ISSN 1680-8789

SERIE

ESTUDIOS ESTADÍSTICOS 105

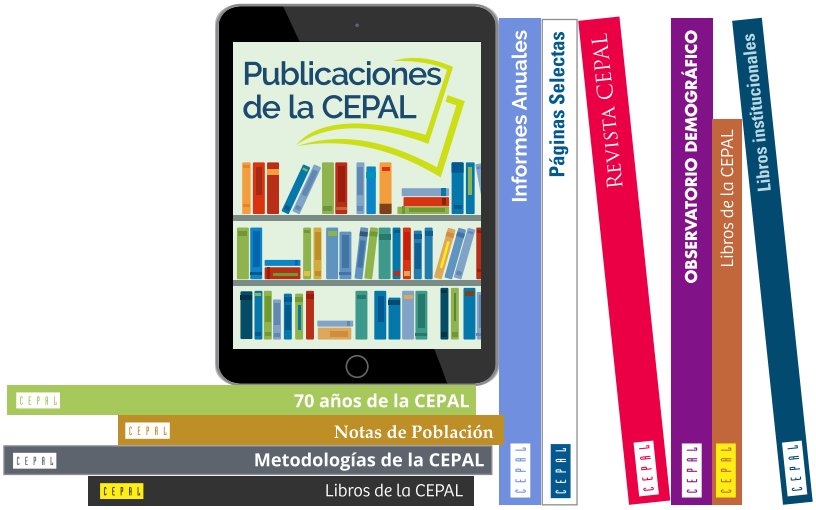
Modelos de unidad para la generación de mapas de pobreza a nivel subnacional

Andrés Gutiérrez
Xavier Mancero
Gabriel Nieto
Felipe Molina
Diego Lemus

NACIONES UNIDAS

CEPAL

Gracias por su interés en esta publicación de la 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

www.cepal.org/es/publications

www.instagram.com/publicacionesdelacepal

www.facebook.com/publicacionesdelacepal

www.issuu.com/publicacionescepal/stacks

www.cepal.org/es/publicaciones/apps

NACIONES UNIDAS

CEPAL

SERIE

ESTUDIOS ESTADÍSTICOS 105

Modelos de unidad para la generación de mapas de pobreza a nivel subnacional

Andrés Gutiérrez
Xavier Mancero
Gabriel Nieto
Felipe Molina
Diego Lemus

NACIONES UNIDAS

CEPAL

Este documento fue preparado por Andrés Gutiérrez, Experto Regional en Estadísticas Sociales, Xavier Mancero, Jefe de la Unidad de Estadísticas Sociales de la División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL), y Gabriel Nieto, Felipe Molina y Diego Lemus, Consultores de dicha División, en el marco de las actividades del decimotercer tramo de la Cuenta de las Naciones Unidas para el Desarrollo.

Las opiniones expresadas en este documento, que no ha sido sometido a revisión editorial, son de exclusiva responsabilidad de los autores y pueden no coincidir con las de la Organización o las de los países que representa.

Los límites y los nombres que figuran en los mapas de esta publicación no implican su apoyo o aceptación oficial por las Naciones Unidas.

Publicación de las Naciones Unidas
ISSN: 1680-8789 (versión electrónica)
ISSN: 1994-7364 (versión impresa)
LC/TS.2022/191
Distribución: L
Copyright © Naciones Unidas, 2023
Todos los derechos reservados
Impreso en Naciones Unidas, Santiago
S.22-00932

Esta publicación debe citarse como: A. Gutiérrez y otros, "Modelos de unidad para la generación de mapas de pobreza a nivel subnacional", *serie Estudios Estadísticos*, Nº 105 (LC/TS.2022/191), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL), 2023.

La autorización para reproducir total o parcialmente esta obra debe solicitarse a la Comisión Económica para América Latina y el Caribe (CEPAL), División de Documentos y Publicaciones, publicaciones.cepal@un.org. Los Estados Miembros de las Naciones Unidas y sus instituciones gubernamentales pueden reproducir esta obra sin autorización previa. Solo se les solicita que mencionen la fuente e informen a la CEPAL de tal reproducción.

Índice

Resumen.....7

Introducción.....9

I. Fuentes de información.....21
A. Encuestas de hogares.....21
B. Censos de población y vivienda.....22
C. Consolidación de las bases de datos..... 24

II. Elementos teóricos de la metodología Pseudo – EBP.....27
A. Covariables utilizadas en las estimaciones.....27
B. Transformación del ingreso.....29
C. Estimación de los coeficientes de regresión.....31
D. Selección de covariables.....35
E. Estimación de la varianza de los efectos aleatorios.....39

III. Estimación del error en el modelo EBP..... 41
A. El estimador de Prasad y Rao.....41
B. Métodos no paramétricos en la estimación del ECM.....43
C. Intervalos de confianza.....45

IV. Criterios de selección y validación de los supuestos del modelo.....49
A. Criterios de selección del modelo..... 50
B. Validación de los supuestos del modelo y medidas de influencia.....53
C. Puntos de balanceo, datos atípicos y/o influyentes.....61

D. Benchmarking con las cifras publicadas....64
E. Comparación con estimaciones directas....65

V. Resultados y mapas de pobreza....70
A. Resultados para Chile....70
B. Resultados para Colombia....73
C. Resultados para Perú....76
D. Mapas de pobreza con información complementaria....78

VI. Conclusiones....81

Bibliografía....83

Anexos....91
Anexo 1....92
Anexo 2....100
Anexo 3....102

Serie Estudios Estadísticos: números publicados....113

Cuadros

Cuadro 1 Listado de variables iniciales consideradas en el proceso de estimación....28
Cuadro 2 Constante óptima de la transformación Log-Shift en los países....31
Cuadro 3 Transformación Log-Shift en los países....32
Cuadro 4 Líneas de pobreza y pobreza extrema empleadas en el estudio....35
Cuadro 5 Resumen métodos de regularización Chile, Colombia y Perú....38
Cuadro 6 Componentes de varianza estimados por país....40
Cuadro 7 Estimaciones pobreza, MSE, CV e IC para Chile, Colombia y Perú....45
Cuadro 8 Valores para diferentes criterios de selección de modelos en los países....53
Cuadro 9 Pruebas de normalidad sobre el efecto aleatorio de los modelos ajustados en los países....55
Cuadro 10 Pruebas de normalidad sobre el término de error en los modelos ajustados en los países....55
Cuadro 11 Pruebas de homocedasticidad sobre los residuos de los modelos ajustados en los países....56
Cuadro 12 Pruebas de homocedasticidad sobre los residuos de los modelos ajustados en los países....57
Cuadro 13 Valores de referencia para indicadores de calidad de estimación....66
Cuadro 14 Porcentaje de comunas por nivel de pobreza....71
Cuadro 15 Chile: 20 comunas con mayores niveles de pobreza....71
Cuadro 16 Chile: 20 comunas con mayores niveles de pobreza extrema....73
Cuadro 17 Porcentaje de municipios por nivel de pobreza....74
Cuadro 18 Colombia: 20 municipios con mayores niveles de pobreza....74
Cuadro 19 Pobreza extrema estimada por municipio, Colombia....75
Cuadro 20 Pobreza estimada por provincia, Perú....76
Cuadro 21 Pobreza extrema estimada por municipio, Perú....77
Cuadro A1 Correspondencia entre el nivel educativo y el número de años de educación....92
Cuadro A2 Descripción de los grupos generados a partir de la Clasificación Internacional Uniforme de Ocupaciones....93
Cuadro A3 Descripción de los grupos generados a partir de la Clasificación Industrial Internacional Uniforme....93

Cuadro A4 Homologación de categorías censo-encuesta de hogares 94
Cuadro A5 Procedimiento de estandarización y homologación de variables en función de la edad 98
Cuadro A6 Variables que generan valores faltantes en el proceso de estandarización y homologación 99
Cuadro A7 Modelo SAE seleccionado para Chile 102
Cuadro A8 Modelo SAE seleccionado para Colombia 105
Cuadro A9 Modelo SAE seleccionado para Perú 108

Gráficos

Gráfico 1 Histograma del ingreso per cápita en Colombia, Chile y Perú 30
Gráfico 2 Histograma del logaritmo del ingreso per cápita en Colombia, Chile y Perú 30
Gráfico 3 Histograma de la transformación Log-Shift del ingreso per cápita en Colombia, Chile y Perú... 31
Gráfico 4 Residuales transformados del modelo ajustado 54
Gráfico 5 Normalidad de los efectos aleatorios para Chile 58
Gráfico 6 Normalidad de los residuales para Chile 58
Gráfico 7 Normalidad de los efectos aleatorios para Colombia. 59
Gráfico 8 Normalidad de los residuales para Colombia. 59
Gráfico 9 Normalidad de los efectos aleatorios para Perú. 60
Gráfico 10 Normalidad de los residuales para Perú 60
Gráfico 11 Residuales internamente estudentizados del modelo ajustado. 61
Gráfico 12 Residuales externamente estudentizados del modelo ajustado 62
Gráfico 13 Residuales internamente estudentizados vs matriz de proyección del modelo ajustado 62
Gráfico 14 Distancias de Cook del modelo ajustado 63
Gráfico A1 Resultados del método Ridge para Chile 100
Gráfico A2 Resultados del método Lasso para Chile 101

Mapas

Mapa 1 Estimación directa en Colombia 67
Mapa 2 Estimación directa en las comunas del norte de Chile 67
Mapa 3 Estimación directa en las comunas del centro de Chile 68
Mapa 4 Estimación directa en las comunas del sur de Chile 68
Mapa 5 Estimación directa en las provincias de Perú 69
Mapa 6 Contraste entre la estimación SAE de la pobreza con la red vial pavimentada en las comunas de Chile 78
Mapa 7 Contraste entre la estimación SAE de la pobreza con la red vial pavimentada y luminosidad de los municipios de Colombia 79
Mapa 8 Contraste entre la estimación SAE de la pobreza con la red vial pavimentada y luminosidad de las provincias de Perú 80

Resumen

Existe un creciente interés por contar con estadísticas sobre diversos grupos de la población con un alto nivel de desagregación geográfica, que generalmente excede la capacidad de las encuestas de hogares para proveer información representativa a dichos niveles. La estimación en áreas pequeñas (SAE, por sus siglas en inglés) es un conjunto de técnicas estadísticas que permiten la estimación de parámetros a nivel subnacional y que buscan mejorar la calidad de las estimaciones directas basadas en encuestas de hogares cuando la desagregación no alcanza los criterios de calidad adecuados para su publicación.

Este documento presenta la metodología y los resultados de la aplicación de un modelo de estimación de áreas pequeñas a nivel de unidad con errores anidados y factores de expansión para la estimación de indicadores de pobreza a nivel provincial en el Perú, comunal en Chile y municipal en Colombia. Al utilizar los censos de población como fuente de información auxiliar, el modelo permite mejorar la precisión de los indicadores de interés en áreas geográficas donde las encuestas no alcanzan la representatividad adecuada. La precisión se evalúa mediante el cálculo de los errores cuadrados medios para cada modelo de estimación en áreas pequeñas, mediante aproximaciones *Bootstrap*.

Este trabajo aprovecha la disponibilidad de censos de población recientes en los tres países, realizados entre 2017 y 2018, así como las encuestas de hogares disponibles en el Banco de Datos de Encuestas de Hogares (BADEHOG) de la División de Estadísticas de la CEPAL para los tres países mencionados. La metodología descrita en este documento puede ser adaptada a cualquier país de la región, para el cual se disponga de los microdatos del censo de población y de una encuesta de hogares realizada en un año cercano que indague por los ingresos de los miembros del hogar y por características auxiliares correlacionadas con el ingreso.

Introducción

Las encuestas de hogares son instrumentos utilizados frecuentemente para generar información sobre un amplio conjunto de indicadores de interés, en temas tan diversos como el empleo, la pobreza y condiciones de vida, el uso del tiempo o la victimización delictiva, entre muchos otros. Dado que las encuestas están basadas en la teoría del muestreo, las estimaciones realizadas están limitadas al alcance determinado por la estrategia de muestreo utilizada (diseño muestral y estimador directo).

Este documento presenta una metodología para la estimación de la pobreza en áreas pequeñas, basada en la combinación de información entre los censos de población y vivienda y las encuestas de hogares, así como los resultados de su aplicación a nivel de división administrativa menor (comuna en Chile, municipio en Colombia y provincia en Perú), en los cuales las estimaciones directas derivadas de la encuesta carecen de suficiente precisión para establecer conclusiones precisas.

Esta sección presenta algunos antecedentes relevantes para la elaboración de los mapas de pobreza. Se describe la metodología de medición de pobreza monetaria de la CEPAL y los indicadores de pobreza habitualmente utilizados. Posteriormente, se revisa el concepto de "área pequeña" y se describe brevemente las principales metodologías de estimación en áreas pequeñas, junto con algunas aplicaciones realizadas en países de la región.

A. Medición de la pobreza monetaria en CEPAL

Las mediciones de pobreza y de pobreza extrema que se utilizan en este documento son efectuadas por la CEPAL sobre la base de una metodología que apunta a lograr la comparabilidad regional, hasta donde la heterogeneidad en las características de las encuestas de hogares lo permiten.

La necesidad de disponer de una medida comparable surge de la diversidad de procedimientos y supuestos utilizados en los países de la región en sus mediciones de pobreza. La construcción de las líneas de

pobreza requiere tomar diversas decisiones (por ejemplo, la estimación de los requerimientos energéticos, el procedimiento de selección de la población de referencia, los criterios nutricionales que satisface la canasta básica de alimentos, la forma de estimar el gasto en bienes no alimentarios, etc.) que son abordadas de maneras diversas en las mediciones nacionales. Los distintos criterios utilizados en dichas estimaciones pueden llevar a que la comparación directa entre ellas genere conclusiones erróneas, al no tener en consideración las diferencias metodológicas.

En la forma habitual de medir la pobreza absoluta se considera a una persona como "pobre extremo" cuando el ingreso por habitante de su hogar es inferior al costo de satisfacer las necesidades básicas de alimentación aun si dedicara todos sus ingresos a la compra de alimentos. Asimismo, una persona se considera como "pobre" cuando el ingreso por habitante de su hogar es inferior al costo de satisfacer las necesidades de alimentación y otras necesidades básicas no alimentarias. Estos costos se conocen como "línea de pobreza extrema" y "línea de pobreza", respectivamente (CEPAL, 2018). El costo de las necesidades de alimentación se estima mediante la construcción de canastas básicas de alimentos, que proveen las cantidades recomendadas de energía y nutrientes, a la vez que reflejan los hábitos de consumo de la población.

Estas recomendaciones vienen de estándares internacionales vigentes para sostener una vida sana, y se calculan tomando en consideración la estructura demográfica de la población y sus niveles de actividad física. Por su parte, los hábitos de consumo son captados mediante las encuestas de ingresos y gastos de los hogares (o encuestas de condiciones de vida, según su disponibilidad en cada país), y corresponden a los de un subconjunto particular de la población, que se adopta como población de referencia con base en criterios establecidos por la metodología.

La ausencia de parámetros normativos claros para determinar la "suficiencia" de los bienes y servicios no alimentarios impide construir una canasta básica de dichos bienes sin incurrir en un ejercicio altamente discrecional y de escasa comparabilidad. Por ello, para determinar el componente no alimentario de la línea de pobreza, se emplea un procedimiento distinto al de los alimentos. La línea de pobreza corresponde a la multiplicación entre el valor de la canasta básica de alimentos y un factor (denominado coeficiente de Orshansky) que corresponde al cociente promedio entre el gasto total y el gasto en alimentos, proporcionado para la misma población de referencia cuyos hábitos se reflejan en la canasta básica de alimentos.

Una vez que se cuenta con las líneas de pobreza extrema y de pobreza, sus valores se comparan con los ingresos de toda la población para determinar qué proporción de ella se encuentra bajo dichos umbrales. El ingreso de los hogares se obtiene sumando los ingresos percibidos por sus miembros (en efectivo y en especie) y abarca ingresos del trabajo, ingresos por jubilaciones, pensiones y otras transferencias, ingresos por la propiedad de activos y otros ingresos (incluido el alquiler imputado). Esta información proviene de las encuestas de hogares que los países llevan a cabo regularmente para propósitos de la medición de pobreza, y que forman parte del banco de datos banco de datos BADEHOG de la CEPAL.

Dado que las líneas de pobreza y de pobreza extrema se estiman para un año base (el año de realización de la encuesta de ingresos y gastos), sus valores se llevan al año de la encuesta que recoge los ingresos de los hogares según la variación observada en los precios. En la metodología de la CEPAL, tanto las líneas de pobreza como los ingresos del hogar están expresados en unidades per cápita. En consecuencia, se asume que el costo por persona de satisfacer las necesidades básicas es el mismo, independientemente de sus características personales y del número de miembros del hogar.

B. Definiciones teóricas de los indicadores de pobreza

La medición de la pobreza involucra generalmente dos etapas, una de "identificación" y otra de "agregación". La identificación consiste en clasificar a las personas como "pobres" o "no pobres". Como se describe en la sección anterior, la metodología utilizada por la CEPAL considera pobres a las personas que residen en hogares cuyos ingresos per cápita están por debajo de la "línea de pobreza". La etapa de "agregación" es

necesaria para generar una medida que indique la extensión y estado actual de la pobreza. Con ello es posible conocer para qué grupo o en qué lugar la pobreza es mayor, cómo esta ha variado a lo largo del tiempo, o cuál ha sido el efecto de una determinada política para su reducción, por ejemplo. Resumir una cantidad amplia de información en un solo número implica que algunas características relevantes del fenómeno pueden no ser reflejadas adecuadamente.

Por ende, es deseable que los indicadores de pobreza cumplan con algunos axiomas básicos. Estos establecen condiciones que determinan el comportamiento esperado de los indicadores y apuntan a capturar la mayor cantidad posible de información en un solo número. Algunos particularmente relevantes para la medición de la pobreza se plantean a continuación (Feres & Mancero, 2001) :

- Axioma focal: el indicador no responde ante cambios en el ingreso de los no-pobres.
- Axioma de monotonocidad: el indicador de pobreza debe aumentar cuando el ingreso de una persona pobre disminuye. Esto quiere decir que debe haber una correspondencia entre la medida de pobreza y la distancia de los pobres respecto de la línea.
- Axioma de transferencia: el indicador debe aumentar ante una transferencia de dinero de un individuo pobre a uno menos pobre. Por lo tanto, este axioma exige que la medida de pobreza sea sensible a la distribución de ingresos bajo la línea de pobreza.
- Axioma de sensibilidad a transferencias: el mencionado axioma requiere que una transferencia de ingresos desde una persona pobre hacia una persona menos pobre incremente la medida de pobreza en mayor grado mientras más pobre sea la persona que entrega sus recursos.
- Axioma de monotonicidad en subgrupos: si el indicador de pobreza aumenta para un subconjunto de personas, el indicador para la población total debe aumentar.

Los indicadores habitualmente utilizados para la medición de la pobreza corresponden a la familia de índices paramétricos propuesta por (Foster, et al., 1984). Estos índices corresponden a la siguiente función:

$$\mathrm{FGT}_{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{z - y_i}{z}\right)^{\alpha} I(y_i < z)$$

Donde N es el tamaño de la población total, y_i es el ingreso per cápita del *i*-ésimo individuo, z es la línea de pobreza[1], I es una función indicadora y α es un parámetro que toma valores mayores o iguales a cero, y que determina las propiedades que cumple el índice. Para todo valor de α, este grupo de medidas es separable aditivamente y cumple con el axioma de monotonicidad en subgrupos.

Incidencia de pobreza

Cuando $\alpha = 0$, el índice de pobreza corresponde al tradicional "índice de recuento" (o "*Head Count Index*", denotado por H), es decir, la proporción de personas que se encuentran debajo de la línea de pobreza:

$$H = \frac{\sum_{i=1}^{N} I(y_i < z)}{N} = \frac{Q}{N}$$

En donde Q denota el total de personas pobres en toda la población. De los axiomas planteados anteriormente, este índice sólo satisface el axioma focal. No cumple con el axioma de monotonicidad, ni de transferencias, ya que una reducción en el ingreso de todos los pobres no afecta al índice.

[1] O pobreza extrema, según corresponda. Con ánimos de ser más breves en la escritura del documento, nos referiremos a pobreza y pobreza extrema, a no ser que se indique lo contrario.

Brecha

Cuando $\alpha = 1$, el índice de pobreza corresponde a la "brecha de pobreza" (o "*Poverty gap*", denotado por PG). Este mide la "profundidad" de la pobreza e indica la distancia promedio de las personas pobres a la línea de pobreza, ponderado por la incidencia de pobreza.

$$PG = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{z - y_i}{z}\right) I(y_i < z) = H\frac{\bar{y}}{z}$$

La brecha de pobreza cumple con el axioma focal y el axioma de monotonicidad: si el ingreso de una persona pobre disminuye, el promedio de ingresos también caerá y el índice PG aumentará. Esta medida no cumple con el axioma de transferencia, ya que un traspaso de ingresos de una persona pobre a una persona menos pobre no se reflejará en el índice, dado que el ingreso promedio entre los pobres permanecerá constante.

Severidad

Cuando $\alpha = 2$, el índice de pobreza resultante satisface el axioma de sensibilidad a transferencias. En este caso, el rol del coeficiente a es asignar una mayor ponderación relativa a las observaciones cuyo ingreso está a una mayor distancia de la línea de pobreza y, por tanto, hace que el indicador sea sensible a la distribución de ingresos entre las personas bajo la línea de pobreza. Mientras mayor sea el valor de α, mayor la ponderación relativa de los ingresos más bajos. Pese a ser el único índice de la familia FGT que cumple con todos los axiomas propuestos, los valores resultantes no tienen una interpretación tan sencilla con la de los indicadores anteriores. Por ello, es común utilizarlo como un indicador complementario a los otros dos, más que como el indicador principal para dar cuenta de los niveles y tendencia de la pobreza.

$$FGT_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{z - y_i}{z}\right)^2 I(y_i < z)$$

C. Definición de área pequeña

Las encuestas de hogares son operaciones estadísticas ampliamente utilizadas con el fin de obtener información de diferentes temáticas de interés público en el tiempo. Son utilizadas por las oficinas nacionales y regionales de estadística, así como por ministerios y gobiernos regionales y locales para obtener estimaciones representativas de indicadores de interés a un nivel de agregación menor determinado sobre la población. Por ejemplo, en Chile, la Encuesta de Caracterización Socioeconómica (CASEN) permite obtener estimaciones a nivel regional, por zonas urbanas y rurales. En tanto, la Gran Encuesta Integrada de Hogares (GEIH) de Colombia proporciona estadísticas a nivel nacional, cabecera-resto, regional, departamental y para cada una de las capitales de los departamentos. En Perú, el diseño muestral de la Encuesta Nacional de Hogares (ENAHO) permite obtener datos representativos a nivel nacional y departamental, por zonas urbanas y rurales.

Bajo estas operaciones estadísticas, un estimador directo (basado en un diseño de muestreo) es aquel que utiliza los factores de expansión y los datos de la encuesta para poder realizar inferencia sobre los parámetros de interés del estudio (Gutiérrez, 2016). La calidad de las estadísticas obtenidas por medio de este tipo de indicadores se evalúa mediante medidas de precisión como el error estándar y el error cuadrático medio. Una estadística se hace oficial si es lo suficientemente confiable para su publicación, es decir, si el error

de muestreo relativo o coeficiente de variación (CV) no supera el umbral de precisión definido por la institución que la genera.

En la práctica, los países de América Latina y el Caribe requieren cada vez más de estadísticas con un mayor nivel de desagregación del originalmente establecido en la planificación de las encuestas de hogares (áreas geográficas menores) o incluso para grupos muy específicos de la población que no estaban considerados inicialmente como dominios de representatividad de las encuestas (por ejemplo, el cruce de etnia con escolaridad), con el fin de implementar y evaluar políticas públicas en diversos ámbitos.

El problema a la hora de emplear estimadores directos radica en que la confiabilidad de su inferencia se empobrece a medida que disminuye el tamaño de muestra (Arima et. al., 2017). Por lo tanto, la inferencia subyacente a la estimación de indicadores en subdivisiones geográficas más pequeñas de las que originalmente estaba planificado producir o en algún dominio socioeconómico de interés resulta inviable por la baja precisión de los indicadores en estos niveles de desagregación. En general, a medida que aumenta la desagregación geográfica o, cuando el tamaño de muestra en el dominio de interés es insuficiente, las estimaciones pierden eficiencia y su confiabilidad disminuye ostensiblemente; en indicadores complejos, este problema genera sesgo en la estimación directa y su error estándar (Pfeffermann, 2013; Morales et. al., 2021).

Según Rao y Molina (2015) un área (dominio) es pequeña si su respectivo tamaño de muestra resulta insuficiente para garantizar una adecuada precisión en la estimación directa del parámetro de interés. Por tanto, el concepto "área pequeña" no está asociado a la dimensión de un agregado geográfico, sino que depende de la fracción de muestreo y del coeficiente de variación. No existe un límite superior universal que delimite el error de muestreo a partir del cual un dominio sea definido como "pequeño". Tampoco existe un tamaño muestral concreto por debajo del cual el área se considere pequeña, pues el error de muestreo varía, no solo dependiendo del tamaño muestral, sino también del indicador que se estima y del estimador utilizado (Bell et. al., 2019). Autores como Molina (2019) y Datta et. al., (2018) afirman que, para un error de muestreo establecido, se requiere un menor tamaño de muestra para la estimación de la media de una variable continua que para calcular la proporción de elementos clasificados respecto a una característica determinada si esta es muy poco o muy común. Por tal motivo, tanto la definición del dominio, el tamaño de muestra y el indicador a estimar resultan importantes a la hora de definir si un área es pequeña o no.

Los dominios usualmente empleados en las encuestas de hogares se pueden clasificar en geográficos, sociodemográficos y mixtos. Dentro de la primera clase se tienen las subpoblaciones generadas por la estructura geográfica de un país (departamentos, estados, provincias, comunas, municipios, entre otros) o por 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). Por otro lado, los dominios sociodemográficos corresponden a las clasificaciones de la población en torno a características como el sexo, la etnia, el nivel educativo, el grupo de edad, entre otras. Los dominios mixtos son aquellos conformados por el cruce entre los geográficos y los sociodemográficos, como por ejemplo la desagregación por grupos de edad y por provincias (Morales et. al., 2021; Rao y Molina, 2015).

En general, se puede considerar, en consonancia con algunas oficinas nacionales de estadística de la región, que una estimación es un dato "no publicable" cuando su error de muestreo relativo (coeficiente de variación) supera el 30% para una subpoblación específica; en este caso, el dominio podría ser considerado como un área pequeña. Por ejemplo, al emplear la Encuesta de Caracterización Socioeconómica Nacional (CASEN), se puede considerar que la mayoría de las comunas de Chile son áreas pequeñas en la estimación de la proporción de individuos con ingresos por debajo del umbral o línea de pobreza. En el mismo orden de ideas, los municipios y las provincias serán las áreas pequeñas para Colombia y Perú, respectivamente. Asimismo, se puede considerar como áreas pequeñas a todas aquellas desagregaciones dadas por el cruce entre agregaciones geográficas y dominios sociodemográficos, como los mencionados anteriormente cuyo error de muestreo relativo sea superior al umbral de publicación establecido.

D. Algunos modelos usados en la construcción de mapas de pobreza

Los modelos de estimación en áreas pequeñas pueden ser clasificados según el nivel de agregación de las variables auxiliares (Morales et al, 2021; Rao y Molina, 2015). Las categorías más conocidas en la literatura relacionada son los modelos de área (o nivel agregado) y los modelos a nivel de unidad. Los primeros enlazan un conjunto de variables agregadas dentro de las áreas pequeñas con sus respectivos estimadores directos. Dichos modelos se hacen necesarios cuando existen restricciones de confidencialidad en los microdatos o cuando estos no están disponibles. Dentro del segundo grupo se encuentran los modelos que permiten enlazar el valor de la variable respuesta con el valor unitario específico de un conjunto de covariables (Hadam et. al., 2020). A continuación, se presenta la descripción de los modelos SAE más usados para la construcción de mapas de pobreza.

Modelo Fay-Herriot (FH)

Uno de los modelos de área más reconocido en la literatura es el propuesto por Fay y Herriot (1979). El modelo FH está compuesto por dos niveles, considerando que el enlace entre un estimador directo $\hat{F}_{\alpha,d}^{Dir}$ a nivel de área, insesgado bajo el diseño de muestreo, y un conjunto de covariables $\boldsymbol{x}_d$ disponibles para todas las áreas que conforman la población. Estas covariables varían de forma regular a través de los dominios (d) mediante el siguiente modelo mixto de regresión:

$$\hat{F}_{\alpha,d}^{Dir} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d,, \qquad d = 1, \dots, D$$

Por definición, $\boldsymbol{\beta}$ corresponde al vector de coeficientes de regresión común a todas las áreas y u_d representa el efecto aleatorio o término de error del modelo específico para el dominio d y simboliza la heterogeneidad en el indicador FGT en los dominios de interés, que no es capaz de ser captada ni cuantificada por las variables auxiliares consideradas. Asimismo, e_d representa el error de muestreo de la estimación directa en el *d*-ésimo dominio, independiente de su respectivo efecto aleatorio. Bajo el modelo FH más simple, se supone que los efectos aleatorios son independientes entre las áreas y siguen una misma distribución con media constante e igual a cero y varianza común σ_u^2, es decir, $u_d \sim^{iid}(0, \sigma_u^2)$. Se asume también que los errores de muestreo son independientes entre sí, tienen media cero y varianzas conocidas ψ_d; es decir, $e_d \sim^{ind}(0, \psi_d)$ (Morales et. al., 2021; Molina, 2019; Rao y Molina, 2015).

En Molina (2019) y Rao y Molina (2015) se afirma que el mejor predictor lineal insesgado empírico (en inglés, *empirical best linear unbiased predictor*, EBLUP) del parámetro de interés $F_{\alpha,d}$ bajo el modelo FH está dado por la siguiente ecuación:

$$F_{\alpha,d}^{\mathrm{FH}} = \boldsymbol{x}_d'\widehat{\boldsymbol{\beta}} + \hat{u}_d, \qquad d = 1, \dots, D$$

En esta expresión $\hat{u}_d$ denota el mejor predictor lineal insesgado empírico (EBLUP) del efecto aleatorio del *d*-ésimo dominio y se puede obtener por medio de la siguiente expresión:

$$\hat{u}_d = \hat{\gamma}_d(\hat{F}_{\alpha,d}^{Dir} - \boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}})$$

donde $\widehat{\boldsymbol{\beta}}$ es el estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$ y

$$\hat{\gamma}_d = \frac{\widehat{\sigma_d^2}}{\widehat{\sigma_d^2} + \widehat{\psi}_d} \in (0,1).$$

Al reemplazar el EBLUP del efecto aleatorio en la ecuación del EBLUP del parámetro de interés $F_{\alpha,d}$ se obtiene que los estimadores resultantes corresponden a una combinación lineal convexa entre el estimador directo y el estimador sintético de regresión, como se puede ver a continuación:

$$\hat{F}_{\alpha,d}^{\mathrm{FH}} = \hat{\gamma}_d\hat{F}_{\alpha,d}^{Dir} + (1 - \hat{\gamma}_{\mathrm{d}})\boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}$$

$\hat{\gamma}_d$ se interpreta como el peso o aporte del estimador directo del EBLUP del parámetro de interés $F_{\alpha,d}$ y define algunas propiedades de este (Datta y Lahiri, 2000). Por ejemplo, $\hat{\gamma}_d \to 1$ cuando $\hat{\psi}_d \to 0$; es decir, cuando el tamaño muestral del d-ésimo dominio es lo suficientemente grande como para garantizar que la variabilidad del estimador directo $\hat{F}_{\alpha,d}^{Dir}$ sea poca en comparación con la heterogeneidad no explicada $\hat{\sigma}_u^2$, el EBLUP le da más peso al estimador directo. En caso contrario, el EBLUP le da más peso al estimador sintético de regresión $\boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}$. Ahora bien, como el estimador $\widehat{\boldsymbol{\beta}}$ utiliza la información de todos los dominios para estimar el parámetro común $\boldsymbol{\beta}$, el EBLUP del parámetro de interés "prestará información" de todas las áreas por medio del estimador sintético de regresión $\boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}$ según lo requerido por la poca o mucha eficiencia del estimador directo.

Modelo con errores anidados de Battese, Harter y Fuller (BHF)

Battese et. al. ,(1988) propusieron un modelo de unidad con errores anidados que relaciona linealmente una variable de interés (Y_{di}) con los valores de p covariables observados para el *i*-ésimo elemento del d-ésimo dominio d. El modelo a nivel de individuo sigue la siguiente relación:

$$Y_{di} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D$$

donde $\boldsymbol{\beta}$ corresponde al vector de coeficientes de regresión común (constante) a todas las áreas, u_d es el efecto aleatorio o término de error del modelo específico para el dominio d (suponiendo que, $u_d \sim^{iid}(0, \sigma_u^2)$) y e_{di} define el error a nivel de individuo. El modelo asume que $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, siendo k_{di} un conjunto de constantes conocidas asociadas a la posible heteroscedasticidad en los individuos dentro de cada área. Al igual que el modelo FH, se asume que los efectos aleatorios u_d son independientes al error a nivel de individuo e_{di} (Molina, 2019; Rao y Molina, 2015).

La metodología de BHF divide la media poblacional respecto a los elementos seleccionados en la muestra de este dominio s_d (valores observados) y otra a los individuos no muestreados (r_d), así $\bar{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} Y_{di}\right)$. En Corral-Rodas et. al. (2021) y Molina (2019) se puede observar que el EBLUP de $\bar{Y}_d$ bajo el modelo BHF se obtiene al ajustar el modelo a los datos de la muestra ($Y_{di} \in s_d$) y pronosticar los valores de la variable de interés en los elementos no muestreados en cada dominio ($Y_{di} \in r_d$), tal como se presenta en la siguiente ecuación:

$$\hat{\bar{Y}}_d^{BHF} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \hat{Y}_{di}^{BHF}\right)$$

Para tal fin, se reemplazan los verdaderos valores de sus componentes de la varianza $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$ por estimadores consistentes $\widehat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ en el siguiente sistema de ecuaciones:

$$\hat{\bar{Y}}_d^{BHF} = \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\beta}} + \hat{u}_d,$$
$$\widehat{u_d} = \hat{\gamma}_d\left(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}{}'\widehat{\boldsymbol{\beta}}\right),$$
$$\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_{d\cdot})$$

En estas ecuaciones, $a_{di} = k_{di}^{-2}$ denota los pesos de las medias muestrales ponderadas de la variable respuesta y las variables auxiliares, $\bar{y}_{da} = \frac{1}{a_{d\cdot}}\sum_{i\in s_d} a_{di}\, Y_{di}$ y $\bar{\boldsymbol{x}}_{da} = \frac{1}{a_{d\cdot}}\sum_{i\in s_d} a_{di}\, \boldsymbol{x}_{di}$, respectivamente. Molina (2019) y Rao y Molina (2015) afirman que, para áreas con fracción muestral despreciable ($n_d/N_d \approx 0$), el EBLUP de la media $\bar{Y}_d$ se puede escribir de la forma:

$$\hat{\bar{Y}}_d^{BHF} \approx \hat{\gamma}_d\{\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widehat{\boldsymbol{\beta}}\} + (1 - \hat{\gamma}_d)\bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}$$

siendo $\bar{\boldsymbol{X}}_d = (\bar{\boldsymbol{X}}_{d1}, \dots, \bar{\boldsymbol{X}}_{dp})$ el vector de medias de las p variables auxiliares en el área d. Bajo un modelo homoscedástico ($k_{di} = 1, \forall i, d$), $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}} \in (0,1)$ define el aporte del estimador "*survey regression*" al EBLUP y define algunas propiedades de este. El EBLUP se acercará al estimador "*survey regression*" ($\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widehat{\boldsymbol{\beta}}$) cuando el tamaño muestral del d-ésimo dominio sea lo suficientemente grande como para garantizar que la variabilidad de éste es poca en comparación con la heterogeneidad no explicada σ_u^2, es decir, cuando $\gamma_d \to 1$ si $n_d \to \infty$. Otro escenario en el que $\gamma_d \to 1$ ocurre cuando las áreas son muy heterogéneas entre sí (σ_u^2 grande comparada con σ_e^2/n_d), o equivalentemente, si las variables auxiliares consideradas no explican adecuadamente la variabilidad del parámetro de interés. En caso contrario, el EBLUP le asignará más peso al estimador sintético de regresión $\boldsymbol{x}_d'\widehat{\boldsymbol{\beta}}$ (que toma información prestada de las demás áreas a través de $\widehat{\boldsymbol{\beta}}$).

Método de Elbers, Lanjouw y Lanjouw (ELL)

Elbers et. al. (2003) propusieron un modelo a nivel de unidad que puede ajustar de forma precisa una gran variedad de indicadores, siempre que estos puedan ser expresados como una función del ingreso de los individuos de la población. Este enfoque ha sido tradicionalmente implementado por el Banco Mundial con el fin de construir mapas de pobreza para los países a nivel mundial (Corral-Rodas et. al., 2021). El método ELL asume que la transformación logarítmica del ingreso en el i-ésimo hogar dentro de cada dominio d está relacionado linealmente con un conjunto de características de este[2]; es decir

$$Y_{di}^* = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D$$

siendo $Y_{di}^* = \log(Y_{di} + c)$ la transformación Log-Shift del ingreso y $c > 0$ es una constante que debe ser estimada por algún método de optimización y que garantice que se genere una variable con distribución aproximadamente normal. Bajo este modelo, u_d es el efecto aleatorio de la ubicación y e_{di} el termino de error asociado a cada hogar de la población. Al igual que el modelo BHF se asume que $u_d \sim^{iid}(0, \sigma_u^2)$ y $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, siendo u_d y e_{di} independientes, y k_{di} constantes conocidas que representan la posible heteroscedasticidad. Finalmente, β es el vector de coeficientes de regresión (Elbers et. al., 2003). El estimador ELL de un parámetro FGT de orden α $\left(F_{\alpha,d}\right)$ se obtiene mediante un procedimiento *Bootstrap*.

Mejor predictor empírico (EBP) con errores anidados

En Molina y Rao (2010) se propone un modelo con errores anidados que permite estimar indicadores no lineales como la incidencia y la brecha de pobreza. Según Molina (2019), esta propuesta permite estimar un indicador FGT de orden α mediante un modelo que asume que la transformación Log-Shift del ingreso (Y_{di}) del i-ésimo hogar dentro del área d está relacionado linealmente con un conjunto de características, a través de la siguiente relación:

$$y_{di}^* = \mathrm{x}_{di}^T\boldsymbol{\beta} + u_d + e_{di}.\ \ i = 1, \dots, N_d,\ d = 1, \dots, D,$$

En donde $Y_{di}^* = \log(Y_{di} + c)$, $\boldsymbol{\beta}$ es el vector de coeficientes de las covariables auxiliares, u_d es el efecto aleatorio del área, tal que $u_d \overset{i.i.d}{\sim} N(0, \sigma_u^2)$ y $e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)$ son los errores a nivel del individuo, los cuales son independientes de los efectos aleatorios. Al igual que en el método ELL, $c > 0$ es una constante que debe ser estimada por algún método de optimización que garantice una distribución aproximadamente normal (Molina et. al. (2014). Como lo afirma Molina (2019), siguiendo la propuesta de Molina y Rao (2010) y Rao y Molina (2015), el estimador del indicador de pobreza FGT se obtiene al desagregarlo en dos partes, la primera asociada a las observaciones dentro de la muestra y otra a las que están por fuera de esta, tal como se presenta a continuación:

---

[2] Siguiendo la recomendación de Marhuenda y otros (2017) de considerar los efectos aleatorios ubicación en el mismo nivel de agregación o dominios donde se desea la estimación.

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

En la práctica el vector de parámetros $\theta = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)$ es desconocido, por lo cual el mejor predictor no se puede calcular de forma analítica. Para remediar el problema, Molina y Rao (2010) y Guadarrama et. al. (2018) proponen procedimientos mediante simulación Montecarlo para estimar indicadores que son funciones del ingreso, como desigualdad y pobreza.

E. Algunas experiencias previas con mapas de pobreza en la región

En la última década, las metodologías de estimación en áreas pequeñas han permitido a países del mundo utilizar mapas de pobreza para el diseño de mejores políticas públicas y focalizar el uso de los recursos públicos con altos niveles de especificidad geográfica (Lange y Pape, 2018). América Latina no ha sido ajena a este proceso y diversos países han implementado mapas de pobreza.

Varios de los desarrollos iniciales en mapas de pobreza fueron impulsados en la región por el Banco Mundial. Tal es el caso de Bolivia, que a inicios de la década de los 2000 desarrolló mapas de pobreza con base en los censos de 1991 y 2001, en un proceso de colaboración entre el Banco Mundial y el Instituto Nacional de Estadísticas (INE) (INE y UDAPE, 2006). La metodología estimó el gasto per cápita de los hogares a través de variables comunes entre el censo y la encuesta, referidas a la estructura y equipamiento de las viviendas, acceso a servicios básicos, características sociodemográficas y de los integrantes del hogar. De esta manera, se generaron 16 modelos, los cuales fueron implementados a partir de los departamentos y su división urbano-rural con el fin de estimar diferentes indicadores de pobreza y desigualdad para cada localidad (como los indicadores FGT, y los coeficientes de Gini, Atkinson, entropía, etc.).

En el caso de Chile, el Ministerio de Desarrollo Social y de Familia ha implementado diferentes métodos y modelos con el fin de generar medidas de pobreza y condiciones de vida a nivel comunal. Los modelos se basan en el método de Fay-Herriot considerado por Jiang et. al. (2002) y Li y Lahiri (2019). La aplicación de los modelos se ha ido adaptando a lo largo del tiempo, tanto debido a mejoras metodológicas como a cambios en el diseño de la encuesta CASEN.

En las dos primeras versiones (2009 y 2011), se utilizó el mismo modelo SAE y las mismas variables auxiliares (actualizadas entre periodos) para la estimación sintética de la pobreza a nivel comunal. Con la encuesta CASEN 2013 se realizó una actualización del modelo SAE y una nueva estimación de las tasas de pobreza comunal para los periodos anteriores, considerando que ahora se basaban en una nueva metodología de medición de pobreza (Ministerio de Desarrollo Social y Familia, 2015).

El diseño de la encuesta CASEN 2015 consideró la generación de estimaciones representativas para un amplio número de comunas (139), por lo que se pudo evaluar y actualizar el modelo de estimación para las comunas que no alcanzaron niveles de precisión adecuados y aquellas sin presencia en la encuesta (Ministerio de Desarrollo Social, 2018). Tanto en este ejercicio como en los anteriores, la información censal fue la del XVII Censo de Población y VI de Vivienda realizado en Chile en el 2002.

Para el periodo 2017 se contó con la información del Censo abreviado de 2017 y una mayor disponibilidad de registros administrativos para el modelo SAE. Con el objetivo de garantizar la comparabilidad con los periodos anteriores se priorizó el uso de las mismas variables censales utilizadas en las estimaciones de años anteriores o proxys de estas. Recientemente, el Ministerio de Desarrollo Social y

Familia, en conjunto con CEPAL, actualizaron la metodología de estimación de pobreza comunal mediante métodos SAE (MDS-CEPAL, 2021)[3].

En el caso de Colombia, la CEPAL colaboró con el Departamento Administrativo para la Prosperidad Social (DPS) para la construcción de un mapa para pobreza monetaria y pobreza monetaria extrema a nivel municipal para todo el territorio nacional. El modelo de errores anidados vinculó el ingreso per cápita de los hogares en la Gran Encuesta Integrada de Hogares (GEIH) para 2018 y 2019 con el Censo Nacional de Población y Vivienda 2018, operaciones implementadas por el Departamento Administrativo Nacional de Estadística (DANE). Los mapas de pobreza generados son útiles para focalizar la oferta social en la población más frágil dentro de una región ya catalogada como vulnerable, evitar duplicidad de programas y lograr que estos tengan una mejor distribución en el territorio.

Ecuador ha tenido diferentes experiencias en la estimación de mapas de pobreza monetaria. En el año 1999, la Oficina de Planificación de la Presidencia de la República publicó mapas de pobreza generados a partir de un modelo de estimación en áreas pequeñas que enlaza la Encuesta de Condiciones de Vida (1995) y el Censo de Población y Vivienda (1990) (Larrea et. al. (1999). En Hentschel et. al. (2000) se presentan los resultados de la implementación del método ELL para la generación de mapas de pobreza a partir de los Censos de Población y Vivienda (1990, 2001) y la Encuesta de Condiciones de Vida (1994, 1999) En el 2007, la Secretaría Técnica del Frente Social del Ecuador en colaboración con el Banco Interamericano de Desarrollo generó el mapa de pobreza en Ecuador empleando la Encuesta de Condiciones de Vida (2006) y el Censo de Población y Vivienda (2001). En Molina et. al. (2015) se presentan los resultados de la vinculación a nivel de unidad de los gastos de los hogares registrados en la Encuesta de Calidad de Vida (ECV) del 2014 y el Censo de Población y Vivienda de 2010 para la estimación de pobreza monetaria a nivel territorial en el país. Al igual que en las experiencias anteriores, se hace uso de la metodología propuesta por Elbers et. al. (2003) y Elbers y Lanjouw (2001).

En el Perú, el uso de modelos de estimación en áreas pequeñas aplicados a pobreza se inició en 2003, cuando el Instituto Nacional de Estadística e Informática (INEI) empleó el método ELL utilizando los datos del Censo de Población y Vivienda de 1993 y de la Encuesta Nacional de Hogares (ENAHO) de ese año (Instituto Nacional de Estadística e Informática (INEI), 2020). La siguiente implementación SAE se realizó en el año 2009, con base en el Censo de Población de 2007 y la encuesta del 2009, con el objetivo de obtener una estimación de la tasa de pobreza monetaria a nivel distrital. En la versión 2015, el INEI empleó nuevamente el método ELL para generar estimaciones de la tasa de pobreza monetaria a nivel distrital e infra distrital cuando el tamaño de muestra era lo suficientemente grande en el distrito. En este periodo se empleó la información del Empadronamiento Distrital de Población y Vivienda (SISFOH) 2012 y 2013 y la ENAHO 2012 y 2013 como fuentes de información. En el 2018 el INEI utilizó la información del Censo de Población de 2017 con el objetivo de generar el mapa de pobreza monetaria bajo los mismos lineamientos usados previamente. Para tal fin, el INEI empleó la ENAHO de 2017 y 2018 y un conjunto de variables auxiliares que provienen de diversos registros administrativos. El Mapa de Pobreza Monetaria Provincial y Distrital 2018 considera el enfoque monetario de la pobreza establecido por la Comisión Consultiva para la Estimación de la Pobreza (Instituto Nacional de Estadística e Informática (INEI), 2020).

Una experiencia fuera de América Latina que constituye un referente útil en la realización de mapas de pobreza monetaria es el programa SAIPE (*Small Area Income and Poverty Estimation*) de los Estados Unidos, que a partir de distintos modelos y métodos de predicción produce diversas estimaciones de pobreza para la población general a nivel de estado, condado y distrito escolar, como también estimaciones de la tasa de pobreza para los niños en edad escolar de estos distritos (Bell y Franco, 2015). Los modelos del programa SAIPE están basados en modelos de área (FH) que vinculan las estimaciones directas obtenidas de distintas encuestas con covariables y registros administrativos de distintas fuentes, las cuales son constantemente

[3] http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/pobreza-comunal/2020/Informe_SAE_2020.pdf.

monitoreadas en términos de su calidad para evitar posibles sesgos (Franco y Bell, 2020; Bell et. al. (2019). Los datos administrativos provienen de fuentes tales como: *U.S. Federal Income Tax Data*, *Supplemental Nutrition Assistance Program Participants Data*, *Census Bureau's Population Estimates Program*, entre otros. Desde el año 2005, la *American Community Survey* provee las estimaciones directas de la pobreza subnacional mientras que la *Current Population Survey* provee las estimaciones a nivel nacional (Pratesi, 2016). De forma similar, el programa SAIPE utiliza la información del Programa de Revisión del Distrito Escolar de la Oficina del Censo (SDRP) para estimar la cantidad de niños que viven en la pobreza dentro de un distrito escolar (Franco, 2020).

I. Fuentes de información

A. Encuestas de hogares

Las encuestas de hogares son una de las principales operaciones estadísticas empleadas por las oficinas nacionales de estadística y otras instituciones públicas para recolectar información actualizada sobre las condiciones socioeconómicas de la población de un país. Esta fuente de información, de gran importancia en el proceso de definición, implementación y seguimiento de políticas públicas, permite obtener estimaciones precisas e insesgadas en los niveles de desagregación que son representativos según el diseño de muestreo establecido (Valliant et al., 2018). En las estimaciones presentadas en este documento se utilizan los microdatos de las encuestas de hogares de Chile, Colombia y Perú, contenidas en el banco de datos BADEHOG de la CEPAL.

Chile

Para el caso de Chile se empleó la Encuesta de Caracterización Socioeconómica Nacional (CASEN) de 2017, a cargo del Ministerio de Desarrollo Social y de Familia (MDS). La finalidad de la encuesta CASEN es medir el nivel de bienestar de la población, especialmente de los grupos pobres y definidos como prioritarios, abordando aspectos demográficos, de educación, salud, vivienda, trabajo e ingresos (Ministerio de Desarrollo Social y Familia, 2015). La encuesta CASEN ha sido implementada desde el año 1987 con una periodicidad que ha cambiado entre dos y tres años; actualmente permite obtener información que es representativa a nivel nacional, regional, nacional urbano y nacional rural.

Para el periodo 2017, en esta operación estadística se alcanzó una muestra 68,466 viviendas y 69,940 hogares. En este periodo se hizo presencia en las mismas 324 comunas que en las versiones anteriores de la encuesta. La muestra de la encuesta ha presentado diversas actualizaciones metodológicas y conceptuales. La primera se derivó de criterios administrativos del Censo 2002, el segundo cambio se generó en el 2008 y en el 2014 con la modificación del marco de muestreo en las áreas urbanas del país y la última se

generó en el 2018 con la modificación tanto el marco de secciones 2002 como el marco de manzanas a partir de los totales de viviendas provenientes del Precenso del año 2016 (Ministerio de Desarrollo Social y Familia, 2018).

Colombia

En Colombia se emplea la Gran Encuesta Integrada de Hogares (GEIH) de 2018, a cargo del Departamento Administrativo Nacional de Estadística (DANE). Su objetivo es proporcionar información básica sobre el tamaño y estructura de la fuerza de trabajo, así como de las características sociodemográficas que permiten caracterizar a la población según sexo, edad, parentesco con el jefe del hogar, nivel educativo, la afiliación al sistema de seguridad social en salud, entre otros (Departamento Administrativo Nacional de Estadística (DANE), 2021).

Según el documento metodológico de la GEIH (Departamento Administrativo Nacional de Estadística (DANE), 2009), esta encuesta da continuidad a un proceso iniciado en la década del setenta con la Encuesta Nacional de Hogares. La cobertura inicial, correspondiente a la zona urbana de las ciudades capitales, pasó a incluir a los municipios circundantes a las siete principales ciudades en los ochenta, para alcanzar una cobertura nacional en los noventa. Desde el año 2006 se inició el diseño y desarrollo de la GEIH como tal, cuyo diseño actualmente permite la desagregación de resultados para el total nacional, total cabeceras, total centros poblados y rural disperso, cada una de las 23 ciudades capitales y áreas metropolitanas, y San Andrés (Departamento Administrativo Nacional de Estadística (DANE), 2021).

Perú

En Perú se emplea la Encuesta Nacional de Hogares (ENAHO), a cargo del Instituto Nacional de Estadística e Informática (INEI). Esta encuesta tiene como finalidad medir el nivel de bienestar de la población peruana respecto a la distribución y características de los hogares y sus ingresos con el fin de cuantificar la población en situación de pobreza. Según el Instituto Nacional de Estadística e Informática (INEI) (2005), esta encuesta ha presentado dos actualizaciones metodológicas y conceptuales: una en 1987 que inició la Encuesta de Hogares de Propósitos Múltiples, y la otra en el 2010 cuando comienza la Encuesta Nacional de Hogares. Actualmente la ENAHO se realiza en julio de cada año, y permite obtener información que es representativa a nivel nacional, regional y para la zonas urbana y rural.

Según el documento técnico de la ENAHO - Instituto Nacional de Estadística e Informática (INEI), que reporta la información en el año 2019 pero los datos empleados corresponden al año 2017, el tamaño de la muestra en este periodo (2017) es de 36,994 viviendas, de las cuales 23346 corresponden a la zona urbana y 13,648 a la parte rural. El tamaño muestral del panel es de 12,514 viviendas particulares mientras que la parte rotativa corresponde a 24,480 unidades. En este periodo, el tamaño muestral de conglomerados fue de 5,359 (3,653 urbanos y 1,706 rurales), de los cuales 1,808 corresponden al panel y 3,551 conglomerados a la parte rotativa.

B. Censos de población y vivienda

Los censos de población son operaciones estadísticas que recolectan información sobre todos los individuos que conforman la población nacional, y permiten obtener el número total y las características principales de las viviendas, los hogares y las personas en todo el territorio de un país. Ello permite contar con información sobre la población referida a sus características sociodemográficas (sexo, edad, parentesco, estado civil, grupo étnico o racial, entre otras), su distribución geográfica y sus condiciones de vida (acceso a servicios públicos, materialidad de la vivienda, nivel educativo alcanzado, tenencia y acceso a productos o servicios en el hogar, etc.), para cualquier nivel deseado de desagregación geográfica.

Los censos de población se realizan aproximadamente cada diez años, pues son una herramienta fundamental en la planeación estratégica y el desarrollo de políticas de carácter nacional y territorial, así como un insumo esencial para la actualización de los marcos de muestreo de las encuestas, las proyecciones de

población y la estimación de diversos parámetros de interés (Naciones Unidas, División de Estadística, 2010). En la práctica los censos en América Latina y el Caribe se han realizado con periodicidad superior, generalmente debido a las restricciones de presupuesto para su implementación. Para la construcción de mapas de pobreza presentada más adelante se utilizan los microdatos de los censos de población de Chile, Colombia y Perú, cuyas características básicas se describen a continuación.

Chile

El Censo de Población y Vivienda 2017 de Chile fue un proceso llevado a cabo por el Instituto Nacional de Estadísticas (INE), en el cual se contó y caracterizó a 17574003 personas y 6499355 viviendas. El Censo de Chile resultó en un censo abreviado que constaba de 21 preguntas para caracterizar de manera resumida a los hogares y las viviendas del territorio nacional. De acuerdo con el Instituto Nacional de Estadísticas (INE), (2018) el censo permitió estimar a la población y generar proyecciones en el mediano plazo, respetando los supuestos de fecundidad, mortalidad y saldo migratorio. Además de las proyecciones, el censo permite caracterizar a las personas y hogares de acuerdo con materialidad de las viviendas, acceso a servicios públicos, datos de hacinamiento, situación ocupacional de las personas y la escolaridad, junto a preguntas de fecundidad y pertenencia étnica.

Colombia

De acuerdo con datos presentados en la ficha metodológica elaborada por el Departamento Administrativo el Censo de Población y Vivienda 2018 permitió contar y caracterizar a las personas residentes de Colombia junto a las viviendas y hogares del territorio. Se registraron 14,242,223 millones de hogares y 44,164,417 personas fueron efectivamente censadas —Departamento Administrativo Nacional de Estadística (DANE), (2021). De lo anterior, se tuvo en cuenta que el 8.5% del total de personas en hogares particulares fueron omitidas, lo que dio un resultado final de 48,258,494 personas en Colombia. Dicho censo permite presentar la distribución de la población por grupos de edad, sexo y características socioeconómicas; además, permite evaluar la inmigración interdepartamental e internacional en los últimos 5 años.

Por otra parte, permitió medir la distribución geográfica de la población junto a indicadores demográficos por departamento, tales como tasa de alfabetismo, migración interna, etc. Presentando además del total de hogares y viviendas, se obtuvo la distribución de los hogares y vivienda por tipo de hogar (Casa, departamento, étnica, etc.), acceso a servicios públicos, tipos de hogar (monoparental, unipersonal, nuclear, etc.) y porcentaje de jefes de hogar según su sexo —Comité Nacional de Expertos (CNE), (2019).

Perú

El XII Censo de Población, VII de Vivienda y III de Comunidades Indígenas o Censo peruano de 2017 fue llevado a cabo por el Instituto Nacional de Estadística e Informática (INEI). Dicho censo permitió enumerar de forma detallada tanto a la población peruana como sus hogares, viviendas y condiciones de vida. De acuerdo con datos de INEI la población efectivamente censada fue de 29,381,884 millones de personas y la población total es de 31,237,385 millones de habitantes contabilizando además 7,698,900 viviendas —Instituto Nacional de Estadística e Informática (INEI), (2019). El censo logro describir el crecimiento, composición, estructura y distribución geográfica de la población del Perú, contempló variables de fecundidad, mortalidad, migración y características sociales de la población tales como: Niveles de educación, alfabetismo, asistencia escolar y acceso a seguros de salud, etnia, religión, etc. De acuerdo con el informe nacional también se recabó las principales características de la población en edad de trabajar (Tenencia en el hogar, lengua aprendida, nivel educacional, etc.) y de los hogares particulares del Perú: tipo de hogar, acceso a tecnologías de información y comunicación, jefatura, etc.

Factores de ponderación en los censos

Autores como Nirel y Glickman, (2009) señalan que, en múltiples países se utilizan las encuestas post-censales de enumeración para evaluar la calidad de los censos o para obtener información sobre futuros problemas de cobertura o sub-registro en el conteo de sus habitantes. Este tipo de problemáticas suele presentarse ya sea porque la persona a encuestar no se encontraba en su hogar o, porque la vivienda registrada no existía al momento de realizar la operación estadística. Las distintas oficinas de estadística emplean ajustes estadísticos para obtener pesos asociados a la falta de cobertura, que permitan calcular la población total del país, considerando la población censada más la población omitida en el censo (Nirel et. al. (2003).

En el caso de Perú, el INEI estimó las probabilidades de omisión censales con base en un modelo de regresión logístico. Dicho modelo utilizó datos integrados externos, junto con la encuesta post-censal (s_p) y el censo de población y vivienda para obtener variables estandarizadas con las cuales se estimó dicha probabilidad. La base de datos del censo contiene una variable con las ponderaciones calibradas, para que totalice la población total estimada con la encuesta postcensal. La calibración se realizó considerando grupos de edad, sexo, área y departamento (INEI, 2019).

En el caso de Chile se empleó el Método de los Componentes Demográficos y la operación de evaluación denominada "Conciliación Censal" para evaluar tanto la cobertura y la composición de la población censada por sexo y edad según región, provincia y comuna. Este método consiste en comparar esta población con aquellas que resultan de tomar en consideración la población de otros censos y la evolución de la mortalidad, la fecundidad y las migraciones durante los períodos intercensales. Esta evaluación se realizó al momento de elaborar las estimaciones y proyecciones de población (Instituto Nacional de Estadísticas (INE), 2018). En Colombia no se utilizaron ajustes post-censales al Censo Nacional de Población y Vivienda 2018 (DANE, 2008).

C. Consolidación de las bases de datos

La metodología presentada en este documento está fundamentada en un modelo de unidad, que estima la relación entre el ingreso per cápita de los hogares y un conjunto de covariables disponibles en las encuestas, para después predecir el ingreso con los microdatos censales respectivos. Para tal fin, es necesario homologar primeramente las categorías de las variables en las encuestas de hogares con las de los censos.

La información disponible en los censos se refiere a características de la vivienda, de los hogares y de las personas. Entre las primeras se encuentran variables como la materialidad de la vivienda, el acceso a servicios públicos o la tenencia de bienes. Dado que el modelo trabaja a nivel de hogares, las variables individuales se agregan a ese nivel, por ejemplo: número de integrantes del hogar, proporción de dependientes en el hogar, nivel educativo del jefe del hogar, tasa de ocupados en el hogar, presencia de discapacitados en el hogar, entre otras variables. Aun cuando las covariables generadas para cada hogar pueden ser continuas o discretas, en su mayoría corresponden a variables dicotómicas; por ejemplo, si el hogar tiene cocina eléctrica o no, si presenta red pública de alcantarillado o no, si el jefe de hogar posee educación superior o no, si algún integrante del hogar trabaja en una determinada actividad económica, entre otras.

De manera previa a la homologación de variables, se realizó un proceso de imputación de variables cuando fue necesario, según las características de la información disponible en cada censo. En Colombia, la base pública disponible contiene registros faltantes para algunas características de los hogares en variables como tenencia de internet, número de dormitorios para dormir, tipo de servicio sanitario, tenencia de internet, lugar donde se preparan los alimentos y tipo de agua para cocinar. Se aplicaron modelos tipo *Naive Bayes* para imputar las variables de los 117,269 hogares con registros faltantes. Autores como Hruschka et. al. (2007) y van Buuren & Groothuis-Oudshoorn, (2011) presentan con mayor detalle el enfoque teórico-práctico de esta metodología de imputación y sus ventajas. En el caso de Perú se realizó una limpieza de las bases de

datos, descartando aquellos hogares no encuestados o censados pero que estaban incluidos en la base de datos. Adicionalmente, fue necesario descartar aquellos hogares donde el jefe de hogar tenía menos de 15 años, pues en la encuesta no existen jefes con esta característica.

Debido a la cantidad de datos y variables a generar en los censos y las encuestas, se utilizaron las librerías *data.table*, *dtplyr* y *dplyr* del software estadístico R, las cuales permitieron un manejo más eficiente de estos grandes volúmenes de datos sin la necesidad de requerir otros programas para su posterior procesamiento. Después de cargar las mencionadas librerías, los datos a utilizar se convierten en formato *lazy_dt*, el cual permite minimizar el tiempo de procesamiento con respecto a otras metodologías en el mismo software. Inicialmente, para todos los países se generaron dos bases de datos por cada operación estadística (censo y encuesta de hogares) empleando el identificador de hogar de la encuesta y del censo. Dentro de estas bases de datos se genera un conjunto de variables indicadoras empleando las categorías estandarizadas definidas para cada variable auxiliar a ser empleada. A partir de estas variables indicadoras homologadas se puede estructurar y realizar adecuadamente el ajuste del modelo SAE considerado en el presente documento.

En la sección 1 de los anexos, más específicamente de los cuadros 22 a la 27 se presentan ejemplos detallados sobre el proceso de homologación de las bases de datos en los países bajo análisis, en las que se detalla la estandarización de algunas de las variables encontradas en los cuestionarios y que debían ser contestadas por el entrevistado, entre ellas se encuentran: años de estudio, ocupación principal de desempeño, actividad del negocio donde trabajó el respondiente, actividad del negocio donde trabajó.

En el caso colombiano, en la misma sección 1 de los anexos, en el Cuadro 25 se presenta, a manera de ejemplo, el proceso de homologación de siete variables empleadas en el estudio. Por otro lado, teniendo en cuenta que en el proceso de homologación de variables se debe estandarizar las preguntas abiertas a ser empleadas en variables categóricas cuyos niveles estén homologados entre las fuentes de información, en el Cuadro 26 se muestra un ejemplo del proceso, tomando como base la homologación de variables en función de la edad. Finalmente, dentro del procedimiento de estandarización y homologación, se puede encontrar situaciones en las que por la estructura del formulario aparezcan opciones de respuesta que no se pueden ser enlazadas entre las distintas fuentes de información, lo cual genera valores faltantes en algunas de las variables de interés, por lo que en el Cuadro 27, se presenta un conjunto de variables que generan valores faltantes al enlazar el censo con la encuesta de hogares.

II. Elementos teóricos de la metodología Pseudo - EBP

La metodología utilizada en el presente documento está fundamentada en un enfoque "basado en el modelo" a nivel de individuo, en el cual se enlaza un indicador general con un conjunto de variables auxiliares para cada elemento de la población. Esta metodología, denominada "pseudo-EBP" corresponde a un modelo con errores anidados que induce un acercamiento del mejor predictor empírico de Molina & Rao (2010), incorporando los factores de expansión de la encuesta, como proponen Guadarrama et. al. (2018). La metodología seleccionada permite aprovechar el acceso a los microdatos del censo de población y así incorporar más información que los modelos a nivel de área.

Asimismo, a diferencia de modelos como BHF o ELL presentados anteriormente, no se requiere conocer de antemano o estimar la varianza de los residuos del modelo. Bajo ciertas condiciones, estos modelos brindan un mayor nivel de desagregación de las estimaciones para cualquier subdominio deseado dentro de las áreas de interés. Es decir, es posible generar estimaciones a nivel de municipio, provincia o comuna, desagregadas, por ejemplo, por características del hogar o individuales (por autorreconocimiento étnico, grupo de edad, sexo, discapacidad, etc., si se trabaja con covariables a nivel de individuo).

Para estimar el error cuadrático medio (ECM) como medida de precisión se utiliza un proceso de simulaciones mediante *Bootstrap*; es decir, a través de múltiples realizaciones del parámetro de interés comparadas respecto a las estimaciones obtenidas en la réplica de la selección aleatoria dentro de cada censo. A continuación, se describe el procedimiento de estimación que permite generar los mapas de pobreza en los niveles de desagregación de interés para cada uno de los países bajo análisis.

A. Covariables utilizadas para las estimaciones

Al tener las bases de datos estructuradas y homologadas (encuestas de hogares y censos), el procedimiento de estimación inicia con la selección inicial de variables auxiliares que estén relacionadas con la pobreza monetaria la cual se mide de forma indirecta a través de los ingresos del hogar. Muchas de las variables disponibles en los censos guardan una relación estrecha con los ingresos del hogar y la condición de pobreza, y son relevantes para el modelo utilizado. Por una parte, la capacidad de generar ingresos en los hogares está

relacionada con las características de educación y participación en el mercado laboral de sus miembros. Esta depende también de la composición del hogar y la proporción de personas dependientes. Por ello resulta importante considerar variables como el nivel educativo del jefe del hogar, la proporción de ocupados e inactivos y la proporción de personas en distintos tramos etarios. Por otra parte, las características de la vivienda y el acceso a bienes y servicios básicos, así como la existencia de privaciones críticas en esos ámbitos, son variables habitualmente usadas para evaluar las condiciones de vida en ámbitos no monetarios y la probabilidad de los hogares de encontrarse en pobreza. Las covariables utilizadas para la aplicación del modelo y su tipo se presentan en el cuadro 1.

Cuadro 1
Listado de variables iniciales consideradas en el proceso de estimación

Descripción	Tipo
Hogar con privación de hacinamiento	Categórica (dicotómica)
Número de personas en el hogar	Categórica
Hogar con privación de trabajo infantil	Categórica (dicotómica)
Sexo del jefe del hogar	Categórica (dicotómica)
Estado civil del jefe del hogar	Categórica
Nivel educativo del jefe del hogar	Categórica
Hogar con privación de analfabetismo	Categórica (dicotómica)
Hogar con privación de inasistencia escolar	Categórica (dicotómica)
Proporción de ocupados en el hogar	Numérica
Proporción de inactivos en el hogar	Numérica
Proporción de niños en el hogar	Numérica
Proporción de adolescentes y jóvenes en el hogar	Numérica
Proporción de adultos mayores en el hogar	Numérica
Proporción de personas con educación superior	Numérica
Tipo de vivienda	Categórica
Material de paredes	Categórica
Material de pisos	Categórica
Hogares con acceso a energía	Categórica (dicotómica)
Estrato socioeconómico	Categórica
Hogar con acceso a energía	Categórica (dicotómica)
Hogar con acceso a acueducto	Categórica (dicotómica)
Hogar con acceso a alcantarillado	Categórica (dicotómica)
Hogar con conexión a gas	Categórica (dicotómica)
Hogar con servicio de recolección de basuras	Categórica (dicotómica)
Hogar con acceso a internet	Categórica (dicotómica)
Tipo de servicio sanitario del hogar	Categórica
Lugar donde se preparan los alimentos	Categórica
Fuente de agua de consumo humano	Categórica
Departamento / Región	Categórica

Fuente: Elaboración propia.

B. Transformación del ingreso

Como paso a seguir, se procede en transformar la variable de interés, el ingreso per cápita de los hogares. Debido a la naturaleza del modelo utilizado, es necesario que la variable a modelar tenga una distribución normal. No obstante, el ingreso per cápita se caracteriza por tener una distribución con profunda asimetría positiva (véase el Gráfico 1), la cual se confirma a partir de las estimaciones del coeficiente de asimetría de Fisher, el cual corresponde a la siguiente expresión (Bickel y Doksum, 2015; Casella y Berger, 2002):

$$\gamma = \frac{\mu_3}{\sigma^3}$$

Siendo μ_3 el tercer momento central y σ es la desviación estándar del logaritmo del ingreso per cápita. Teóricamente, si $\gamma > 0$, la distribución de la variable transformada presenta sesgo a la derecha o asimetría positiva; por el contrario, si $\gamma < 0$, la distribución de la variable transformada presenta sesgo a la izquierda o asimetría negativa. Finalmente, la distribución es simétrica cuando $\gamma = 0$.

Para identificar una transformación óptima del ingreso per cápita en los hogares que resulta en una variable con distribución aproximadamente normal se hicieron pruebas con las familias de transformaciones Box-Cox y Log-Shift. La familia de transformaciones de Box-Cox conforman una metodología ampliamente utilizada para corregir sesgos en la distribución de una variable continua con problemas de no normalidad (Venables y Ripley, 2002; Cryer y Chan, 2008). El proceso iterativo consta de los siguientes pasos:

i) Seleccionar el rango de valores de transformación de potencia (γ) que será empleado en cada uno de los países. El procedimiento considera una secuencia de valores en el intervalo $-2 \leq \gamma \leq 2$.

ii) Para cada valor de lambda se realiza la transformación dada por:

$$f(y_i) = \begin{cases} \dfrac{y_i^{\gamma} - 1}{\gamma} & \text{para } \gamma \neq 0 \\ \log(y_i) & \text{para } \gamma = 0 \end{cases}$$

para todos los datos de ingreso en los hogares ($i = 1, \cdots, n$) en los países.

iii) Se selecciona el valor de lambda que garantice un coeficiente de asimetría de Fisher cercano a cero.

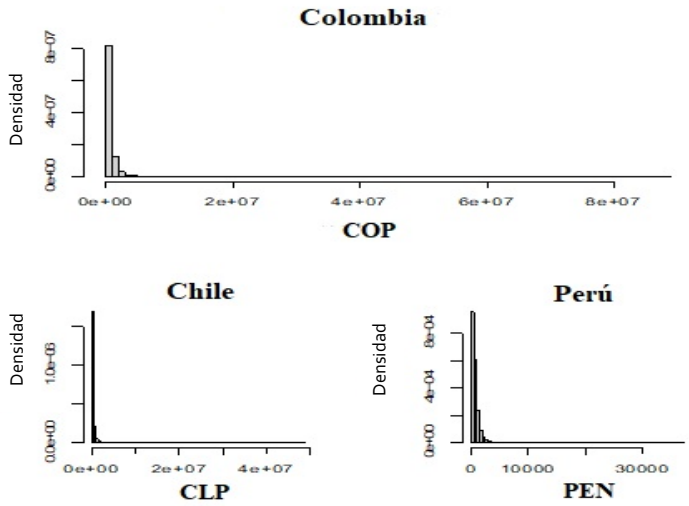
En el procedimiento secuencial considerado para la transformación del ingreso per cápita, el valor de γ resultó ser muy cercano a cero; es decir, se utilizó la transformación $f(y_i) = \log(y_i)$ para los tres países. En el gráfico 2 se presenta la densidad de la transformación logaritmo del ingreso per cápita en los tres países, donde se puede observar un apreciable aumento en la simetría de las distribuciones. El análisis confirmatorio se realiza empleando diferentes pruebas de normalidad sobre la variable transformada. Las pruebas de normalidad evalúan el siguiente contraste de hipótesis:

$$H_0 \coloneqq y_1, \ldots, y_n \sim N(\mu, \sigma^2);\ \sigma^2 > 0, \mu \in \mathbb{R} \text{ desconocidos}$$
$$H_1 \coloneqq y_1, \ldots, y_n \text{ no sigue distribución normal}$$

Se emplearon las pruebas de normalidad tradicionales de la librería *fBasics* del paquete estadístico R como lo son la prueba de Lilliefors, la prueba de Jarque-Bera y la de D'Agostino-Pearson. Los resultados de las pruebas de normalidad rechazan el supuesto de normalidad de las variables transformadas del ingreso[4]; es decir, la transformación Box-Cox no logra generar una variable con distribución aproximadamente normal.

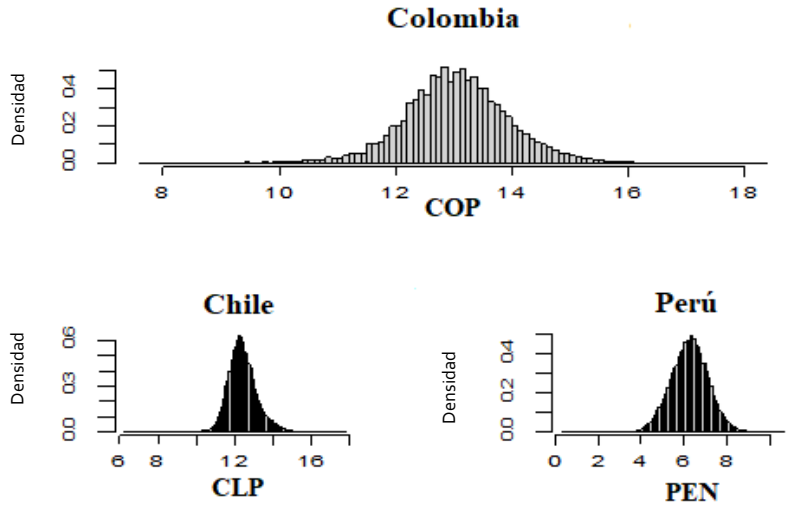
[4] Se consideró un nivel de significancia nominal del 1% para todas las pruebas empleadas.

Gráfico 1
Histograma del ingreso per cápita en Colombia, Chile y Perú



Fuente: Elaboración propia.

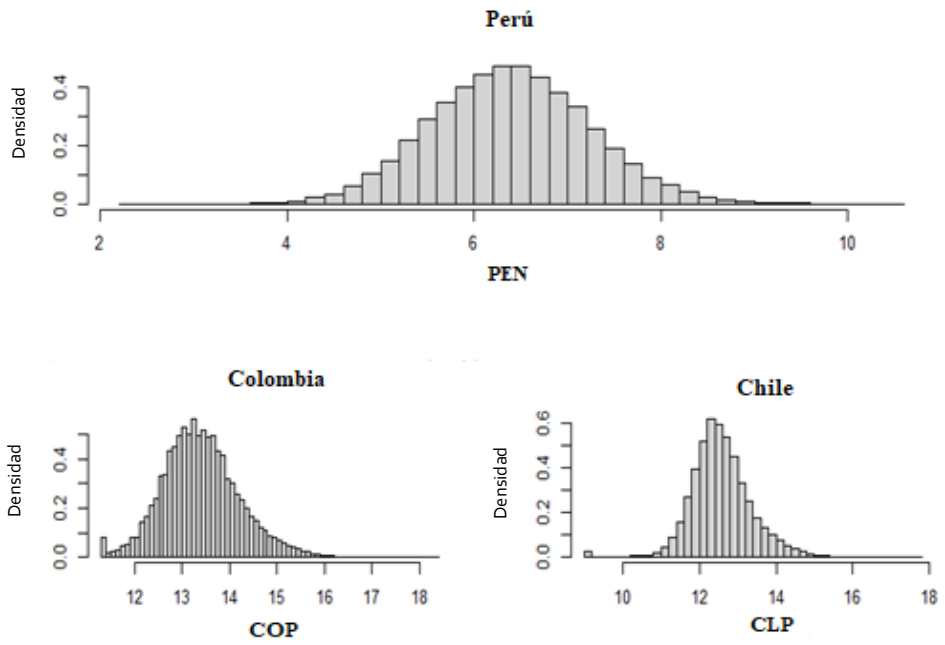
Gráfico 2
Histograma del logaritmo del ingreso per cápita en Colombia, Chile y Perú



Fuente: Elaboración propia.

Como alternativa, se utilizó la familia de transformaciones Log-Shift. Este procedimiento iterativo parte de la creación de la variable $y_{di}^{*} = \log(y_{di} + c)$ para una grilla de valores c predefinidos, de modo que esta transformación permita generar normalidad en la predicción de un modelo lineal de la forma $Y = X^{T}\beta + \epsilon$. La constante seleccionada c es aquella que garantice la normalidad de la variable transformada, junto con un coeficiente de asimetría de Fisher más cercano a cero. En el gráfico 3 se presenta la densidad de la transformación Log-Shift del ingreso per cápita en los países, donde se puede observar una mayor simetría de las distribuciones.

Gráfico 3
Histograma de la transformación Log-Shift del ingreso per cápita en Colombia, Chile y Perú



Fuente: Elaboración propia.

El valor óptimo de la constante c para cada uno de los países, resultante del proceso iterativo, se presenta en el cuadro 2. Estos garantizan valores cercanos a cero para el coeficiente de asimetría de Fisher y resultados aceptables de las pruebas de normalidad.

Cuadro 2
Constante óptima de la transformación Log-Shift en los países

País	Constante
Chile	$ 8 600 CLP
Colombia	$ 81 958 COP
Perú	$ 10,96 PEN

Fuente: Elaboración propia.

C. Estimación de los coeficientes de regresión

En el proceso de cuantificar los ingresos promedio de los hogares latinoamericanos se propone un modelo de unidad que permite realizar una conciliación con el diseño de muestreo complejo. En Guadarrama et. al. (2018) se presenta un modelo con errores anidados que induce un acercamiento del mejor predictor empírico de Molina & Rao, (2010) denominado Pseudo-EBP. El modelo que se presenta a continuación considera una transformación Log-Shift del ingreso $y^*_{di} = \log(y_{di} + c)$ para relacionarla linealmente con un predictor lineal, un efecto aleatorio de área (u_d) y su respectivo error a nivel de hogar (e_{di}):

$$\boldsymbol{y}^*_{di} = \mathrm{x}^T_{di}\boldsymbol{\beta} + u_d + e_{di}.\ \ i = 1, \dots, N_d,\ d = 1, \dots, D,$$

En el modelo presentado anteriormente, el efecto aleatorio de área $u_d \overset{i.i.d}{\sim} N(0, \sigma^2_u)$ es independiente al error a nivel individuo $e_{di} \overset{iid}{\sim} N(0, \sigma^2_e)$, mientras que $\boldsymbol{\beta}$ representa el vector de coeficientes de las covariables. Las variables transformadas por país se resumen en el cuadro 3.

Cuadro 3
Transformación Log-Shift en los países

País	Constante
Chile	$y_{di}^{*} = \log(\, y_{di}^{*} + 8\,600)$
Colombia	$y_{di}^{*} = \log(\, y_{di}^{*} + 81\,958)$
Perú	$y_{di}^{*} = \log(\, y_{di}^{*} + 10{,}96)$

Fuente: Elaboración propia.

Bajo el modelo con errores anidados, los vectores de la variable transformada $\boldsymbol{y}_d = (y_{d1}^{*}, \ldots, y_{d,N_d}^{*})'$ son independientes entre los dominios de cada uno de los países ($d = 1, 2, \ldots, D$) y siguen una distribución normal de media $\boldsymbol{\mu}_d = \mathrm{X}_d\boldsymbol{\beta}$ y matriz de covarianzas $\mathrm{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}^{T} + \sigma_e^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}'$[5], es decir, $\boldsymbol{y}_d \sim^{ind} N(\boldsymbol{X}_d\boldsymbol{\beta}, \boldsymbol{V}_d)$ —Rao & Molina, (2015). El vector de la variable transformada se puede clasificar en la parte observada por la encuesta (y_{ds}) y la que no (y_{dr}), es decir, $\boldsymbol{y_d} = \begin{pmatrix} y_{ds}' \\ y_{dr}' \end{pmatrix}$. Bajo esta descomposición, la matriz de covariables y la de varianzas-covarianzas se puede estructurar de la siguiente manera:

$$\boldsymbol{X_d} = \begin{pmatrix} X_{ds} \\ X_{dr} \end{pmatrix}, \quad \boldsymbol{V_d} = \begin{pmatrix} V_{ds} & V_{dsr} \\ V_{drs} & V_{dr} \end{pmatrix}$$

Dado que los individuos y_{di} sigue una distribución normal, entonces las condicionales también siguen una distribución normal para los hogares no observados en la encuesta ($i \in r_d$), la cual se encuentra definida por:

$$y_{di}|\mathrm{y_{ds}} \sim N(\mu_{di|s}, \sigma_{di|s}^2)$$

donde la media y la varianza condicional se define como

$$\mu_{di|s} = \mathrm{x}_{di}^{T}\boldsymbol{\beta} + \gamma_d(\bar{y}_d - \bar{\mathrm{x}}_d\boldsymbol{\beta})$$

$$\sigma_{di|s}^2 = \sigma_u^2(1 - \gamma_d) + \sigma_e^2$$

Donde $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_d}}$, $\bar{y}_d = \frac{1}{n_d}\sum_{i=1}^{n_d} y_{di}$ y $\bar{\mathrm{x}}_d = \frac{1}{n_d}\sum_{i=1}^{n_d} \mathrm{x}_{di}$. Según Molina, (2019) para aquellos indicadores FGT que puedan ser definidos como una función de $\boldsymbol{y_d}$ - es decir, $\boldsymbol{\delta_d} = \boldsymbol{f}(\boldsymbol{y_d})$ - el mejor predictor lineal es aquel que minimiza el Error Cuadrático Medio (ECM) y está dado por:

$$\tilde{\delta}_d^{B}(\boldsymbol{\theta}) = E_{\mathrm{y}_{dr}}\left[\delta_d\left(\mathrm{y}_d\middle|\mathrm{y}_{ds}, \boldsymbol{\theta}\right)\right]$$

siendo $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2)$ el vector de parámetros. En esta expresión se puede observar que el valor esperado de la variable de interés para elementos no seleccionados en la muestra dentro de la subdivisión de interés d (denotados $\boldsymbol{y_{dr}}$) está condicionado a los valores observados para los elementos seleccionados en la muestra y al verdadero valor del vector de parámetros del modelo, los cuales usualmente se desconocen. Por lo tanto, al reemplazar $\boldsymbol{\theta}$ por un estimador consistente $\widetilde{\boldsymbol{\theta}}$ se obtiene el mejor predictor lineal empírico $\widetilde{\boldsymbol{\delta}}_{\boldsymbol{d}}^{\boldsymbol{EB}} = \widetilde{\boldsymbol{\delta}}_{\boldsymbol{d}}^{\boldsymbol{B}}(\widetilde{\boldsymbol{\theta}})$. En la ecuación anterior, en algunas ocasiones, no se puede estimar la esperanza de forma analítica; una forma de afrontar de forma empírica esta situación consiste en emplear simulación Montecarlo. Considerando la notación mencionada, la expresión para el estimador mínimo cuadrático ponderado de $\boldsymbol{\beta}$ está dada por:

[5] $\mathbf{1}_j$ denota un vector unitario de longitud j.

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \boldsymbol{X}_{ds}\, \boldsymbol{V}_{ds}^{-1} \boldsymbol{X}_{ds}{}'\right)^{-1} \sum_{d=1}^{D} \boldsymbol{X}_{ds}\, \boldsymbol{V}_{ds}^{-1} \boldsymbol{y}_{ds}$$

Empleando este estimador, Molina (2019) y Rao y Molina (2015) afirman que para aquellos dominios donde la fracción muestral es despreciable ($n_d/N_d \approx 0$) el mejor estimador lineal insesgado ("best linear unbiased predictor" o BLUP) del ingreso medio de los hogares ($\bar{Y}_d$) y del efecto aleatorio (u_d) están dados por el siguiente sistema de ecuaciones:

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d\{\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widetilde{\boldsymbol{\beta}}\} + (1-\gamma_d)\bar{\boldsymbol{X}}_d{}'\widetilde{\boldsymbol{\beta}}$$

$$\tilde{u}_d = \gamma_d\left(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}{}'\widetilde{\boldsymbol{\beta}}\right)$$

$$\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{d\cdot})$$

En estas ecuaciones se definen tanto la media muestral ponderada de la variable transformada como de las variables auxiliares están dadas por

$$\bar{y}_{da} = a_{d\cdot}^{-1} \sum_{i \in s_d} a_{di} \boldsymbol{y}_{di}^{*}$$

$$\bar{x}_{da} = a_{d\cdot}^{-1} \sum_{i \in s_d} a_{di} x_{di}$$

En estas las anteriores expresiones, el ponderador $a_{di} = k_{di}^{-2}$, está conformado por las constantes k_{di}, que se suponen conocidas y que representan la posible heteroscedasticidad y donde $a_{d\cdot} = \sum_{i \in s_d} a_{di}$. El parámetro de ponderación por dominio $\gamma_d \in (0,1)$ permite describir el BLUP en función de dos estimadores, el estimador sintético de regresión dado por $\bar{\boldsymbol{X}}_d{}'\widetilde{\boldsymbol{\beta}}$ y el estimador de regresión dado por $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widetilde{\boldsymbol{\beta}}$; como estimador directo. Los valores del parámetro de compresión dependen tanto del tamaño de muestra del dominio como de la heterogeneidad entre los dominios y se pueden interpretar así:

- Si $\gamma_d \to 0$, el BLUP se aproxima al estimador sintético. Este caso ocurre cuando el tamaño de muestra del dominio es pequeño ($n_d \to 0$) o de forma equivalente, cuando las variables auxiliares consideradas explican gran parte de la variabilidad del ingreso ($\sigma_u^2 << \sigma_e^2/n_d$).
- Si $\gamma_d \to 1$, el BLUP se aproxima al estimador de regresión. Este caso ocurre cuando el tamaño de muestra del dominio es grande ($n_d \to N_d$) o de forma equivalente, cuando las variables auxiliares consideradas no son capaces de explicar la variabilidad del ingreso ($\sigma_u^2 >> \sigma_e^2/n_d$).

El parámetro de compresión depende directamente del verdadero valor de las componentes de la varianza del modelo $\boldsymbol{\theta} = \left(\widetilde{\boldsymbol{\beta}}, \sigma_u^2, \sigma_e^2\right)'$ que son desconocidas en la práctica (Datta, et. al. (2018). Con el fin de poder obtener una estimación de ingreso promedio de los hogares se debe reemplazar la varianza del componente asociado al efecto aleatorio de área (σ_u^2) y la varianza asociada al termino de error a nivel individuo (σ_e^2) por un estimador consistente $\widehat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$. A partir de estos estimadores se obtiene el mejor lineal estimador insesgado empírico (EBLUP) dado por:

$$\hat{Y}_{di}^{EBLUP} = \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\beta}} + \hat{u}_d,$$

$$\hat{u}_d = \hat{\gamma}_d\left(\bar{y}_{dw} - \bar{\boldsymbol{x}}_{dw}{}'\widehat{\boldsymbol{\beta}}\right),$$

$$\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_{d\cdot}).$$

Para evitar el sesgo inducido al ignorar el diseño de muestreo en el modelo, los parámetros de la anterior distribución se estiman consistentemente incluyendo los factores de expansión (w_{kd}) de la encuesta en su forma funcional (Guadarrama et. al.(2018). De esta forma, se tiene que:

$$\hat{\mu}_{dr|s} = X_{dr}\widehat{\boldsymbol{\beta}} + \hat{\gamma}_d\left(\bar{y}_{dw} - \bar{x}'_{dw}\widehat{\boldsymbol{\beta}}\right)\mathbf{1}_{N_d-n_d}$$

$$\widehat{\boldsymbol{V}}_{dr|s} = (\hat{\sigma}_e^2 + \hat{\sigma}_u^2(1-\gamma_d))\mathbf{1}_{N_d-n_d}\mathbf{1}'_{N_d-n_d}$$

donde $\bar{y}_{dw}$ y $\bar{x}_{dw}$ denota, respectivamente, la media muestral expandida de la variable transformada y de las variables auxiliares, las cuales se definen mediante las siguientes ecuaciones:

$$\bar{y}_{dw} = \frac{\sum_{j\in s_d} w_{kd}y_{kd}}{\sum_{j\in s_d} w_{kd}}$$

$$\bar{x}_{dw} = \frac{\sum_{j\in s_d} w_{kd}x_{kd}}{\sum_{j\in s_d} w_{kd}}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2\delta_d^2}$$

$$\delta_d^2 = \frac{\sum_{j\in s_d} w_{kd}^2}{\left(\sum_{j\in s_d} w_{kd}\right)^2}$$

En estas ecuaciones $\hat{\beta}, \hat{\sigma}_e^2$ y $\hat{\sigma}_u^2$ son estimadores consistentes de los coeficientes de regresión del modelo de unidad, así como de las varianzas de los errores y efectos aleatorios, respectivamente; estimadores que se presentarán en las siguientes subsecciones. Lo anterior implica que el proceso de estimación se realiza con toda la muestra de la encuesta, la cual suele generalmente ser de gran tamaño. En este escenario, la distribución condicional de un hogar que no fue seleccionado en la muestra ($i \in r_d$) está dado por:

$$y_{dr}|y_{ds} \sim N(\mu_{di|s}, \sigma^2_{di|s}) \text{ con } d = 1, \dots, D$$

Ahora bien, como el interés puntual de la metodología presentada en este documento consiste en estimar el indicador de pobreza FGT de orden α ($\delta_d = F_{\alpha d}$), la ecuación de estimación se obtiene al reescribir el indicador como una función de las variables respuesta y_{di}. Sin embargo, dado que con la información disponible no es posible identificar y vincular las unidades de las muestras con las unidades de los censos, entonces el acercamiento que se utiliza es del tipo Census-EB, el cual asume que todos los elementos del censo están asociados a las observaciones fuera de la muestra, tal como se presenta a continuación:

$$\tilde{F}^B_{\alpha d}(\theta) = \frac{1}{N_d}\left(\sum_{i\in r_d}\tilde{F}^B_{\alpha,di}(\theta)\right)$$

Resulta importante aclarar que, en todas las encuestas de hogares de la región, y en particular en CASEN, GEIH y ENAHO, la fracción entre las unidades seleccionadas en las muestras y el número de personas del país es muy cercano a cero, y por ende el predictor tendrá un desempeño bastante similar al Pseudo-EBP. Asimismo, se considera un procedimiento de simulación de Montecarlo para estimar los indicadores de pobreza, pues como es bien sabido, la esperanza resulta ser el mejor predictor y cuando no se puede calcular los indicadores de forma analítica se aproxima siguiendo los pasos del siguiente algoritmo:

i) Ajustar el modelo basado en errores anidados presentado en la sección D. a los datos de la encuesta $(\boldsymbol{y}_s, \boldsymbol{X}_s)$ con el fin de obtener estimadores consistentes $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2, \hat{\sigma}_e^2)^T$ del vector de parámetros desconocidos $\boldsymbol{\theta}$.

ii) Estimar el vector de medias $\mu_{di|s}$ y de varianzas $\boldsymbol{\sigma}^2_{di|s}$ empleando la estimación $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \hat{\sigma}^2_u, \hat{\sigma}^2_e)^T$ del paso 1 en las ecuaciones definidas para la distribución condicional $Y_{di}|\boldsymbol{y}_{ds} \sim N\left(\boldsymbol{\mu}_{di|s}, \boldsymbol{\sigma}^2_{di|s}\right)$ para $d = 1, 2, \dots, D$.

iii) Generar $a = 1, \dots, A$ simulaciones de Montecarlo del vector de la variable de interés para los hogares por fuera de la muestra de la subdivisión de interés d, $\boldsymbol{y}^{(a)}_{dr}$.

iv) Calcular el indicador de interés $\widetilde{F^{(a)}_{\alpha d}} = F_{\alpha d}\left(\boldsymbol{y}^{(a)}_{d}\right)$ para cada simulación de Montecarlo $a = 1, \dots, A$.

v) La estimación de Montecarlo del Census-EB del indicador FGT ($\hat{F}^{\text{EB}}_{\alpha d}$) se obtiene promediando los indicadores para los A censos simulados, es decir,

$$\hat{F}^{\text{EB}}_{\alpha d} = \frac{1}{A}\sum_{a=1}^{A} \widetilde{F^{(a)}_{\alpha d}}$$

Es importante señalar que, para el caso de las tasas e incidencias de pobreza, se consideraron las líneas de pobreza y de pobreza extrema establecidas por la CEPAL para cada país analizado. En el algoritmo de estimación previamente descrito se puede observar que la estimación Montecarlo del indicador FGT depende de la estimación del vector de coeficientes de regresión del efecto fijo del modelo anidado, es decir, depende de $\widehat{\boldsymbol{\beta}}^T$. Por lo anteriormente descrito, una parte importante de esta etapa consiste en identificar la capacidad predictiva de las variables auxiliares empleadas.

Este proceso fue implementado en el software R para estimar por provincia los ingresos medios, la tasa de pobreza (para pobreza y pobreza extrema), totales de pobreza (Pobreza e indigencia) Para el caso de las tasas e incidencias de pobreza, se consideraron las líneas de pobreza tanto para pobreza como para pobreza extrema. Estas líneas de pobreza se definen en cada uno de los países según el área de residencia del hogar (Urbano - Rural) y están dadas en el cuadro 4.

Cuadro 4
Líneas de pobreza y pobreza extrema empleadas en el estudio
(Moneda local)

País	Tipo	Urbano	Rural
Chile	Pobreza	113 958	84 538
	Pobreza extrema	51 309	46 705
Colombia	Pobreza	296 845	200 76
	Pobreza extrema	147 169	127 346
Perú	Pobreza	316,76	210,97
	Pobreza extrema	144,6	120,3

Fuente: CEPAL (2018), Medición de la pobreza por ingresos: actualización metodológica y resultados.

D. Selección de covariables

En la actualidad, en la era del *Big Data*, se realizan múltiples tipos de estudios para los cuales, gracias a los avances de la tecnología, es posible encontrar y consolidar una gran cantidad de información, tanto así, que es posible llegar a tener cientos e incluso miles de variables asociadas a un solo individuo.

Esto es altamente beneficioso para los investigadores ya que esta información puede ser juntada y relacionada con el fin de explicar algún fenómeno o problema, enriqueciendo así los estudios o investigaciones; pero a su vez, también se ha convertido en un gran desafío para encontrar nuevas formas matemáticas, que permitan darle sentido, orden, coherencia y confiabilidad al volumen de información que

hoy día, es posible recolectar y en el caso puntual de las técnicas de regresión, encontrar nuevas formas que permitan desarrollar modelos que se ajusten a las necesidades modernas y puedan explicar de manera precisa los interrogantes del estudio puntual que se esté llevando a cabo, eliminando sesgos, colinealidad y sobreajuste, entre otros.

Dicho lo anterior, como primera medida debe tenerse en cuenta que en la regresión clásica es necesario que el número de variables sea ostensiblemente menor al número de observaciones disponibles, ya que de ocurrir lo contrario, el modelo simplemente no ajustará apropiadamente. A esto se le conoce como el "problema de alta dimensión" ya que el disponer de multitud de información, no repercute necesariamente en mejores resultados y predicciones del modelo, debido a que es posible que gran parte de la información o variables disponibles, no aporten información relevante al modelo, generando ruido. Ahora bien, cuando se tiene un gran número de variables disponibles y estas son incluidas en su totalidad en el modelo planteado, surge otro problema, el cual se relaciona con la interpretación de los coeficientes, ya que se convierte en un gran reto el darles sentido a los resultados obtenidos y llegar así, a una correcta interpretación de los coeficientes de regresión, en el que sea posible definir el aporte individual de cada uno de ellos, sobre la variable de interés en el modelo final.

Dado que las encuestas de hogares tienen tamaños de muestra generalmente grandes a nivel nacional, y que el número de variables está acotado por las que están disponibles en el censo y en la encuesta, los modelos de unidad en SAE generalmente deben buscar, más que la eliminación de variables, la parsimonia en el modelo. Es decir, que las variables involucradas en el modelamiento aporten significativamente en términos del poder predictivo del modelo y dejar de lado las variables que no aportan en ese sentido. En la búsqueda de la parsimonia, una primera opción consiste en generar modelos lineales distintos a partir de diferentes combinaciones de las covariables (con y sin intercepto) y comparar sus respectivas medidas de bondad de ajuste. Adicional a lo anterior, el número de variables significativas, y las medidas de bondad de ajuste, como los criterios de información AIC o BIC, R^2 y R^2 ajustado son algunos de los elementos utilizados para el análisis y comparabilidad de los modelos.

Por otro lado, como la selección de variables es uno de los mayores problemas en todas las técnicas de análisis de datos, ya que, en el trasfondo, el problema real al que se enfrenta el investigador es a un problema de combinatoria, es decir, todo se resumen en elegir la combinación óptima de variables disponibles, que explique de mejor manera la variable de interés. De este modo, con el fin de dar solución a algunos de estos problemas, se implementaron metodologías de penalización o regularización, las cuales son una modificación de los métodos clásicos de regresión lineal principalmente para la selección de variables. Entre los métodos de penalización más conocidos, se encuentran la regresión Ridge y la regresión Lasso. Ahora bien, en el algoritmo de estimación previamente descrito se puede observar que la estimación Montecarlo del indicador FGT depende de la estimación del vector de coeficientes de regresión del efecto fijo del modelo anidado, es decir, depende de $\widehat{\boldsymbol{\beta}}$. Por lo anteriormente descrito, una parte importante de esta etapa consiste en identificar la capacidad predictiva de las variables auxiliares empleadas.

Cabe mencionar que este tipo de métodos de regularización son más empleados en estudios asociados a modelos de área, ya que representan una herramienta útil y potente cuando se tiene un volumen importante de datos, además, Burgard, et. al.(2021). menciona la importancia de aplicar estas técnicas para producir estimaciones estadísticas de área confiables, a la hora de enfrentar errores desconocidos de medición de covariables, siendo sus resultados altamente efectivos.

Regresión Ridge

En los modelos de regresión clásica se emplea el método de mínimos cuadrados, lo que consiste básicamente en minimizar la raíz del error cuadrático realizando una resta entre el valor real y la predicción. En el caso de la regresión Ridge se introduce una función de penalización o regularización, que según Amat

(2016), penaliza la suma de los coeficientes elevados al cuadrado y su propósito principal es el de reducir o contraer el valor de todos estos coeficientes disponibles en el modelo sin llegar a eliminarlos, dicho de otro modo, esta penalización suma al error de la sumatoria de los coeficientes de regresión, lo que resultaría en que a mayor valor de los coeficientes de regresión, mayor será el valor del error. Por ello es necesario, introducir una función que minimice este efecto, el cual se representa por λ, el cual se usa como un parámetro de calibración que controla el efecto relativo del término de penalización, es decir, que cuando $\lambda = 0$, la penalización es nula, suprimiendo la variable respectiva, lo que resulta finalmente, en una regresión clásica por mínimos cuadrados, por otro lado, cuanto mayor peso se le asigna a este λ produce un efecto que hace que los coeficientes se contraigan.

En otro contexto, como primer paso para establecer la viabilidad de un conjunto de covariables, es posible utilizar modelos de regresión Ridge y Lasso adaptado, los cuales constituyen un buen punto de partida para analizar el ajuste de las covariables. Según Fan y Li, (2006) estas metodologías pertenecen a una familia de modelos gaussianos donde, suponiendo que se tiene un vector de observaciones $\boldsymbol{x_j} \in \mathbb{R}^p$ y respuestas $y_i \in \mathbb{R}, i = 1, \dots, n$, y en el que Castro, (2013) establece que la restricción del estimador Ridge se expresa de la siguiente manera:

$$\hat{\beta}^{*(n)} = \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\| \boldsymbol{y} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{x_j}\, \beta_j \right\|^2 \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 \leq s$$

En donde s, representa $\lambda \geq 0$, el cual es el parámetro que denota el grado de penalización del modelo y es tal que a mayor valor más predictores quedan excluidos ($\lambda \geq 0$). El procedimiento de estimación en la regresión Ridge tiene como objetivo identificar el λ óptimo, empleando los coeficientes de regresión como factor de penalización para las covariables y en especial aquellas que menos información aportan al modelo.

Con esto, una gran ventaja de la regresión Ridge, es que al aplicar un valor óptimo de λ se puede lograr una reducción importante en la varianza, sin aumentar de manera significativa el sesgo, lo que traduce en un error total más bajo, lo que no siempre sucede en un modelo de regresión clásico, en el que según Amat (2016), en ocasiones, al existir una relación aproximadamente lineal entre los predictores y la variable de respuesta, existe mínimo sesgo, pero alta varianza, debido a pequeños cambios en los datos, impactando significativamente el modelo resultante. Sin embargo, entre algunas de las desventajas de emplear esta metodología, hace referencia al hecho que el método puede contraer los coeficientes de manera eficiente, pero nunca logra fijarlos en cero (0), con la excepción en el que $\lambda = \infty$, lo cual, sería visto como un modelo nulo. Esto, da como resultado que siempre será generado un modelo en el que se incluyan todas las variables predictoras disponibles originalmente, resolviendo así, el problema de precisión o sobreajuste, pero no el de interpretación.

Regresión Lasso

Al igual que en la regresión Ridge, según Amat (2016), se emplea el parámetro λ, el cual cuando su valor es igual a cero el resultado final corresponde a un modelo de regresión clásico, pero cuando el valor de este parámetro λ aumenta, la penalización sobre los coeficientes es mayor lo que resulta en que un número mayor de variables predictoras sean excluidas del modelo En el procedimiento de estimación se considera la propuesta metodológica de Zou, (2006) que modifica la regresión Lasso al utilizar un conjunto de pesos $w_j \in \mathbb{R}^p$ que permite penalizar aquellos coeficientes garantizando su convergencia a cero.

El objetivo de esta adaptación es resolver la siguiente ecuación:

$$\hat{\beta}^{*(n)} = \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\| \boldsymbol{y} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{x_j}\, \beta_j \right\|^2 + \lambda_n \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \widehat{w_j} |\beta_j|$$

En el caso de la regresión Lasso, el proceso de penalización es diferente ya que la regularización no se hace sobre los coeficientes al cuadrado, sino lo hace sobre la sumatoria del valor absoluto de los coeficientes de regresión o coeficientes, efecto que logra forzar a que los coeficientes lleguen a cero (0). De este modo, se sabe que un predictor cuyo coeficiente es igual a cero (0) su influencia sobre el modelo final es nula, razón por la cual, este método en su proceso de regularización va descartando los coeficientes de regresión que no aportan información relevante al modelo, es decir, variables que, en lugar de sumar valor al resultado, aportan ruido en el resultado final y complican la interpretación final de los resultados, como sucede en el caso de la regresión Ridge.

El procedimiento utilizado en este documento considera un conjunto de factores de penalización que suman uno; puntualmente, $\widehat{w}_j = \frac{|\beta_j|^{-1}}{\sum_{i=1}^{p}|\beta_j|^{-1}}$, para evaluar las variables y poder determinar el mejor modelo lineal. Por las características y definiciones usadas para obtener los modelos respectivos, se incluye el intercepto como variable predictora.

El procedimiento consiste en un proceso iterativo de selección de variables (*stepwise*) en el cual se descartan las covariables con $\beta_i = 0$ $(i = 1, \dots, p)$ al comparar el modelo reducido con aquellos previamente ajustados. La función *glmnet* de la librería homónima en R permite ajustar un modelo lineal generalizado en el proceso de penalización de la regresión Ridge/Lasso empleando el criterio de máxima verosimilitud penalizada para una cuadrícula de valores del parámetro de regularización λ. Así que, el valor óptimo de lambda se selecciona mediante validación cruzada, en el cual se compara el modelo no regularizado con el regularizado en función del error cuadrático medio. Por el contrario, el coeficiente de determinación R^2 y R^2 ajustado se puede estimar de forma específica mediante su respectiva definición y su análisis es similar.

De acuerdo a lo anterior, una ventaja de los modelos Lasso frente a los modelos Ridge, es que los primeros son capaces de identificar variables que no aportan, eliminándolas del modelo, lo que reduce así el tamaño final de este, mejorando el problema en la interpretabilidad, sobre todo en el evento de tener un número muy amplio de variables, mientras que Ridge conserva el conjunto completo de covariables en el resultado final del modelo, en la sección 2 de los anexos, en las figuras 24 y 25, se presenta un ejemplo de la representación gráfica de estos modelos para el país de Chile, en el que se explica la forma correcta de interpretar sus resultados.

Cuadro 5
Resumen métodos de regularización Chile, Colombia y Perú

País	Método	λ_{min}	$Log(\lambda_{min})$	λ_{opt}	$Log(\lambda_{opt})$	R^2	R^2_{adj}	N° Final de variables
Chile	Ridge	0,0377	-32 781	0,1669	-17 904	0,5179	0,51	75
	Lasso	0,0025	-59 915	0,0954	-23 496	0,517	0,499	
Colombia	Ridge	0,0371	-32 941	0,0858	-24 557	0,5237	0,521	81
	Lasso	0,0015	-65 023	0,0344	-33 697	0,5241	0,520	
Perú	Ridge	0,0516	-29 664	0,1435	-19 414	0,6762	0,669	85
	Lasso	0,0012	-67 254	0,029	-35 405	0,6775	0,626	

Fuente: Elaboración propia.

En el cuadro 5, se presentan los valores obtenidos del coeficiente de determinación y el coeficiente de determinación ajustado de los modelos de regresión ajustados, también se incluyen los valores del parámetro de penalización de los modelos estimados. Cabe mencionar, que uno de los objetivos principales al ajustar estos modelos, entre otras cosas, es evitar el sobreajuste del modelo, atenuar el efecto de la correlación entre

predictores y minimizar y/o eliminar la influencia de los predictores menos relevantes y así, de identificar el modelo con el mayor poder predictivo.

Ahora bien, en esta tabla, se tiene para cada uno de los países observados, el resumen de los valores obtenidos para cada valor calculado de λ, el valor mínimo y el valor óptimo sobre cada uno de los métodos de regularización. Ahora bien, en dicha tabla no se observan diferencias realmente significativas en lo que respecta a los coeficientes de determinación, al comparar los valores finales de los modelos Ridge y Lasso para los tres países analizados. En este mismo contexto, el modelo que obtuvo mejor ajuste fue el calculado para Perú obteniendo un R^2 de 0.67. Por otro lado, se resalta la potencia del modelo Lasso, ya que, en medio de su proceso de penalización logró identificar un gran número de variables que no aportaban información relevante al modelo, excluyéndolas de este, y al final del proceso en el caso de Chile, permanecieron 75 variables, en Colombia 81 variables y en el caso del Perú permanecieron 85 variables que logran explicar la variable de interés, pobreza monetaria.

La regresión Lasso, cobra mucho valor cuando es aplicada en estudios asociados a Big Data y políticas públicas en los que es posible integrar una gran cantidad de información proveniente de diversas fuentes y se convierte en un trabajo dispendioso el tratar con situaciones asociadas a la colinealidad y sobreajuste de la información, determinando que covariables pueden aportar información valiosa a la variable respuesta del estudio y cuáles no lo hacen. Por lo cual, con estos métodos, es posible llegar a tener modelos más precisos, óptimos y parsimoniosos, en el que su interpretabilidad sea más sencilla y práctica.

Finalmente, la implementación de estos métodos no debe tomarse como un paso obligatorio o estándar en el ajuste de todo modelo, ya que el uso de estos dependerá principalmente del tipo de datos disponibles, de la cantidad de covariables y el total de observaciones o individuos que se tengan, entre algunas otras cosas. Es por esta razón que en el presente estudio no se empleó ninguno de los métodos de regularización mencionados, entre otras cosas, a que no se cuenta con un número demasiado grande de variables, por lo que no representa un problema que deba ser solucionado. Además, como interés principal se optó por el enfoque predictivo del modelo, el cual se consiguió finalmente, mediante el ajuste de un modelo saturado con el total de variables disponibles provenientes de las encuestas, alcanzando un R^2 más que aceptable. Por otro lado, cabe señalar que el valor obtenido para el coeficiente de determinación fue superior con el modelo saturado que el obtenido con los modelos Ridge y Lasso.

E. Estimación de la varianza de los efectos aleatorios

En la práctica existen diferentes metodologías que permiten estimar los componentes de varianza del término de error (σ_e^2) y de los efectos aleatorios (σ_u^2). Stukel & Rao (1997) proponen el uso de dos regresiones por mínimos cuadrados ordinarios y luego el método de momentos para obtener estimadores insesgados de estos parámetros. Fuller & Battese (1973) proponen una metodología que permite realizar la estimación bajo un muestreo aleatorio simple ($k_{di} = 1$, para todo i y d), en el cual la suma de cuadrados de residuales (SSE_1) con v_1 grados de libertad se obtiene al regresar (por el origen) las desviaciones $k_{di}^{-1}(\boldsymbol{y}_{di} - \overline{\boldsymbol{y}}_{da})$ respecto a las desviaciones no nulas $k_{di}^{-1}(\boldsymbol{x}_{di} - \overline{\boldsymbol{x}}_{da})$ para aquellas áreas con tamaño de muestra $n_d > 1$. Esto conduce al estimador insesgado de σ_e^2 dado por

$$\hat{\sigma}_e^2 = v_1^{-1} SSE_1$$

siendo $v_1 = n - d - p$; en donde p es el número de $\boldsymbol{x}$-desviaciones distintas de cero. En el siguiente paso, se ajusta un modelo de regresión de $\boldsymbol{y}_{di}/k_{di}$ sobre $\boldsymbol{x}_{di}/k_{di}$ con el fin de estimar la suma de cuadrados de residuales SSE_2. Bajo este modelo, un estimador insesgado de σ_u^2 está dado por:

$$\tilde{\sigma}_u^2 = \eta^{-1}[SSE_2 - (n-p)\hat{\sigma}_e^2]$$

Siendo

$$\eta = \sum_{d=1}^{D} a_{d*}\left[1 - a_{d*}\bar{x}'_{da}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} a_{di}x_{di}x'_{di}\right)^{-1}\bar{x}_{da}\right]$$

Los estimadores presentados $\tilde{\sigma}_u^2$ y $\hat{\sigma}_e^2$ resultan equivalentes a los obtenidos por el método atribuido a Henderson, (1953), el cual requiere una regresión por mínimos cuadrados ordinarios sobre $p + d$ variables en contraste con el método de transformación que es implementado sobre p variables, y por lo tanto se vuelve computacionalmente más lento a medida que aumenta el número de áreas pequeñas. Esta metodología tiene el problema de que la estimación de la varianza de los efectos aleatorios $\hat{\sigma}_u^2$ puede tomar valores negativos y una posible solución consiste en truncar este valor a cero, es decir, $\hat{\sigma}_u^2 = max\,(\tilde{\sigma}_u^2\,, 0)$. El estimador truncado $\hat{\sigma}_u^2$ es consistente al aumentar el número de áreas, pero a su vez pierde la propiedad de insesgamiento. En Battese, et. al. (1988) se presentan estimadores alternativos para la varianza del término de error (σ_e^2) y de los efectos aleatorios (σ_v^2) que resultan ser aproximadamente insesgados.

Los métodos de estimación habituales, concretamente máxima verosimilitud (ML), máxima verosimilitud residual (REML) y el método de Henderson III, satisfacen estas condiciones. En el Cuadro 6 se presentan los componentes de varianza estimados por los modelos SAE ajustados para cada uno de los países.

Cuadro 6
Componentes de varianza estimados por país

Componente	Chile	Colombia	Perú
σ_v^2	0,0096	0,0191	0,0144
σ_e^2	0,2966	0,2985	0,2975

Fuente: Elaboración propia.

Como primera medida, es pertinente recordar que la varianza es una medida de dispersión empleada principalmente para describir la variación de un conjunto de datos. Para el caso de un modelo de regresión, la varianza se emplea para determinar cuánto de la variación existente en la variable de interés puede explicarse por cada uno de los coeficientes del modelo. De esta manera, a menor valor de la varianza se obtiene mejor poder predictivo con el modelo y, por el contrario, a mayor valor de la varianza las predicciones del modelo pueden llegar a ser menos precisas.

Dicho lo anterior, se observa que la varianza del modelo para cada uno de los tres países observados es baja, lo que indica que el ajuste de este modelo es bueno, pero cabe resaltar, que los resultados obtenidos con el modelo para Chile frente a los resultados de Colombia y Perú es mejor, dado que el valor obtenido de la varianza es más bajo, lo que finalmente, resulta en mejores predicciones. Por otra parte, si bien el modelo construido en Colombia ajustó bien, es posible que las predicciones en los municipios no sean tan precisas, como las predicciones realizadas en las provincias de Perú o las comunas de Chile, ya que la varianza de este modelo es mayor, que la obtenida para los otros dos modelos.

En el siguiente capítulo se presenta una extensión del proceso de estimación de los componentes de varianza del modelo de errores anidados presentado en este capítulo.

III. Estimación del error en el modelo EBP

Por definición, el error cuadrático medio (ECM) es una medida de desempeño de un estimador que mide el promedio de los errores al cuadrado, en términos de la diferencia entre el parámetro y la estimación. Esta medida es ampliamente utilizada para validar la calidad estadística de las estimaciones e identificar aquellas áreas sobre las cuales se podrá realizar una inferencia apropiada. Las áreas que no pasen los criterios de calidad mínimos deberían ser excluidas de cualquier conclusión, hipótesis o inferencia y no deberían ser incluidas en el mapa de pobreza (Gutiérrez et. al. (2020).

A. El estimador de Prasad y Rao

En el capítulo anterior se presenta la fundamentación teórica de la metodología Pseudo – EBP con la cual se generaron las mejores estimaciones lineales insesgadas empíricas de los ingresos medios (EBLUP), para poder dar cuenta de la tasa de pobreza y de pobreza extrema bajo el modelo de regresión lineal de errores anidados. Bajo las condiciones de regularidad definidas por Prasad & Rao, (1990) donde las constantes que representan la posible heteroscedasticidad (k_{di}) y el tamaño de muestra en el área (n_d) están uniformemente acotados y los elementos de X_i están uniformemente acotados de manera que $\boldsymbol{A_i} := \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$ es $[O(1)]_{p\times p}$.

Por lo tanto, bajo el supuesto de normalidad sobre el efecto aleatorio $\boldsymbol{u_d} \overset{\boldsymbol{i.i.d}}{\sim} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_u^2})$ y el término de error a nivel de individuo $\boldsymbol{e_{di}} \overset{\boldsymbol{iid}}{\sim} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_e^2})$, los autores obtuvieron una aproximación del error cuadrático medio del estimador FH que, según Corral-Rodas, et. al., (2021) y Molina, (2019) puede ser generalizado para el modelo y los parámetros mencionados en este documento. Esta aproximación se reduce a

$$MSE(\hat{F}_{ad}^{\mathrm{EB}}) \approx g_{1\mathrm{d}}(\boldsymbol{\delta}) + g_{2\mathrm{d}}(\boldsymbol{\delta}) + g_{3\mathrm{d}}(\boldsymbol{\delta})$$

donde $\boldsymbol{\delta} = (\sigma_u^2, \sigma_e^2)'$ corresponde al vector de componentes de varianza del modelo. Los términos de la ecuación anterior son, respectivamente:

$$g_{1d}(\boldsymbol{\delta}) = \gamma_d\left(\frac{\sigma_e^2}{a_d.}\right)$$

El cual es un término que, al aumentar el número total de dominios D, es de orden $O(1)$ y está asociado al error debido a la predicción del efecto aleatorio del área u_d. El segundo término $g_{2d}(\sigma_u^2)$ es el error debido a la estimación del vector de coeficientes de regresión $\boldsymbol{\beta}$

$$g_{2d}(\boldsymbol{\delta}) = (\overline{\mathbf{X}}_{\mathrm{d}} - \gamma_d \overline{\boldsymbol{x}}_{da})' \left(\sum_{d=1}^{D} A_i \right)^{-1} (\overline{\mathbf{X}}_{\mathrm{d}} - \gamma_d \overline{\boldsymbol{x}}_{da})$$

Mientras que $g_{3d}(\sigma_u^2)$ es el error debido a la estimación de la varianza σ_u^2

$$g_{3d}(\boldsymbol{\delta}) = a_{d*}^{-2}(\sigma_u^2 + \sigma_e^2/a_{d*})^{-3}\, h\,(\sigma_u^2, \sigma_e^2)$$

donde

$$a_{d*} = \sum\nolimits_{i \in s_d} k_{di}^{-2}$$

siendo

$$h(\boldsymbol{\delta}) = \sigma_e^4 \bar{V}_{uu}(\boldsymbol{\delta}) + \sigma_e^4 \bar{V}_{ee}(\boldsymbol{\delta}) - 2\sigma_e^2\, \sigma_v^2 \bar{V}_{ue}(\boldsymbol{\delta})$$

En donde $\bar{V}_{ee}(\boldsymbol{\delta})$, $\bar{V}_{uu}(\boldsymbol{\delta})$ y $\bar{V}_{ue}(\boldsymbol{\delta})$ denotan, respectivamente, las varianzas y covarianza asintótica entre los estimadores $\hat{\sigma}_e^2$ y $\hat{\sigma}_v^2$. Según Molina, (2019), los dos últimos términos de la ecuación -$g_{2d}(\sigma_u^2)$ y $g_{3d}(\sigma_u^2)$- son de orden $O(D^{-1})$; es decir, que para un gran número de áreas D estos términos tienden a ser nulos. Por el contrario, para un número de áreas moderado es necesario tener en cuenta los tres términos para evitar la subestimación del ECM. En Molina, et. al., (2019), se puede identificar que $\bar{V}_{uu}(\boldsymbol{\delta})$ es un término que depende del método de estimación utilizado. Bajo el método de Henderson III, se tienen las siguientes expresiones para las varianzas asintóticas

$$\bar{V}_{uu}(\boldsymbol{\delta}) = 2\eta_1^{-2}[\nu_1^{-1}\ (n - p - \nu_1)\,(n - p)\,\sigma_e^4 + \eta_2 \sigma_u^4 + 2\eta_1 \sigma_e^2\, \sigma_u^2]$$

$$\bar{V}_{ee}(\boldsymbol{\delta}) = 2\nu_1^{-1}\sigma_e^4$$

Mientras que la covarianza entre los estimadores:

$$\bar{V}_{ue}(\boldsymbol{\delta}) = \bar{V}_{eu}(\boldsymbol{\delta}) = -2\eta_1^{-1}\nu_1^{-1}\,(n - p - \nu_1)\sigma_e^4$$

Donde η_1 y η_2 se definen a partir de las ecuaciones

$$\eta_1 = \sum_{d=1}^{D} a_{d*}[1 - a_{d*}\bar{x}'_{da}A_1^{-1}\bar{x}_{da}]$$

$$\eta_2 = \sum_{d=1}^{D} a_{d*}^2(1 - 2a_{d*}\overline{\boldsymbol{x}}'_{da}A_1^{-1}\overline{\boldsymbol{x}}_{da}) + tr\left[\left(A_1^{-1}\sum_{i=1}^{m} 2a_{d*}\overline{\boldsymbol{x}}_{da}\overline{\boldsymbol{x}}'_{da}\right)^2\right]$$

Con

$$A_1 = \sum_{d=1}^{D}\sum_{i=1}^{n_d} a_{di}x_{di}x'_{di}$$

Para el caso de errores homocedásticos, donde $k_{di} = 1$ y $a_{di} = k_{di}^{-2} = 1$, las varianzas asintóticas $\bar{V}_{ee}(\boldsymbol{\delta})$, $\bar{V}_{uu}(\boldsymbol{\delta})$ y la covarianza asintótica convergen a las expresiones obtenidas por Prasad & Rao, (1990). Por el contrario, si se utiliza el método de máxima verosimilitud (ML) o de máxima verosimilitud restringida (REML) para estimar σ_u^2 y σ_e^2, es decir, en el caso donde $\hat{\delta} = \hat{\delta}_{ML}$ o $\hat{\delta} = \hat{\delta}_{RE}$ se tiene una matriz de varianzas-covarianzas asintótica $\bar{V}(\delta) = \bar{V}_{ML}(\delta) = \bar{V}_{RE}(\delta)$ dada por la inversa de la matriz de información de Fisher $I(\boldsymbol{\delta})$ cuya diagonal principal está conformada por:

$$\mathrm{I}_{uu}(\boldsymbol{\delta}) = \frac{1}{2}\sum_{d=1}^{D} tr\left[\left(V_d^{-1}\frac{\partial V_d}{\partial \sigma_u^2}\right)^2\right],$$

$$\mathrm{I}_{ee}(\boldsymbol{\delta}) = \frac{1}{2}\sum_{d=1}^{D} tr\left[\left(V_d^{-1}\frac{\partial V_d}{\partial \sigma_e^2}\right)^2\right]$$

Mientras que, por fuera de la diagonal, los términos están dados por la siguiente ecuación

$$\mathrm{I}_{ue}(\boldsymbol{\delta}) = \mathrm{I}_{eu}(\boldsymbol{\delta}) = \frac{1}{2}\sum_{d=1}^{D} tr\left[V_d^{-1}\left(\frac{\partial V_d}{\partial \sigma_u^2}\right)V_i^{-1}\left(\frac{\partial V_d}{\partial \sigma_e^2}\right)\right],$$

Donde

$$\frac{\partial V_d}{\partial \sigma_u^2} = \mathbf{1}_{n_d}\mathbf{1}'_{n_d}$$

$$\frac{\partial V_d}{\partial \sigma_e^2} = \sigma_e^{-2}R_i.$$

Ahora bien, como

$$V_d^{-1} = \frac{1}{\sigma_e^2}\left[diag_{1\le i\le n_d}(a_{di}) - \frac{\gamma_{\mathrm{d}}}{a_{d*}}\mathbf{a}_d\mathbf{a}_d^T\right]$$

Entonces, la matriz de información de Fisher se reduce al siguiente sistema de ecuaciones

$$\mathrm{I}_{uu}(\boldsymbol{\delta}) = \frac{1}{2}\sum_{d=1}^{D} a_{d*}^2\alpha_d^{-2}$$

$$\mathrm{I}_{ee}(\boldsymbol{\delta}) = \frac{1}{2}\sum_{d=1}^{D}[(n_d-1)\sigma_e^{-4} + \alpha_i^{-2}]$$

$$\mathrm{I}_{ue}(\delta) = \frac{1}{2}\sum_{i=1}^{m} a_{d*}^2\alpha_i^{-2}$$

Donde

$$\alpha_i = \sigma_e^2 + a_i.\sigma_v^2.$$

Nuevamente, para el caso de errores homocedásticos, donde $k_{di} = 1$ y $a_{di} = k_{di}^{-2} = 1$, las expresiones para la matriz de información de Fisher $\mathrm{I}_{uu}(\delta)$ y $\mathrm{I}_{ue}(\delta)$ converge a las expresiones teóricas obtenidas por Datta & Lahiri, (2000).

B. Métodos no paramétricos en la estimación del ECM

Para los indicadores calculados mediante simulaciones de Montecarlo se utilizó el método *Bootstrap* paramétrico descrito por Molina & Rao, (2010) para realizar las estimaciones de los errores cuadráticos medios basado en el método *Bootstrap* para poblaciones finitas de González-Manteiga, et. al., (2008). En este documento se considera el método *Bootstrap* paramétrico para estimar el ECM de predictor Census-EB del parámetro FGT de orden α $\left(F_{\alpha,d}\right)$ asociado al ingreso medio de los hogares. El algoritmo considerado describe una ligera modificación al publicado en (Molina, 2019) y tiene los siguientes pasos:

- Ajustar el modelo de errores anidados descrito anteriormente y obtener el vector de estimadores
$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2, \hat{\sigma}_e^2\right)'$$
- Generar efectos *Bootstrap* para los dominios de interés d a partir de la densidad normal dada por
$$u_d^{*(b)} \sim_{iid} N(0, \hat{\sigma}_u^2), \qquad d = 1, 2, \dots, D$$
- Generar los errores *Bootstrap* de forma similar pero independiente a $u_d^{*(b)}$, empleando la densidad normal dada por
$$e_{di}^{*(b)} \sim_{iid} N(0, \hat{\sigma}_e^2), i = 1, 2, \dots, N_d; d = 1, 2, \dots, D$$
- Generar la población o pseudocenso *Bootstrap* de la variable respuesta empleando los hogares del censo ubicados en las provincia, comuna y municipios seleccionados en las encuestas de hogares de cada país y la ecuación del modelo
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}^T \widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}$$
En donde, $u_d^{*(b)}$ $_{iid} N(0, \hat{\sigma}_u^2)$ fue generado en el paso 2 y $e_{di}^{*(b)}$ $_{iid} N(0, \hat{\sigma}_e^2)$ en el paso 3.
- Calcular el indicador FGT de interés a partir del censo *Bootstrap* $\tilde{F}_{\alpha d}^{*(b)} = F_{\alpha d}\left(\boldsymbol{y}_d^{*(b)}\right)$ para todo dominio de interés d. Nótese que $\boldsymbol{y}_d^{*(b)} = \left(Y_{d1}^{*(b)}, \dots, Y_{dN_d}^{*(b)}\right)^T$ es el vector censal de la variable respuesta para el área d .
- Para los mismos dominios seleccionados originalmente en las encuestas de hogares de cada país ($s = s_1 \cup \dots \cup s_D$) se hace una selección de hogares dentro del vector censal de la variable respuesta $\boldsymbol{y}_d^{*(b)}$ del mismo tamaño de muestra establecido para la encuesta de hogares. La selección garantiza siempre que los mismos índices de los elementos en la muestra original sean incluidos en la muestra *Bootstrap*; es decir, se sigue el mismo comportamiento que la muestra en la encuesta de hogares en cada una de las subdivisiones de interés. Esta selección
$$\boldsymbol{y}_s^{*(b)} = \left[\left(\boldsymbol{y}_{1s}^{*(b)}\right)^T, \dots, \left(\boldsymbol{y}_{Ds}^{*(b)}\right)^T\right]^T$$
se conoce como pseudomuestra.
- Sobre la muestra de hogares seleccionados se realiza la estimación del modelo de efectos mixtos, de la misma forma como el usado en el paso 1 de este algoritmo. Al ajustar de nuevo el modelo de errores anidados descrito en la sección D. sobre los datos de la pseudomuestra, se obtienen los predictores EB *Bootstrap* para el indicador FGT de interés en el área d, es decir, $\hat{F}_{\alpha d}^{EB*(b)}$ para $d = 1, 2, \dots, D$.
Repetir los pasos del 2 al 7, $b = 1,2, \dots B,$ veces para completar el proceso de simulación *Bootstrap*. El estimador del error cuadrático medio está dado por:

$$MSE_B\left(\hat{F}_{\alpha d}^{EB}\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{F}_{\alpha d}^{EB*(b)} - \tilde{F}_{\alpha d}^{*(b)}\right)^2, \qquad d = 1, \dots, D$$

En la anterior ecuación, B corresponde al número de simulaciones *Bootstrap* realizadas, D es el número de provincias, comunas y municipios seleccionados en las encuestas de hogares de cada país, $\hat{F}_{\alpha d}^{EB*(b)}$ corresponde a los estimados por municipio del proceso de montecarlo y $\tilde{F}_{\alpha d}^{*(b)}$ los indicadores FGT estimados en el pseudocenso por dominio de interés.

De este modo, con el error cuadrático medio estimado se puede calcular una medida de variación relativa, similar al coeficiente de variación, con el fin de tener un criterio objetivo que permita definir la calidad de las estimaciones. El coeficiente de variación se define como sigue a continuación:

$$\widehat{CV} = \frac{\sqrt{MSE_B\left(\hat{F}_{\alpha d}^{EB}\right)}}{\hat{F}_{\alpha d}^{\text{EB}}} * 100$$

Se recomienda que las provincias, comunas y municipios con un $\widehat{CV}$ mayor a 30% sean excluidas de los mapas de pobreza al considerarse que no tienen la precisión requerida.

C. Intervalos de confianza

Es posible considerar dos metodologías diferentes para la estimación por intervalo de los indicadores FGT de interés en los dominios considerados para cada uno de los países bajo análisis en este documento. El primer enfoque consiste en un intervalo de confianza tipo Wald establecido a partir del enfoque de Datta & Lahiri, (2000) y Prasad & Rao, (1990) para la estimación del error cuadrático medio del predictor Census-EB del parámetro FGT de orden α asociado al ingreso medio de los hogares, el cual se obtiene a partir de la distribución normal teórica del estimador como se presenta a continuación:

$$\hat{F}_{\alpha d}^{EB} \pm Z_{\alpha/2} \times RMSE(\hat{F}_{\alpha d}^{\text{EB}})$$

En donde $RMSE(\hat{F}_{\alpha d}^{\text{EB}})$ es la raíz del error cuadrado medio obtenido analíticamente. Por el contrario, el segundo enfoque considera el método de *Bootstrap* paramétrico presentado en la sección B del presente capítulo para estimar el ECM de predictor Census-EB del parámetro FGT de orden α $\left(F_{\alpha,d}\right)$. El intervalo de confianza del $(1-\alpha)\%$ bajo este enfoque está dado por medio de la siguiente expresión

$$\hat{F}_{\alpha d}^{EB} \pm Z_{\alpha/2} \times RMSE_B(\hat{F}_{\alpha d}^{\text{EB}})$$

Donde $Z_{\alpha/2}$ es el cuantil $\alpha/2$ superior de una distribución normal estándar y $\hat{F}_{\alpha d}^{EB}$ es el EBLUP del indicador FGT de orden α, mientras que $RMSE_B(\hat{F}_{\alpha d}^{\text{EB}})$ es la raíz del error cuadrado medio obtenido mediante *Bootstrap*.

Por otro lado, es muy importante que los conceptos anteriormente descritos, se entiendan e interpreten de la mejor manera, ya que una correcta interpretación de los resultados arrojados por el modelo ajustado, concluye un proceso, dispendioso y complejo, que permita llegar a conclusiones e inferencias acertadas. Por esta razón en el Cuadro 7 se extrajo una porción de las tablas con las salidas de los modelos para cada uno de los países observados, en la cual se especifican las estimaciones de pobreza obtenidas para las unidades geográficas correspondientes a provincias en el caso de Perú, comunas en el caso de Chile y municipios en el caso de Colombia; las cuales van acompañadas del respectivo coeficiente de variación, error cuadrático medio y su respectivo intervalo de confianza, los cuales se explican columna a columna a continuación.

Cuadro 7
Estimaciones pobreza, MSE, CV e IC para Chile, Colombia y Perú

País	Departamento o provincia	Municipio, provincia o comuna	Estimación pobreza	MSE pobreza	CV pobreza	LI	LS
Chile	Limarí	Monte Patria	18,45%	3,35E-04	9,9%	14,9%	22,0%
	Santiago	La Pintana	15,82%	2,12E-04	9,2%	13,0%	18,7%
	Valparaíso	Valparaíso	8,46%	2,49E-05	5,9%	7,5%	9,4%
	Arica	Arica	8,12%	1,58E-05	4,9%	7,3%	8,9%
	Cordillera	Puente Alto	7,34%	2,78E-05	7,2%	6,3%	8,4%

País	Departamento o provincia	Municipio, provincia o comuna	Estimación pobreza	MSE pobreza	CV pobreza	LI	LS
	Iquique	Iquique	5,89%	1,69E-05	7,0%	5,1%	6,7%
Colombia	Chocó	Alto Baudó	80,64%	5,26E-04	2,8%	76,2%	85,1%
	Nariño	Magüí Payán	77,25%	1,54E-03	5,1%	69,6%	84,9%
	La Guajira	Uribia	75,21%	2,40E-04	2,1%	72,2%	78,2%
	Cundinamarca	Chía	6,24%	1,78E-05	6,8%	5,4%	7,1%
	Antioquia	Sabaneta	3,78%	7,18E-06	7,1%	3,3%	4,3%
	Antioquia	Envigado	2,83%	1,77E-06	4,7%	2,6%	3,1%
Perú	Loreto	Datem del Marañón	42,73%	5,34E-04	5,4%	38,2%	47,3%
	Cusco	Paucartambo	42,31%	6,12E-04	5,9%	37,5%	47,2%
	Pasco	Daniel Alcides Carrión	40,96%	6,39E-04	6,2%	36,0%	45,9%
	San Martín	San Martín	15,82%	1,06E-04	6,5%	13,8%	17,8%
	Tumbes	Tumbes	8,67%	2,30E-04	17,5%	5,7%	11,6%
	Áncash	Santa	13,95%	5,00E-05	5,1%	12,6%	15,3%

Fuente: Elaboración propia.

Como primera medida, en el cuadro 7 se observan las columnas "País" y "Departamento o provincia", las cuales no hacen parte de la salida particular de este modelo, pero se agrega con el fin de identificar con mayor precisión a qué lugar en particular se refiere la estimación realizada. Luego se observa la columna "Municipio, Provincia o Comuna", la cual hace referencia a la unidad de observación sobre la cual se están calculando lo valores de la estimación asociado a la variable de interés, que en este caso es pobreza. La cuarta columna hace referencia a la estimación calculada sobre los niveles de pobreza para cada una de las unidades de observación de la lista, que para este caso particular se ajustaron para tres países de Sur América, Chile, Colombia y Perú.

Luego, se tiene la columna que hace referencia al error cuadrático medio, el cual es una medida de dispersión que se aplica sobre el error del pronóstico, es decir, que su propósito es el de establecer la distancia de separación entre el pronóstico hecho y el valor real. A partir de este valor se calcula el coeficiente de variación en el que se toma su raíz cuadrada y se divide entre la estimación obtenida. La siguiente columna que se relaciona, hace referencia al coeficiente de variación, el cual es calculado partiendo del error cuadrático medio, sacando su raíz cuadrada y luego dividiéndola sobre la estimación calculada. Esta medida es empleada para correlacionar la desviación estándar de un conjunto de datos respecto a su media aritmética. Es decir, que este coeficiente, es la variación media esperada del conjunto de datos con respecto a la media aritmética, razón por lo cual, se esperan valores pequeños del CV (ojalá por debajo del 15%) de esta manera las estimaciones se consideran confiables debido a su poca variabilidad y mayor precisión, a su vez, valores grandes de CV gran variabilidad en las estimaciones, asumiendo sesgo y poca precisión.

Las últimas dos columnas detallan los limites superior e inferior que pueden tomar las estimaciones calculadas, lo que se conoce como intervalo de confianza, el cual hace referencia a un rango de valores entre los que se encuentra el valor real de la variable que se está estimando; de este modo, a valores más pequeños del coeficiente de variación el intervalo de confianza se hace más estrecho.

Dicho esto, de acuerdo a lo observado en la tabla anterior, se estima que la comuna de Monte Patria en la provincia de Limarí que hace parte de la región de Coquimbo en Chile es de 18.4% con un coeficiente de variación del 9,93% y un intervalo de confianza entre el 14,86% y 22,03%, es decir que el valor real de habitantes en condición de pobreza monetaria en la comuna oscila entre los porcentajes mencionados. La estimación de pobreza monetaria para la comuna de Arica, capital de la provincia homónima en la región de

Arica y Parinacota es del 8.1% con un coeficiente de variación del 4,90% y un intervalo de confianza entre el 7,34% y el 8,90%, es decir que el valor real del porcentaje de pobreza de la comuna se encuentra entre estos porcentajes. Ahora bien, este intervalo de confianza es más estrecho que el anterior, debido a que el coeficiente de variación a su vez es más pequeño, lo que implica, una menor variación existente en la estimación calculada.

IV. Criterios de información y validación de los supuestos del modelo

En esta metodología para la construcción de mapas de pobreza en los tres países de interés, la selección y validación de los supuestos de un modelo se consolida como una actividad importante y recurrente en el proceso de estimar indicadores FGT a partir del ingreso de los hogares en los dominios de interés en cada país. Puntualmente, se considera que la selección inadecuada de un modelo invalida todos los esfuerzos metodológicos en el proceso de estimación de parámetros y generación de pronósticos. El procedimiento ilustrado en el presente capítulo busca garantizar un equilibrio entre la bondad de ajuste del modelo y su complejidad; en otras palabras, un enfoque que permite identificar el modelo con mejor aproximación a los datos recolectados (aquel que capture sus regularidades subyacentes) en los países, su complejidad y el grado de validez de los resultados que genera respecto a los supuestos intrínsecos sobre los que se soporta teóricamente.

En la primera parte del capítulo se presentan los diferentes criterios de selección de modelos empleados en el proceso metodológico del enfoque pseudo-EBP. Según Kohavi, (1995) y Arlot & Celisse, (2010) los enfoques más empleados en la selección de modelos son los métodos de validación cruzada, el *Bootstrapping,* el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). En el procedimiento implementado en este documento no se considera el uso de los métodos de validación cruzada al ser metodologías computacionalmente costosas en la generación y validación del error del modelo a la luz de la magnitud de los datos que deben ser procesados; por el contrario, los criterios AIC y BIC no requieren validación y son computacionalmente eficientes.

Tampoco se consideran pruebas formales de bondad de ajuste puesto que en algunos casos estas resultan inapropiadas según la estructura del componente de varianza de los modelos contrastados; por el contrario, Bishop, (2006) sugiere que las pruebas de bondad pueden ser poco informativas a la hora de medir la capacidad de los modelos para capturar la regularidad en los datos e identificar el más adecuado y parsimonioso - Burnham & Anderson, (2002) y Konishi & Kitagawa, (2008)-. En la literatura relacionada se han establecido un conjunto de criterios de información ampliamente utilizados en la práctica debido a sus propiedades teóricas y facilidad de ser implementadas. En este capítulo se describen aquellas que fueron

empleadas para seleccionar el modelo más parsimonioso y así capturar la relación entre el indicador FGT y el conjunto de variables auxiliares consolidado para cada uno de los países bajo análisis.

Adicional a lo previamente expresado, se presenta la secuencia de pasos para validar los supuestos del modelo pseudo-EBP e identificar puntos de balanceo y valores atípicos. Finalmente, se realiza un contraste entre las estimaciones directas y los pronósticos del modelo seleccionado en cada uno de los países,

A. Criterios de selección de modelos

En la actualidad existen diferentes enfoques metodológicos para la selección de modelos lineales mixtos parsimoniosos; es decir, aquel con la menor cantidad de variables dependientes que den cuenta de la mayor parte del componente de variabilidad de la variable respuesta - Stroup, (2013)-. Según Rao & Molina, (2015), los métodos más empleados en la práctica para la inclusión de efectos fijos, es decir, selección de variables en los modelos de estimación en áreas pequeñas incluyen el Criterio de Información de Akaike (AIC, por sus siglas en inglés) propuesto Akaike, (1974), el Criterio de Información Bayesiano (BIC, por sus siglas en inglés) propuesto por Schwarz, (1978), el Criterio de Información de Akaike corregido ($AICc$) propuesto por Burnham y Anderson, (2002); Burnham y Anderson, (2004) realiza una corrección del criterio AIC para tamaño de muestra pequeños, los métodos secuenciales de selección de variables ("hacia adelante", "hacia atrás" y "paso a paso") y la estadística C_p de Mallows —Buscemi y Plaia, (2020). Las extensiones de estos métodos al modelo lineal mixto se presentan con mayor detalle en Müller, et. al., (2013) y Stroup, (2013). En esta sección se presenta una breve descripción de estos enfoques.

Criterio de Información de Akaike

El origen del AIC se remonta a la propuesta de Kullback y Leibler, (1951) en la cual se presenta una medida de discrepancia entre dos modelos: el modelo "verdadero" que depende sólo de los datos $f(\boldsymbol{y})$ y un modelo "aproximado" que depende de los datos y los parámetros a estimar $g(\boldsymbol{y}|\boldsymbol{\theta})$. Konishi y Kitagawa, (2008), definen la medida de información $K-L$ para el caso de variables continuas, mediante la siguiente expresión:

$$I(f,g) = E_f\left[log\left(\frac{f(\boldsymbol{y})}{g(\boldsymbol{y}|\boldsymbol{\theta})}\right)\right]$$

Argumentando que esta medida es teóricamente similar a una estadística de razón de verosimilitud. Las propiedades de esta medida son directas; en el caso de tener un modelo aproximado perfecto $f(\boldsymbol{y}) = g(\boldsymbol{y}|\boldsymbol{\theta})$ y, así generar una razón logarítmica nula $I(f,g) = 0$. La discrepancia entre los modelos parte de establecer $E_f\left[\log\left(f(\boldsymbol{y})\right)\right] = c$, donde c es una constante por definir, en la ecuación:

$$I(f,g) = c - E_f\left[\log\left(g(\boldsymbol{y}|\boldsymbol{\theta})\right)\right]$$

Al despejar se obtiene $-E_f\left[\log\left(g(\boldsymbol{y}|\boldsymbol{\theta})\right)\right] = I(f,g) - c$, la distancia $K-L$ que corresponde a la expresión $I(f,g) - c$. Akaike, (1974) demostró que $-2\ell(\widehat{\boldsymbol{\theta}}) + 2k$ proporciona una buena aproximación a la distancia $K-L$ para el modelo de regresión lineal simple; donde $\ell(\widehat{\boldsymbol{\theta}})$ y k, denotan la log-verosimilitud alcanzada por el modelo ajustado y el número de parámetros en el vector $\boldsymbol{\theta}$, respectivamente. En la literatura relacionada, a la expresión:

$$AIC = -2(\ell\left(\widehat{\boldsymbol{\theta}}\right) - k)$$

se le conoce como el criterio de información de Akaike. Por el contrario, en los modelos lineales generalizados mixtos (GLMM) el vector de parámetros incluye aquellos asociados al componente de varianzas y los asociados a la parte fija del modelo. Rao & Molina, (2015) argumentan que para el caso del modelo lineal mixto general (sobre el cual reposa la teoría SAE) se puede emplear la log-verosimilitud completa $\ell(\boldsymbol{\beta}, \boldsymbol{\delta})$ o la log-verosimilitud restringida, $\ell_R(\boldsymbol{\delta})$, para estimar el Criterio de Información de Akaike; en estas expresiones $\boldsymbol{\beta}$ el vector de coeficientes de regresión de efectos fijos y $\boldsymbol{\delta} = (\sigma_u^2, \sigma_e^2)'$ el vector de componentes

de varianza del modelo. El Criterio de Información de Akaike bajo normalidad de los efectos aleatorios u_d y los errores e_{di} en el modelo GLMM empleando la log-verosimilitud completa está dado por la siguiente ecuación —Stroup, (2013):

$$AIC = -2\ell\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML}\right) + 2(q^* + p)$$

Siendo $\ell\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML}\right)$ la log-verosimilitud completa evaluada en $\widehat{\boldsymbol{\beta}}_{ML}$, $\widehat{\boldsymbol{\delta}}_{ML}$ que son los estimadores de máxima verosimilitud de $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$, respectivamente, p denota la longitud del vector de efectos fijos $\boldsymbol{\beta}$ y q^* es el número efectivo de parámetros del componente de varianzas $\boldsymbol{\delta}$. En el escenario donde se emplea la log-verosimilitud restringida $\ell_R(\boldsymbol{\delta})$, el criterio AIC está dado por

$$AIC = -2\ell_{\mathrm{R}}\left(\widehat{\boldsymbol{\delta}}_{ML}\right) + 2(q^*)$$

El AIC es un criterio que penaliza el ingreso de nuevos parámetros a un modelo si estos no generan un aumento considerable en su log-verosimilitud. Bajo este criterio, se prefieren los modelos con los valores AIC más pequeños.

Criterio de Información de Akaike corregido

En las aplicaciones prácticas donde se tengan tamaños de muestra pequeños, Burnham & Anderson, (2002) recomiendan emplear la siguiente corrección sobre el criterio AIC

$$AICc = -2\ell(\widehat{\boldsymbol{\theta}}) + \frac{2(q^* + p)n^*}{n^* - (q^* + p) - 1}$$

donde

$$n^* = \begin{cases} n, & \text{si } \ell\left(\widehat{\boldsymbol{\theta}}\right) = \ell\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML}\right) \\ n - p, & si\ \ell\left(\widehat{\boldsymbol{\theta}}\right) = \ell_{\mathrm{R}}\left(\widehat{\boldsymbol{\delta}}_{ML}\right) \end{cases}$$

Criterio de Información de Akaike condicional

Rao y Molina, (2015) afirman que, para los modelos de estimación en áreas pequeñas, donde el proceso de estimación se enfoca en el efecto aleatorio por dominio y los parámetros de regresión $\boldsymbol{\beta}$, el AIC condicional tiene un mejor desempeño que el AIC tradicional y el AIC corregido. El AIC condicional (usualmente denotado $cAIC$), se define a partir de la log-verosimilitud condicionada sobre el efecto aleatorio $\ell(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{u})$ empleando la siguiente expresión:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{u}) = c - \frac{1}{2}[\log|\boldsymbol{R}| + (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z}\mathbf{u})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z}\mathbf{u})]$$

Para un valor particular del vector de componentes de varianza $\boldsymbol{\delta}$ se puede definir una matriz de proyección ortogonal ($\boldsymbol{H_1}$) a partir de las soluciones de las ecuaciones del modelo mixto ($\boldsymbol{\beta}^*$ y $\mathbf{u}^*$)

$$\boldsymbol{H_1 y} = \boldsymbol{X\beta}^* + \boldsymbol{Z}\mathbf{u}^*$$

las cuales según Stroup, (2013) están dadas por:

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{u}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y} \end{bmatrix}$$

Siendo $\boldsymbol{X}$ e $\boldsymbol{Z}$ la matriz de diseño de la parte fija y el efecto aleatorio de un modelo lineal mixto tradicional, $\boldsymbol{G}$ y $\boldsymbol{R}$ las matrices de covarianza asociadas al vector de componentes de varianza $\boldsymbol{\delta}$. Según Müller, et. al., (2013), los grados de libertad efectivos a ser empleados en el proceso de estimación de $\boldsymbol{\beta}$ y $\boldsymbol{u}$ se definen por medio de la siguiente expresión

$$\varrho(\boldsymbol{\delta}) = tr[\boldsymbol{H}_1(\boldsymbol{\delta})] = tr[(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}RV^{-1}X}] + n - tr(\boldsymbol{RV^{-1}})$$

El AIC condicional está dado por

$$\boldsymbol{cAIC} = -2l\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML} | \, \widehat{\mathbf{u}}^{\boldsymbol{H}}\right) + 2\left[\varrho\left(\widehat{\boldsymbol{\delta}}_{ML}\right) + q^{*}\right]$$

Donde $\widehat{\mathbf{u}}^{\boldsymbol{H}} = \widetilde{\mathbf{u}}(\widehat{\boldsymbol{\delta}}_{ML})$ y $\varrho(\widehat{\boldsymbol{\delta}}_{ML})$ son los estimadores EBLUP del efecto aleatorio y los grados de libertad $\varrho(\boldsymbol{\delta})$, respectivamente. En Vaida y Blanchard, (2005) se propone la siguiente actualización metodológica del término de penalización en el $cAIC$ bajo el supuesto de homocedasticidad, es decir, para el caso donde $\boldsymbol{R} = \boldsymbol{\sigma}^2 I_n$

$$\alpha_{n,VB}\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML}\right) = \frac{2n}{n-p-2}\left[\varrho\left(\widehat{\boldsymbol{\delta}}_{ML}\right) + 1 - \frac{\varrho\left(\widehat{\boldsymbol{\delta}}_{ML}\right) - p}{n-p}\right]$$

En esta ecuación $\varrho\left(\widehat{\boldsymbol{\delta}}_{ML}\right) + 1$ define los grados de libertad efectivos para estimar $\boldsymbol{\beta}$, $\boldsymbol{u}$, y el escalar $\boldsymbol{\sigma}^{\boldsymbol{2}}$. Las propiedades estadísticas del AIC condicional se describen en Han, B. (2013), quien determinó que la propuesta de Vaida & Blanchard, (2005) funciona mejor al estar asociado al método de estimación de σ_u^2, la varianza del efecto aleatorio u_d bajo el modelo de Fay-Herriot.

Criterio de información bayesiano

Schwarz, (1978) propuso un criterio de información alternativo al AIC bajo un enfoque bayesiano. El autor, bajo el marco de trabajo del modelo de regresión lineal y una distribución previa equiprobable para todos los posibles modelos comparables determinó que el criterio de información bayesiano (BIC) está dado por:

$$BIC = -2\ell\left(\widehat{\boldsymbol{\theta}}\right) + \; k\log(n)$$

donde n es el tamaño de la muestra y k es el número de parámetros estimados por el modelo. Al igual que para el AIC, se puede emplear la log-verosimilitud completa $\ell(\boldsymbol{\beta}, \boldsymbol{\delta})$ o la log-verosimilitud restringida, $\ell_R(\boldsymbol{\delta})$ para estimar el BIC. Según Müller, et. al., (2013) y Stroup, (2013), el BIC bajo normalidad en los efectos aleatorios u_d y los errores e_{di} para el modelo GLMM y considerando la log-verosimilitud completa está dado por la siguiente ecuación

$$BIC = -2\ell\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML}\right) + log(n)(q^{*} + p)$$

Nuevamente, $\ell\left(\widehat{\boldsymbol{\beta}}_{ML}, \widehat{\boldsymbol{\delta}}_{ML}\right)$ denota la log-verosimilitud completa evaluada en $\widehat{\boldsymbol{\beta}}_{ML}$, $\widehat{\boldsymbol{\delta}}_{ML}$ que son los estimadores de máxima verosimilitud de $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$, respectivamente, p denota la longitud del vector de efectos fijos $\boldsymbol{\beta}$ y q^{*} es el número efectivo de parámetros del componente de varianzas $\boldsymbol{\delta}$ y n es el tamaño de la muestra. Si se mantienen constantes los efectos fijos del modelo, las diferencias entre los criterios de información de los modelos comparados estarán dadas por los parámetros del componente de varianza. En el escenario donde se emplea la log-verosimilitud restringida $\ell_R(\boldsymbol{\delta})$, el criterio AIC está dado por:

$$BIC = -2\ell_{\mathrm{R}}\left(\widehat{\boldsymbol{\delta}}\right) + log(n^{*})(q^{*})$$

Nuevamente, se prefieren los modelos con BIC más pequeños.

Limitaciones de los criterios de información

Autores como Jiang, et. al., (2008) afirman que los criterios de información tienen varias limitaciones al ser implementados en los modelos mixtos, como señalan:

- Al tener observaciones correlacionadas se debe reemplazar el tamaño de muestra n por el tamaño de muestra efectivo n^{*}. El problema radica en la correcta elección de n^{*}.
- Dependen de una función de log-verosimilitud conocida, la cual no está disponible si no se supone una familia paramétrica.
- El desempeño de los criterios en muestras finita depende del término de penalización seleccionado.

A su vez, se propone un enfoque alternativo que no tiene estas limitaciones. Por ejemplo, el método de "la grilla" consiste en un algoritmo que, en la primera etapa, delimita un subgrupo de modelos "correctos" a partir de una "cerca o valla" estadística que elimina los modelos incorrectos. Posteriormente, se identifica el modelo "óptimo" del subgrupo generado empleando un criterio adecuado. Según Nishii, (1984) el desempeño de los métodos clásicos de selección de variables se evalúa en dos vías, teóricamente, a partir de una prueba de consistencia donde el tamaño de la muestra tiende a infinito y, empíricamente, a través de estudios de simulación para tamaños de muestra finitas.

Resulta importante señalar que, después de un adecuado proceso de selección de variables en el cual se consideraron múltiples escenarios en cada uno de los países, se obtuvieron los siguientes resultados de los criterios de selección de modelos que se resumen en el cuadro 8.

Cuadro 8
Valores para diferentes criterios de selección de modelos en los países

Criterio	Chile	Colombia	Perú
AIC	116 648,8	541 139,3	128 894,1
AICc	116 648,8	541 139,3	128 894,1
cAIC	117 138,5	541 615,4	129 377,4
BIC	117 804,2	542 091,6	129 947,9

Fuente: Elaboración propia.

B. Validación de los supuestos del modelo y medidas de influencia

En esta subsección se presenta un resumen metodológico del proceso de validación de los supuestos del modelo, los diagnósticos basados en residuos y medidas de influencia para el caso particular del modelo lineal mixto que considera la transformación Log-Shift del ingreso $y_{di}^{*} = \log(y_{di} + c)$ y su relación con el predictor lineal, los efectos aleatorios de área (u_d) y su respectivo error a nivel de hogar (e_{di}):

$$\boldsymbol{y}_{di}^{*} = \mathrm{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}.\ \ i = 1, \dots, N_d,\ d = 1, \dots, D,$$

Como se ha mencionado anteriormente, en este modelo se asume que el efecto aleatorio de área $u_d \overset{i.i.d}{\sim} N(0, \sigma_u^2)$ es independiente al error a nivel individuo $e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)$. En McCulloch, et. al., (2008) se ilustran dos procedimientos para el análisis de los residuos del modelo con estructura de covarianza diagonal de bloques. Bajo la primera metodología se obtienen los residuales de Pearson del modelo lineal mixto transformado en un modelo de regresión lineal estándar, tal como se presenta a continuación:

$$\widetilde{y_d} = \tilde{x}_d'\boldsymbol{\beta} + \tilde{\epsilon}_d,\ \mathrm{d} = 1, \dots, D,$$

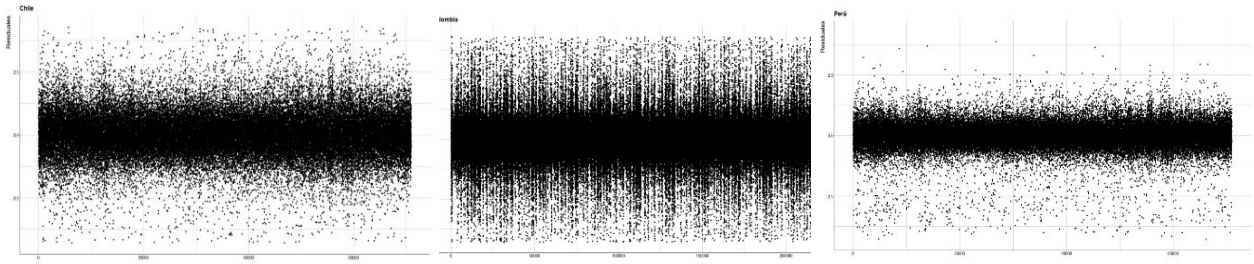
donde $\widetilde{y_d} = A_d^{-\frac{1}{2}} y_d$, $\tilde{x}_d' = A_d^{-1/2} X_d$, $\tilde{\epsilon}_d = A_d^{-1/2}(\boldsymbol{Z}_d\boldsymbol{u}_d + \boldsymbol{e}_d)$ y $V_d = \sigma_e^2 A_d$ (nótese que σ_e^2 es la varianza del término de error).

El modelo transformado se puede expresar como un modelo de regresión lineal estándar con errores independientes e idénticamente distribuidos normal de media nula y matriz de covarianzas $\sigma_e^2\, I_n$, con el fin de implementar los métodos tradicionales de selección y validación de los supuestos del modelo, análisis de residuales e identificación de valores atípicos e influyentes. Esta metodología es ampliamente utilizada porque generan residuales no correlacionados y, por lo tanto, sus gráficos se analizan de forma similar que en los modelos de regresión lineal.

La desventaja del enfoque radica en que la transformación genera residuales suavizados que pueden enmascarar los valores atípicos e influyentes de los datos originales. Adicional a lo anterior, los residuales transformados se centran en los efectos fijos por lo que, si hay alguna desviación del modelo, no se puede identificar si esta se debe a los efectos aleatorios o los errores del modelo. En el gráfico 4, se presentan los residuales transformados para los modelos ajustados en Chile, izquierda, Colombia, centro, y Perú, derecha, en los cuales se observa un comportamiento homogéneo, sin tendencias determinísticas ni patrones en forma de embudo que impliquen un problema de ajuste en los países.

Gráfico 4
Residuales transformados del modelo ajustado

A. Chile B. Colombia C. Perú



Fuente: Elaboración propia.

La segunda metodología está fundamentada en los residuos condicionales definidos por Jacqmin-Gadda, et. al., (2007), los cuales a pesar de estar correlacionados utilizan de forma explícita los efectos aleatorios estimados por lo que permiten verificar adecuadamente las desviaciones en la distribución de los errores del modelo. De acuerdo con lo previamente mencionado, los supuestos a validar están definidos sobre el efecto aleatorio del área y los errores a nivel individuo de la siguiente manera:

$$u_d \overset{i.i.d}{\sim} N(0, \sigma_u^2) \text{ y } e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)$$

Por lo cual se considera que verificar los supuestos normalidad de los efectos aleatorios y los residuos es la primera etapa del proceso de validación del modelo. En McCulloch, et. al., (2008), se describe el procedimiento propuesto por Calvin, (1993) para verificar el supuesto de normalidad en la distribución de efectos aleatorios del modelo lineal mixto. Los autores emplean los estimadores BLUP de los efectos aleatorios u_d adecuadamente estandarizados por sus desviaciones estándar estimadas en el desarrollo de gráficas Q-Q normal entre las variables $z_d = c'\tilde{\boldsymbol{u}}_d^B$ y $\phi^{-1}[F_m^*(z_i)]$ donde ϕ es la función de distribución acumulada de la distribución normal estándar y $F_m^*(z)$ es una función de distribución acumulada empírica z_d definido como

$$F_D^*(z) = \frac{\sum_{d=1}^{D} w_d I(z_d \leq z)}{\sum_{d=1}^{D} w_d}$$

siendo $\boldsymbol{w_d} = \boldsymbol{c'V(\tilde{u}_d^B)c}$ un vector de pesos y $\boldsymbol{I(z_d \leq z)}$ es una variable indicadora que será igual a 1 si $\boldsymbol{z_d \leq z}$ e igual a 0 en caso contrario. Claeskens y Hart, (2009) estudian el uso de las pruebas formales de normalidad para los efectos aleatorios y los errores del modelo. Bajo este escenario, se consideran las siguientes pruebas para detectar posibles violaciones a los supuestos de normalidad del modelo: prueba de D'Agostino, Lilliefors, Jarque-Bera y Weisberg-Bingham.

En el cuadro 9 y el cuadro 10 se presentan, respectivamente, las estadísticas de prueba y los valores p de las pruebas de normalidad consideradas para validar los supuestos sobre el efecto aleatorio y los residuos del modelo ajustado para cada uno de los países.

Cuadro 9
Pruebas de normalidad sobre el efecto aleatorio de los modelos ajustados en los países

País	Prueba		Estadística	Valor P
Chile	D'Agostino	Omnibus	234,14	< 2,2e-16
		Sesgo	10,96	< 2,2e-16
		Curtosis	10,67	< 2,2e-16
Colombia	D'Agostino	Omnibus	10,03	0,01
		Sesgo	1,76	0,08
		Curtosis	2,63	0,01
Perú	D'Agostino	Omnibus	9,50	0,01
		Sesgo	-1,64	0,10
		Curtosis	2,61	0,01
Chile	Lilliefors		0,04	0,01
Colombia			0,03	0,07
Perú			0,05	0,02
Chile	Jarque-Bera		2587,20	< 2,2e-16
Colombia			12,56	0,01
Perú			13,12	0,01
Chile	Spiegelhalter		1,37	< 2,2e-16
Colombia			1,28	0,0055
Perú			1,32	< 2,2e-16
Chile	Weisberg-Bingham		0,91	< 2,2e-16
Colombia			1,00	0,01
Perú			0,99	0,01

Fuente: Elaboración propia.

Cuadro 10
Pruebas de normalidad sobre el término de error en los modelos ajustados en los países

País	Prueba		Estadística	Valor P
Chile	D'Agostino	Omnibus	654,64	< 2,2e-16
		Sesgo	-111,56	< 2,2e-16
		Curtosis	12,99	< 2,2e-16
Colombia	D'Agostino	Omnibus	546,06	< 2,2e-16
		Sesgo	-8,17	2,20E-16
		Curtosis	5,02	< 2,2e-16

País	Prueba		Estadística	Valor P
Perú	D'Agostino	Omnibus	719,91	< 2,2e-16
		Sesgo	-45,33	< 2,2e-16
		Curtosis	7,17	< 2,2e-16
Chile	Lilliefors		0,07	< 2,2e-16
Colombia			0,04	< 2,2e-16
Perú			0,03	< 2,2e-16
Chile	Jarque-Bera		315426	< 2,2e-16
Colombia			39319	< 2,2e-16
Perú			76854	< 2,2e-16
Chile	Weisberg-Bingham		0,90	< 2,2e-16
Colombia			0,98	< 2,2e-16
Perú			0,96	< 2,2e-16

Fuente: Elaboración propia.

En el cuadro 9 y el cuadro 10 se puede observar que tanto los residuos como el efecto aleatorio no se ajustan adecuadamente a una distribución normal, pero al observar la simetría de la distribución parece ser que al menos este supuesto se cumple.

A continuación, en el cuadro 11 se presentan los resultados de las pruebas de White y de Breusch-Pagan para analizar si los residuos de los modelos ajustados tienen varianza constante. La prueba de Breusch-Pagan es el más simple de los enfoques ya que sólo detecta formas lineales de heterocedasticidad. Por tal motivo, se hace uso de la prueba de White con el fin de contrastar no linealidades utilizando los cuadrados y los productos cruzados de todos los regresores. En esta tabla se puede observar que la prueba de White afirma que los residuos de los modelos ajustados en los países tienen varianza constante.

Cuadro 11
Pruebas de homocedasticidad sobre los residuos de los modelos ajustados en los países

País	Prueba	Estadística	Valor P
Chile	Breusch-Pagan	322,45	0,00
	White	1 043,95	0,47
Colombia	Breusch-Pagan	1 389,81	0,00
	White	2 781,96	0,24
Perú	Breusch-Pagan	222,81	0,00
	White	10 198,44	0,57

Fuente: Elaboración propia.

Finalmente, en el cuadro 12 se consideran las pruebas de autocorrelación serial de Durbin-Watson y de Breusch-Godfrey las cuales coinciden en que los residuos de los modelos ajustados en los países no están correlacionadados.

Cuadro 12
Pruebas de homocedasticidad sobre los residuos de los modelos ajustados en los países

País	Prueba	Rezago	Estadística	Valor P
Chile	Durbin-Watson	5	1,97	0,04
		10	1,97	0,09
		15	1,99	0,21
	Breusch-Godfrey	5	12,67	0,03
		10	17,87	0,06
		15	23,59	0,07
Colombia	Durbin-Watson	5	1,99	0,00
		10	1,99	0,06
		15	2,00	0,95
	Breusch-Godfrey	5	13,54	0,02
		10	16,62	0,08
		15	22,52	0,09
Perú	Durbin-Watson	5	1,99	0,16
		10	1,98	0,15
		15	2,01	0,46
	Breusch-Godfrey	5	14,33	0,01
		10	15,87	0,10
		15	21,10	0,13

Fuente: Elaboración propia.

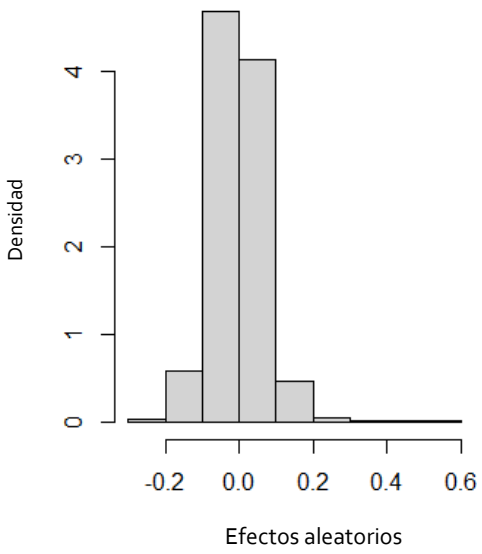
A continuación, se presentan los resultados de los gráficos Q-Q normal para los efectos aleatorios y los residuos de los modelos implementados en Chile, Colombia y Perú.

En general, los efectos aleatorios se comportan de manera simétrica y se ajustan a los cuantiles de una distribución normal como se aprecia en el gráfico 5, gráfico 7 y gráfico 9, respectivamente. En el caso de Chile (gráfico 5), se aprecia una ligera distorsión en la cola derecha del Q-Q normal de los efectos aleatorios la cual es generada por algunas comunas que poseen los ingresos más altos respecto al respectivo nivel de desagregación, puntualmente, las comunas La Reina, Providencia, las Condes y Vitacura). De forma similar, en el gráfico 9 se puede observar que algunas provincias de Perú reportan ingresos per cápita relativamente más altos respecto a las demás provincias del país; puntualmente, las provincias de Lima, el Callao y Cajamarca). Finalmente, en Colombia, (gráfico 7) no se observan ningún municipio que tenga un comportamiento particular que distorsione la distribución normal del efecto aleatorio.

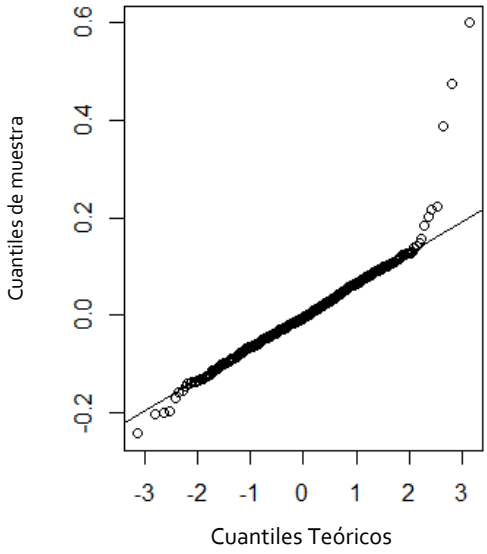
Por el contrario, al analizar el gráfico 6, el gráfico 8 y el gráfico 10 se puede observar que, incluso después de la transformación logarítmica, los residuales todavía muestran una distribución sesgada a izquierda lo que ocurre en presencia de valores de bienestar muy pequeños, que la transformación log-Shift transforma en valores negativos muy pequeños. A pesar de que los residuales presentan colas pesadas, el error cuadrado medio y las estimaciones del modelo son robustas bajo desviaciones de la normalidad.

Gráfico 5
Normalidad de los efectos aleatorios para Chile

A. Histograma de los efectos aleatorios



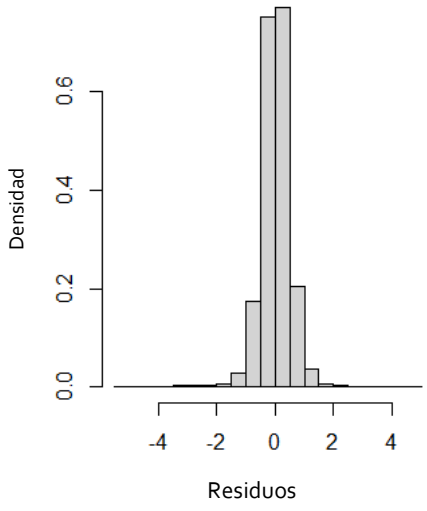
B. Q-Q de normalidad para los efectos aleatorios



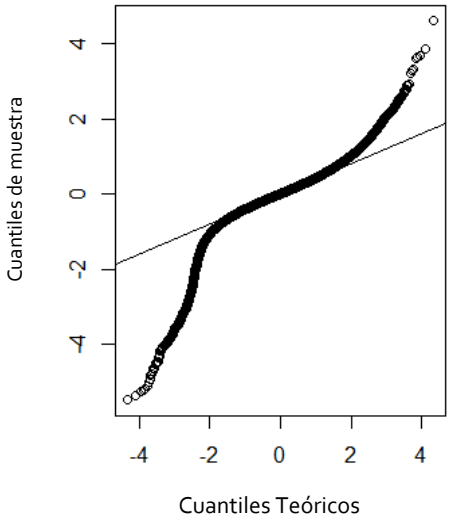
Fuente: Elaboración propia.

Gráfico 6
Normalidad de los residuales para Chile

A.Histograma de los residuales



B. Q-Q de normalidad para los residuales



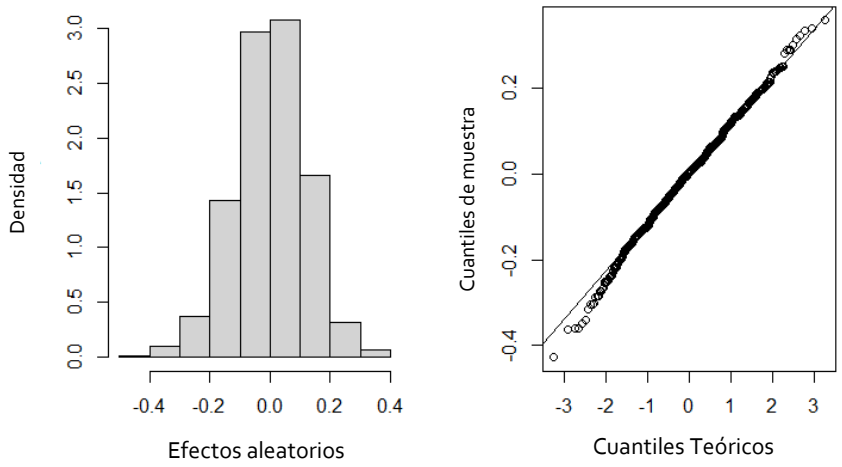
Fuente: Elaboración propia.

Por el contrario, al analizar el gráfico 6, el gráfico 8 y el gráfico 10 se puede observar que, incluso después de la transformación logarítmica, los residuales todavía muestran una distribución sesgada a izquierda lo que ocurre en presencia de valores de bienestar muy pequeños, que la transformación log-Shift transforma en

valores negativos muy pequeños. A pesar de que los residuales presentan colas pesadas, el error cuadrado medio y las estimaciones del modelo son robustas bajo desviaciones de la normalidad. En CEPAL se consideran los enfoques alternativos propuestos por Diallo y Rao (2018) que consideran de forma explícita una distribución normal sesgada para los residuos y la propuesta de Graf, et. al., (2019) para el estimador EB bajo distribuciones sesgadas generales.

Gráfico 7
Normalidad de los efectos aleatorios para Colombia

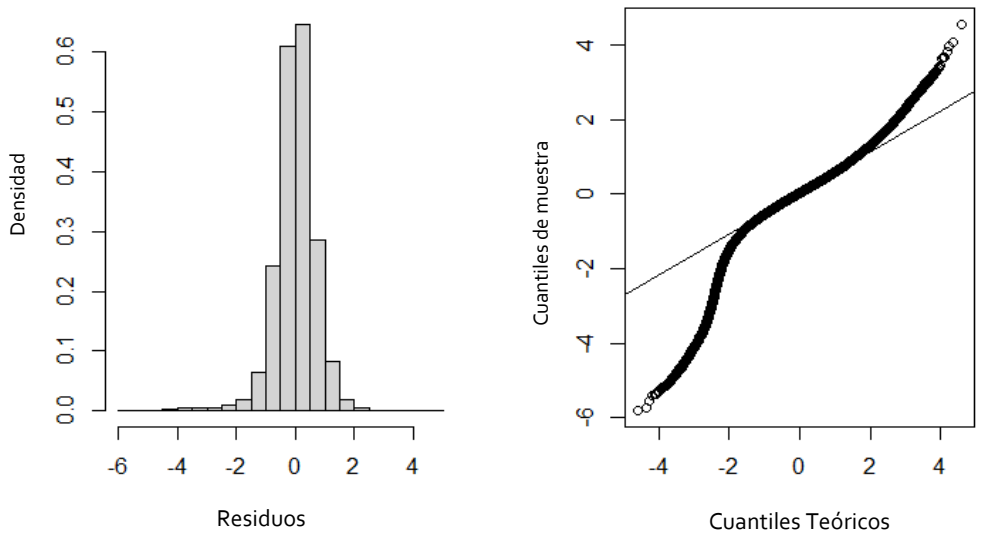
A. Histograma de los efectos aleatorios B. Q-Q de normalidad para los efectos aleatorios



Fuente: Elaboración propia.

Gráfico 8
Normalidad de los residuales para Colombia

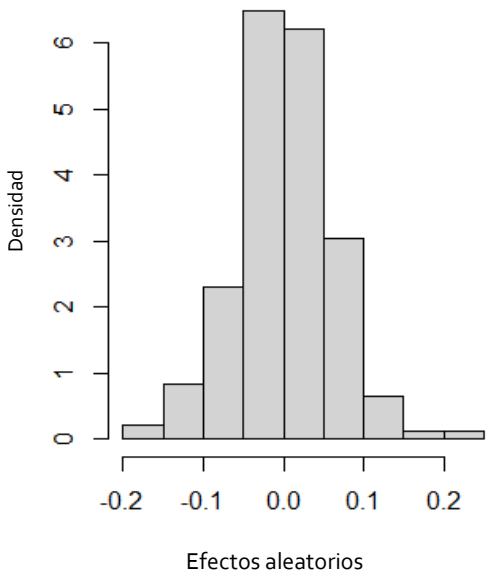
A.Histograma de los residuales B. Q-Q de normalidad para los residuales



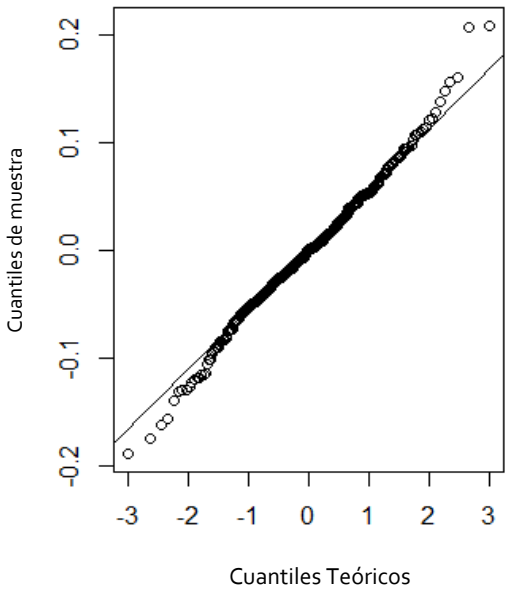
Fuente: Elaboración propia.

Gráfico 9
Normalidad de los efectos aleatorios para Perú

A. Histograma de los efectos aleatorios



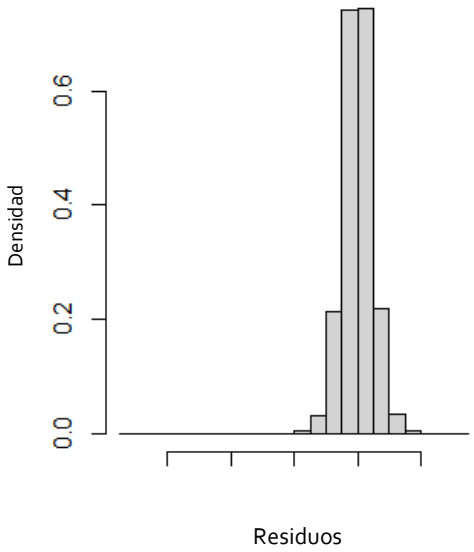
B. Q-Q de normalidad para los efectos aleatorios



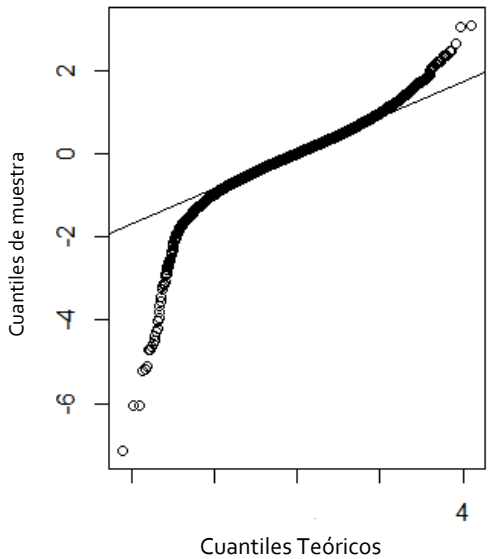
Fuente: Elaboración propia.

Gráfico 10
Normalidad de los residuales para Perú

A. Histograma de los residuales



B. Q-Q de normalidad para los residuales



Fuente: Elaboración propia.

C. Puntos de balanceo, datos atípicos y/o influyentes

Desde un punto de vista conceptual, los puntos de balanceo, datos atípicos y/o influyentes en el modelo considerado para la generación de mapas de pobreza tienen la misma definición que para el modelo lineal mixto. Inicialmente se considera la propuesta de Zewotir & Galpin, (2007) para detectar atípicos y puntos de alto apalancamiento en modelos lineales mixtos con más de un efecto aleatorio ($\boldsymbol{k = 1, 2, \ldots, r}$). Los autores, bajo el supuesto de razones de varianza $\left(\boldsymbol{\sigma_k^2/\sigma_e^2}\right)$ conocidas, determinaron que los residuos EBLUP $(\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X\hat{\beta}} - \boldsymbol{Z\hat{u}})$, son tales que, $\hat{\boldsymbol{e}} = \hat{\boldsymbol{S}}\boldsymbol{y} \sim \boldsymbol{N}(0, \hat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\boldsymbol{2}}\hat{\boldsymbol{S}})$, siendo

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\boldsymbol{2}}[\hat{\boldsymbol{V}}^{-\boldsymbol{1}} - \hat{\boldsymbol{V}}^{-\boldsymbol{1}}\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-\boldsymbol{1}}\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-\boldsymbol{1}}]$$

Bajo esta consideración, definen dos tipos de residuales, los internamente estudentizados ($\boldsymbol{t_i}$) y los externamente estudentizados ($\boldsymbol{t_i^*}$), definidos por medio de las siguientes expresiones:

$$\widehat{t_i} = \frac{\hat{e}_i}{\hat{\sigma}_e\sqrt{\widehat{s_{ii}}}} \quad \text{y} \quad \widehat{t_i^*} = \frac{\hat{e}_i}{\hat{\sigma}_{e(i)}\sqrt{\widehat{s_{ii}}}}$$

$\boldsymbol{s_{ii}}$ denota el i-esimo elemento de la diagonal principal de la matriz $\boldsymbol{S}$ y $\hat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\boldsymbol{2}}$ representa el estimador de la varianza del término de error, el cual está dado por

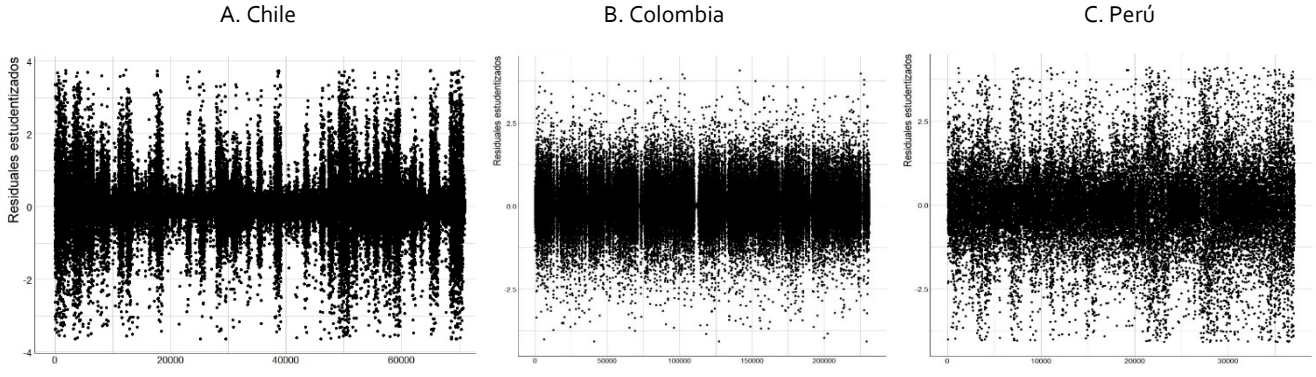
$$\hat{\sigma}_e^2 = \frac{1}{n}\,(\boldsymbol{y} - \boldsymbol{X\hat{\beta}})'\left(I + \sum_{k\,=\,1}^{r}\hat{\lambda}_k\boldsymbol{Z}_k\boldsymbol{Z}_k'\right)^{-1}(\boldsymbol{y} - \boldsymbol{X\hat{\beta}})$$

Finalmente, $\hat{\sigma}_{e(i)}^2$ corresponde al estimador $\hat{\sigma}_e^2$ eliminando la i-ésima observación de los datos. Los autores demostraron que, al emplear los estimadores de máxima verosimilitud del componente de varianza, si el i-ésimo residual internamente estudentizado supera el umbral dado por

$$\widehat{t_i} > \sqrt{\frac{4n}{n\,-\,p\,-\,3}}$$

Luego, este dato puede ser considerado como un valor atípico. También determinaron que los datos y_i con un valor pequeño s_{ii} tienen un alto apalancamiento en el modelo. A continuación, en el gráfico 11, se presentan los residuales internamente estudentizados del modelo ajustado para Chile, izquierda, Colombia, centro y Perú derecha; en estos gráficos no se observa ningún patrón anómalo o extraño, tampoco se observan datos atípicos.

Gráfico 11
Residuales internamente estudentizados del modelo ajustado

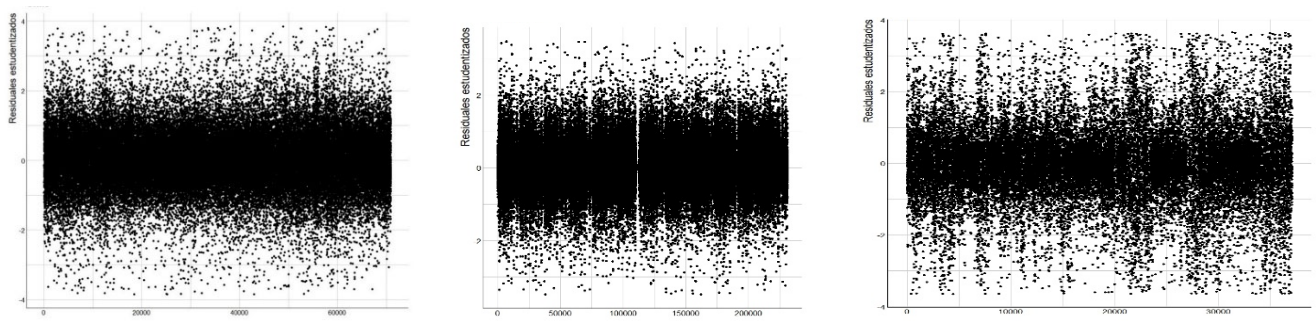


Fuente: Elaboración propia.

Posteriormente, en el gráfico 12, se presentan los residuales externamente estudentizados del modelo ajustado para Chile, izquierda, Colombia, centro y Perú derecha; en estos gráficos no se observa ningún patrón anómalo o extraño, tampoco se observan datos influyentes.

Gráfico 12
Residuales externamente estudentizados del modelo ajustado

A. Chile B. Colombia C. Perú



Fuente: Elaboración propia.

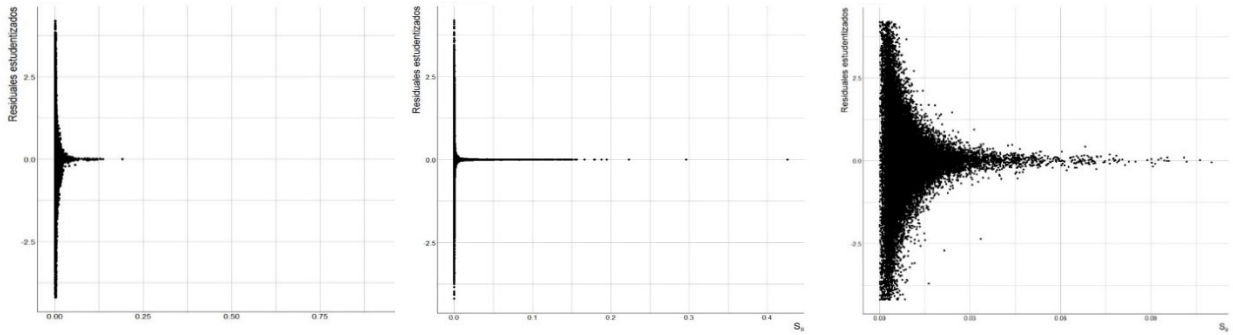
Asimismo, los autores demostraron que si

$$\widehat{s_{ii}} < \left(\frac{1-2p}{n}\right)\left\{1 - \sum_{k=1}^{r}\left[\frac{\sigma_e^2}{\sigma_k^2 + n/(2h_k)}\right]^{-1}\right\}$$

Entonces, esta observación se considera un punto de balanceo. Una forma simultánea para identificar los valores atípicos y puntos de balanceo consiste en emplear la gráfica de s_{ii} versus $\tilde{e}_i^2 \ / \ \tilde{e}'\tilde{e}$. En este gráfico, los puntos que se alejan de la nube principal y caen en la esquina inferior izquierda (s_{ii} pequeño) se consideran puntos de balanceo mientras que aquellos valores que se alejan de esta nube al lado derecho se pueden considerar valores atípicos. En el gráfico 13 se puede observar el resultado obtenido para el caso chileno a la izquierda, en el cual se pueden apreciar algunos puntos de balanceo que serán revisados con más detalle para una próxima implementación del modelo.

Gráfico 13
Residuales internamente estudentizados vs matriz de proyección del modelo ajustado

A. Chile B. Colombia C. Perú



Fuente: Elaboración propia.

En la misma gráfica en el centro se puede observar la ilustración para el caso colombiano en el cual se pueden apreciar algunos puntos de balanceo que serán revisados con más detalle para una próxima implementación del modelo. Finalmente, en la derecha de la gráfica se presenta el resultado obtenido en Perú, donde no se observa ningún dato influyente.

Zewotir y Galpin (2007) afirman que los estimadores de los efectos aleatorios pueden interpretarse como residuales que pueden ser empleados como medidas de diagnóstico para detectar observaciones (bloques) que tienen un comportamiento diferente a los demás en el conjunto de datos, al estandarizar los elementos de $\boldsymbol{u}_k$ por la raíz cuadrada de los elementos de la diagonal principal de la matriz de covarianzas $V(\widehat{\boldsymbol{u}}_k) = \hat{\sigma}_e^2 \hat{\lambda}_k^2 \widehat{\boldsymbol{Z}}'_k \widehat{\boldsymbol{S}} \widehat{\boldsymbol{Z}}_k$. Por ende, otro enfoque ampliamente utilizado como medida diagnóstica es la distancia propuesta por Cook, (1977). Zhu, et. al., (2012) proponen una generalización de esta medida para los modelos lineales mixtos con una estructura de covarianza diagonal de bloques, en la cual se mide el efecto de eliminar un dominio en la estimación del vector de coeficientes de regresión $\boldsymbol{\beta}$ siguiendo la propuesta de Banerjee y Frees, (1997). Siguiendo la notación de Stroup, (2013), $\widehat{\boldsymbol{\beta}}_{(i)}$ denota el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ después de eliminar el i-ésimo dominio en el proceso de estimación. McCulloch et. al., (2008) definen la influencia del dominio i en la estimación de

$$\hat{\beta} = \left(\sum_{i=1}^{D} X_i^T V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{D} X_i^T V_i^{-1} y_i$$

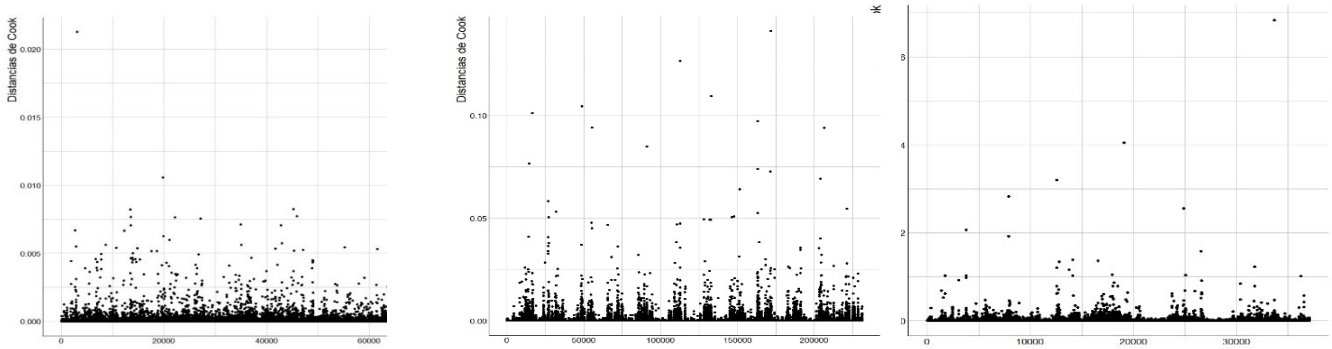
Por medio de la distancia de Mahalanobis dada por

$$B_i(\hat{\beta}) = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(i)})' \left(\sum_{i=1}^{D} X_i^T V_i^{-1} X_i\right)(\hat{\beta} - \hat{\beta}_{(i)})$$

En la práctica se emplea una gráfica con las distancias de Cook $B_i(\hat{\beta})$ para $i = 1, \ldots, m$ para para identificar los dominios influyentes. En Kelly y Mathew (1994) y Christensen et. al. , (1992) se proponen generalizaciones de la distancia de Cook para el caso de los modelos lineales mixtos con más de un efecto aleatorio, la cual permite medir la influencia de una observación en la estimación de los efectos fijos y los componentes de la varianza. Christensen et. al., (1992) proponen un algoritmo computacional que hace más eficiente el proceso de estimación de esta medida en los modelos a nivel de unidad. En el Gráfico 16 se presentan las distancias de Cook obtenidos al ajustar el modelo SAE en Chile, al lado izquierdo en el cual se pueden apreciar algunos puntos influyentes que serán revisados con más detalle para una próxima implementación del modelo.

Gráfico 14
Distancias de Cook del modelo ajustado

A. Chile B. Colombia C. Perú



Fuente: Elaboración propia.

En el mismo gráfico 14 se presentan las distancias de Cook generadas después de ajustar el modelo SAE en Colombia, en el centro, en esta ilustración se puede observar algunos puntos influyentes los cuales serán revisados con más detalle en una próxima implementación del modelo para el país.

Finalmente, al lado derecho de la misma gráfica, las distancias de Cook generadas después de ajustar el modelo SAE en Perú, en esta ilustración se puede observar algunos puntos influyentes los cuales serán revisados con más detalle en una próxima implementación del modelo para el país.

D. Benchmarking con las cifras publicadas

Una de las ventajas del modelo desarrollado en este documento es su capacidad para generar predicciones de los indicadores a los distintos niveles de desagregación que permitan los datos del censo. Ahora bien, es importante garantizar que las desagregaciones obtenidas sean consistentes con las estimaciones directas a nivel nacional y en los niveles en los que la encuesta provee resultados representativos, que son la base de las estadísticas publicadas. Para mantener la consistencia con la literatura especializada, las estimaciones que proveen el "punto de referencia" para ajustar las cifras desagregadas se denominan con la palabra inglesa *benchmark*.

Para realizar el ajuste de las cifras desagregadas se empleó el método de calibración de los factores de expansión, usado frecuentemente en el análisis de encuestas por muestreo. Esta técnica es usada cuando se tiene presencia de sesgo o poca precisión en las estimaciones a causa de la no respuesta o de tamaños de muestra reducidos. El método consiste en ajustar los factores de expansión de la encuesta de manera que las estimaciones obtenidas sean consistentes con las variables auxiliares utilizadas como referencia, que pueden ser estimaciones o parámetros poblacionales obtenidos a partir de otra encuesta o registro administrativo. Para ajustar las estimaciones provenientes de la metodología pseudo-EBP es necesario encontrar un conjunto de pesos w_k que satisfacen la siguiente restricción:

$$\hat{\mathrm{R}}_{cal} = \left(\hat{R}_{1,cal}, \dots, \hat{R}_{Q,cal}\right)' = \left(R_1, \dots, R_Q\right)' = \mathrm{R}$$

El algoritmo consta de los siguientes pasos (Gutiérrez, 2016):

i) Suponga que se tiene acceso a un vector de información $X_k = \left(x_{1k}, x_{2k}, \dots, x_{pk}\right)$ de p variables auxiliares y conocido para cada uno de los individuos de la muestra.
ii) A través de registros administrativos o datos oficiales se conoce el valor del total del vector x para el total de la población, $T_X = \sum_{k \in U} x_k$
iii) El propósito es estimar el total de x usando la información en la muestra, $x_k \; k \in S$
iv) Buscando que la estimación cumpla con la siguiente restricción, $\sum_{k \in S} w_k x_k = T_x$
v) La idea consiste en buscar pesos w_k tan cercanos como sea posible al inverso de la probabilidad de inclusión del k-ésimo elemento $d_k = 1/\pi_k$

En este caso en particular el punto de referencia para la variable a calibrar no es un total sino una razón o proporción (la tasa de pobreza), por lo que se realiza la búsqueda del vector w_k que cumpla con la siguiente restricción (Gutiérrez et. al., 2016):

$$\hat{R}_{q,cal} = \frac{\sum_{k \in s} w_k \, y_{qk}}{\sum_{k \in s} w_k \, x_{qk}}$$

Si la razón poblacional es conocida, resulta posible imponer la siguiente restricción en las ecuaciones de calibración: $\hat{\mathrm{R}}_{cal} = \mathrm{R}_U$. Es decir, para $q = 1, \dots, Q$.

$$z_{qk} = \begin{cases} y_{qk} - R_q x_{qk} & \text{si } k \in s_h \\ 0 & \text{en otro caso} \end{cases}$$

donde

$$t_{z_q} = \sum_{k \in U} z_{qk} = \sum_{k \in U} y_{qk} - R_q x_{qk} = 0$$

Si las medias de los subgrupos son conocidas, la restricción queda definida como

$$\bar{y}_{cal} = \left(\bar{y}_{1,cal}, \ldots, \bar{y}_{Q,cal}\right)' = \left(\bar{y}_1, \ldots, \bar{y}_Q\right)' = \bar{y}$$

De esta forma se obtiene la restricción para las ecuaciones de calibración: $\bar{y}_{cal} = \bar{y}$. Donde, para $q = 1, \ldots, Q$,

$$z_{qk} = \begin{cases} y_{qk} - \bar{y}_q & \text{si } k \in s_h \\ 0 & \text{en otro caso} \end{cases}$$

El procedimiento de calibración de las estimaciones se realizó a nivel urbano, rural y departamental para el caso de Perú, urbano, rural y regional para el caso de Chile, y Urbano, y departamental para el caso de Colombia. Como resultado de este ejercicio de calibración, en cada paso del método de Monte Carlo, se obtuvo un vector de ponderadores w_k para cada hogar k, que se aplica en el cálculo final de las incidencias de pobreza por municipio y al conformar las agregaciones de interés por país se obtienen los datos publicados regularmente por la CEPAL.

E. Comparación con las estimaciones directas

La estimación directa es aquella estadística calculada usando únicamente la información proveniente de la encuesta y la cual tiene asociado una medida de error muestral. Para determinar la calidad de las estimaciones directas, se utilizaron los criterios planteados por Gutiérrez et. al. (2020), y así descartar aquellas comunas, provincias o municipios que no alcanzaron los umbrales mínimos para que sus estimaciones sean consideradas confiables. Cabe señalar que las encuestas de hogares fueron procesadas teniendo en cuenta los criterios del diseño de muestreo definidos para cada una de ellas.

Los criterios utilizados incluyen:

- Grados de libertad (GL): corresponden a la resta entre el número de unidades primarias de muestreo (UPM) y los estratos en determinada subpoblación o área (comunas, sexo, entre otras).
- Coeficiente de variación (CV): este se define, como la razón entre el error estándar y la estimación puntual.
- Coeficiente de variación logarítmico (CVL): supera las limitaciones del coeficiente de variación para proporciones cercanas a cero.
- Efecto de diseño (DEFF): razón entre la varianza de la estimación bajo el diseño complejo y la varianza bajo muestreo aleatorio simple de esa misma estimación.
- Tamaño de muestra (n): cantidad de observaciones disponibles en el área o estrato de interés.
- Tamaño de muestra efectivo (n_{eff}): cociente entre el tamaño definido en el diseño complejo y el DEFF, corresponde a una aproximación de la información que realmente aporta la muestra alcanzada.
- Casos no ponderados (y): corresponde a los casos en la muestra en donde se observa la variable de interés; en el presente estudio corresponde al número de personas en situación de pobreza o pobreza extrema.

En el cuadro 13 se presenta un resumen de los umbrales definidos por la literatura para determinar si una cifra cumple con los requisitos para ser publicada como estadística oficial según el criterio de calidad empleado.

Cuadro 13
Valores de referencia para indicadores de calidad de estimación

Indicador	Valor	Referencia
Coeficiente de Variación CV	20%	(Gutiérrez, 2018)
Coeficiente de Variación Logarítmico (CVL)	17,5%	(Barnett-Walker, et al., 2003)
Efecto diseño (DEFF)	-	-
Tamaño muestra (n)	100	(Barnett-Walker, et al., 2003)
Tamaño muestra efectiva (n_{eff})	68	(Barnett-Walker, et al., 2003)
Grados de libertad (gl)	8	(Parker, et al., 2017)
Casos no ponderados	50	(National Research Council, 2015)

Fuente: Criterios de calidad en la estimación de indicadores usando encuestas de hogares. Una aplicación a la migración internacional (Serie estudios estadísticos N101. Cepal, junio 2020).

Empleando los criterios de calidad mencionados se generaron los mapas de pobreza con las estimaciones directas que los cumplen con el fin de identificar el porcentaje de cada país que puede ser analizado en el nivel de desagregación de interés.

En el mapa 1 se presentan los mapas con la estimación directa a nivel municipal de la tasa de pobreza y de pobreza extrema en Colombia, respectivamente. En la parte izquierda de la gráfica se presentan las estimaciones directas de los municipios que cumplen los requisitos establecidos por los criterios de calidad en una escala de amarillo a rojo según la tasa estimada; por el contrario, si el municipio está sombreado de color gris es porque su estimación directa no cumple alguno de los criterios de calidad considerados en el estudio. En las gráficas mencionadas se puede observar que gran parte del país no puede tener una estimación directa de calidad para los indicadores de interés.

En el caso de Chile se presenta una división del país en las comunas en del norte, centro y sur con el objetivo de generar gráficos más ilustrativos. En el mapa 2, el mapa 3 y el mapa 4 se presentan los mapas con la estimación a nivel comunal de la tasa de pobreza y la distribución geográfica de la tasa de pobreza extrema del mencionado país.

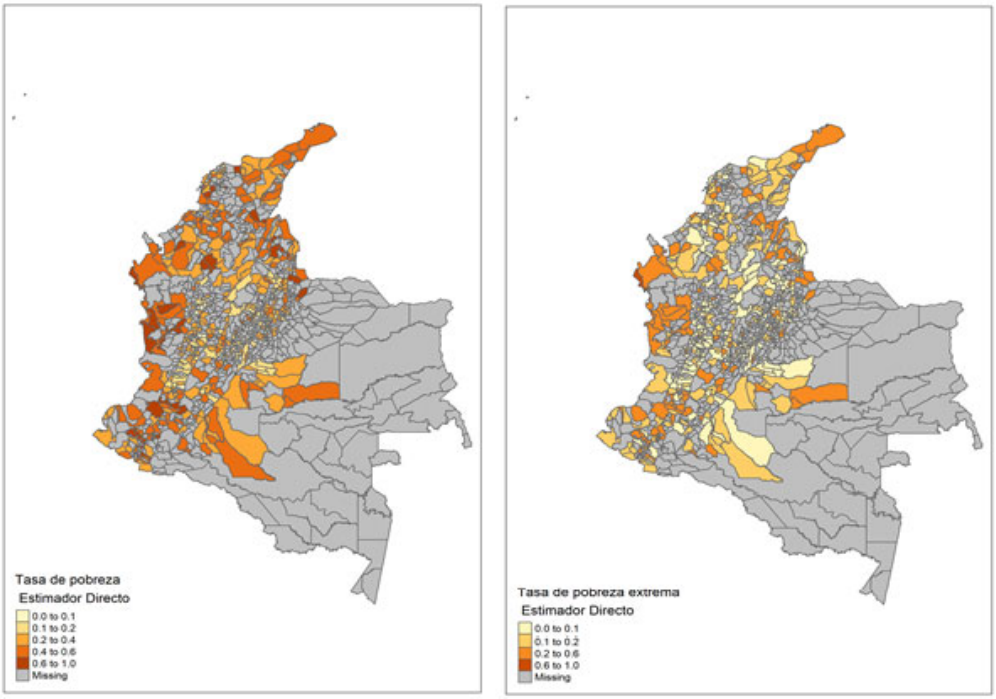
A diferencia de Colombia, la encuesta CASEN permite obtener un mayor número de comunas con estimaciones directas que cumplen los criterios de calidad, las cuales son ilustradas gráficamente en una escala de amarillo a rojo para cada una de las tasas estimadas; por el contrario, si la es sombreado de color gris es porque su estimación directa no cumple alguno de los criterios de calidad considerados en el estudio. En Chile, las comunas que no permiten obtener estimaciones directas confiables se encuentran, en su mayoría, al sur del país.

En el mapa 5, se presentan los mapas con la estimación a nivel de provincia de la tasa de pobreza (izquierda) y de pobreza extrema (derecha) en Perú, respectivamente. En los mapas se presentan, en una escala de amarillo a rojo, las estimaciones directas de la tasa de pobreza (izquierda) y de pobreza extrema (derecha) en aquellas provincias que cumplen los requisitos establecidos por los criterios de calidad según la tasa estimada; por el contrario, si la provincia está sombreada de color gris es porque su estimación directa no cumple alguno de los criterios de calidad considerados en el estudio.

En las gráficas mencionadas se puede observar que la gran parte del país no puede tener una estimación directa de calidad para los indicadores de interés.

Mapa 1
Estimación directa en Colombia

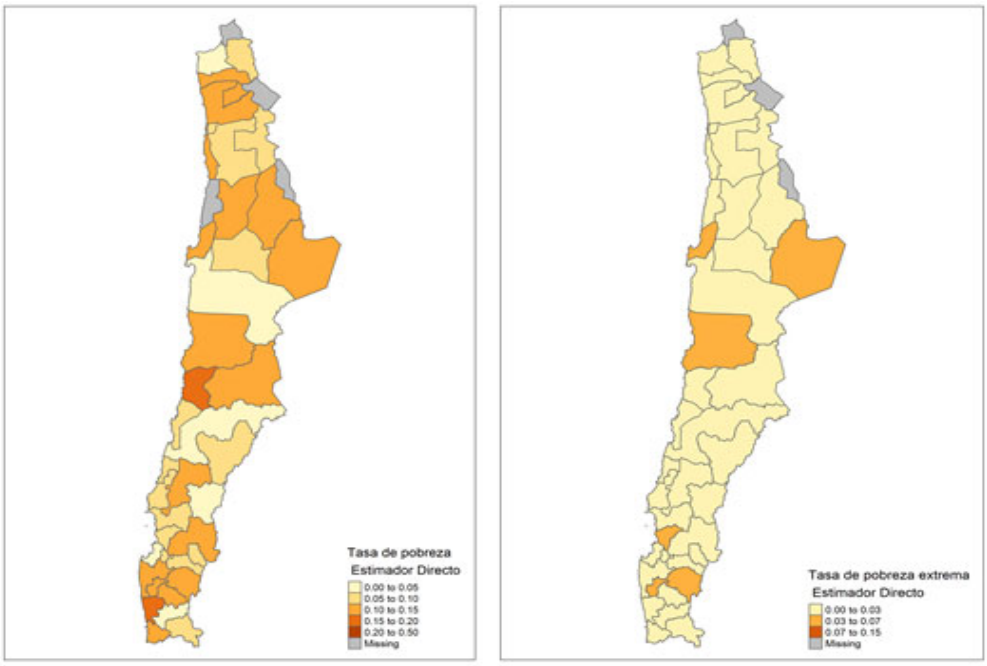
A. Tasa de pobreza B. Tasa de pobreza extrema



Fuente: Elaboración propia.

Mapa 2
Estimación directa en las comunas del norte de Chile

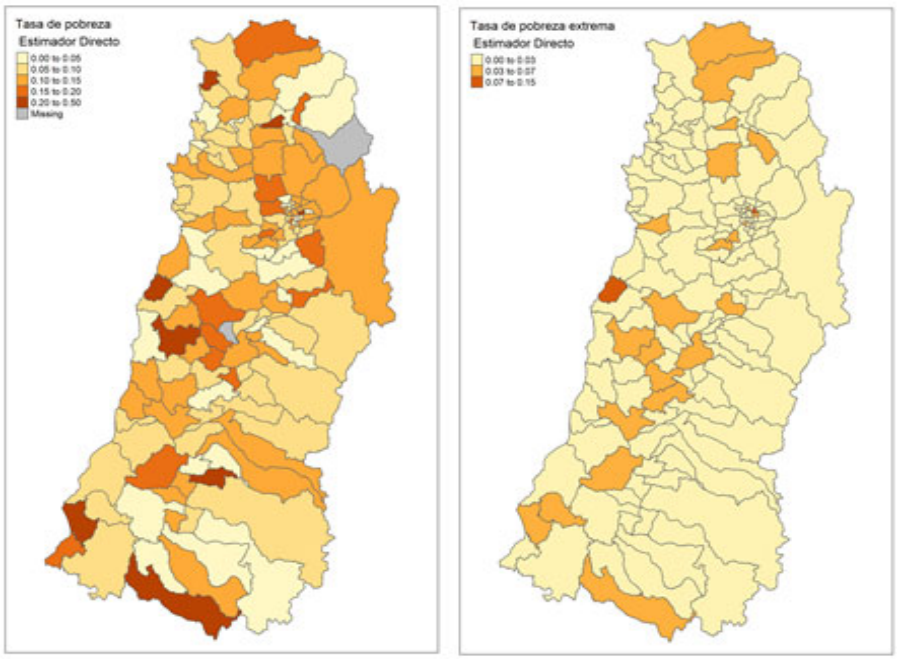
A. Tasa de pobreza B. Tasa de pobreza extrema



Fuente: Elaboración propia.

Mapa 3
Estimación directa en las comunas del centro de Chile

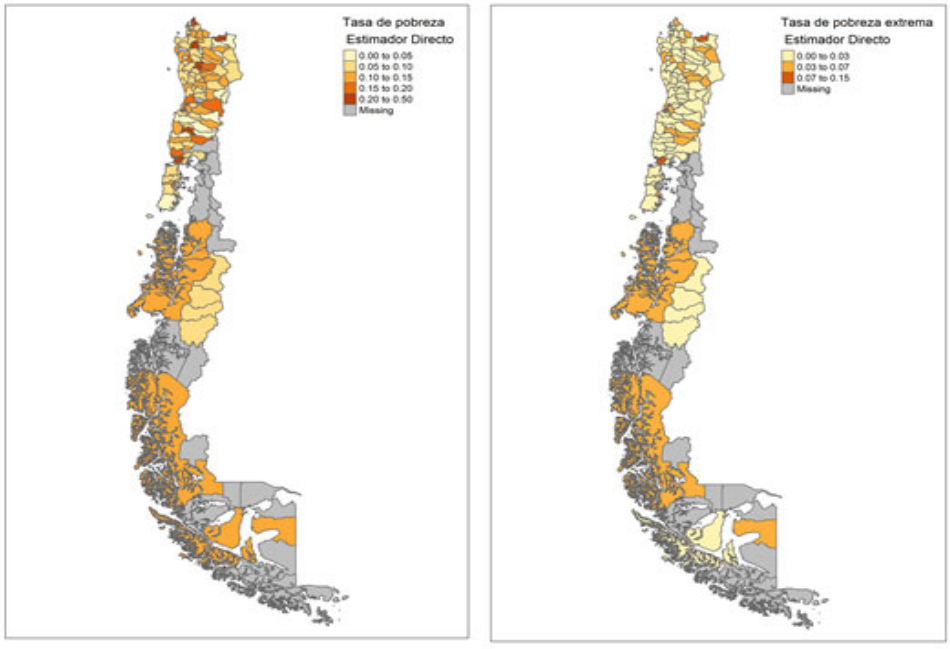
A. Tasa de pobreza B. Tasa de pobreza extrema



Fuente: Elaboración propia.

Mapa 4
Estimación directa en las comunas del sur de Chile

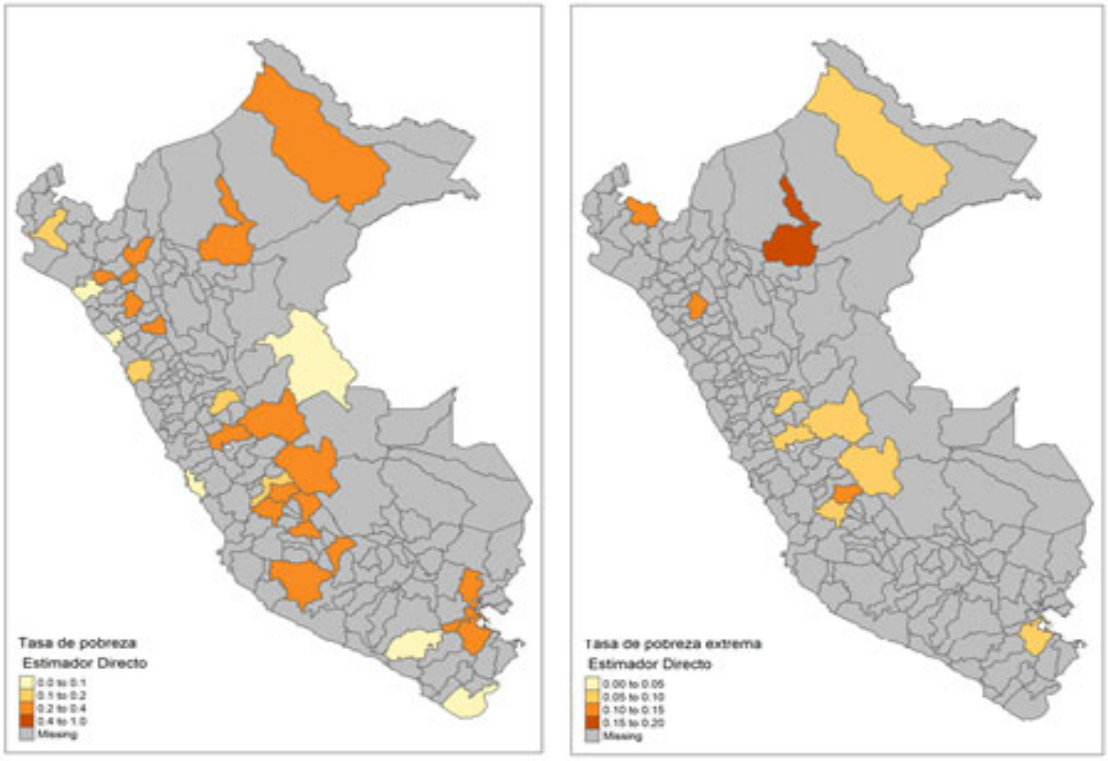
A. Tasa de pobreza B. Tasa de pobreza extrema



Fuente: Elaboración propia.

Mapa 5
Estimación directa en las provincias de Perú

A. Tasa de pobreza B. Tasa de pobreza extrema



Fuente: Elaboración propia.

V. Resultados y mapas de pobreza

En este último capítulo se presentan los resultados del modelo SAE seleccionado para estimar indicadores FGT a partir del ingreso de los hogares en los dominios de interés en cada país. Después de un meticuloso proceso de selección de variables, los modelos identificados en los países garantizan un equilibrio entre la bondad de ajuste del modelo y su complejidad; en otras palabras, el modelo con mejor aproximación a los datos recolectados (aquel que capture sus regularidades subyacentes) en los países, su complejidad y el grado de validez de los resultados que genera respecto a los supuestos intrínsecos sobre los que se soporta teóricamente.

En la sección 4 de los anexos, se presenta el resumen de los modelos seleccionados para cada uno de los países observados, en el cuadro 29, se encontrarán los resultados del modelo ajustado para Chile, en el cuadro 30, los resultados obtenidos para Colombia y finalmente, el cuadro 31 que contiene los resultados del modelo ajustado para Perú. En estas tablas, se puede apreciar que algunas de las variables tenidas en cuenta en el modelo, resultan ser no significativas desde el punto de vista estadístico, pero se mantienen en el análisis ya que se consideran relevantes desde el punto de vista técnico e interpretativo. Finalmente, en la sección 3 de los anexos se incluyen los mapas por separado de infraestructura en redes viales y contaminación lumínica, específicamente las figuras 26 y 27 para Chile, las figuras 28 y 29 para Colombia y las figuras 30 y 31 para Perú.

A. Resultados para Chile

De acuerdo con las estimaciones obtenidas con el modelo Pseudo-EBP, es posible clasificar a las comunas de Chile en tres grupos según sus niveles de pobreza: tasas iguales o inferiores al 7%, tasas superiores al 7% e inferiores al 15%, y tasas iguales o superiores al 15%.

La baja incidencia de la pobreza en Chile, en términos comparativos con la región, se manifiesta también en las tasas de pobreza por comuna. Solamente un 13% de las comunas presentan niveles de pobreza superiores al 15%, umbral por sobre el que se encuentran las tasas de pobreza de todos los países de América Latina, con la excepción de Uruguay, además de Chile (ver la tabla 14).

Cuadro 14
Porcentaje de comunas por nivel de pobreza
(En porcentajes)

Grupo	Nivel pobreza	Comunas por nivel	Participación
1	Menor a 7%	107	30,9%
2	Entre 7% y 15%	194	56,1%
3	Mayor a 15%	45	13,0%

Fuente: Elaboración propia.

De acuerdo al cuadro 15, Las comunas de Chile, que concentran los mayores niveles de pobreza monetaria son aquellas ubicadas principalmente en el sur del país, específicamente en las regiones de Maule, Biobío y Araucanía. En la provincia de Biobío, en la región del Biobío, 8 de 14 comunas pertenecen a este grupo. Las provincias de Cautín y Malleco, en la región de la Araucanía, tienen el 48% y 73% de sus comunas en este grupo (ver el cuadro 15).

Cuadro 15
Chile: 20 comunas con mayores niveles de pobreza
(En porcentajes)

Código Provincia	Provincia	Comuna	Estimación pobreza	LI	LS
8314	Biobío	Alto Biobío	23,3%	15,8%	30,7%
9121	Cautín	Cholchol	21,3%	15,6%	27,0%
9106	Cautín	Galvarino	20,7%	14,3%	27,1%
4303	Limarí	Monte Patria	18,5%	14,9%	22,0%
9202	Malleco	Collipulli	17,8%	13,4%	22,3%
8302	Biobío	Antuco	17,6%	13,3%	21,9%
9119	Cautín	Vilcún	17,6%	13,9%	21,3%
8310	Biobío	San Rosendo	17,6%	12,6%	22,5%
8307	Biobío	Negrete	17,6%	12,5%	22,6%
16103	Ñuble	Cobquecura	17,6%	12,1%	23,1%
16108	Ñuble	Ninhue	17,4%	12,6%	22,2%
9109	Cautín	Loncoche	17,4%	12,4%	22,4%
8312	Biobío	Tucapel	17,1%	10,7%	23,5%
16106	Ñuble	Chillán Viejo	17,1%	12,7%	21,5%
8206	Arauco	Los Álamos	17,0%	10,6%	23,4%
4304	Limarí	Punitaqui	17,0%	13,3%	20,6%
4103	Elqui	Andacollo	16,9%	10,9%	22,9%
9208	Malleco	Purén	16,9%	12,6%	21,1%
9205	Malleco	Lonquimay	16,8%	10,8%	22,7%
9102	Cautín	Carahue	16,7%	12,1%	21,3%

Fuente: Elaboración propia.

Las comunas con mayores niveles de pobreza monetaria se caracterizan por bajas tasas de ocupación, escaso acceso a servicios básicos de agua y saneamiento, escasa infraestructura para atención en salud, bajo desempeño escolar, entre otros factores que limitan las posibilidades de las personas de contar con mejores ingresos. Estas comunas cuentan con una alta participación de comunidades indígenas, por lo que se evidencia una asociación entre el grupo étnico y la situación de pobreza.

La región del Biobío, con una población rural del 84%, tiene los niveles de analfabetismo más altos del país y la mayor tasa promedio de desempleo (Biblioteca del Congreso Nacional de Chile, BCN (2016). En esta región se encuentra la comuna de Alto Biobío, con los mayores niveles de pobreza y pobreza extrema monetaria del país, con tasas del 23,2% y 8,7% respectivamente (ver la tabla 14). En esta comuna, cerca del 86% de la población se identifica como perteneciente a un pueblo indígena, más específicamente a la nación Mapuche, según datos del censo nacional de población en Chile 2017, INE (2017). Se estima que el 77% de la población carece de servicios básicos, superando el promedio de la región de 17%. En lo referente al acceso a la salud, en la comuna se registra un consultorio general rural y 8 postas de salud rural (establecimientos de atención ambulatoria), respecto de los 37 consultorios generales y los 175 posta de salud rural establecidos en la región del Biobío, según datos del MINSAL.

A su vez, según datos del MINSAL, la comuna presenta una tasa de natalidad de 24,7 nacidos vivos por cada 1 000 habitantes, muy por encima de las tasas de la región (12,7) y del país (13,6). En lo que respecta al rendimiento académico, los resultados en las diferentes pruebas del SIMCE para lectura y matemáticas, son ostensiblemente más bajos que los resultados obtenidos por los estudiantes de la región y del país, según las estadísticas territoriales de la Biblioteca del Congreso Nacional de Chile BCN (2016). Finalmente, según datos del Servicio de Impuestos Internos SII (2016), en la comuna se encuentran registradas un total de 90 empresas, lo que representa el 0,08% de las empresas registradas en la región, dando cuenta de un bajo dinamismo productivo.

Por su parte, la región de la Araucanía es una de las regiones con mayores niveles de pobreza, una tasa de ocupación 8 puntos porcentuales por debajo del promedio nacional según las estadísticas territoriales de la Biblioteca del Congreso Nacional de Chile BCN (2016) y con un promedio de años de escolaridad un año menor al promedio nacional. En esta región se encuentra Cholchol, la comuna con la segunda mayor tasa de pobreza (21,3%) y con una pobreza extrema del 4,6%. En la provincia de Cautín, donde se encuentra esta comuna, el 61% de la población no cuenta con acceso a servicios básicos (agua, drenaje, electricidad), porcentaje muy por encima del promedio en la región de la Araucanía (27%) según las estadísticas territoriales de la Biblioteca del Congreso Nacional de Chile BCN (2015). Se estima que el 77% de la población se identifica como perteneciente a pueblos originarios, en su mayoría correspondiente al pueblo Mapuche.

En lo que respecta al acceso a la salud, la comuna cuenta un consultorio general rural y 3 postas de salud rural (establecimientos de atención ambulatoria), frente a los 23 consultorios generales con los que cuenta la región de la Araucanía y las 193 postas de salud rural. La comuna cuenta con 232 empresas, que representan el 0,44% de las empresas registradas en la región de la Araucanía. En esta misma región está la comuna de Galvarino, también en la provincia de Cautín, con una tasa estimada de 20,7% de pobreza y de 4,5% de pobreza extrema. Según informes de la Biblioteca del Congreso Nacional de Chile BCN, el 68% de los habitantes de la comuna no cuentan con servicios básicos de agua, electricidad o alcantarillado y el porcentaje de hogares en situación de hacinamiento alcanza el 20%.

Por otra parte, las comunas de Torres del Paine en la provincia de última Esperanza; San Gregorio, Laguna Blanca y Río Verde en la provincia de Magallanes; Primavera y Timaukel en la provincia de Tierra del Fuego y las comunas de La Reina, Ñuñoa, Providencia, Las Condes y Vitacura en la provincia de Santiago, presentan un estimado de pobreza inferior al 1%, siendo las comunas de Las Condes y Vitacura las que de menor tasa de pobreza monetaria del país con el 0,2% y 0,1% respectivamente. Cabe mencionar, que La Antártica Chilena presenta una tasa de pobreza de cero, ya que la población del lugar corresponde

principalmente a personal militar y sus familias, con un número no superior a las 160 personas, las cuales viven en bases bajo la responsabilidad del gobierno nacional.

En lo que respecta a la tasa de pobreza extrema, en el Cuadro 16, se presentan las 20 comunas con mayores incidencias de este indicador. De manera similar a lo descrito para la pobreza, las comunas con los mayores niveles de pobreza extrema se concentran en las regiones del Biobío, la Araucanía, Atacama y Arica, aunque en este caso también figura una comuna de la Región Metropolitana y La Pintana.

Cuadro 16
Chile: 20 comunas con mayores niveles de pobreza extrema
(En porcentajes)

Código provincia	Provincia	Comuna	Estimación pobreza extrema	LI	LS
8314	Biobío	Alto Biobío	8,73%	3,9%	13,6%
15202	Parinacota	General Lagos	7,03%	0,8%	13,2%
3103	Copiapó	Tierra Amarilla	5,24%	3,8%	6,7%
3302	Huasco	Alto del Carmen	4,74%	2,8%	6,6%
3303	Huasco	Freirina	4,62%	3,3%	6,0%
9121	Cautín	Cholchol	4,61%	1,9%	7,3%
9106	Cautín	Galvarino	4,50%	1,3%	7,7%
2301	Tocopilla	Tocopilla	4,44%	3,3%	5,5%
3201	Chañaral	Chañaral	4,17%	3,1%	5,3%
8302	Biobío	Antuco	4,03%	2,0%	6,1%
2102	Antofagasta	Mejillones	3,91%	2,8%	5,0%
4303	Limarí	Monte Patria	3,85%	2,4%	5,3%
1403	Tamarugal	Colchane	3,79%	-0,3%	7,9%
16108	Ñuble	Ninhue	3,67%	1,7%	5,7%
13112	Santiago	La Pintana	3,65%	3,0%	4,3%
8207	Arauco	Tirúa	3,65%	0,6%	6,7%
7202	Cauquenes	Chanco	3,60%	1,2%	6,0%
3102	Copiapó	Caldera	3,56%	2,8%	4,3%
9205	Malleco	Lonquimay	3,55%	0,6%	6,5%
7203	Cauquenes	Pelluhue	3,54%	1,4%	5,7%

Fuente: Elaboración propia.

B. Resultados para Colombia

A partir del modelo SAE presentado se genera estimaciones de pobreza para cada uno de los 1.122 municipios de Colombia. Estos se pueden clasificar en tres grupos, según sus niveles de pobreza, tal como se observa en el Cuadro 17. De esta manera, el grupo uno (1) contempla nivel de pobreza igual o inferior al 15%; el grupo 2, mayor a 15% y menor a 40%; finalmente, el tercer grupo, nivel de pobreza igual o superior al 40%.

Así, La tercera parte de los municipios de Colombia tiene niveles de pobreza superiores al 40%. En este grupo resaltan los departamentos del Chocó, Guainía y Vaupés, con la totalidad de sus municipios (30, 6 y 9 respectivamente) clasificados en este grupo. A su vez, los departamentos de Magdalena, Nariño, Cauca y Amazonas tienen más del 80% de sus municipios están clasificados en esta categoría. Por otra parte, solo 64 municipios del país presentan tasas de pobreza iguales o inferiores al 15%, lo que equivale al 5,7% del total de municipios. Estos resultados muestran una distribución concentrada de los recursos

económicos del país en pocas regiones, esencialmente en las ciudades principales y zonas urbanas intermedias. En este grupo, los departamentos de San Andrés y Cundinamarca presentan el mayor número de municipios con tasas de pobreza inferiores al 15%. Los municipios que concentran los mayores niveles de pobreza monetaria son aquellos pertenecientes a los departamentos ubicados en la periferia del país, entre los que se encuentran los departamentos de la región Pacífico (Chocó, Cauca y Nariño), región Caribe (La Guajira, Magdalena, Bolívar, Sucre, Córdoba y Cesar), departamentos fronterizos (Norte de Santander y Arauca), y la región Orinoquía y Amazonía (Vichada, Guainía, Vaupés, Amazonas y Caquetá).

Cuadro 17
Porcentaje de municipios por nivel de pobreza
(En porcentajes)

Grupo	Nivel pobreza	Municipios por nivel	Participación
1	Menor o igual a 15%	385	34,3%
2	Entre 16% y 39%	64	5,7%
3	Igual o mayor a 40%	673	60,0%

Fuente: Elaboración propia.

Cuadro 18
Colombia: 20 municipios con mayores niveles de pobreza
(En porcentajes)

Código municipio	Departamento	Municipio	Estimación Pobreza	LI	LS
27025	Chocó	Alto Baudó	80,64%	76,15%	85,13%
52427	Nariño	Magüí Payán	77,25%	69,56%	84,94%
44847	La Guajira	Uribia	75,21%	72,17%	78,24%
19785	Cauca	Sucre	75,02%	65,44%	84,61%
27250	Chocó	El Litoral del San Juan	72,70%	60,65%	84,75%
97511	Vaupés	Pacoa	72,06%	62,00%	82,12%
27077	Chocó	Bajo Baudó	71,57%	66,62%	76,52%
27073	Chocó	Bagadó	71,07%	65,29%	76,86%
27600	Chocó	Río Quito	70,76%	63,15%	78,37%
94888	Guainía	Morichal	70,03%	55,82%	84,25%
27372	Chocó	Juradó	69,96%	63,57%	76,36%
27425	Chocó	Medio Atrato	69,29%	63,36%	75,23%
27099	Chocó	Bojayá	68,70%	57,76%	79,64%
27413	Chocó	Lloró	68,26%	61,01%	75,51%
27580	Chocó	Río Iró	68,01%	58,99%	77,04%
94887	Guainía	Pana Pana	66,52%	49,32%	83,72%
52565	Nariño	Providencia	66,38%	59,31%	73,44%
27150	Chocó	El Carmen del Darién	66,15%	56,56%	75,74%
27430	Chocó	Medio Baudó	65,96%	55,98%	75,94%

Fuente: Elaboración propia.

De acuerdo con el cuadro 18, el municipio Alto Baudó presenta los mayores niveles de pobreza monetaria del país, con una tasa estimada del 80,6%. Según el DANE (2018), este municipio cuenta con uno

de los indicadores de analfabetismo más altos del país (47%), un 40% de la población sin acceso a servicios de salud y una tasa del 84% de la población en desocupación. Según el Instituto Colombiano de Bienestar Familiar ICBF (2018), distintos factores que, sumados a índices elevados de violencia, desnutrición infantil, MINSALUD (2015) desescolarización y embarazo adolescente entre otros, influyen en la elevada tasa pobreza monetaria; DANE Resultados Censo Nacional de Población y Vivienda (2018).

El municipio de Magüi Payán alcanza una tasa del 77,3% de pobreza monetaria. En este municipio, conformado en un 97% por población afrodescendiente y 3% de población indígena, entre un 3% y un 5% de la población está en condición de desplazamiento forzado. Este municipio presenta el mayor porcentaje de rezago escolar (53%) y presenta una alta prevalencia de violencia perpetrada por grupos armados ilegales, reclutamiento forzado (infantil y adultos), trabajo infantil, cultivos de uso ilícito, entre otros factores determinantes de la elevada pobreza, Según informe MIRA (2018).

Otro de los municipios con altos niveles de pobreza (75,2%) es Uribia, en el departamento de La Guajira, conformado principalmente por el pueblo indígena Wayúu, con un 78% de su población perteneciente a dicha etnia. Según la información del DANE en la toma de decisiones regionales Riohacha, Uribia (2021), es el municipio con el mayor crecimiento poblacional del departamento, con un crecimiento del 41% de la población Wayúu entre el 2005 y el 2018. El municipio presenta déficits importantes en el suministro de agua potable, altos niveles de desnutrición infantil, altas tasas de analfabetismo, inasistencia escolar y hacinamiento.

En el caso de los municipios de menor incidencia de pobreza, estos se concentran en los departamentos del interior del país, correspondientes a Antioquia, Santander, Boyacá, Cundinamarca, Eje Cafetero y Valle del Cauca. Los municipios de El Retiro, Sabaneta y Envigado en el departamento de Antioquia, y Tabio y Chía en el departamento de Cundinamarca tienen niveles de pobreza inferiores al 8%, consistentes con altos niveles de alfabetismo, escolarización y acceso a salud y servicios básicos.

De acuerdo al cuadro 19, el panorama de la pobreza extrema en Colombia es similar al descrito para la pobreza monetaria. Las mayores concentraciones se presentan en el departamento del Chocó, con 13 municipios en la lista de municipios con mayores niveles de pobreza y de pobreza extrema. Los municipios de Alto Baudó y Uribia en La Guajira se encuentran entre los tres primeros con la estimación de pobreza más alta y se repiten tanto en pobreza como en pobreza extrema, lo cual confirma la gravedad de la situación en dichas regiones.

Cuadro 19
Pobreza extrema estimada por municipio, Colombia
(En porcentajes)

Código municipio	Departamento	Municipio	Estimación pobreza extrema	LI	LS
27025	Chocó	Alto Baudó	66,12%	61,6%	70,6%
27250	Chocó	El Litoral del San Juan	55,86%	43,8%	67,9%
44847	La Guajira	Uribia	55,29%	52,3%	58,3%
27077	Chocó	Bajo Baudó	55,08%	50,1%	60,0%
27073	Chocó	Bagadó	54,90%	49,1%	60,7%
19785	Cauca	Sucre	52,56%	43,0%	62,2%
27600	Chocó	Río Quito	52,44%	44,8%	60,1%
27099	Chocó	Bojayá	52,00%	41,1%	62,9%
27425	Chocó	Medio Atrato	51,83%	45,9%	57,8%
27372	Chocó	Juradó	51,52%	45,1%	57,9%
97511	Vaupés	Pacoa	51,46%	41,4%	61,5%

Código municipio	Departamento	Municipio	Estimación pobreza extrema	LI	LS
94888	Guainía	Morichal	50,05%	35,8%	64,3%
52427	Nariño	Magüí Payán	49,85%	42,2%	57,5%
27413	Chocó	Lloró	49,17%	41,9%	56,4%
27430	Chocó	Medio Baudó	48,93%	39,0%	58,9%
27580	Chocó	Río Iró	48,68%	39,7%	57,7%
27150	Chocó	El Carmen del Darién	48,58%	39,0%	58,2%
27495	Chocó	Nuquí	47,41%	42,1%	52,7%
94887	Guainía	Pana Pana	45,70%	28,5%	62,9%
91430	Amazonas	La Victoria	42,99%	28,6%	57,4%

Fuente: Elaboración propia.

C. Resultados para Perú

En el Cuadro 20 se presentan las 20 provincias con mayores tasas de pobreza del Perú. La provincia de Datem del Marañón ubicada en el departamento de Loreto presenta la mayor tasa de pobreza (42,7%). Según el Banco Central de Reserva del Perú, en el departamento de Loreto habitan al menos 11 de las 13 familias lingüísticas existentes de la Amazonía, y el 47% del total de los habitantes del departamento hacen parte de alguna comunidad indígena, que se concentran mayormente en las provincias de Datem del Marañón y Maynas. Por otro lado, según el MINSA, en el departamento de Loreto se presenta una tasa promedio de desnutrición en menores de 5 años cercana al 25%, superior al promedio nacional, y que en la provincia Datem del Marañón alcanza un 35%.

Cuadro 20
Pobreza estimada por provincia, Perú
(En porcentajes)

Código provincia	Departamento	Provincia	Estimación pobreza	LI	LS
1607	Loreto	Datem del Marañón	42,73%	38,20%	47,26%
811	Cusco	Paucartambo	42,31%	37,46%	47,16%
1902	Pasco	Daniel Alcides Carrión	40,96%	36,00%	45,92%
2102	Puno	Azángaro	38,87%	35,67%	42,08%
613	Cajamarca	Santa Cruz	38,80%	34,46%	43,15%
505	Ayacucho	La Mar	38,49%	34,82%	42,16%
1605	Loreto	Requena	38,09%	33,27%	42,90%
2103	Puno	Carabaya	37,84%	34,01%	41,67%
2110	Puno	San Antonio de Putina	37,51%	33,62%	41,41%
1004	Huánuco	Huacaybamba	37,26%	32,35%	42,17%
805	Cusco	Canas	37,20%	32,59%	41,80%
1603	Loreto	Loreto	37,11%	32,40%	41,82%
602	Cajamarca	Cajabamba	37,09%	33,23%	40,96%
1602	Loreto	Alto Amazonas	37,05%	33,26%	40,85%
2113	Puno	Yunguyo	36,90%	33,52%	40,27%

Código provincia	Departamento	Provincia	Estimación pobreza	LI	LS
1604	Loreto	Mariscal Ramón Castilla	36,82%	32,42%	41,23%
1606	Loreto	Ucayali	36,45%	32,59%	40,30%
2108	Puno	Melgar	35,97%	32,22%	39,71%
611	Cajamarca	San Miguel	35,70%	31,85%	39,54%
104	Amazonas	Condorcanqui	35,40%	31,69%	39,12%

Fuente: Elaboración propia.

Según información del INEI, el 54% de la población de Loreto no cuenta con acceso a servicio de agua y el 64% carece de servicio de saneamiento, valores que superan ampliamente el promedio nacional (33% y 41% respectivamente). Los porcentajes de población sin acceso a agua y saneamiento alcanzan 95% y 90%, respectivamente, en la provincia de Datem del Marañón. Asimismo, se estima que el 14% de las personas de esta provincia están en condición de analfabetismo. Al tener en cuenta los datos proporcionados por las diversas fuentes citadas, se evidencia una relación entre los altos niveles de pobreza y las malas condiciones de vida, y que afectan principalmente a la población de las comunidades indígenas.

La provincia de Pucartambo en el departamento de Cusco presenta una tasa de pobreza del 42,3%, lo cual la ubica en el segundo lugar de las provincias con mayores niveles de pobreza. Según datos proporcionados por el INEI, es una de las provincias con las mayores tasas de desnutrición infantil del país y la mayor parte de su población carece de los servicios públicos básicos. Por su parte, la provincia Daniel Alcides Carrión, en el departamento de Pasco, tiene un nivel estimado de pobreza del 41,0%. En este departamento, el 79% de la población no cuenta con servicio de agua y el 83% no cuenta con acceso a servicio de saneamiento; además, se estima que el 10% de la población mayor de 15 años es analfabeta y el 21% de los niños menores de 5 años presenta desnutrición crónica.

En el extremo opuesto, las provincias de Ica, Chincha, Pisco y Nazca en la provincia de Ica, presentan los menores niveles de pobreza monetaria del país, situándose por debajo del 2%. Según el INEI, el departamento de Ica, presenta una tasa de alfabetismo del 97% y es uno de los departamentos con mayor cobertura de servicios públicos del país, con el 84% de los hogares con servicio de agua, el 94% con acceso a servicios de saneamiento y el 98% con servicio de luz eléctrica.

En lo que respecta a la pobreza extrema, en el cuadro 21 se repiten Datem del Marañón y Paucartambo como las provincias con las tasas más altas de este indicador.

Cuadro 21
Pobreza extrema estimada por municipio. Perú
(En porcentajes)

Código provincia	Departamento	Provincia	Estimación pobreza extrema	LI	LS
1607	Loreto	Datem del Marañón	17,70%	16,43%	18,96%
811	Cusco	Paucartambo	15,44%	13,90%	16,98%
1604	Loreto	Mariscal Ramón Castilla	13,52%	12,44%	14,60%
1602	Loreto	Alto Amazonas	13,11%	12,22%	13,99%
1603	Loreto	Loreto	12,96%	11,76%	14,15%
1004	Huánuco	Huacaybamba	12,94%	11,80%	14,07%
2207	San Martín	Picota	12,56%	11,82%	13,29%
2110	Puno	San Antonio de Putina	12,45%	11,30%	13,61%
1605	Loreto	Requena	12,45%	11,17%	13,73%

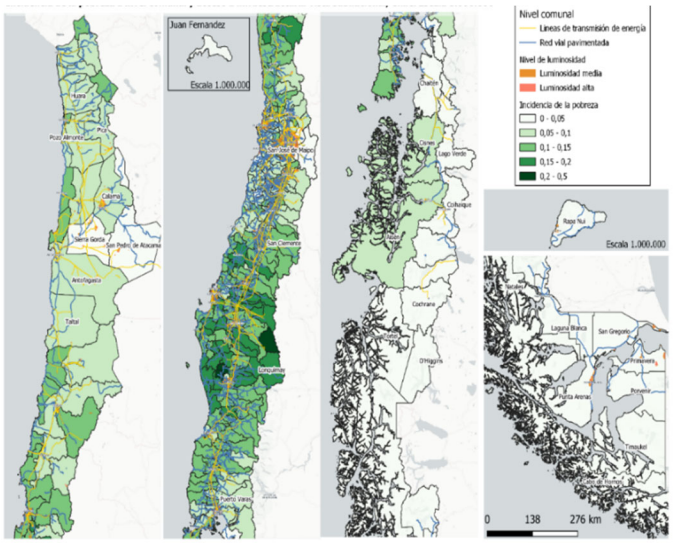
Código provincia	Departamento	Provincia	Estimación pobreza extrema	LI	LS
2113	Puno	Yunguyo	12,31%	11,46%	13,15%
1606	Loreto	Ucayali	12,28%	11,38%	13,18%
2102	Puno	Azángaro	12,16%	11,21%	13,10%
805	Cusco	Canas	12,09%	10,76%	13,42%
613	Cajamarca	Santa Cruz	11,47%	10,36%	12,58%
2109	Puno	Moho	11,25%	10,47%	12,03%
1902	Pasco	Daniel Alcides Carrión	11,19%	9,62%	12,77%
1306	La Libertad	Otuzco	10,93%	10,24%	11,63%
2103	Puno	Carabaya	10,86%	9,67%	12,06%
1309	La Libertad	Sánchez Carrión	10,72%	9,94%	11,49%
2104	Puno	Chucuito	10,42%	9,65%	11,19%

Fuente: Elaboración propia.

D. Mapas de pobreza con información complementaria

Una de las ventajas de mostrar las cifras sobre pobreza en un mapa es la posibilidad de relacionar estos datos con información geográfica relevante. En los gráficos 20, 21 y 22 se presentan los mapas de pobreza de Chile, Colombia y Perú incluyendo información sobre la infraestructura de redes viales pavimentadas y líneas de transmisión de energía eléctrica.

Mapa 6
Contraste entre la estimación SAE de la pobreza con la red vial pavimentada en las comunas de Chile



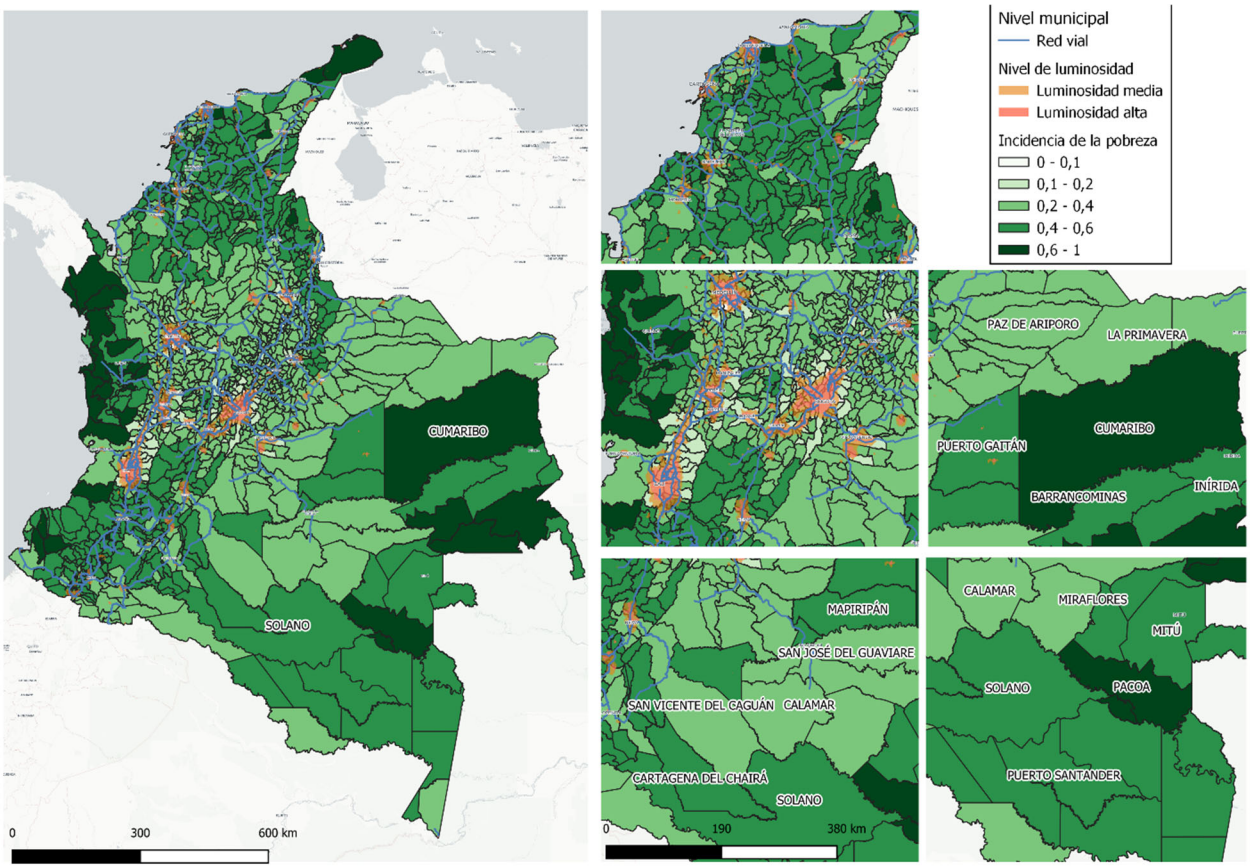
Fuente: Elaboración propia. para la elaboración de este mapa se usaron los datos de la red eléctrica que proporciona Frontel Chile en su página web[6] y también, fueron usados los datos de la red vial nacional que proporciona el Ministerio de Obras Públicas de Chile en la página web[7] de infraestructura de datos geoespaciales de Chile IDE.

6 https://web.gruposaesa.cl/web/frontel/tendido-electrico.

7 https://www.ide.cl/index.php/transporte/item/1708-red-vial-nacional.

En los tres países, se observa una relación inversa entre los niveles de pobreza y la disponibilidad de infraestructura. Cabe esperar que la escasa disponibilidad de infraestructura impacte negativamente sobre las posibilidades de acceso a recursos económicos y mercados productivos, haciendo más difícil para la población encontrar formas de generar ingresos suficientes para satisfacer sus necesidades básicas. En ese sentido, una mayor inversión en infraestructura vial y de redes eléctricas en los países analizados contribuiría a aliviar la pobreza en las zonas donde estos servicios son deficientes, no solo a través de un mayor acceso a medios productivos y de intercambios, sino facilitando el acceso a establecimientos educativos y de salud. No obstante, cabe reconocer que las posibilidades de avance en este ámbito se ven afectados por el contexto geográfico, ya que en los países analizados varias de las zonas de alta pobreza son también áreas de difícil accesibilidad (selvas o montañas) o de alta dispersión poblacional (zonas desérticas), en las que el desarrollo de infraestructura enfrenta desafíos particulares.

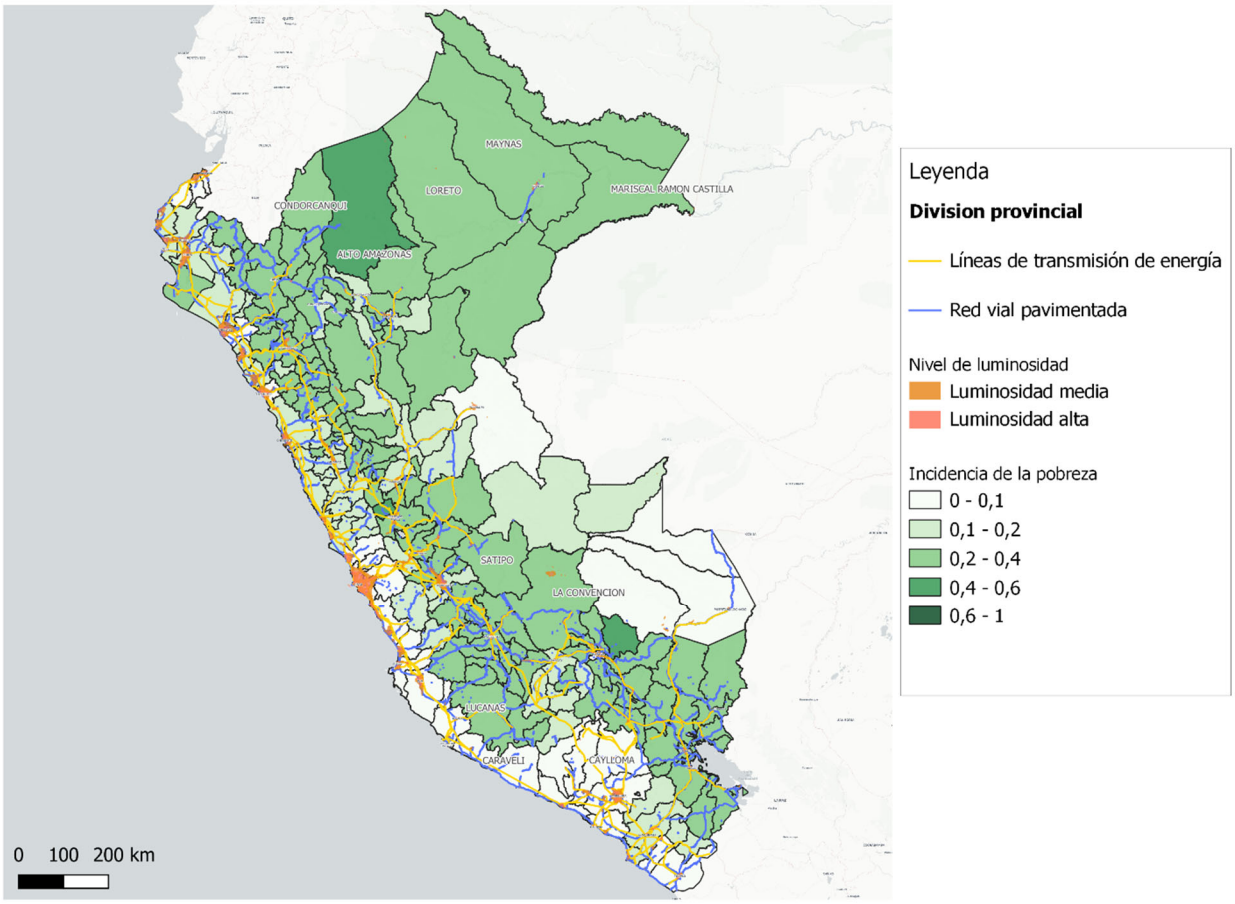
Mapa 7
Contraste entre la estimación SAE de la pobreza con la red vial pavimentada y luminosidad de los municipios de Colombia



Fuente: Elaboración propia. Para la elaboración de este mapa se usaron los datos que suministra el Instituto Nacional de Vías, INVIAS en la página web[8] en la que el gobierno Colombiano dispone para proporcionar información de carácter público.

8 Red vial: https://www.datos.gov.co/Transporte/Red-Vial/dkuf-xn5d.

Mapa 8
Contraste entre la estimación SAE de la pobreza con la red vial pavimentada y luminosidad de las provincias de Perú



Fuente: Elaboración propia. Para la elaboración de este mapa se usaron los datos que proporciona Geo GPS del Perú en su página web[9], y además, también fueron usados los datos que proporciona el Ministerio de Transporte y Comunicaciones del Perú en su página web[10].

9 https://www.geogpsperu.com/2020/09/mapa-de-electricidad-lineas-de.html.
10 Red vial: https://portal.mtc.gob.pe/estadisticas/descarga.html.

VI. Conclusiones

La información recolectada mediante las encuestas de hogares es valiosa, pero tiene una capacidad limitada de desagregación. Los modelos de estimación en áreas pequeñas proveen una vía útil para obtener estimaciones desagregadas con mayores niveles de precisión y confiabilidad que los de las estimaciones directas, especialmente en condiciones en los que se cuenta con un tamaño de muestra deficiente o incluso cuando llega a ser inexistente. El método presentado en el presente documento es un modelo de unidad, que combina la información de las encuestas con la de los censos de población para estimar la pobreza extrema, la pobreza y los ingresos medios de la población a nivel de cada hogar, para después agregar los resultados a nivel de división administrativa menor. Si bien en este documento el método se aplica a la información de tres países, su extensión a todos los países de la región que cuenten con encuestas y censos recientes es totalmente viable.

Existe un margen amplio para mejorar las potencialidades de la combinación de los datos de censos y encuestas. En la medida que las metodologías de SAE se incorporen en los procesos de generación de estadísticas oficiales, es posible tener en cuenta la complementariedad de estas fuentes tanto en el tipo de preguntas y categorías de respuesta en los cuestionarios, como en el diseño de la muestra.

Los resultados permiten contrastar las tasas de pobreza estimadas por métodos SAE con otras fuentes de información disponibles en cada país. En general, en los tres países se constata que las áreas con mayores niveles de pobreza se caracterizan por un menor acceso a bienes y servicios básicos y cuentan una infraestructura deficitaria, que se manifiesta en una menor cantidad de establecimientos educativos y centros de salud, así como menor disponibilidad de redes viales. Esta información puede servir como insumo en estudios que aborden el impacto de la inversión pública para el desarrollo de infraestructura sobre la reducción de la pobreza monetaria y la mejora de las condiciones de vida.

Si se considera la característica de comparabilidad regional que tienen las estimaciones de pobreza de la CEPAL, los resultados obtenidos con este estudio pueden ser utilizados para realizar algunas comparaciones a nivel de áreas pequeñas entre países. Por ejemplo, es posible constatar que mientras en Chile el 90% de las comunas tienen niveles de pobreza inferiores al 15%, en Perú el 28% de las provincias está en el mismo rango, y en Colombia, solo el 5% de los municipios. A su vez, en Chile ninguna comuna presenta niveles de pobreza

superiores al 40% y en Perú solo el 2% de las provincias lo hace, mientras que en Colombia un 34% de los municipios excede dicho umbral. No obstante, debe reconocerse que la posibilidad de realizar este tipo de comparaciones está limitada por la falta de especificidad de las líneas de pobreza para reflejar el costo de vida en áreas pequeñas cuando estas enfrentan niveles de precios o estructuras de consumo que difieren ostensiblemente de los promedios urbanos o rurales.

Bibliografía

Akaike, H. (1974), A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, 19(6), pp. 716-723.

Amat, J. (2016), Selección de predictores, regularización ridge, lasso, elastic net y reducción de dimensionalidad.

Arima, S. et. al. (2017), Multivariate Fay-Herriot Bayesian Estimation of Small Area Means Under Functional Measurement Error. Journal of the Royal Statistical Society - Series A, 180(4), pp. 1191-1209.

Arlot, S. & Celisse, A. (2010), A Survey of Cross Validation Procedures for Model Selection. Statistics Surveys, Volumen 4, pp. 40-79.

Banerjee, M. & Frees, E. W. (1997), Influence Diagnostics for Linear Longitudinal Models. Journal of the American Statistical Association, 92(439), pp. 999-1005.

Barnett-Walker, K. C. et. al. (2003), 2001 National Household Survey on Drug Abuse, s.l.: s.n.

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988), An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), pp. 28-36.

Bedi, T., Coudouel, A. & Simler, K (2007), More than a pretty picture: using poverty maps to design better policies and interventions. 20433 ed. Washington, D.C.: The World Bank.

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.

Beresovsky V, Burt CW, Parsons V, Schenker N & Mutter R. (2011), Application of hierarchical Bayesian nodels with poststratification for small-area estimation from complex survey data In Proceedings of Section on Survey Research Methods, pp. 4745–4756. Alexandria: American Statistical Association.

Biblioteca del Congreso Nacional de Chile, BCN (2016), Reporte regional. https://www.bcn.cl/siit/reportes regionales/reporte_final.html?anno=2021&cod_region= 8.

_____(2015), Estadística territoriales. https://www.bcn.cl/siit/reportescomunales/comunas_v.html?anno =2015& idcom=8314.

Bickel, P. J. & Doksum, K. A. (2015), Mathematical Statistics: Basic Ideas and Selected Topics, Volume I. 2 ed. New York, New York: Chapman and Hall/CRC.

Bishop, C. M. (2006), Pattern recognition and machine learning. 1 ed. New York, New York: Springer.

Brown G, Chambers R, Heady P & Heasman D. (2001), Evaluation of small area estimation methods an application to unemployment estimates from the UK LFS In Proceedings of Statistics Canada Symposium.

Burgard J.P., Krause. J., Kreber D., & Morales D. (2021), The generalized equivalence of regularization and min–max robustification in linear mixed models. Statistical Papers volume 62, pages2857–2883 (2021),

Burnham, K. P. & Anderson, D. R. (2002), Model selection and multimodel inference: a practical information-theoretic approach. 2 ed. New York, New York: Springer.

Burnham, K. P. & Anderson, D. R. (2004), Multimodel Inference: understanding AIC and BIC in model selection. Sociological Methods & Research, 33(2), p. 261–304.

Buscemi, S. & Plaia, A. (2020), Model selection in linear mixed-effect models. AStA Advances in Statistical Analysis, 104(4), p. 529–575.

Calvin, J. A. (1993), REML Estimation in Unbalanced Multivariate Variance Components Models Using an EM Algorithm. Biometrics, 49(3), pp. 691-701.

Casella, G. & Berger, R. L. (2002), Statistical Inference. 2 ed. Pacific Grove: Duxbury Press.

Castro S. (2013), Estimación y selección de variables en grandes dimensiones, Regresión Ridge, GNN, Lasso, Elastic Net, SCAD.

Christensen, R., Pearson, L. M. & Johnson, W. (1992), Case-Deletion Diagnostics for Mixed Models. Technometrics, 34(1), pp. 38-45.

Claeskens, G. & Hart, J. D. (2009), Goodness-of-Fit Tests in Mixed Models. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, 18(2), pp. 213-239.

Comisión Económica para América Latina y el Caribe (CEPAL), (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), Panorama Social de América Latina. LC/PUB.2019/3-P ed. Santiago: Comisión Económica para América Latina y el Caribe (CEPAL).

_____(2018), Medición de la pobreza por ingresos: actualización metodológica y resultados. Santiago de Chile: Metodologías de la CEPAL, N° 2, (LC/PUB.2018/22-P).

Comité Nacional de Expertos (CNE) (2019), Evaluación del Censo Nacional de Población, Bogotá, D.C.: Comité Nacional de Expertos.

Cook, R. D. (1977), Detection of influential observation in linear regression. Technometrics, 19(1), pp. 15-18.

Corral, P., Himelein, K., McGee, K. & Molina, I. (2021), A Map of the Poor or a Poor Map?. Policy Research Working Paper, World Bank, Volumen 9620.

Corral-Rodas, P., Molina, I. & Nguyen, M. (2021), Pull your small area estimates up by the Bootstraps, Journal of Statistical Computation and Simulation, 91(16), pp. 3304-3357.

Cryer, J. D. & Chan, K.-S. (2008), Time Series Analysis with Applications in R. 2 ed. New York, New York: Springer.

Danafar, S., Fukumizu, K. & Gomez, F. (2014), Kernel-Based Information Criterion. Computer and Information Science, 8(1), pp. 1-15.

Das, S. & Chambers, R. (2017), Robust mean-squared error estimation for poverty estimates based on the method of elbers, lanjouw and lanjouw. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(4), pp. 1137-1161.

Datta, G. S. & Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica, 10(2), pp. 613-627.

Datta, G. S., Rao, J., Torabi, M. & Liu, B. (2018), Small Area Estimation with Multiple Covariates Measured with Errors: A Nested Error Linear Regression Approach of Combining Two Surveys. Journal of Multivariate Analysis, Volumen 167, pp. 49-59.

Departamento Administrativo Nacional de Estadística (DANE) (2021), Diseño DSO Ficha Metodológica Censo Nacional de Población y Vivienda 2018, Bogotá, D.C.: Dirección de Censos y Demografía (DCD), Departamento Administrativo Nacional de Estadística.

_____(2021), Gran Encuesta Integrada de Hogares - GEIH - 2020, Bogotá, D.C.: Dirección de Metodología y Producción Estadística - DIMPE.

_____(2021), Gran Encuesta Integrada de Hogares GEIH - 2020, Bogotá: DANE.

_____(2020), La información del DANE en la toma de decisiones de los departamentos.
_____(2020), La información del DANE en la toma de decisiones regionales Riohacha, Uribia – La Guajira.
_____(2018), Resultados Censo Nacional de Población y Vivienda 2018. Riosucio, Quibdó, Chocó.
_____(2009), Metodología Gran Encuesta Integrada de Hogares. 1 ed. Bogotá: DANE, Dirección de Metodología y Producción Estadística (DIMPE), Imprenta Nacional de Colombia.
_____(2008), Especificaciones del proceso de conciliación censal como herramienta para medición de la omisión censal a nivel nacional Censo 2005, Bogotá, D.C.: Departamento Administrativo Nacional de Estadística.
Diallo, M. S. & Rao, J. N. K. (2018), Small area estimation of complex parameters under unit-level models with skew-normal errors. Scandinavian Journal of Statistics, 45(4), pp. 1092-1116.
División de Recursos Naturales e Infraestructura, CEPAL (2004), Desarrollo de infraestructura y crecimiento económico: revisión conceptual. 1 ed. Santiago: Comisión Económica para América Latina y el Caribe (CEPAL).
Dorfman, A.H. (2018), Towards a Routine External Evaluation Protocol for Small Area Estimation, Int Stat Rev., 86(2), p. 259-274.
Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2003), Micro–level estimation of poverty and inequality. Econometrica, 71(1), p. 355–364.
Elbers, C. & Lanjouw, P. (2001), Intersectoral Transfer, Growth, and Inequality in Rural Ecuador. World Development, 29(3), pp. 481-496.
Engstrom, R., Hersh, J. & Newhouse, D. (2017), Poverty from space using high-resolution satellite imagery for estimating economic well-being. Policy Research Working Paper, World Bank, Washington, DC., Volumen 8284.
Falorsi PD & Righi P. (2008), A balanced sampling approach for multi-way stratification designs for small area estimation. Survey Methodology, 34(2), pp. 223–234.
Fan, J. & Li, R. (2006), Statistical challenges with high dimensionality: Feature selection in knowledge discovery. Madrid, España, Proceedings of the Madrid International Congress of Mathematicians.
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).
Filgueira, C. & Peri, A. (2004), América Latina: los rostros de la pobreza y sus causas determinantes. Serie CEPAL Nº 54 ed. Santiago de Chile: Población y Desarrollo.
Filgueira, F. (2009), El desarrollo maniatado en América Latina: estados superficiales. 1 ed. Buenos Aires: Consejo Latinoamericano de Ciencias Sociales - CLACSO.
Foster, J., Greer, J. & Thorbecke, E. (1984), A class of decomposable poverty measures. Econometrica, 52(3), pp. 761-766.
Franco, C. (2020), Comparison of small area models for estimation of U.S. county poverty rates of school-aged children using an artificial population and a design-based simulation, Washington, D.C.: Census Bureau Research Report Series RRS2020/04.
Franco, C. & Bell, W. (2020), Using American Community Survey Data to Improve Estimates from Smaller Surveys through Bivariate Small Area Estimation Models, Washington, D.C.: Center for Statistical Research & Methodology, Research and Methodology Directorate, U.S. Census Bureau.
Fuller, W. A. & Battese, G. E. (1973), Transformations for Estimation of Linear Models with Nested-Error Structure. Journal of the American Statistical Association, 68(343), pp. 626-632.
González-Manteiga, W. et. al. (2008), Bootstrap mean squared error of a small-area EBLUP. Journal of Statistical Computation and Simulation, 78(5), pp. 443-462.
Graf, M., Marín, J. M. & Molina, I. (2019), A generalized mixed model for skewed distributions applied to small area estimation. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, 28(2), pp. 565-597.
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. (2018), Limitaciones de las encuestas de hogares en la medición de indicadores sociales. Santiago de Chile, Presentación en «Taller regional sobre desagregación de estadísticas sociales mediante metodologías de SAE.

Gutiérrez, A. et. al. (2020), Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional. Serie Estudios Estadísticos, Nº 101 (LC/TS.2020/52) ed. Santiago de Chile: Comisión Económica para América Latina y el Caribe (CEPAL),

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

Gutiérrez, H. A., Zhang, H. & Rodríguez, N. A. (2016), The Performance of Multivariate Calibration on Ratios, Means and Proportions. Revista Colombiana de Estadística, 39(2), pp. 281-305.

Hadam, S., Würz, N. & Kreutzmann, A. K. (2020), Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany. Working paper, Freie Universität.

Henderson, C. R. (1953), Estimation of Variance and Covariance Components. Biometrics, 9(2), pp. 226-252.

Hentschel, J., Lanjouw, J. O., Lanjouw, P. & Poggi, J. (1998), Combining Census and Survey Data to Study Spatial Dimensions of Poverty. The World Bank Economic Review, 14(1), pp. 147-165.

Hentschel, J., Lanjouw, J. O., Lanjouw, P. & Poggi, J. (2000), Combined household data with census data to construct a disaggregated poverty map: a case study of Ecuador. World Bank Economic Review, 14(1), pp. 147-166.

Hentschel, Jesko & Lanjouw, Jean & Lanjouw, Peter & Poggi, Javier (2000), Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador. The World Bank Economic Review. 14. 10.1093/wber/14.1.147.

Hruschka, E. R., Hruschka, E. R. & Ebecken, N. F. F. (2007), Bayesian networks for imputation in classification problems. Journal of Intelligent Information Systems, 29(3), pp. 231-252.

INE; UDAPE (2006), Pobreza y desigualdad en municipios de Bolivia: estimación del gasto combinando el Censo 2001 y las encuestas de hogares. 2 ed. La Paz: Instituto Nacional de Estadística, Unidad de análisis de políticas sociales y económicas.

Instituto Colombiano de Bienestar Familiar ICBF. (2018), Rendición Pública de Cuentas ICBF 2018

Instituto Nacional de Estadística e Informática (INEI) (2019), Censos Nacionales 2017: XII de Población, VII de Vivienda y III de Comunidades Indígenas, Lima: Instituto Nacional de Estadística e Informática - PCM.

_____(2019), Encuesta Nacional de Hogares 2019: diccionario, Lima: Instituto Nacional de Estadística e Informática.

_____(2006), Encuesta Nacional de Hogares (ENAHO), condiciones de vida y pobreza, Lima: INEI.

_____(2005), Encuesta Nacional de Hogares (ENAHO) 2005: diseño muestral, Lima: Instituto Nacional de Estadística e Informática (INEI).

Instituto Nacional de Estadística y Geografía (2021), Encuesta Nacional de Ingresos y Gastos de los Hogares 2020: ENIGH nueva serie: diseño muestral, México: Instituto Nacional de Estadística y Geografía, INEGI.

Instituto Nacional de Estadísticas (INE) (2018), Memoria del Censo 2017, Santiago de Chile: Instituto Nacional de Estadísticas.

_____(2017), Resultados Censo 2017 por país, regiones y comunas.

Jacqmin-Gadda, H. et. al. (2007), Robustness of the linear mixed model to misspecified error distribution. Computational Statistics and Data Analysis, 51(10), pp. 5142-5154.

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

Jean, N. et. al. (2016), Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), pp. 790-794.

Jiang, J., Nguyen, T. & Rao, J. S. (2009), A simplified adaptive fence procedure. Statistics and Probability Letters, 79(5), pp. 625-629.

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).
Kelly, R. J. & Mathew, T. (1994), Improved Nonnegative Estimation of Variance Components in Some Mixed Models with Unbalanced Data. Technometrics, 36(2), pp. 171-181.
Kohavi, R. (1995), A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Volumen 2, pp. 1137-1145.
Konishi, S. & Kitagawa, G. (2008), Information Criteria and Statistical Modeling. 1 ed. New York, New York: Springer.
Kullback, S. & Leibler, R. A. (1951), On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), pp. 79-86.
Lange, S. & Pape, U. J. P. P. (2018), Small area estimation of poverty under structural change. World Bank Policy Research Working Paper, Volumen 8472.
Larrea, C., Carrasco, F., Cervantes, J. & Viedma, N. (1999), Desarrollo social y gestión municipal en el Ecuador: jerarquización y tipología. 1 ed. Quito: Editorial Abya Yala.
Li, Y., & Lahiri P. (2019), A Simple Adaptation of Variable Selection Software for Regression Models to Select Variables in Nested Error Regression Models, The Indian Journal of Statistics, 2019, vol. 81, issue 2, No 6, 302-317.
Longford, N.T. (2006), Sample size calculation for small-area estimation. Survey Methodology, 32(1), pp. 87-96.
Lumley, T. (2010), Complex Surveys: A Guide to Analysis Using R. 1 ed. Hoboken, New Jersey: Wiley.
Maples, J. J. (2011), Using small area modeling to Improve Design-Based estimates of variance for county level poverty rate estimates in the American Community Survey, Washington, D.C.: Research Report RRS2011/02. Center for Statistical Research and Methodology, U.S. Census Bureau.
Marhuenda, Y., Molina, I., Morales, D. & Rao, J. (2017), Poverty mapping in small areas under a twofold nested error regression model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(4), pp. 1111-1136.
Marker, D.A. (2001), Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. Survey Methodology, Vol. 27, pp. 183–188.
Masaki, T. et. al. (2020), Small Area Estimation of non-monetary poverty with geospatial data. World Bank Policy Research Working Paper, Vol 9383.
McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008), Linear, Generalized, and Mixed Models. 2 ed. New York, New York: Wiley Series in Probability and Statistics.
Ministerio de Desarrollo Social (MDS), (2018), CASEN 2017: metodología de diseño muestral, Santiago de Chile: Ministerio de Desarrollo Social.
_____(2018), CASEN 2017: metodología de diseño muestral, Santiago de Chile: MDS.
Ministerio de Desarrollo Social y Familia (2018), Estimaciones de la tasa de pobreza por ingresos y multidimensional a nivel comunal, año 2017, Santiago de Chile: Ministerio de Desarrollo Social y Familia.
_____(2015), Nueva Metodología de Medición de la Pobreza por Ingresos y Multidimensional, Santiago de Chile: Ministerio de Desarrollo Social y Familia.
Ministerio de salud de Colombia MINSALUD (2015), Chocó y sus Determinantes Sociales, Mira, 2018 Informe Final Municipio de Magüí Payan (Nariño, Colombia).
Molefe, W.B. (2011), Sample Design for Small Area Estimation, Ph.D Thesis, University of Wollongong (available online).
Molefe, W.B. & Clark, R.G. (2015), Model-assisted optimal allocation for planned domains using composite estimation. Surv. Method, 27, 183–188. 41, 377–387.
Molina, A. et. al. (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. (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, 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.
Müller, S., Scealy, J. L. & Welsh, A. H. (2013), Model selection in linear mixed models. Statistical Science, 28(2), pp. 135-167.
MDSF-CEPAL (2021), Estimaciones Comunales de Pobreza por ingresos en Chile Mediante Métodos de Estimación en Áreas Pequeñas.
Naciones Unidas, División de Estadística (2010), Principios y recomendaciones para los censos de población y habitación. 2 ed. Nueva York: Naciones Unidas.
National Research Council (2015), Realizing the Potential of the American Community Survey: Challenges, Tradeoffs, and Opportunities, Washington, D.C.: National Academies Press.
Nirel, R. & Glickman, H. (2009), Cap 21. Sample Surveys and Censuses. En: D. Pfeffermann & C. Rao, edits. Handbook of Statistics 29A: sample surveys, design, methods and applications. Amsterdam: Elsevier, pp. 539-565.
Nirel, R., Glickman, H. & Ben-Hur, D. (2003), A strategy for a system of coverage samples for an integrated Census. Ontario, Canada, Proceedings of Statistics Canada Symposium.
Nishii, R. (1984), Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. The Annals of Statistics, 12(2), pp. 758-765.
Parker, J. D. et. al. (2017), National Center for Health Statistics Data Presentation Standards for Proportions. Vital Health Statistics, 2(175), pp. 1-22.
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.
Rao, J. & Molina, I. (2015), Small Area Estimation. 2 ed. Hoboken, New Jersey: Wiley.
Santillán-Pizarro, M. M. & Laplante, B. (2009), La dinámica de la pobreza y las variables de población en la Argentina: un análisis longitudinal a partir de la encuesta permanente de hogares (1995-2003), Notas de población, Centro Latinoamericano y Caribeño de Demografía (CELADE) - División de Población de la CEPAL, Volumen 89, pp. 13-55.
Schwarz, G. (1978), Estimating the dimension of a model. The annals of statistics, 6(2), pp. 461 - 464.
Secretaría Técnica del Frente Social; Banco Mundial (1996), La geografía de la pobreza en el Ecuador, Quito: Secretaría Técnica del Frente Social.
Servicio de Impuestos Internos SII (2016), Estadísticas De Empresas Por Región. https://www.sii.cl/estadisticas/empresas_region.htm.
Singh, M.P., Gambino, J. & Mantel, H.J. (1994), Issues and strategies for small area data. Survey Methodology, Vol 20, pp. 3–22.
Stroup, W. W. (2013), Generalized linear mixed models: modern concepts, methods and applications. 1 ed. Boca Raton, Florida: CRC Press.
Stukel, D. M. & Rao, J. N. K. (1997), Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. Statistics & Probability Letters, 35(4), pp. 401-407.
United Nations (2021), Guidelines on data disaggregation for SDG: Indicators using survey data., Roma, Italia: Food and Agriculture Organization of the United Nations.
Vaida, F. & Blanchard, S. (2005), Conditional Akaike information for mixed-effects models. Biometrika, 92(2), pp. 351-370.
Valliant, R., Dever, J. A. & Kreuter, F. (2018), Practical Tools for Designing and Weighting Survey Samples. 2 ed. New York, New York: Springer.

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011), Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), pp. 1-67.

Venables, W. N. & Ripley, B. D. (2002), Modern Applied Statistics with S. 4 ed. New York, New York: Springer.

Watmough, G. R., Atkinson, P. M., Saikia, A. & Hutton, C. W. (2016), Understanding the evidence base for poverty-environment relationships using remotely sensed satellite data: An example from assam. World Development, Volumen 78, pp. 188-203.

World Bank (2004), Ecuador Poverty Assessment. 1 ed. Washington, D.C.: World Bank.

Zewotir, T. T. & Galpin, J. (2007), A unified approach on residuals, leverages and outliers in the linear mixed model. Test, 16(1), pp. 58-75.

Zhu, H., Ibrahim, J. G. & Cho, H. (2012), Perturbation and scaled Cook's distance. The Annals of Statistics, 40(2), pp. 785-811.

Zou, H. (2006), The adaptive Lasso and its Oracle properties. Journal of the American Statistical Association, 101(476), pp. 1418-1429.

Anexos

Anexo 1
Proceso de Homologación de las bases de datos

Estandarización de la variable años de estudio

La variable años de estudio fue estandarizada utilizando los procesos definidos por el banco de encuestas de hogares (BADEHOG) de CEPAL, pero de forma específica para atender la estructura del censo con el cual se trabajaba para establecer los mismos parámetros. La variable creada para el número de años de estudio para cada persona del hogar se ajusta a la estructura presentada en el cuadro A1, en la cual los niveles educativos "Primaria" y "Secundaria" dependen del último año cursado por la persona y el número de años de estudio por culminar algún nivel educativo específico. A partir de este punto, a los miembros del hogar que tenían nivel educativo superior no universitario se le asignaron 13.5 años de estudio, si aún no habían obtenido el título y de 15 años para quienes sí habían completado sus estudios. De forma similar, a las personas con nivel educativo superior universitario se les asignaron 14.5 años si aún no habían culminado sus estudios y 17 años para quienes ya habían obtenido el título. Finalmente, para los miembros del hogar que poseían magister o doctorado, se imputaron 19 años de estudio.

Cuadro A1
Correspondencia entre el nivel educativo y el número de años de educación

Nivel educativo	Años imputados
Sin nivel, inicial	0
Primaria	0 + último año cursado
Secundaria	6 + último año cursado
Superior -no universitaria incompleta	12 + 1.5
Superior-no universitaria completa	12 + 3
Superior-universitaria incompleta	12 + 2.5
Superior-universitaria completa	12 + 4
Maestría/ doctorado	17 + 2

Fuente: Elaboración propia.

Las particularidades de los Censos por país se listan a continuación: para la Encuesta Nacional de Hogares (ENAHO) se creó la variable número de años de educación para todos los niveles de estudios aprobados empleando el procedimiento previamente mencionado, en cambio, para el XII Censo de Población, VII de Vivienda y III de Comunidades Indígenas el procedimiento de asignación del número de años de escolaridad se realizó hasta el nivel educativo de secundaria. En el caso de Chile y Colombia la variable de años de estudio se encontraba implícita en los microdatos, tanto de las encuestas de hogares como en los censos, por lo cual solo se prosiguió con la categorización respectiva de las variables.

Ocupación principal de desempeño

En Perú, la variable que asociada a la ocupación principal de desempeño está codificada según la Clasificación Internacional Uniforme de Ocupaciones (CIOU-88). En este país tanto para el censo como para la encuesta de hogares, se extrajo el primer dígito de cada respuesta y se formaron diez grupos a partir de estos, tal como se presenta en el cuadro A2. Para cada hogar se generó una variable indicadora que permite clasificar si en este reside una persona cuya ocupación pertenece a dicho grupo.

Para Chile y Colombia, no se definió una pregunta que permitiera establecer la ocupación principal de desempeño de las personas en los módulos censales establecidos, por lo cual, en los microdatos del censo no se pudo realizar la clasificación de cada integrante del hogar en los grupos de ocupación previamente listados. Para estos países la variable asociada a la Ocupación de los miembros del hogar no fue empleada en el estudio al no poder ser homogenizada entre las dos fuentes de información consideradas.

Cuadro A2
Descripción de los grupos generados a partir de la Clasificación Internacional Uniforme de Ocupaciones

Grupo	Descripción
Grupo 1	Miembros del poder ejecutivo y de los cuerpos legislativos y personal directivo de la administración pública y de empresas
Grupo 2	Profesionales científicos e intelectuales
Grupo 3	Técnicos profesionales de nivel medio
Grupo 4	Empleados de oficina
Grupo 5	Trabajadores de los servicios y vendedores de comercio y mercados
Grupo 6	Agricultores y trabajadores calificados agropecuarios y pesqueros
Grupo 7	Oficiales, operáticos y artesanos de artes mecánicas y de otros oficios
Grupo 8	Operadores de instalaciones y máquinas y montadores
Grupo 9	Trabajadores no calificados
Grupo 0	Fuerzas armadas

Fuente: Elaboración propia.

Actividad del negocio donde trabajó

En los microdatos del censo y las encuestas de hogares en Perú y Chile se tiene una variable que permite identificar la actividad económica principal del negocio donde trabaja cada uno de los miembros del hogar, la cual está codificada según la Clasificación Industrial Internacional Uniforme (CIIU revisión 4). El procedimiento de estandarización de los microdatos de ambos países consistió en extraer el primer y segundo dígito de cada respuesta y conformar 21 grupos con la respectiva clasificación económica asociados (cuadroA3).

Cuadro A3
Descripción de los grupos generados a partir de la Clasificación Industrial Internacional Uniforme

Grupo	Descripción
Grupo A	Agricultura, ganadería, silvicultura y pesca
Grupo B	Explotación de minas y canteras
Grupo C	Industrias manufactureras
Grupo D	Suministro de electricidad, gas, vapor y aires acondicionado
Grupo E	Suministro de agua, evacuación de aguas residuales, gestión de desechos y descontaminación
Grupo F	Construcción
Grupo G	Comercio al por mayor y al por menor, reparación de vehículos automotores y motocicletas
Grupo H	Transporte y almacenamiento
Grupo I	Actividades de alojamiento y de servicio de comidas
Grupo J	Información y comunicaciones
Grupo K	Actividades financieras y de seguros
Grupo L	Actividades inmobiliarias
Grupo M	Actividades profesionales, científicas y técnicas
Grupo N	Actividades de servicios administrativos y de apoyo
Grupo O	Administración pública y defensa, planes de seguridad social de afiliación obligatoria
Grupo P	Enseñanza
Grupo Q	Actividades de atención de la salud humana y de asistencia social
Grupo R	Actividades artísticas, de entretenimiento y recreativas
Grupo S	Otras actividades de servicios
Grupo T	Actividades de los hogares como empleadores, actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio
Grupo U	Actividades de organizaciones y órganos extraterritoriales

Fuente: Elaboración propia.

Para cada hogar de Chile y Perú se conformó una variable indicadora que permitía clasificar los hogares en aquellos donde al menos uno de sus integrantes labora en una empresa cuya actividad económica pertenece a dicho grupo.

En el caso de Colombia, en los módulos del censo no se definió una pregunta que permitiera establecer la actividad económica principal del negocio donde trabaja cada uno de los miembros del hogar, por lo cual, no se pudo realizar la clasificación de los hogares respecto a las variables indicadoras definidas por los 21 grupos listados previamente; por tal motivo, esta variable no pudo ser empleada en el estudio al no poder ser homogenizada entre las dos fuentes de información consideradas.

Homologación de las opciones de respuesta para las variables categóricas

Como se indicó anteriormente, el proceso de homologación de las opciones de respuesta entre las encuestas de hogares y los censos es un procedimiento específico a cada uno de los países al estar vinculado a la estructura que definan los Institutos Nacionales de Estadística en los formularios de los módulos correspondientes. Resulta importante aclarar que el proceso de homologación debe realizarse cuando las opciones de respuesta difieren considerablemente entre las fuentes de información consideradas. En el cuadro A4 se presenta, a manera de ejemplo, el proceso de homologación de siete variables empleadas en el estudio del caso colombiano.

Cuadro A4
Homologación de categorías Censo – Encuesta de hogares

Variable	Opciones de respuesta Censo	Opciones de respuesta encuesta	Opciones de respuesta homologadas	Regla para la homologación
Tipo de vivienda	1. Casa	1. Casa	1. Casa	1. Las opciones 3 y 4 de la encuesta se colapsaron en la categoría “3. Tipo cuarto”, categoría asociada a la opción de respuesta 3 del censo.
	2. Apartamento	2. Apartamento	2. Apartamento	2. Las opciones 4 y 5 del censo se unieron en la categoría “4. Vivienda étnica”, mientras que la opción 5 de la encuesta se recodificó de la misma manera.
	3. Tipo cuarto	3. Cuarto(s) en inquilinato	3. Tipo cuarto	3. La opción 6 del censo y de la encuesta se recodificaron en la categoría unificada “Otro (contenedor, carpa, embarcación, vagón, cueva, refugio natural, etc.)”
	4. Vivienda tradicional indígena	4. Cuarto (s) en otro tipo de estructura	4. Vivienda étnica	
	5. Vivienda tradicional étnica (afrocolombiana, isleña, ROM	5. Vivienda indígena	5. Otro (contenedor, carpa, embarcación, vagón, cueva, refugio natural, etc.)	
	6. Otro (contenedor, carpa, embarcación, vagón, cueva, refugio natural, etc.)	6. Otra vivienda (carpa, vagón, embarcación, cueva, refugio natural, etc.)		
¿Cuál es el material que predomina en las paredes exteriores de la vivienda en la que usted reside?	1. Bloque, ladrillo, piedra, madera pulida	1. Ladrillo, bloque, material prefabricado, piedra	1. Ladrillo, bloque, material prefabricado, piedra, madera pulida, concreto vaciado	1. En el censo se colapsaron las opciones de respuesta 1, 2 y 3, al igual que las opciones 1 y 2 de la encuesta en la categoría unificada “1. Ladrillo, bloque, material prefabricado, piedra, madera pulida, concreto vaciado”
	2. Concreto vaciado	2. Madera pulida	2. Adobe, tapia pisada o Bahareque	2. En la encuesta se unieron las opciones de respuesta 3 y 4 en la categoría unificada “2. Adobe, tapia pisada o Bahareque”; en el censo se recodificó la opción 5 por esta categoría

Variable	Opciones de respuesta Censo	Opciones de respuesta encuesta	Opciones de respuesta homologadas	Regla para la homologación
	3. Material prefabricado	3. Adobe o tapia pisada	3. Madera burda, tabla, tablón	3. La opción de respuesta 6 del censo y 5 de la encuesta se recodificaron en la categoría "3. Madera burda, tabla, tablón"
	4. Guadua	4. Bahareque	4. Guadua	4. La opción 5 de la encuesta se recodificó en la categoría "4. Guadua"
	5. Tapia pisada, bahareque, adobe	5. Madera burda, tabla, tablón	5. Caña, esterilla, otro tipo de material vegetal	5. La opción de respuesta 7, tanto del censo como de la encuesta, se recodificó en la categoría "5. Caña, esterilla, otro tipo de material vegetal"
	6. Madera burda, tabla, tablón	6. Guadua	6. Zinc, tela, cartón, latas, desechos, plástico	6. La opción de respuesta 8, tanto del censo como de la encuesta, se recodificó en la categoría "6. Zinc, tela, cartón, latas, desechos, plástico"
	7. Caña, esterilla, otros vegetales	7. Caña, esterilla, otro tipo de material vegetal	7. Sin paredes	7. La opción de respuesta 9, tanto del censo como de la encuesta, se recodificó en la categoría "7. Sin paredes"
	8. Materiales de desecho (zinc, tela, cartón, latas, plásticos, otros)	8. Zinc, tela, cartón, latas, desechos, plástico		
	9. No tiene paredes	9. Sin paredes		
¿Cuál es el material que predomina en los pisos de la vivienda que usted reside?	1. Mármol, parqué, madera pulida y lacada	1. Tierra, arena	1. Mármol, parqué, madera pulida y lacada	1. En la encuesta se unieron las opciones de respuesta 5 y 6 en la categoría unificada "1. Mármol, parqué, madera pulida y lacada"
	2. Baldosa, vinilo, tableta, ladrillo, laminado	2. Cemento, gravilla	2. Baldosa, vinilo, tableta, ladrillo, laminado	2. La opción 4 de la encuesta se recodificó en la categoría "2. Baldosa, vinilo, tableta, ladrillo, laminado"
	3. Alfombra	3. Madera burda, tabla, tablón, otro vegetal	3. Alfombra	3. La opción 7 de la encuesta se recodificó en la categoría "3. Alfombra"
	4. Cemento, gravilla	4. Baldosín, ladrillo, vinisol, otros materiales sintéticos	4. Cemento, gravilla	4. La opción 2 de la encuesta se recodificó en la categoría "4. Cemento, gravilla"
	5. Madera burda, tabla, tablón, otro vegetal	5. Mármol	5. Madera burda, tabla, tablón, otro vegetal	5. La opción 3 de la encuesta se recodificó en la categoría "5. Madera burda, tabla, tablón, otro vegetal"
	6. Tierra, arena, barro	6. Madera pulida	6. Tierra, arena, barro	6. La opción 1 de la encuesta se recodificó en la categoría "6. Tierra, arena, barro"
		7. Alfombra o tapete de pared a pared		
¿De dónde obtiene principalmente su hogar el agua para preparar los alimentos?	1. ¿Acueducto público?	1. De acueducto por tubería	1. ¿Acueducto público?	1. En la encuesta se unieron las opciones de respuesta 1 y 2 en la categoría unificada "1. ¿Acueducto público?"
	2. ¿Acueducto veredal?	2. De otra fuente por tubería	2. ¿Acueducto veredal o red de distribución comunitaria?	2. En el censo se unieron las opciones de respuesta 2 y 3 en la categoría unificada "2. ¿Acueducto veredal o red de distribución comunitaria?"
	3. ¿Red de distribución comunitaria?	3. De pozo con bomba	3. ¿Pozo con bomba?	3. La opción de respuesta 4 del censo y 3 de la encuesta se recodificaron en la categoría "3. ¿Pozo con bomba?"
	4. ¿Pozo con bomba?	4. De pozo sin bomba, aljibe, jagüey o barreno	4. ¿Pozo sin bomba, aljibe, jaguey o barreno?	4. La opción de respuesta 5 del censo y 4 de la encuesta se recodificaron en la categoría

Variable	Opciones de respuesta Censo	Opciones de respuesta encuesta	Opciones de respuesta homologadas	Regla para la homologación
				"4. ¿Pozo sin bomba, aljibe, jaguey o barreno?"
	5. ¿Pozo sin bomba, aljibe, jaguey o barreno?	5. Aguas lluvias	5. ¿Agua lluvia?	5. La opción de respuesta 6 del censo y 5 de la encuesta se recodificaron en la categoría "5. ¿Agua lluvia?"
	6. ¿Agua lluvia?	6. Río, quebrada, nacimiento o manantial	6. ¿Río, quebrada, manantial, nacimiento?	6. La opción de respuesta 7 del censo y 6 de la encuesta se recodificaron en la categoría "6. ¿Río, quebrada, manantial, nacimiento?"
	7. ¿Rio, quebrada, manantial, nacimiento?	7. De pila pública	7. ¿Pila pública?	7. La opción de respuesta 8 del censo y 7 de la encuesta se recodificaron en la categoría "7. ¿Pila pública?"
	8. ¿Pila pública?	8. Carro tanque	8. ¿Carrotanque?	8. La opción de respuesta 9 del censo y 8 de la encuesta se recodificaron en la categoría "8. ¿Carrotanque?"
	9. ¿Carrotanque?	9. Aguatero	9. ¿Aguatero?	9. La opción de respuesta 10 del censo y 9 de la encuesta se recodificaron en la categoría "9. ¿Aguatero?"
	10. ¿Aguatero?	10. Agua embotellada o en bolsa	10. ¿Agua embotellada o en bolsa?	10. La opción de respuesta 11 del censo y 10 de la encuesta se recodificaron en la categoría "10. ¿Agua embotellada o en bolsa?"
	11. ¿Agua embotellada o en bolsa?		11. NS/NR	11. En el censo se unieron las opciones de respuesta 11 y 12 en la categoría unificada "11. NS/NR"
	12. no preparan alimentos			
	99. No informa			
¿Cuál es el nivel educativo más alto alcanzado por ... y el último año o grado aprobado en ese nivel?	1. Preescolar	1. Ninguno	1. Ninguno	1. La opción de respuesta 10 del censo se recodificó en la categoría "1. Ninguno".
	2. Básica primaria	2. Preescolar	2. Preescolar	2. La opción de respuesta 1 del censo se recodificó en la categoría "2. Preescolar".
	3. Básica secundaria	3. Básica primaria (1o - 5o)	3. Básica primaria (1o - 5o)	3. La opción de respuesta 2 del censo se recodificó en la categoría "3. Básica primaria (1o - 5o)"
	4. Media académica o clásica	4. Básica secundaria (6o - 9o)	4. Básica secundaria (6o - 9o)	4. La opción de respuesta 3 del censo se recodificó en la categoría "4. Básica secundaria (6o - 9o)"
	5. Media técnica	5. Media (10o - 13o)	5. Media (10o - 13o)	5. En el censo se unieron las opciones de respuesta 4, 5 y 6 en la categoría unificada "5. Media (10o - 13o)"
	6. Normalista	6. Superior o universitaria	6. Superior o universitaria	6. En el censo se unieron las opciones de respuesta 7, 8 y 9 en la categoría unificada "Superior o universitaria"
	7. Técnica profesional o Tecnológica	9. No sabe, no informa	7. No sabe, no informa	7. Las opciones de respuesta 99 y 9 del censo y la encuesta, respectivamente, se recodificaron en la categoría "No sabe, no informa"
	8. Universitario			
	9. Especialización, maestría, doctorado			
	10. Ninguno			
	99. No Informa			
	No Aplica			

Variable	Opciones de respuesta Censo	Opciones de respuesta encuesta	Opciones de respuesta homologadas	Regla para la homologación
¿En qué actividad ocupó ... la mayor parte del tiempo la semana pasada?	1. Trabajó por lo menos una hora en una actividad que le generó algún ingreso?	1. Trabajando	1. Trabajando	1. En el censo se unieron las opciones de respuesta 1, 2 y 3 en la categoría unificada “1. Trabajando”
	2. ¿Trabajó o ayudó en un negocio por lo menos una hora sin que le pagaran?	2. Buscando trabajo	2. Buscando trabajo	2. La opción de respuesta 4 en el censo se recodificó en la categoría “2. Buscando trabajo”
	3. ¿No trabajó, pero tenía un empleo, trabajo o negocio por el que recibe ingresos?	3. Estudiando	3. Estudiando	3. La opción de respuesta 6 en el censo se recodificó en la categoría “3. Estudiando”
	4. ¿Busco trabajo?	4. Oficios del hogar	4. Oficios del hogar	4. La opción de respuesta 7 del censo se recodificó en la categoría “4. Oficios del hogar”
	5. ¿Vivió de jubilación, pensión o renta?	5. Incapacitado permanente para trabajar	5. Incapacitado permanente para trabajar	5. La opción de respuesta 8 del censo se recodificó en la categoría “5. Incapacitado permanente para trabajar”
	6. ¿Estudió?	6. Otra actividad	6. Otra actividad	6. En el censo se unieron las opciones de respuesta 5 y 9 en la categoría unificada “6. Otra actividad”
	7. ¿Realizó oficios del hogar?	7. No informa	7. No informa	7. La opción de respuesta 0 del censo se recodificó en la categoría “7. No informa, No aplica”
	8. ¿Está incapacitado(a) permanentemente para trabajar?	No aplica	No aplica	
	9. ¿Estuvo en otra situación?			
	10. No informa			
	No aplica			
¿Actualmente el estado civil de … es?:	1. ¿Unión libre?	1. No está casado(a) y vive en pareja hace menos de dos años	1. Unión libre	1. En la encuesta, se unieron las opciones de respuesta 1 y 2 en la categoría unificada “1. Unión libre”
	2. ¿Casado(a)?	2. No está casado (a) y vive en pareja hace dos años o más	2. Casado(a)	2. En la encuesta, la opción de respuesta 3 se recodificó en la categoría “2. Casado(a)”
	3. ¿Divorciado(a)?	3. Está casado (a)	3. Está separado (a) o divorciado (a)	3. En el censo se unieron las opciones de respuesta 3, 4 y 5 en la categoría unificada “3. Está separado (a) o divorciado (a)”; en la encuesta, se recodificó la opción de respuesta 4 en esta categoría unificada.
	4. ¿Separado(a) de unión libre?	4. Está separado (a) o divorciado (a)	4. Está viudo (a)	4. Las opciones de respuesta 6 y 5 del censo y la encuesta, respectivamente, se recodificaron en la categoría unificada “4. Está viudo (a)”
	5. ¿Separado(a) de matrimonio?	5. Está viudo (a)	5. Está soltero (a)	5. Las opciones de respuesta 7 y 6 del censo y la encuesta, respectivamente, se recodificaron en la categoría unificada “Está soltero (a)”
	6. ¿Viudo(a)?	6. Está soltero (a)	6. No Informa	6. Las demás opciones del censo y la encuesta se clasificaron en la categoría “6. No Informa, No Aplica”

Variable	Opciones de respuesta Censo	Opciones de respuesta encuesta	Opciones de respuesta homologadas	Regla para la homologación
	7. ¿Soltero(a)? (Nunca se ha casado, ni ha vivido en unión libre)		No Aplica	
	9. No Informa			
	No Aplica			

Fuente: Elaboración propia.

Estructura del formato de las variables de pregunta abierta en las dos fuentes

En el proceso de homologación de variables se debe estandarizar las preguntas abiertas a ser empleadas en variables categóricas cuyos niveles estén homologados entre las fuentes de información. Un ejemplo de este procedimiento de estandarización se tiene con la variable edad, la cual es una pregunta abierta en la encuesta de hogares de Colombia, pero para el censo se presenta en rangos quinquenales a causa de las condiciones de confidencialidad de los microdatos del DANE. En el cuadro A5 se encuentra el procedimiento de estandarización y homologación de algunas variables que están asociadas a la edad de los integrantes del hogar.

Cuadro A5
Procedimiento de estandarización y homologación de variables en función de la edad

Variable	Filtro de población de interés
¿Sabe leer y escribir?	Se ajustó el rango de edad a 5 años o más en la GEIH por ser el rango quinquenal empleado en el Censo
¿actualmente ... Asiste a la escuela, colegio o universidad?	Se ajustó el rango de edad a 5 años o más en la GEIH por ser el rango quinquenal empleado en el Censo
¿cuál es el nivel educativo más alto alcanzado por ... ¿Y el último año o grado aprobado en este nivel?	Se ajustó el rango de edad a 5 años o más en la GEIH por ser el rango quinquenal empleado en el Censo
¿En qué actividad ocupó ... la mayor parte del tiempo la semana pasada?	Se ajustó el rango de edad en Censo y en la encuesta, definiendo la población en edad de trabajar en 10 años o más para la parte rural y de 15 años o más para la parte urbana (ya que en el censo la edad está en rangos quinquenales)

Fuente: Elaboración propia.

Imputación de valores faltantes

Dentro del procedimiento de estandarización y homologación, se puede encontrar situaciones en las que por la estructura del formulario aparezcan opciones de respuesta que no se pueden enlazar entre las fuentes de información, lo cual genera valores faltantes en algunas de las variables de interés dentro hogares. En el A6 se presenta conjunto de variables que generan valores faltantes al enlazar el censo con la encuesta de hogares.

Para la imputación de variables solo se revisaron las variables a nivel de vivienda y hogar, dado que, para el caso de las variables a nivel de persona, se deben agrupar a nivel hogar y en ese caso se pueden excluir los valores faltantes de la agregación. El modelo que se usó para la imputación fue un modelo supervisado de Naive Bayes, el cual es un modelo de clasificación de aprendizaje automático (Machine Learning) basado en el teorema de Bayes. Para este algoritmo se requiere de una variable discreta con dos o más categorías que indican la pertenencia a cada una de las clases y un conjunto de variables independientes que sirvan como predictoras para el proceso de clasificación en cada una de las clases de la variable a imputar. La imputación

se aplicó sobre los hogares con valores faltantes en el censo, los cuales se identifican con la opción de respuesta "No sabe/No Responde (NS/NR)" en cada una de las variables identificadas en el cuadro A6.

Cuadro A6
Variables que generan valores faltantes en el proceso de estandarización y homologación

Variable	Categorías	
Acceso a conexión de gas en el hogar	1	Sí
	2	No
	3	NS/NR
Acceso al servicio de recolección de basuras	1	Sí
	2	No
	3	NS/NR
Acceso al servicio de internet	1	Sí
	2	No
	3	NS/NR
Incluyendo sala-comedor ¿de cuántos cuartos en total dispone este hogar?	1	Número de cuartos
	2	NS/NR
¿En cuántos de esos cuartos duermen las personas de este hogar?	1	Número de cuartos
	2	NS/NR
¿En cuál de los siguientes lugares, preparan los alimentos las personas de este hogar?	1	En un cuarto usado solo para cocinar
	2	En un cuarto usado también para dormir
	3	En una sala comedor con lavaplatos
	4	En una sala comedor sin lavaplatos
	5	En un patio, corredor, enramada, al aire libre
	6	En ninguna parte, no preparan alimentos
	9	NS/NR

Fuente: Elaboración propia.

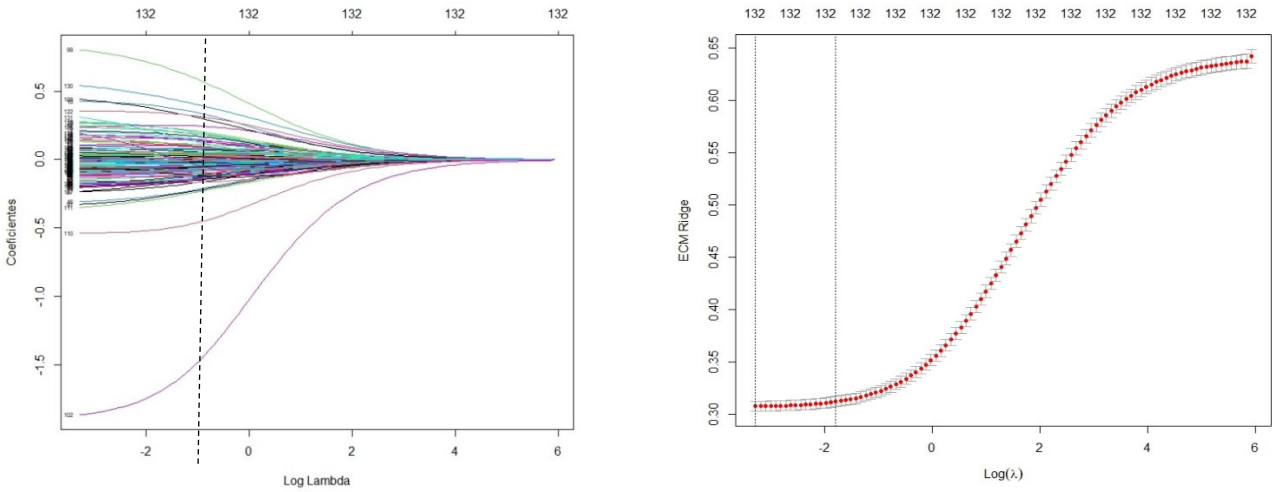
Anexo 2
Representación gráfica modelos Ridge y Lasso

Con el fin de conseguir una correcta interpretación de las gráficas que arroja cada una de estas regresiones, se explicará de manera sencilla su interpretación, tomando como ejemplo las gráficas obtenidas únicamente para Chile, no se incluye Colombia y Perú, ya que las gráficas se interpretan de la misma manera solo cambian los valores de λ. De este modo, en el eje Y se encuentran todos los coeficientes de regresión disponibles en el modelo y en el eje X se encuentra el logaritmo de λ donde se marca el valor óptimo de este. Al iniciar el proceso de regularización en Chile, se debe evaluar el grado de convergencia a cero de los coeficientes de regresión del modelo, así como la evolución del error de validación cruzada a medida que se incrementa el valor del grado de penalización del modelo. Así que, en el gráfico A1 al lado izquierdo se puede observar la convergencia a cero de los coeficientes de regresión del método Ridge, a medida que se identifica el parámetro λ óptimo, el cual es aproximadamente cercano a -1.79, como se confirma en el cuadro 5 "Resumen métodos de regularización Chile, Colombia y Perú" en la pág. 36 de este documento.

Gráfico A1
Resultados del método Ridge para Chile

A. Coeficientes del modelo en función de la regularización del método Ridge para Chile

B. Error de validación cruzada en función de la regularización Ridge para Chile



Fuente: Elaboración propia.

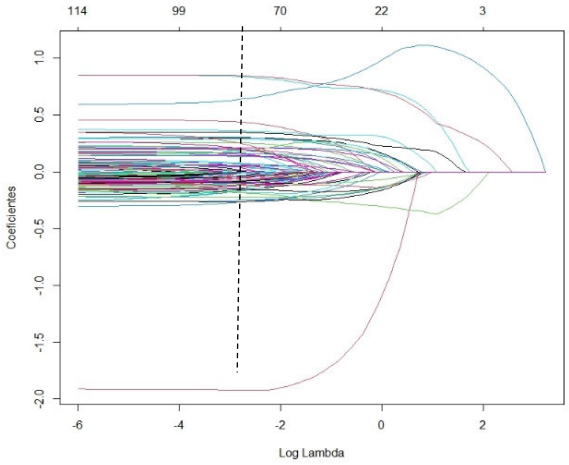
Ahora bien, en el gráfico A1 en la parte derecha se puede observar el cambio en el error cuadrático medio en función de la regularización para Chile. Como el objetivo del método Ridge es reducir varianza sin aumentar el sesgo en la estimación se debe elegir el grado de penalización que minimice el error total. En el caso de Chile se presenta el valor de λ, que minimiza el error que comete el modelo al predecir observaciones del conjunto de validación y del conjunto de prueba y el valor óptimo, el cual no se aleja más de una desviación estándar del mínimo valor del conjunto de prueba.

Como se mencionó anteriormente, el método Ridge fuerza que los coeficientes de regresión tiendan a cero, pero realmente nunca alcanzan este valor, dejando al final el número total de covariables con los que se contaba en un principio, por lo que finalmente, se ajusta un modelo con las 132 variables iniciales. Bajo esta

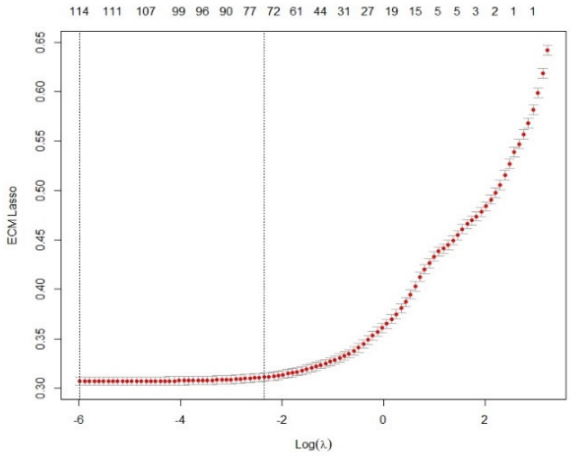
consideración resulta de gran relevancia implementar la regularización Lasso adaptado que tiene la propiedad de forzar que los coeficientes de los predictores tiendan a cero y en el proceso elimina las covariables que no aportan información relevante al modelo.

Gráfico A2
Resultados del método Lasso para Chile

A. Coeficientes del modelo en función de la regularización del método Lasso para Chile



B. Error de validación cruzada en función de la regularización Lasso para Chile



Fuente: Elaboración propia.

Respecto a lo anterior , en la gráfico A2 en el lado izquierdo se puede observar la convergencia a cero de los coeficientes de regresión Lasso a medida que se identifica el parámetro $\boldsymbol{\lambda}$ óptimo, Por otra parte, complementando con el mismo gráfico al lado derecho, se observa el cambio del error cuadrático medio al aumentar el grado de penalización y disminuir el número de variables presentes en el modelo; además, se puede apreciar el rango de valores en el que se ubica el valor de $\boldsymbol{\lambda}$ desde su valor mínimo (-5.99) hasta su valor óptimo de (-2.349).

Por otro lado, se observa como al final del ajuste, solo quedan un aproximado de 75 covariables, no 132 variables como se tenía originalmente. De esta manera, se obtiene un modelo más eficiente (sobre todo en lo que respecta a interpretabilidad), que pueda explicar de manera más precisa y práctica la situación de pobreza en Chile, demostrando con ello, la potencia de esta metodología en la selección del mejor modelo.

Anexo 3
Cuadros de resultados de modelos finales

En la presente sección se presentan los Cuadros con los valores de los coeficientes arrojados por el modelo escogido para cada uno de los tres países observados. Así, el cuadro A7 corresponde al modelo ajusta para Chile, el cuadro A8 corresponde al modelo ajustado para Colombia, y, finalmente, el cuadro A9 que contiene el modelo SAE elegido para el Perú.

Modelo Chile

Cuadro A7
Modelo SAE seleccionado para Chile

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
Intercepto		12,336	0,2032	60,706	0
Provincia	Prov_11	-0,2438	0,0971	-2,51	0,006
	Prov_14	-0,2721	0,0923	-2,949	0,0016
	Prov_21	-0,1177	0,0935	-1,259	0,104
	Prov_22	0,0211	0,0984	0,215	0,3898
	Prov_23	-0,281	0,1038	-2,707	0,0034
	Prov_31	-0,2611	0,0916	-2,85	0,0022
	Prov_32	-0,2585	0,1028	-2,515	0,006
	Prov_33	-0,3149	0,0884	-3,561	0,0002
	Prov_41	-0,4411	0,0838	-5,262	0
	Prov_42	-0,3917	0,0871	-4,496	0
	Prov_43	-0,4604	0,0855	-5,386	0
	Prov_51	-0,2012	0,0846	-2,378	0,0087
	Prov_53	-0,4164	0,0881	-4,726	0
	Prov_54	-0,3853	0,0873	-4,413	0
	Prov_55	-0,3526	0,085	-4,15	0
	Prov_56	-0,2295	0,0854	-2,687	0,0036
	Prov_57	-0,3531	0,0838	-4,216	0
	Prov_58	-0,2528	0,0863	-2,93	0,0017
	Prov_61	-0,384	0,0783	-4,904	0
	Prov_62	-0,4298	0,0841	-5,113	0
	Prov_63	-0,4362	0,0805	-5,417	0
	Prov_71	-0,3857	0,0805	-4,792	0
	Prov_72	-0,6473	0,089	-7,273	0
	Prov_73	-0,4622	0,0807	-5,726	0
	Prov_74	-0,5196	0,0814	-6,381	0
	Prov_81	-0,3817	0,0796	-4,794	0
	Prov_82	-0,4658	0,0822	-5,666	0

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
	Prov_83	-0,4749	0,0791	-6,007	0
	Prov_91	-0,4661	0,0777	-5,998	0
	Prov_92	-0,453	0,0799	-5,667	0
	Prov_101	-0,3223	0,0812	-3,968	0
	Prov_102	-0,3302	0,0809	-4,083	0
	Prov_103	-0,3556	0,0824	-4,314	0
	Prov_111	-0,0922	0,1047	-0,881	0,1892
	Prov_112	-0,0902	0,0987	-0,914	0,1804
	Prov_113	-0,0847	0,1223	-0,693	0,2442
	Prov_114	-0,2232	0,1071	-2,084	0,0186
	Prov_123	0,15	0,122	1,23	0,1872
	Prov_124	-0,0704	0,113	-0,623	0,2666
	Prov_131	-0,1363	0,0778	-1,752	0,0399
	Prov_132	-0,0932	0,0944	-0,988	0,1616
	Prov_133	-0,1315	0,0888	-1,481	0,0693
	Prov_134	-0,2254	0,0877	-2,571	0,0051
	Prov_135	-0,3325	0,0867	-3,837	0,0001
	Prov_136	-0,2248	0,0852	-2,639	0,0042
	Prov_141	-0,4048	0,0808	-5,008	0
	Prov_142	-0,3867	0,0859	-4,503	0
	Prov_151	-0,2932	0,0986	-2,974	0,0015
	Prov_152	-0,3658	0,1191	-3,072	0,0011
	Prov_161	-0,5399	0,0811	-6,66	0
	Prov_162	-0,5324	0,0827	-6,438	0
	Prov_163	-0,4912	0,0844	-5,819	0
Área		0,0124	0,0117	1,057	0,2282
Tipo de vivienda	Casa	0,0434	0,1375	0,316	0,3795
	Depto	-0,0399	0,138	-0,289	0,3863
	Ruka	-0,1452	0,2157	-0,673	0,2505
	Conv	-0,1798	0,1409	-1,276	0,101
	MedAg	-0,0779	0,1353	-0,575	0,2826
Material de paredes	MP1	0,3066	0,0954	3,212	0,0023
	MP2	0,2612	0,0953	2,741	0,0093
	MP3	0,1942	0,0953	2,038	0,05
	MP4	0,148	0,096	1,541	0,1217
	MP5	0,2218	0,0961	2,308	0,0278
	MP6	0,0534	0,1357	0,393	0,3693
Material de paredes	MT1	0,0081	0,067	0,121	0,396
	MT2	-0,0334	0,0673	-0,496	0,3099

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
	MT3	-0,1158	0,0668	-1,734	0,0415
	MT4	-0,0114	0,1317	-0,087	0,4653
	MT5	0,086	0,1893	0,454	0,3599
	MT6	-0,0739	0,1371	-0,539	0,2949
Materiales de piso	MPiso1	0,0688	0,1173	0,587	0,3358
	MPiso2	-0,0446	0,1176	-0,379	0,3523
	MPiso3	-0,0173	0,1177	-0,147	0,4416
	MPiso4	-0,1078	0,1248	-0,864	0,1938
Hacinamiento		-0,1385	0,0103	-13,506	0
Agua	Tipo1	0,1193	0,0414	2,883	0,0063
	Tipo2	0,1081	0,0422	2,56	0,0151
	Tipo3	0,0244	0,0445	0,549	0,3431
	Tipo4	0,061	0,0433	1,408	0,1481
Cohabitación		-0,1099	0,0086	-12,801	0
Jefe del hogar sin conyugue		0,1532	0,0072	21,159	0
Tamaño del hogar	tamhog2	-0,1933	0,0075	-25,719	0
	tamhog3	-0,3254	0,0102	-31,813	0
	tamhog4	-0,4575	0,0165	-27,78	0
Tipo de hogar	Monoparental (H)	-0,1988	0,0663	-2,997	0,0014
	Monoparental (F)	-0,0559	0,0142	-3,948	0
	biparental	-0,1447	0,0659	-2,195	0,0141
Número de hijos en el hogar	Nhijos_Op1	0,1055	0,0652	1,619	0,1076
	Nhijos_Op2	-0,0099	0,0649	-0,153	0,4392
	Nhijos_Op3	-0,0095	0,0405	-0,234	0,4075
	Nhijos_Op4	-0,0926	0,1596	-0,58	0,281
Edad del jefe del hogar	Edad_jefe1	-20,349	0,5484	-3,711	0,0001
	Edad_jefe2	-0,3443	0,0079	-43,36	0
	Edad_jefe3	-0,1373	0,0063	-21,647	0
Sexo jefe del hogar - Mujer		-0,0445	0,0061	-7,344	0
Proporción mujeres		-0,0956	0,0089	-10,693	0
Servicio doméstico		0,7016	0,0404	17,383	0
Migración	Mediano plazo	-0,1592	0,0197	-8,074	0
	Largo plazo	-0,1172	0,0148	-7,901	0
Población indígena		-0,0611	0,0062	-9,779	0
Dependencia económica	Ocupados	0,8416	0,0108	77,929	0
	Desocupados	-0,1631	0,0163	-10,024	0
Actividad del negocio donde trabajó	RamaA	0,0284	0,0075	3,762	0,0003
	RamaB	0,2943	0,0127	23,188	0
	RamaC	0,0462	0,0075	6,167	0

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
	RamaF	0,0601	0,0073	8,227	0
	RamaG	-0,0105	0,0061	-1,702	0,0444
	RamaI	0,0367	0,0096	3,838	0,0003
	RamaO	0,1943	0,0089	21,894	0
	RamaP	0,0821	0,0078	10,463	0
	RamaQ	0,1629	0,0091	17,974	0
	RamaT	-0,0992	0,009	-11,046	0
	RamaU	0,1917	0,1645	1,165	0,2024
Nivel educativo del jefe de Hogar	Primaria	0,0082	0,0125	0,657	0,3215
	Secundaria	0,0268	0,0113	2,374	0,0238
	Técnico	0,1	0,0163	6,139	0
	Superior	0,2079	0,0151	13,807	0
Nivel educativo Hogar	Primaria	-0,2599	0,0156	-16,649	0
	Secundaria	-0,0886	0,0133	-6,686	0
	Técnico	0,1279	0,021	6,098	0
	Superior	0,5348	0,0184	29,002	0
Prop_jubilados		0,3747	0,0113	33,212	0
Prop_estudiantes		0,2447	0,0174	14,059	0

Fuente: Elaboración propia.

Modelo Colombia

Cuadro A8
Modelo SAE seleccionado para Colombia

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
Intercepto		12,336	0,2032	60,706	0
Provincia	Prov_11	-0,2438	0,0971	-2,51	0,006
	Prov_14	-0,2721	0,0923	-2,949	0,0016
	Prov_21	-0,1177	0,0935	-1,259	0,104
	Prov_22	0,0211	0,0984	0,215	0,3898
	Prov_23	-0,281	0,1038	-2,707	0,0034
	Prov_31	-0,2611	0,0916	-2,85	0,0022
	Prov_32	-0,2585	0,1028	-2,515	0,006
	Prov_33	-0,3149	0,0884	-3,561	0,0002
	Prov_41	-0,4411	0,0838	-5,262	0
	Prov_42	-0,3917	0,0871	-4,496	0
	Prov_43	-0,4604	0,0855	-5,386	0
	Prov_51	-0,2012	0,0846	-2,378	0,0087
	Prov_53	-0,4164	0,0881	-4,726	0
	Prov_54	-0,3853	0,0873	-4,413	0

Coeficientes	Estimador	Error estándar	Estadístico de prueba	Valor-p
Prov_55	-0,3526	0,085	-4,15	0
Prov_56	-0,2295	0,0854	-2,687	0,0036
Prov_57	-0,3531	0,0838	-4,216	0
Prov_58	-0,2528	0,0863	-2,93	0,0017
Prov_61	-0,384	0,0783	-4,904	0
Prov_62	-0,4298	0,0841	-5,113	0
Prov_63	-0,4362	0,0805	-5,417	0
Prov_71	-0,3857	0,0805	-4,792	0
Prov_72	-0,6473	0,089	-7,273	0
Prov_73	-0,4622	0,0807	-5,726	0
Prov_74	-0,5196	0,0814	-6,381	0
Prov_81	-0,3817	0,0796	-4,794	0
Prov_82	-0,4658	0,0822	-5,666	0
Prov_83	-0,4749	0,0791	-6,007	0
Prov_91	-0,4661	0,0777	-5,998	0
Prov_92	-0,453	0,0799	-5,667	0
Prov_101	-0,3223	0,0812	-3,968	0
Prov_102	-0,3302	0,0809	-4,083	0
Prov_103	-0,3556	0,0824	-4,314	0
Prov_111	-0,0922	0,1047	-0,881	0,1892
Prov_112	-0,0902	0,0987	-0,914	0,1804
Prov_113	-0,0847	0,1223	-0,693	0,2442
Prov_114	-0,2232	0,1071	-2,084	0,0186
Prov_123	0,15	0,122	1,23	0,1872
Prov_124	-0,0704	0,113	-0,623	0,2666
Prov_131	-0,1363	0,0778	-1,752	0,0399
Prov_132	-0,0932	0,0944	-0,988	0,1616
Prov_133	-0,1315	0,0888	-1,481	0,0693
Prov_134	-0,2254	0,0877	-2,571	0,0051
Prov_135	-0,3325	0,0867	-3,837	0,0001
Prov_136	-0,2248	0,0852	-2,639	0,0042
Prov_141	-0,4048	0,0808	-5,008	0
Prov_142	-0,3867	0,0859	-4,503	0
Prov_151	-0,2932	0,0986	-2,974	0,0015
Prov_152	-0,3658	0,1191	-3,072	0,0011
Prov_161	-0,5399	0,0811	-6,66	0
Prov_162	-0,5324	0,0827	-6,438	0
Prov_163	-0,4912	0,0844	-5,819	0
Área	0,0124	0,0117	1,057	0,2282

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
Tipo de vivienda	Casa	0,0434	0,1375	0,316	0,3795
	Depto	-0,0399	0,138	-0,289	0,3863
	Ruka	-0,1452	0,2157	-0,673	0,2505
	Conv	-0,1798	0,1409	-1,276	0,101
	MedAg	-0,0779	0,1353	-0,575	0,2826
Material de paredes	MP1	0,3066	0,0954	3,212	0,0023
	MP2	0,2612	0,0953	2,741	0,0093
	MP3	0,1942	0,0953	2,038	0,05
	MP4	0,148	0,096	1,541	0,1217
	MP5	0,2218	0,0961	2,308	0,0278
	MP6	0,0534	0,1357	0,393	0,3693
Material de paredes	MT1	0,0081	0,067	0,121	0,396
	MT2	-0,0334	0,0673	-0,496	0,3099
	MT3	-0,1158	0,0668	-1,734	0,0415
	MT4	-0,0114	0,1317	-0,087	0,4653
	MT5	0,086	0,1893	0,454	0,3599
	MT6	-0,0739	0,1371	-0,539	0,2949
Materiales de piso	MPiso1	0,0688	0,1173	0,587	0,3358
	MPiso2	-0,0446	0,1176	-0,379	0,3523
	MPiso3	-0,0173	0,1177	-0,147	0,4416
	MPiso4	-0,1078	0,1248	-0,864	0,1938
Hacinamiento		-0,1385	0,0103	-13,506	0
Agua	Tipo1	0,1193	0,0414	2,883	0,0063
	Tipo2	0,1081	0,0422	2,56	0,0151
	Tipo3	0,0244	0,0445	0,549	0,3431
	Tipo4	0,061	0,0433	1,408	0,1481
Cohabitación		-0,1099	0,0086	-12,801	0
Jefe del hogar sin conyugue		0,1532	0,0072	21,159	0
Tamaño del hogar	tamhog2	-0,1933	0,0075	-25,719	0
	tamhog3	-0,3254	0,0102	-31,813	0
	tamhog4	-0,4575	0,0165	-27,78	0
Tipo de hogar	Monoparental (H)	-0,1988	0,0663	-2,997	0,0014
	Monoparental (F)	-0,0559	0,0142	-3,948	0
	biparental	-0,1447	0,0659	-2,195	0,0141
Número de hijos en el hogar	Nhijos_Op1	0,1055	0,0652	1,619	0,1076
	Nhijos_Op2	-0,0099	0,0649	-0,153	0,4392
	Nhijos_Op3	-0,0095	0,0405	-0,234	0,4075
	Nhijos_Op4	-0,0926	0,1596	-0,58	0,281

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
Edad del jefe del hogar	Edad_jefe1	-20,349	0,5484	-3,711	0,0001
	Edad_jefe2	-0,3443	0,0079	-43,36	0
	Edad_jefe3	-0,1373	0,0063	-21,647	0
Sexo jefe del hogar - Mujer		-0,0445	0,0061	-7,344	0
Proporción mujeres		-0,0956	0,0089	-10,693	0
Servicio doméstico		0,7016	0,0404	17,383	0
Migración	Mediano plazo	-0,1592	0,0197	-8,074	0
	Largo plazo	-0,1172	0,0148	-7,901	0
Población indígena		-0,0611	0,0062	-9,779	0
Dependencia económica	Ocupados	0,8416	0,0108	77,929	0
	Desocupados	-0,1631	0,0163	-10,024	0
Actividad del negocio donde trabajó	RamaA	0,0284	0,0075	3,762	0,0003
	RamaB	0,2943	0,0127	23,188	0
	RamaC	0,0462	0,0075	6,167	0
	RamaF	0,0601	0,0073	8,227	0
	RamaG	-0,0105	0,0061	-1,702	0,0444
	RamaI	0,0367	0,0096	3,838	0,0003
	RamaO	0,1943	0,0089	21,894	0
	RamaP	0,0821	0,0078	10,463	0
	RamaQ	0,1629	0,0091	17,974	0
	RamaT	-0,0992	0,009	-11,046	0
	RamaU	0,1917	0,1645	1,165	0,2024
Nivel educativo del jefe de Hogar	Primaria	0,0082	0,0125	0,657	0,3215
	Secundaria	0,0268	0,0113	2,374	0,0238
	Técnico	0,1	0,0163	6,139	0
	Superior	0,2079	0,0151	13,807	0
Nivel educativo Hogar	Primaria	-0,2599	0,0156	-16,649	0
	Secundaria	-0,0886	0,0133	-6,686	0
	Técnico	0,1279	0,021	6,098	0
	Superior	0,5348	0,0184	29,002	0
Prop_jubilados		0,3747	0,0113	33,212	0
Prop_estudiantes		0,2447	0,0174	14,059	0

Fuente: Elaboración propia.

Modelo Perú

Cuadro A9
Modelo SAE seleccionado para Perú

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
Intercepto		54,565	0,0681	80,069	0
Departamento	Depto01	-0,02	0,0571	-0,35	0,3752
	Depto02	-0,0287	0,0527	-0,544	0,3441
	Depto03	-0,042	0,058	-0,724	0,307
	Depto04	0,0824	0,0567	1,454	0,1386
	Depto05	-0,1619	0,0546	-2,964	0,0049
	Depto06	-0,1623	0,0522	-3,107	0,0032
	Depto07	-0,0858	0,1035	-0,829	0,2829
	Depto08	-0,1431	0,0529	-2,706	0,0103
	Depto09	-0,2459	0,058	-4,241	0
	Depto10	-0,0796	0,0544	-1,463	0,1368
	Depto11	0,0016	0,0606	0,026	0,3988
	Depto12	-0,1102	0,0536	-2,057	0,0481
	Depto13	-0,062	0,0538	-1,151	0,2057
	Depto14	0,0322	0,063	0,511	0,3501
	Depto15	-0,0128	0,0551	-0,232	0,3883
	Depto16	-0,0955	0,0556	-1,718	0,0912
	Depto17	0,1147	0,075	1,529	0,124
	Depto18	-0,0665	0,0725	-0,917	0,262
	Depto19	-0,2788	0,0642	-4,341	0
	Depto20	-0,0385	0,0553	-0,695	0,3133
	Depto21	-0,3224	0,0529	-6,095	0
	Depto22	-0,0248	0,0534	-0,464	0,3582
	Depto23	-0,0623	0,0722	-0,863	0,2749
	Depto24	0,0311	0,0743	0,419	0,3654
Área		0,1325	0,0149	8,904	0
Población	pob_2	-0,0121	0,033	-0,366	0,3731
	pob_3	-0,0244	0,0336	-0,726	0,3065
	pob_4	-0,0263	0,0347	-0,758	0,2993
	pob_5	0,0227	0,0392	0,58	0,3372
Etnia		0,0108	0,0069	1,563	0,1176
Tamaño del hogar	tamhog	0,0005	0,0044	0,103	0,3968
	logtamhog	-0,4713	0,0175	-26,964	0
Dependencia económica		-0,108	0,0112	-9,637	0
Afiliación al sistema de salud	afi_sis	-0,0843	0,0067	-12,607	0
	afi_essalud	0,1304	0,0072	18,193	0

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
	afi_arm	0,2345	0,0197	11,887	0
	afi_priv	0,2385	0,014	17,096	0
Nivel Educativo	Primaria	-0,1035	0,0298	-3,477	0,0009
	Superior	0,1247	0,025	4,992	0
Analfabetismo		-0,0074	0,0115	-0,638	0,3255
Años de educación jefe	Hombre	-0,022	0,0025	-8,714	0
	Mujer	0,0014	0,0001	10,683	0
Años de educación conyugue	Hombre	-0,0135	0,0019	-7,074	0
	Mujer	0,0007	0,0001	5,688	0
Años de educación	hog	0,0507	0,0038	13,479	0
	hog2	-0,0012	0,0002	-5,965	0
Promedio edad_hog		0,0043	0,0003	14,978	0
Tasa de ocupación		0,384	0,0123	31,129	0
Tasa de indigencia		-0,1728	0,007	-24,638	0
Discapacidad		-0,0092	0,0078	-1,182	0,1984
Alcantarillado		-0,0983	0,0093	-10,52	0
Combustible cocina	Comboc1	0,062	0,0097	6,418	0
	Comboc2	0,1119	0,0114	9,815	0
	Comboc3	0,0518	0,016	3,236	0,0021
	Comboc5	-0,0367	0,008	-4,581	0
	Comboc6	-0,0803	0,0165	-4,877	0
	Comboc8	0,1793	0,0201	8,92	0
Tenencia de vivienda	TenViv1	0,0287	0,0063	4,593	0
	TenViv2	0,0377	0,0083	4,523	0
	TenViv3	-0,0672	0,0109	-6,164	0
	TenViv4	0,0328	0,0077	4,263	0
	TenViv5	0,0493	0,0078	6,295	0
	TenViv6	0,0404	0,0083	4,878	0
	TenViv7	0,0548	0,0073	7,495	0
	TenViv8	0,0396	0,0071	5,568	0
	TenViv9	0,0214	0,0077	2,774	0,0085
	TenViv10	0,0689	0,0105	6,573	0
	TenViv11	0,0612	0,0086	7,089	0
	TenViv12	0,083	0,0066	12,555	0
	TenViv13	0,0613	0,0084	7,313	0
	TenViv14	0,1536	0,0092	16,758	0
	TenViv15	0,094	0,0093	10,155	0
Núm_hab_dormir		0,0466	0,0026	17,914	0
Hacinamiento		0,0203	0,0047	4,343	0
Agua	agviv1	0,0328	0,0143	2,288	0,0291

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
	agviv3	0,0145	0,0241	0,603	0,3326
	agviv4	0,0546	0,0238	2,289	0,0291
	agviv5	0,1025	0,0234	4,388	0
	agviv6	0,0594	0,0238	2,501	0,0175
	agviv7	0,0628	0,0226	2,777	0,0084
Número de servicios públicos	Servhig1	0,0031	0,0154	0,203	0,3908
	Servhig2	0,0211	0,0225	0,938	0,257
	Servhig3	-0,0051	0,0169	-0,305	0,3808
	Servhig4	0,0099	0,0196	0,505	0,3512
	Servhig5	0,0044	0,0166	0,262	0,3855
	Servhig6	0,0473	0,0275	1,719	0,091
	Servhig7	-0,0484	0,0191	-2,54	0,0158
Alumbrado		0,0159	0,0144	1,108	0,2159
Material de paredes	MP1	0,057	0,0097	5,878	0
	MP2	0,0396	0,035	1,129	0,2109
	MP3	0,0212	0,0318	0,666	0,3196
	MP4	0,028	0,0131	2,14	0,0404
	MP5	0,0377	0,022	1,715	0,0917
Material de techo	Mtecho1	0,0714	0,0224	3,182	0,0025
	Mtecho2	0,0775	0,0273	2,839	0,0071
	Mtecho3	0,0093	0,0231	0,403	0,3678
	Mtecho4	0,0353	0,0206	1,709	0,0926
	Mtecho5	0,0553	0,0257	2,151	0,0395
	Mtecho6	0,0178	0,0392	0,455	0,3597
Material de pisos	MPisos1	0,2611	0,0169	15,406	0
	MPisos2	0,1723	0,0156	11,026	0
	MPisos3	0,1421	0,0126	11,291	0
	MPisos4	0,0784	0,013	6,016	0
	MPisos5	0,0549	0,0087	6,284	0
Ocupación principal	Grupo1	0,2924	0,0286	10,225	0
	Grupo2	0,1605	0,0104	15,434	0
	Grupo3	0,0951	0,0093	10,27	0
	Grupo5	0,0641	0,0088	7,315	0
	Grupo6	-0,0347	0,0131	-2,653	0,0118
	Grupo8	0,0658	0,0103	6,419	0
	Grupo9	-0,0111	0,0071	-1,561	0,118
	Grupo0	-0,001	0,0328	-0,032	0,3987
Actividad del negocio	GrupoA	0,0973	0,0125	7,762	0
	GrupoB	0,3996	0,02	20,004	0

Coeficientes		Estimador	Error estándar	Estadístico de prueba	Valor-p
	GrupoC	0,1046	0,0084	12,453	0
	GrupoD	0,2098	0,0599	3,5	0,0009
	GrupoE	-0,023	0,0299	-0,771	0,2964
	GrupoF	0,2093	0,0099	21,098	0
	GrupoG	0,0931	0,0087	10,685	0
	GrupoH	0,1033	0,0113	9,14	0
	GrupoI	0,1296	0,0097	13,326	0
	GrupoJ	0,0512	0,0214	2,394	0,0227
	GrupoK	0,1662	0,0207	8,026	0
	GrupoL	0,1356	0,046	2,949	0,0052
	GrupoM	0,0812	0,0155	5,23	0
	GrupoN	0,0271	0,0127	2,124	0,0418
	GrupoO	0,1426	0,0115	12,43	0
	GrupoQ	0,087	0,0138	6,323	0
	GrupoR	0,0425	0,0212	2	0,054
	GrupoS	0,0378	0,0122	3,105	0,0032
	GrupoT	0,1539	0,0136	11,339	0
Ratio económicamente activos hogar	ratec1	0,1272	0,0148	8,586	0
	ratec2	0,0908	0,0128	7,091	0
	ratec3	0,1595	0,0172	9,266	0
Nivel Educativo	Primaria	0,0465	0,0125	3,732	0,0004
	Superior	-0,0531	0,0112	-4,747	0
Analfabetismo		0,0097	0,0136	0,715	0,309

Fuente: Elaboración propia.

NACIONES UNIDAS

CEPAL

Serie

Estudios Estadísticos

Números publicados

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

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, 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.
99. Non monetary indicators to monitor SDG targets 1.2 and 1.4: Standards, availability, comparability and quality, María Emma Santos (LC/TS.2019/4), 2018.
98. Una propuesta de estimación del producto interno bruto trimestral de América Latina y el Caribe, Giannina López, Alberto Malmierca (LC.TS.2018/88), 2018.
97. Desagregación de datos en encuestas de hogares. Metodologías de estimación en áreas pequeñas, Isabel Molina (LC/TS.2018/82), 2018.
96. ¿Cuál es el alcance de las transferencias no contributivas en América Latina?: discrepancias entre encuestas y registros, Pablo Villatoro, Simone Cecchini (LC/TS.2018/46), 2018.

ESTUDIOS ESTADÍSTICOS

Números publicados:

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

104 Cambio de año de referencia de los agregados regionales anuales en las cuentas nacionales
Claudia de Camino, Giannina López, Patricia Marchant, Ana Victoria Vega y Alfonso Zerón

103 Predicciones agregadas de pobreza con información a escala micro y macro
Evaluación, diagnóstico y propuestas
Walter Sosa Escudero y Magdalena Cornejo

102 La medición de la discriminación en base al autorreporte
Estado de situación y desafíos
Pablo Villatoro

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

LC/TS.2022/191