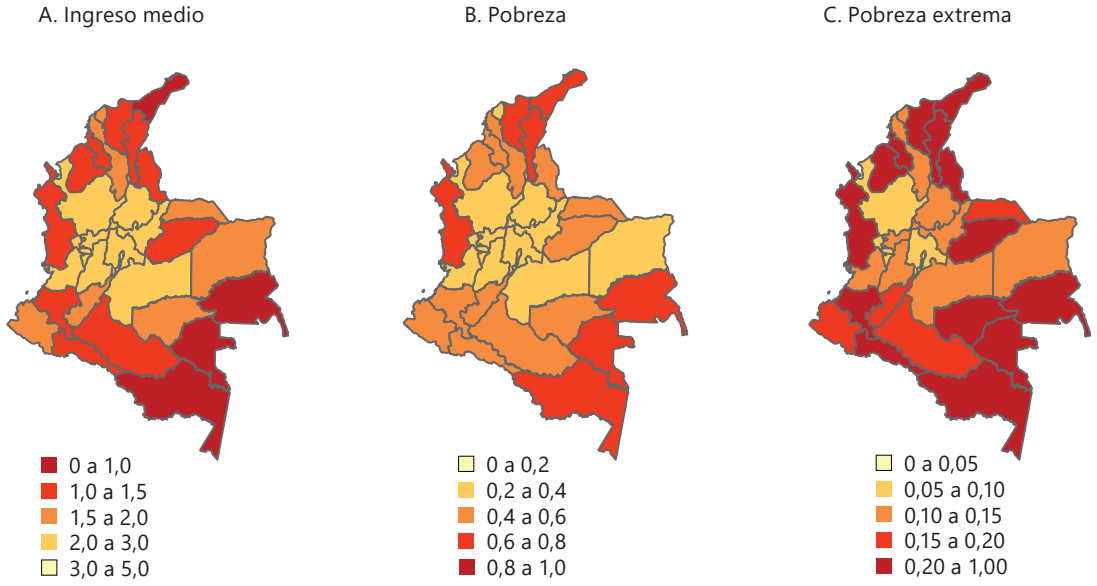
Fuente: Elaboración propia.

E. Colombia

Como se evidencia en el mapa A1.5, se resalta una gran desigualdad en los niveles de ingresos y pobreza entre las diferentes regiones de Colombia, con algunas zonas más afectadas por la pobreza y la indigencia que otras. En este país, los niveles de ingreso medio varían significativamente entre las divisiones administrativas. Bogotá presenta el ingreso más alto, equivalente a 3.29 veces la línea de pobreza, seguido de Antioquia con 2.89 veces. En el otro extremo, Bolívar y Atlántico muestran los ingresos más bajos, con valores cercanos a 1.6 y 1.9 veces la línea de pobreza, respectivamente, lo que indica disparidades económicas entre las regiones.

En términos de pobreza, la proporción de personas por debajo de la línea de indigencia varía considerablemente. Bolívar tiene la mayor proporción de personas en indigencia (13.8%), mientras que Antioquia y Bogotá presentan tasas más bajas, alrededor del 9%. Respecto a la pobreza general, Bolívar y Atlántico son las más afectadas, con casi el 44.5% y 39.3% de sus habitantes por debajo de la línea de pobreza, en comparación con Antioquia y Bogotá, que presentan tasas más moderadas (22.8% y 23.6%).

Mapa A1.5
Colombia 2021: ingreso medio, pobreza y pobreza extrema, por departamento



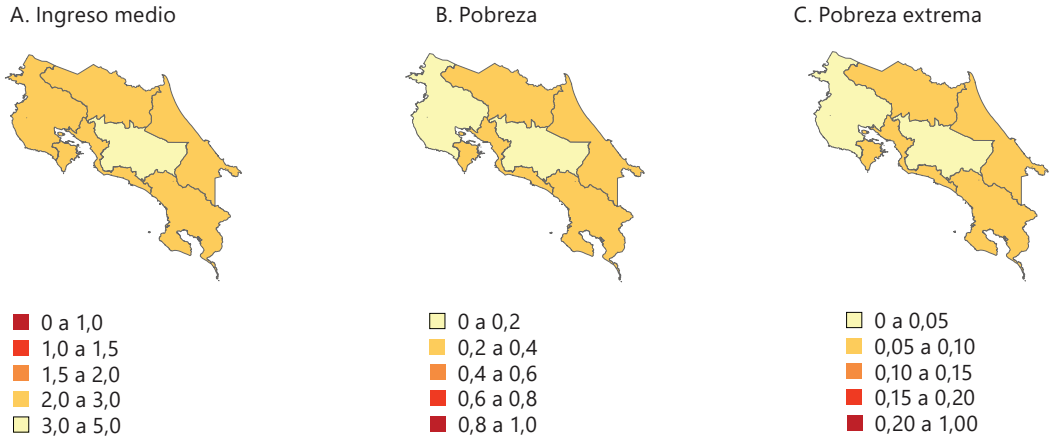
Fuente: Elaboración propia.

F. Costa Rica

En Costa Rica, el ingreso medio varía significativamente entre las regiones. Como lo muestra el mapa A1.6, la región Central presenta el mayor ingreso medio, equivalente a 3.87 veces la línea de pobreza, mientras que la región Huetar Caribe tiene el ingreso más bajo, con 2.22 veces la línea de pobreza. Este contraste refleja desigualdades económicas entre las distintas áreas del país.

En términos de pobreza, las tasas de indigencia son relativamente bajas en todas las regiones, siendo la región Central la menos afectada (1.8%) y la región Huetar Caribe la más vulnerable (6.4%). En cuanto a la pobreza relativa, la región Huetar Caribe también es la más afectada, con un 31.2% de su población bajo esta línea, mientras que la región Central.

Mapa A1.6
Costa Rica 2022: ingreso medio, pobreza y pobreza extrema, por provincia



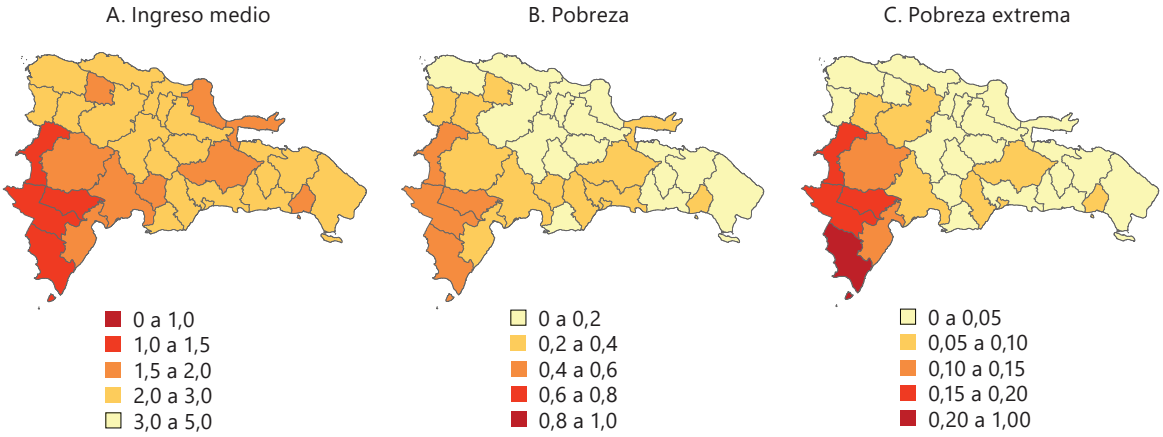
Fuente: Elaboración propia.

G. República Dominicana

Según el mapa A1.7, en la República Dominicana los niveles de ingreso medio varían entre las divisiones administrativas. El Distrito Nacional y Dajabón tienen los ingresos medios más altos, con valores cercanos a 2.48 y 2.37 veces la línea de pobreza, mientras que Bahoruco muestra el ingreso más bajo, con solo 1.39 veces la línea de pobreza, lo que refleja desigualdades económicas considerables en el país.

En términos de pobreza, Bahoruco es la región más afectada, con un 16.4% de su población viviendo en situación de indigencia y un 46.3% bajo la línea de pobreza. En contraste, Dajabón y el Distrito Nacional tienen las tasas más bajas de indigencia (4-5%) y pobreza general (20-21%). Esto muestra una concentración significativa de pobreza en las áreas más rurales, mientras que las zonas urbanas presentan mejores condiciones socioeconómicas.

Mapa A1.7
República Dominicana 2022: ingreso medio, pobreza y pobreza extrema, por departamento



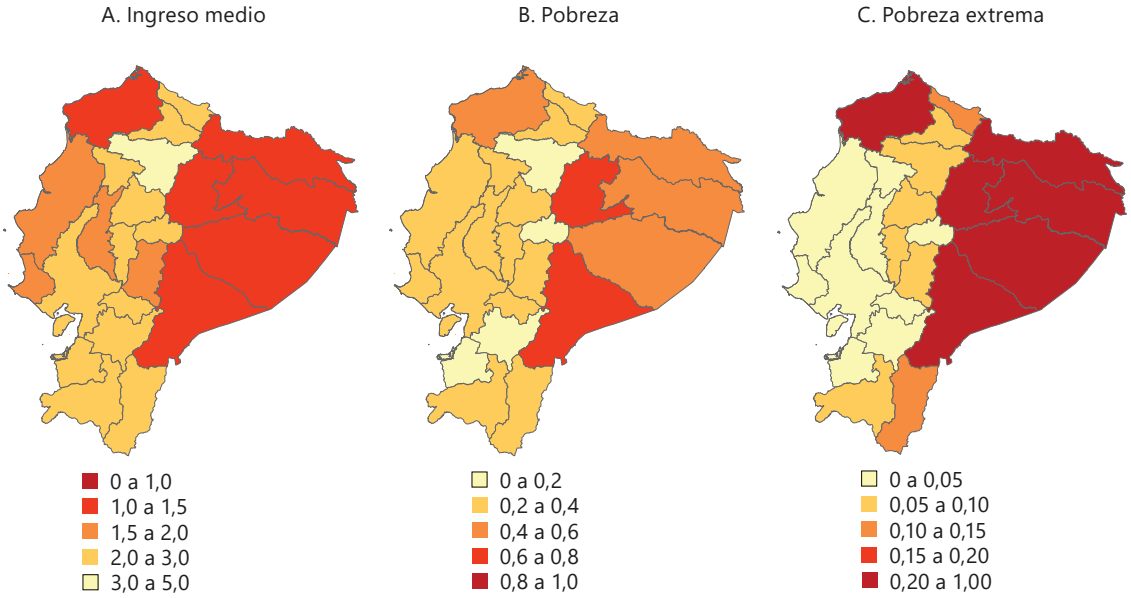
Fuente: Elaboración propia.

H. Ecuador

En Ecuador, los niveles de ingreso medio son heterogéneos entre las divisiones administrativas. Como se observa en el mapa A1.8, Azuay tiene el ingreso medio más alto, equivalente a 2.92 veces la línea de pobreza, mientras que Bolívar muestra un ingreso más bajo, alrededor de 2.08 veces la línea de pobreza. Estas diferencias indican disparidades económicas notables entre las diferentes provincias.

En cuanto a la pobreza, Carchi tiene la mayor proporción de personas en situación de indigencia (13.3%) y también una tasa de pobreza general elevada (34.6%). En contraste, Azuay tiene una baja incidencia tanto de indigencia (3%) como de pobreza (12.4%), lo que sugiere mejores condiciones socioeconómicas en esta provincia. En provincias como Bolívar y Cotopaxi, la situación es intermedia, con tasas de pobreza de aproximadamente 28-29%.

Mapa A1.8
Ecuador 2022: ingreso medio, pobreza y pobreza extrema, por provincia



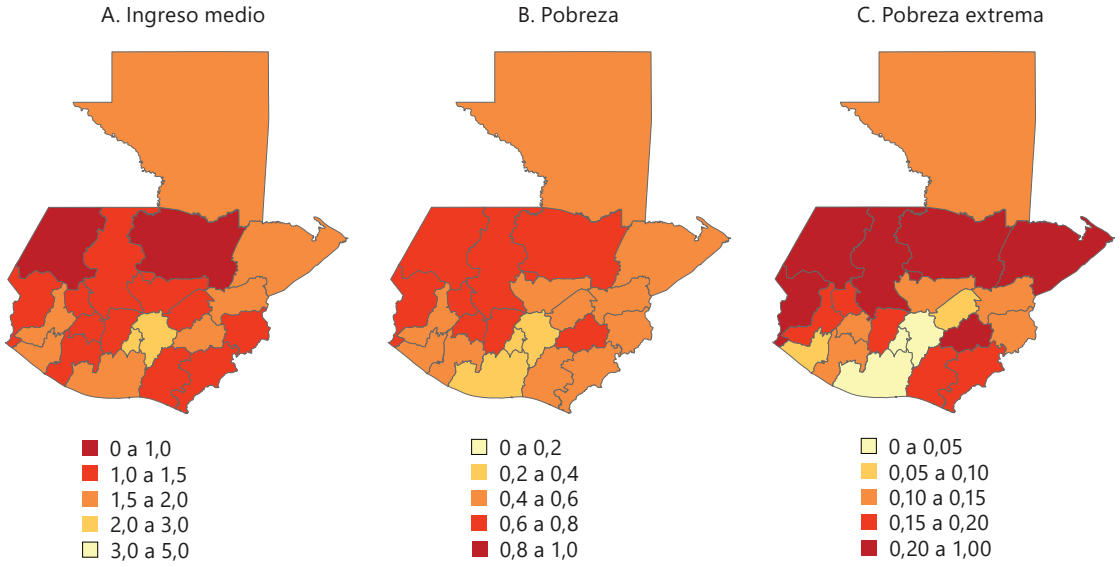
Fuente: Elaboración propia.

I. Guatemala

En Guatemala, el ingreso medio muestra variaciones significativas entre las regiones. En el mapa A1.9 se observa que el departamento de Guatemala tiene el ingreso más alto, con un valor de 2.84 veces la línea de pobreza, mientras que departamentos como Chimaltenango y El Progreso tienen ingresos medios más bajos, alrededor de 1.22 y 1.48 veces la línea de pobreza, respectivamente. Estas disparidades reflejan diferencias económicas marcadas dentro del país.

En cuanto a la pobreza, Chimaltenango es la región más afectada, con un 16.8% de la población en situación de indigencia y un 62.3% bajo la línea de pobreza. En comparación, el departamento de Guatemala presenta tasas mucho más bajas, con solo un 2.6% de indigencia y un 20.3% en situación de pobreza. Departamentos como Escuintla y Sacatepéquez muestran niveles intermedios de pobreza, con alrededor del 32-33% de la población viviendo bajo la línea de pobreza.

Mapa A1.9
Guatemala 2014: ingreso medio, pobreza y pobreza extrema, por departamento



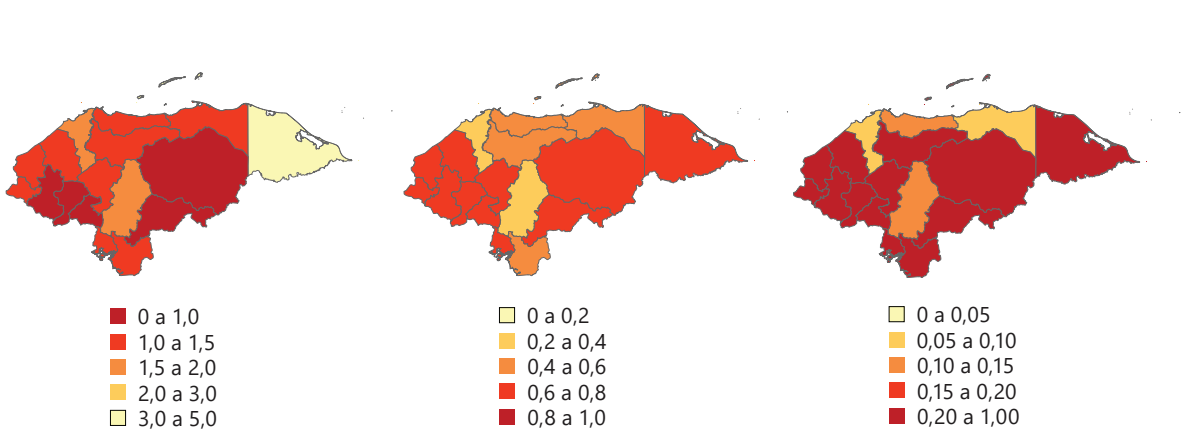
Fuente: Elaboración propia.

J. Honduras

En Honduras, el ingreso medio varía de manera significativa entre los departamentos. El mapa A1.10 evidencia que Cortés presenta el ingreso medio más alto, con un valor de 1.76 veces la línea de pobreza, mientras que Copán tiene el ingreso más bajo, con apenas 1.06 veces la línea de pobreza. Este patrón evidencia la existencia de desigualdades económicas dentro del país.

En cuanto a la pobreza, Copán es la región más afectada, con un 30.7% de su población en situación de indigencia y un 62.7% bajo la línea de pobreza. En comparación, Cortés, aunque también presenta desafíos, tiene una menor proporción de personas en indigencia (6.9%) y pobreza (37.2%). Otros departamentos, como Comayagua y Atlántida, muestran tasas intermedias de pobreza, con entre el 48-62% de la población bajo la línea de pobreza.

Mapa A1.10
Honduras 2019: ingreso medio, pobreza y pobreza extrema, por departamento



Fuente: Elaboración propia.

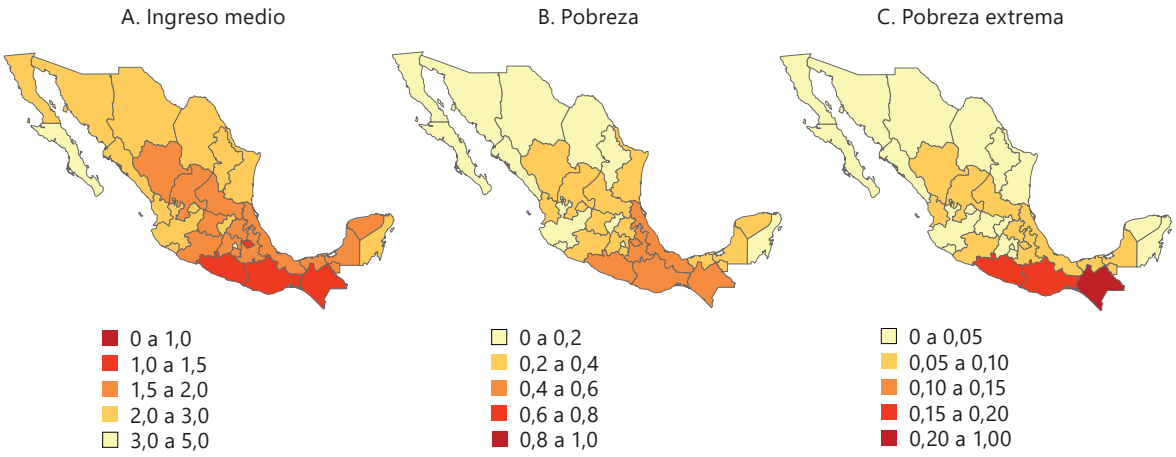
K. México

Como se muestra en el mapa A1.11, México destaca por tener una considerable disparidad entre las regiones del norte y el sur, con algunos estados mostrando niveles de ingresos significativamente más altos y menores índices de pobreza en comparación con los más vulnerables.

En México, los niveles de ingreso medio muestran importantes variaciones entre los estados. Baja California Sur y Baja California destacan con ingresos medios altos, de 3.26 y 2.99 veces la línea de pobreza, respectivamente. En el otro extremo, estados como Campeche tienen un ingreso medio más bajo, de 1.86 veces la línea de pobreza, lo que refleja diferencias considerables en los niveles de ingreso a nivel regional.

En cuanto a la pobreza, Baja California Sur y Baja California también tienen las tasas más bajas de indigencia y pobreza, con apenas un 1% de indigencia y alrededor del 7-8% de pobreza general. Por otro lado, Campeche muestra una proporción más alta de personas en situación de indigencia (7.2%) y pobreza (33.6%). Aguascalientes y Coahuila presentan situaciones intermedias, con tasas de pobreza cercanas al 18-22%, y bajas tasas de indigencia (2.6-2.7%).

Mapa A1.11
México 2022: ingreso medio, pobreza y pobreza extrema, por estado



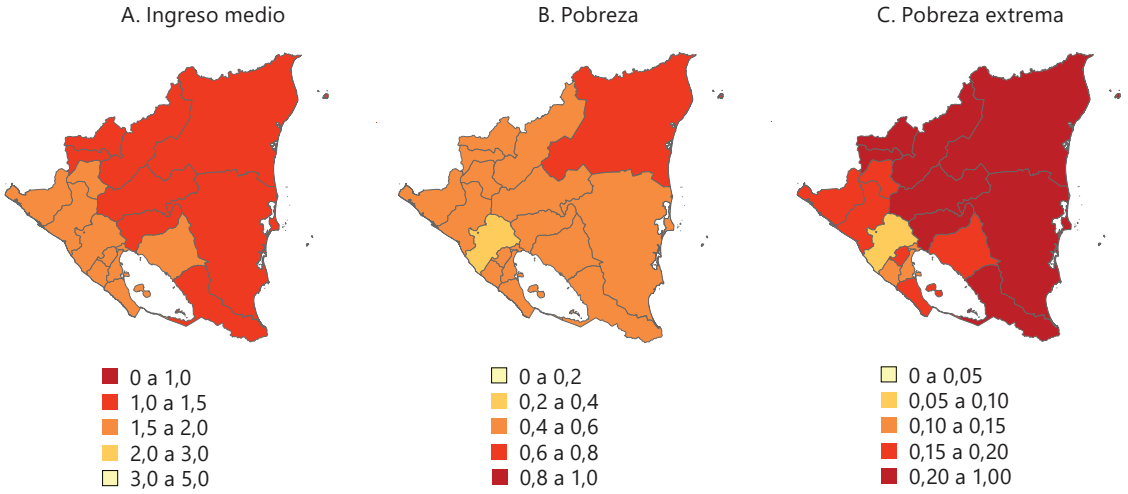
Fuente: Elaboración propia.

L. Nicaragua

En Nicaragua se resalta una alta prevalencia de la pobreza, especialmente en regiones rurales, como Jinotega y Madriz, que enfrentan desafíos socioeconómicos más agudos. Según el mapa A1.12, los niveles de ingreso medio son relativamente bajos en comparación con otros países de la región. Jinotega y Madriz tienen los ingresos medios más bajos, de aproximadamente 1.23 y 1.27 veces la línea de pobreza, mientras que Estelí y Chinandega presentan ingresos medios algo más altos, cercanos a 1.63 y 1.58 veces la línea de pobreza. Esto refleja una economía con limitaciones en varias regiones del país.

En cuanto a la pobreza, Jinotega y Madriz son las áreas más afectadas, con un 28.7% y 29.5% de su población en situación de indigencia, y más del 55% bajo la línea de pobreza. En contraste, Estelí y Chinandega, aunque siguen siendo vulnerables, tienen tasas de pobreza más bajas, con alrededor del 43-45% de la población viviendo bajo la línea de pobreza y un 15% en indigencia.

Mapa A1.12
Nicaragua 2014: ingreso medio, pobreza y pobreza extrema, por departamento



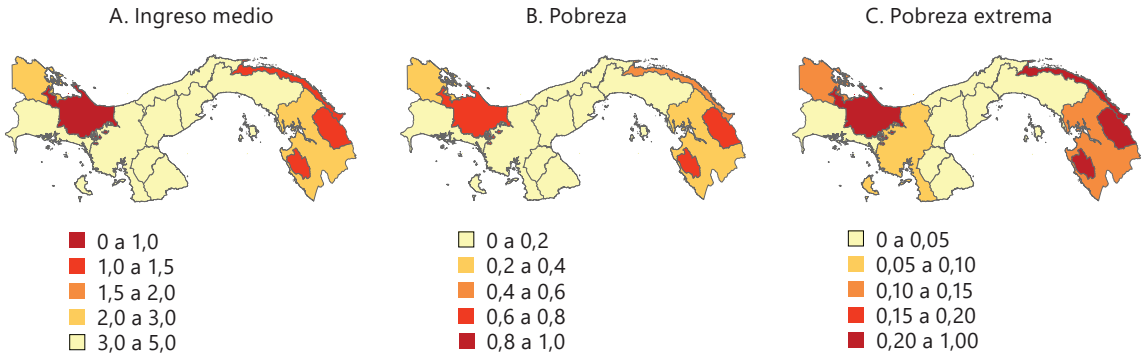
Fuente: Elaboración propia.

M. Panamá

Según el mapa A1.13, aunque Panamá en general presenta ingresos medios altos, algunas áreas rurales, como Darién y Bocas del Toro, aún enfrentan desafíos importantes en términos de pobreza. En Panamá, el ingreso medio varía significativamente entre las provincias. Chiriquí y Colón presentan los ingresos más altos, con valores de 3.67 y 3.61 veces la línea de pobreza, respectivamente, mientras que Darién y Bocas del Toro tienen ingresos más bajos, en torno a 2.66 y 2.83 veces la línea de pobreza. Este rango refleja una distribución de ingresos favorable en algunas regiones, pero con disparidades en otras.

En cuanto a la pobreza, Bocas del Toro y Darién tienen las tasas más altas de pobreza, con el 12.5% y el 10.2% de la población en indigencia, y más del 28% bajo la línea de pobreza. En cambio, Colón y Coclé muestran mejores condiciones socioeconómicas, con tasas de pobreza más bajas (9-15%) y menos del 5% de la población en situación de indigencia.

Mapa A1.13
Panamá 2022: ingreso medio, pobreza y pobreza extrema, por provincia



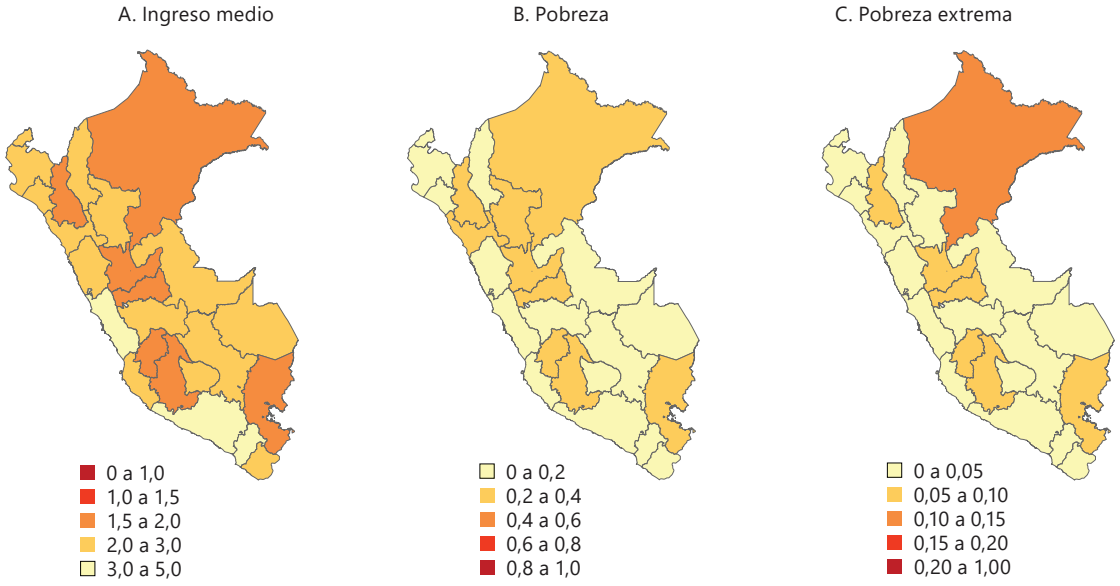
Fuente: Elaboración propia.

N. Perú

En Perú, los niveles de ingreso medio muestran variaciones importantes entre las regiones. Según el mapa A1.14, Arequipa destaca con el ingreso más alto, equivalente a 3.10 veces la línea de pobreza, mientras que Ayacucho presenta el ingreso más bajo, con 1.80 veces la línea de pobreza. Este patrón evidencia disparidades significativas en los niveles de ingreso entre las distintas regiones del país.

En cuanto a la pobreza, Ayacucho tiene una alta proporción de personas en situación de indigencia (7.8%) y un 29.4% de su población vive por debajo de la línea de pobreza. En contraste, regiones como Apurímac y Ancash tienen tasas de indigencia muy bajas (alrededor del 1-2%) y menores proporciones de pobreza general (14-19%), lo que refleja mejores condiciones socioeconómicas en estas áreas. Amazonas tiene una situación intermedia, con un 19.6% de su población bajo la línea de pobreza.

Mapa A1.14
Perú 2022: ingreso medio, pobreza y pobreza extrema, por departamento



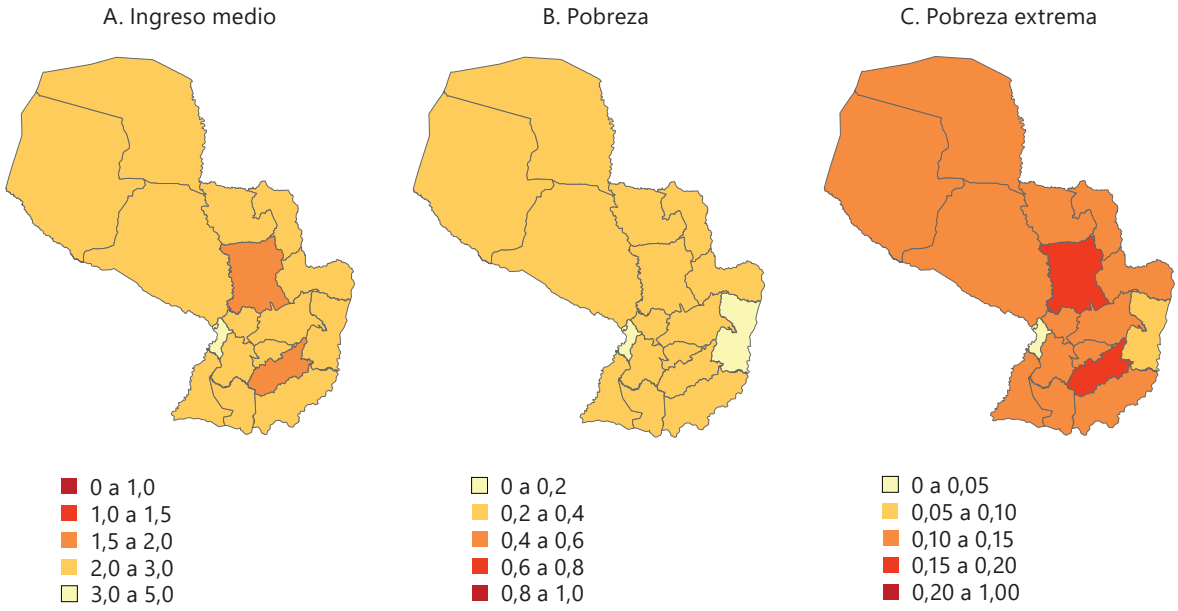
Fuente: Elaboración propia.

O. Paraguay

Según el mapa A1.15, Asunción se destaca con un ingreso medio muy alto, equivalente a 4.54 veces la línea de pobreza, mientras que otras regiones como Caazapá y San Pedro presentan ingresos más bajos, de alrededor de 1.78 y 1.97 veces la línea de pobreza. Estas diferencias reflejan una marcada disparidad económica entre la capital y las áreas rurales del país.

En términos de pobreza, San Pedro y Caazapá son las regiones más afectadas, con un 17-18% de la población en situación de indigencia y más del 37% bajo la línea de pobreza. En contraste, Asunción muestra una tasa muy baja de pobreza general (8%) y una indigencia mínima (1.8%), lo que indica mejores condiciones socioeconómicas en la capital. Otras regiones como Caaguazú e Itapúa se encuentran en una situación intermedia, con tasas de pobreza de aproximadamente 30-34%.

Mapa A1.15
Paraguay 2022: ingreso medio, pobreza y pobreza extrema, por departamento



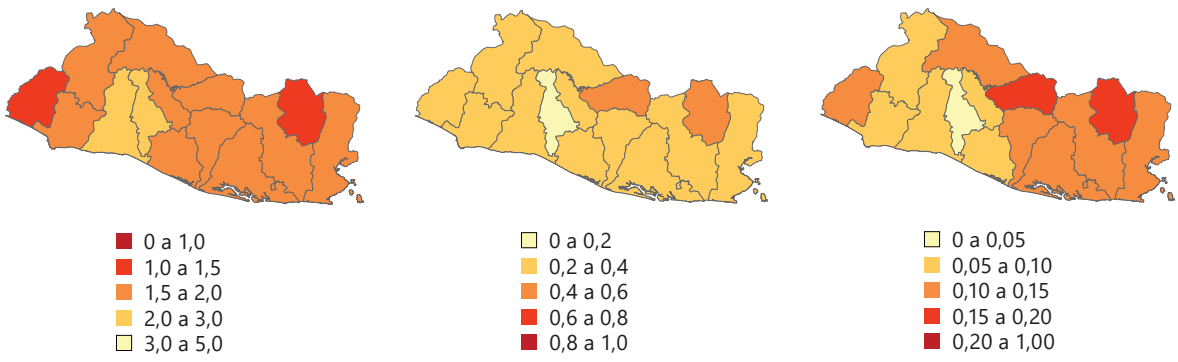
Fuente: Elaboración propia.

P. El Salvador

Según el mapa A1.16, La Libertad se destaca como el departamento con el ingreso más alto, equivalente a 2.05 veces la línea de pobreza, mientras que Ahuachapán tiene el ingreso más bajo, con solo 1.45 veces la línea de pobreza. Otras regiones como Santa Ana y Sonsonate muestran ingresos intermedios, cercanos a 1.6 veces la línea de pobreza.

En términos de pobreza, Ahuachapán es la región más afectada, con un 13.7% de la población en situación de indigencia y un 39.3% bajo la línea de pobreza. En contraste, La Libertad tiene mejores condiciones, con una tasa de pobreza general del 23.2% y solo un 5.7% de su población en situación de indigencia. Las otras regiones presentan niveles intermedios de pobreza, con tasas de pobreza general entre el 33% y el 35%.

Mapa A1.16
El Salvador 2022: ingreso medio, pobreza y pobreza extrema, por departamento



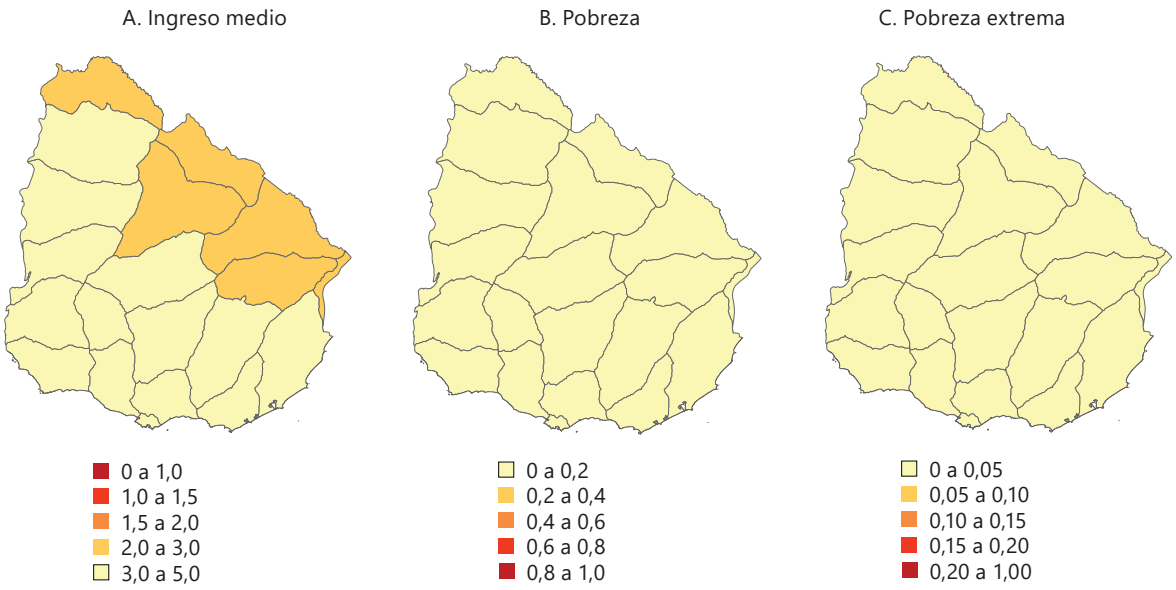
Fuente: Elaboración propia.

Q. Uruguay

Según el mapa A1.17, Uruguay goza de una situación económica sólida, con bajos niveles de pobreza y un ingreso medio elevado, aunque las regiones rurales presentan desafíos menores en comparación con la capital. En este país, los niveles de ingreso medio son notablemente altos en comparación con otros países de la región. Montevideo, la capital, tiene el ingreso más alto, equivalente a 5.36 veces la línea de pobreza, mientras que otras regiones como Canelones y Colonia también muestran ingresos elevados, alrededor de 4 veces la línea de pobreza. Las regiones más rurales, como Artigas y Cerro Largo, tienen ingresos medios más bajos, cercanos a 2.33 y 2.75 veces la línea de pobreza, respectivamente.

En términos de pobreza, Uruguay muestra tasas muy bajas en todas sus regiones. Montevideo tiene una tasa de pobreza del 3.3% y una tasa mínima de indigencia (0.27%). Regiones como Colonia y Canelones mantienen tasas de pobreza menores al 5%, con casi inexistente indigencia. Incluso en las zonas más rurales, como Cerro Largo y Artigas, la pobreza es moderada, con tasas de alrededor del 7-12%.

Mapa A1.17
Uruguay 2022: ingreso medio, pobreza y pobreza extrema, por departamento



Fuente: Elaboración propia

NAÇÕES UNIDAS

CEPAL

Serie Estudios Estadísticos

Números publicados

Un listado completo así como los archivos pdf están disponibles en www.cepal.org/publicaciones

109. Estimación en áreas pequeñas de los indicadores de pobreza en América Latina: una aplicación basada en modelos de regresión multinivel con posestratificación, Andrés Gutiérrez, Xavier Mancero y Stalyn Guerrero (LC/TS.2024/137), 2025.
108. Servicios de intermediación financiera medidos indirectamente: medición en un contexto de tasas de interés subsidiadas y existencia de encajes bancarios, Federico Dorin, Lourdes Erro, Salvador Marconi y Juan Carlos Propatto (LC/TS.2024/17), 2024.
107. Índice de mala calidad del empleo: una exploración de la última década en América Latina, Mauricio Apablaza, Pablo Villatoro, Pablo González, Kirsten Sehnbruch y Xavier Mancero (LC/TS.2023/199), 2024.
106. Efectos de diseño para indicadores sociales en América Latina: función generalizada de varianza para estimadores directos provenientes de encuestas de hogares, Andrés Gutiérrez y Giovany Babativa-Márquez (LC/TS.2023/95), 2023.
105. Modelos de unidad para la generación de mapas de pobreza a nivel subnacional, Andrés Gutiérrez, Xavier Mancero, Gabriel Nieto, Felipe Molina y Diego Lemus (LC/TS.2022/191), 2023.
104. Cambio de año de referencia de los agregados regionales anuales en las cuentas nacionales, C. de Camino y otros (LC/TS. 2022/158), 2022.
103. Predicciones agregadas de pobreza con información a escala micro y macro: evaluación, diagnóstico y propuestas, Walter Sosa Escudero y Magdalena Cornejo (LC/TS.2022/95), 2022.
102. La medición de la discriminación en base al autorreporte: estado de situación y desafíos, Pablo Villatoro (LC/TS. 2021/87), 2021.
101. Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional, Andrés Gutiérrez, Xavier Mancero, Álvaro Fuentes, Felipe López y Felipe Molina (LC/TS.2020/52), 2020.
100. Desafíos en el diseño de medidas de pobreza multidimensional, María Emma Santos (LC/TS.2019/5), 2019.

ESTUDIOS ESTADÍSTICOS

Números publicados:

109 Estimación en áreas pequeñas de los indicadores de pobreza en América Latina
Una aplicación basada en modelos de regresión multinivel con posestratificación
Andrés Gutiérrez, Xavier Mancero y Stalyn Guerrero

108 Servicios de intermediación financiera medidos indirectamente
Medición en un contexto de tasas de interés subsidiadas y existencia de encajes bancarios
Federico Dorin, Lourdes Erro, Salvador Marconi y Juan Carlos Propatto

107 Índice de mala calidad del empleo
Una exploración de la última década en América Latina
Mauricio Apablaza, Pablo Villatoro, Pablo González, Kirsten Sehnbruch y Xavier Mancero

106 Efectos de diseño para indicadores sociales en América Latina
Función generalizada de varianza para estimadores directos provenientes de encuestas de hogares
Andrés Gutiérrez y Giovany Babativa-Márquez

CEPAL
Comisión Económica para América Latina y el Caribe (CEPAL)
Economic Commission for Latin America and the Caribbean (ECLAC)
www.cepal.org

LC/TS.2024/137