ISSN: 2788-5836

United Nations

ECLAC

ECLAC Statistical Briefings

No. 4 | December 2021

Subnational Poverty Estimates for Latin America

Contents

Subnational poverty maps are used by governments to design, implement and monitor development policies more effectively by targeting them to the places or population groups that need them most urgently. This note describes the poverty mapping methodology based on small-area estimation methods, as used by the Statistics Division of ECLAC.

The methodology makes it possible to combine census information with household survey data to make estimations at the provincial, communal, or municipal levels, for which direct estimates from household surveys are generally too imprecise. This note illustrates the results obtained with this methodology for Chile, Colombia, and Peru.

1. Introduction

Household surveys are designed and implemented by national statistical offices to generate representative statistics at a predefined level of aggregation; generally based on large geographic subdivisions, sex, or socioeconomic groups of the population. However, when direct estimations of different indicators are needed in smaller subdivisions than those envisaged initially (for example, at the provincial or municipal level), the inference resulting from the surveys is not very precise or accurate. In general, the higher the disaggregation of the estimations, the less efficient they become, and their reliability declines ostensibly. In the case of some complex indicators, this can generate bias problems in the direct estimation and its standard error.

Small area estimation (SAE) is a set of statistical techniques that serve to obtain disaggregated estimates of population parameters to improve inference quality when the disaggregation of household surveys does not meet the quality criteria required for publication. This briefing outlines a series of steps for applying the unit-level nested error model to estimate indicators of interest relating to household income. These include the incidence, gap, and intensity of poverty and extreme poverty —at the provincial level in Peru, the district (comuna) level in Chile, and the municipal level in Colombia. The results show a gain in precision for indicators in smaller geographic areas where surveys do not attain adequate levels of representativeness. This is achieved by calculating the mean squared errors (MSEs) for each of the established models.

2. Background of poverty measurements in ECLAC

ECLAC periodically produces estimates of poverty and extreme poverty for 18 Latin American countries, using a methodology that aims to achieve regional comparability. The general methodology for measuring absolute poverty classifies a person as poor when their household per capita income is below the poverty line. This is defined as the cost of covering their food and other, non-food basic needs.[1]

The cost of basic food needs is estimated by constructing basic food baskets, which provide the recommended amounts of energy and nutrients while also reflecting the consumption habits of the population in question. The corresponding nutritional requirements are obtained from current international recommendations for maintaining a healthy lifestyle. Consumption habits are captured through household income and expenditure surveys and correspond to those of a particular population subset, which is adopted as the reference population based on the criteria established by the methodology. The monthly cost of the basic food basket is known as the "extreme poverty line." The poverty line itself is calculated by multiplying the extreme poverty line by the quotient between the total expenditure and the expenditure on food of the same reference population used to define the basic food basket.

The indicators commonly used to measure poverty belong to the family of parametric indices proposed by Foster, Greer, and Thorbecke (1984).[2] These indices (known as FGT) are based on the following function:

$$F_{\alpha d} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z - y_{di}}{z}\right)^{\alpha} I(y_{di} < z)$$

1 Economic Commission for Latin America and the Caribbean (ECLAC), "Income poverty measurement: Updated methodology and results", Eclac Methodologies, No. 2 (LC/PUB.2018/22-P), Santiago, 2019.

2 J. Foster, J. Greer and E. Thorbecke, "A class of decomposable poverty measures", *Econometrica*, vol. 52, No. 3, 1984.

where N_d is the total population of the group of interest d; y_{di} is household per capita income n; z is the poverty line; and α is a parameter greater than or equal to 0 that determines the properties that the index fulfills. When $\alpha = 0$ the FGT index corresponds to the traditional "headcount index" (denoted by H); that is, the proportion of the population living below the poverty line. If $\alpha = 1$, the FGT index corresponds to the "poverty gap," which measures the average distance between the income of the poor and the poverty line, weighted by the incidence of poverty. If $\alpha = 2$, the coefficient assigns a greater relative weight to observations for which income is further from the poverty line; as a result, the indicator is sensitive to the distribution of income among people below the poverty line.

3. Data sources

Implementation of the SAE unit-level model presented in this briefing requires two sources of data. The first is national household surveys. In this case, data are obtained from the Household Survey Data Bank (BADEHOG), a repository of household surveys from 18 Latin American countries maintained by the ECLAC Statistics Division. The second data source consists of national population censuses, which were accessed through the websites of the corresponding national statistical offices.

In the case of Chile, the 2017 National Social and Economic Survey (CASEN survey), which corresponds to a representative sample at the national, regional, national urban, and national rural levels, was used in conjunction with the 2017 Population and Housing Census. For Colombia, the Comprehensive Survey of Households of 2018, which is representative of the national, national urban, national rural, regional, departmental, and for the capitals of the country's departments, was used together with the 2018 National Population and Housing Census. In the case of Peru, the 2017 National Household Survey (ENAHO), which is representative at the national, urban, rural, and departmental levels, was used together with the twelfth Population Census, seventh Housing Census, and third Census of Indigenous Communities of 2017.

4. Small area estimation model

A unit-level model with adjustment for the complex sampling design is used to estimate average income. This model gives an approximation to the best empirical predictor (pseudo-EBP) based on the nested-error model,[3] as proposed by Guadarrama, Molina, and Rao (2018).[4] The method assumes that the transformed income variables, $y^*_{di} = log(y_{di} + c)$, follow the model described below (for simplicity, y_{di} will be referred to as the transformed variable):

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

where β is the vector of coefficients of the covariates, u_d is the random area effect, such that $u_d \overset{iid}{\sim} N(0,\sigma^2_u)$ and $e_{di} \overset{iid}{\sim} N(0,\sigma^2_e)$ are the individual-level errors, independent of the random effects. According to Molina (2019),[5] for FGT indicators that can be defined as a function of y_d the best linear predictor is the one that minimizes the

3 Molina, I. y Rao, J.N.K. (2010). Small Area Estimation of Poverty Indicators. *Canadian Journal of Statistics*, 38, 369-385.

4 M. Guadarrama, I. Molina and J. N. K. Rao, "Small area estimation of general parameters under complex sampling designs". *Computational Statistics & Data Analysis*, No. 121, 2017.

5 I. Molina, "Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas", *Statistical Studies series*, No. 97 (LC/TS.2018/82/Rev.1), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC), 2019.

mean squared error (MSE). It is given by the expected value of the elements that are not selected in the sample within the subdivision of interest d, conditional on the observed values of the selected elements.

Since the available data do not make it possible to identify and link sample units with census units, a "census-EB" type of approach is used. This assumes that all census items are associated with the out-of-sample observations, as follows:

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

In the region's household surveys, particularly those used in this briefing, the ratio of the number of units selected in the samples relative to the country's population is very close to zero; so the census-EB predictor performs quite similarly pseudo-EBP.

5. Procedure for generating poverty maps

This paper aims to produce geographically disaggregated poverty indicators and maps to visualize the resulting estimations. This helps afford policymakers a clear view of the incidence of the estimated indicator in different geographic domains, using different shades or colours to represent the magnitude of the income and poverty indicators. The poverty mapping process involves the following stages:

Stage 1:

- Standardization and harmonization of the databases
- Preparation of inequality, income, and poverty indicators

Stage 2:

- Estimation of the SAE model for income, inequality, or poverty indicators
- Definition of interactions and selection of auxiliary variables

Stage 3:

- Parametric bootstrap simulation to estimate MSE

Stage 4:

- Validation of model assumptions
- Benchmarking with survey estimates

Stage 5:

- Generation of maps based on the estimation of the FGT indicator and respective MSE

In the first stage, the unit-level models fitted to the survey data are replicated using the respective census microdata. Therefore, it is necessary to standardize the relevant variables by applying homogeneous definitions and categories in both data sources. This rules out possible biases induced by different measures in the covariates or prediction errors owing to different variables with similar names. For this purpose, standardized structures are generated, together with a dictionary of variables describing the categories and other specifications required in each case.

For example, to construct the variable "Years of study, "Peru's ENAHO survey identifies the last year of approved studies passed at all education levels. By contrast, the census in Peru identifies this very specific disaggregation only up to secondary school. Thereafter, the response options are much more general —referencing only complete or incomplete higher education, but not the number of years completed. In the cases of Chile and Colombia, however, the years of study variable was implicit in the microdata, so only the respective categorization process was carried out.

In the second part of this stage, the variable of interest is transformed to ensure the structure of the nested-error model defined in section 4. For example, the model considers a transformation of the per capita household income variable to ensure an approximately normal distribution; to this end, the Box-Cox and Log-Shift families of transformations were reviewed. The latter was chosen to perform the income transformation in the models of the three countries, although the parameters associated with each transformation turned out to be different.

A new variable $f(y_{di}) = log(y_{di} + c)$ is created for a grid of predefined values of c. The selected value of c. guarantees the normality of the transformed variable, together with a Fisher's coefficient of skewness closer to zero. Adjusting the process to the countries under analysis yielded the following local currency values for the constant c: Ch$ 8,600 (Chile), Col$ 81,958 (Colombia), and S/. 10.96 (Peru).

In the second stage of the methodology, a Monte Carlo simulation procedure is used to estimate the poverty indicators. It is often impossible to calculate analytically what may define the best predictor. Moreover, an essential part of this stage consists of identifying the predictive capacity of the auxiliary variables used.

A first alternative is to generate different linear models from various combinations of the covariates (with and without intercept) and compare their diagnosis measures. Then, the number of significant variables and the goodness-of-fit measures, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC), are among the elements used to analyse and compare the models. In addition, as a first step to establishing the feasibility of a set of covariates, ridge and lasso regressions were used, adapted to analyse the fit of the covariates.

The third stage of the proposed methodology uses the parametric bootstrap method to estimate the MSE of the census-EB predictor. This consists of generating entire populations using the census households located in the provinces, districts, and municipalities selected in each country's household survey and the model equation. A sample is selected with the same characteristics as the original sample, and the FGT indicator of interest is estimated from each entire population for all subdivisions of interest. After repeating this procedure many times, the estimated MSE will be obtained. At this stage, it is recommended that provinces, districts, and municipalities with a coefficient of variation above 30% be excluded from the map because they are considered insufficiently precise.

In the fourth stage of the procedure, benchmarking is performed with the ECLAC estimates obtained from the survey of the FGT indicator under analysis. This process is carried out at the level at which the survey estimates are representative (unbiased and precise); in other words, at the national, urban, rural, and departmental levels. This process aims to bring the aggregations of provinces, districts and municipalities to the figure reported for the different levels of disaggregation, in order to: (i) eliminate the bias produced by poor model specification; (ii) improve the estimates in the provinces, districts and municipalities based on the unbiased and consistent official estimations; and (iii) make it possible to construct a poverty map so as to be comparable with the published figures.

Subsequently, the assumptions of the nested error model described in section 4 need to be validated. In particular, normality tests are performed (Kolmogorov-Smirnov and Jarque-Bera tests); and graphical diagnostics in the form of kernel density histograms and quantile-quantile plots are also used. Next, heteroscedasticity tests are performed (White and Breusch-Pagan tests), and any outliers or influential values are identified (Cook's distances and dfbetas).

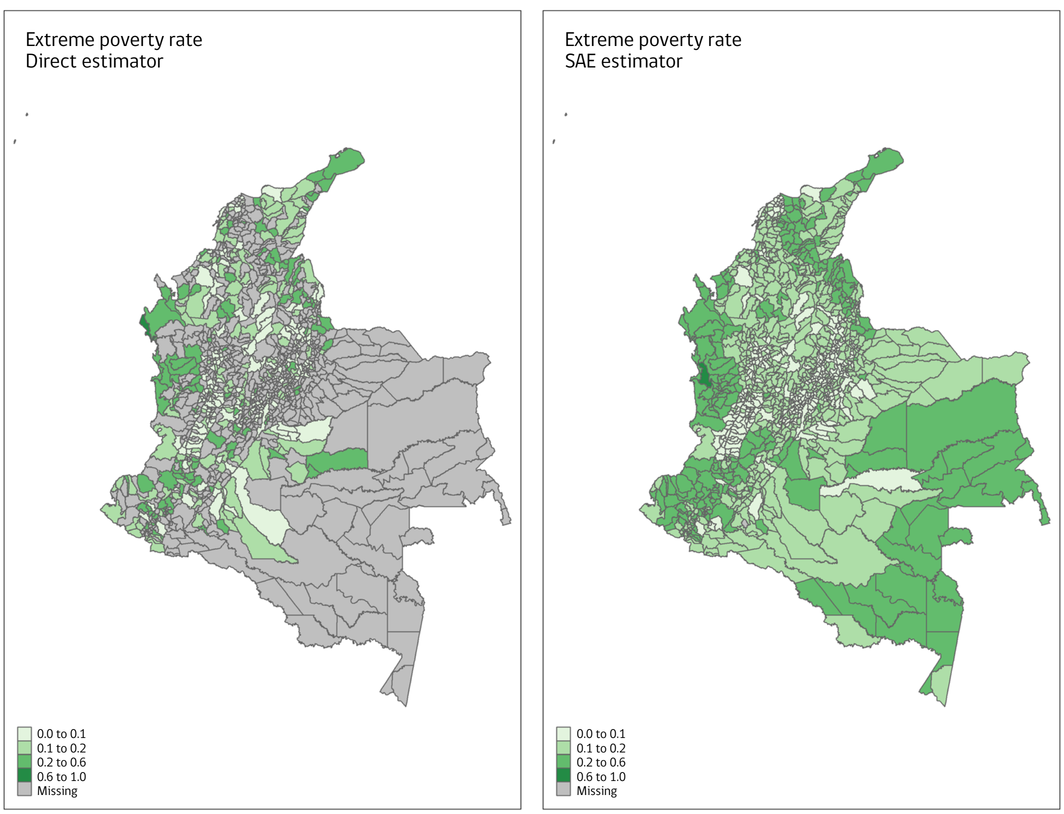
The last stage of the procedure uses geographic information systems (GIS) and the strata of interest in each country at the province, district and municipality levels to generate the maps presented in the appendix of this document.

6. Results

Poverty maps are instruments that serve to condense a vast amount of data on cities or municipalities into a single image. The visualization of poverty is useful as a communication tool and facilitates the analysis of the spatial relations between different indicators to enable a better understanding of poverty in the countries of the region. It also helps to identify priority areas and the geographic targeting of public expenditure and improve the coverage of social programmes, among other uses.

A selection of the poverty maps generated for Chile, Colombia, and Peru are shown below. All of them were prepared using four cut-off points and a colour scale ranging from light green (lower poverty) to red (higher poverty), to illustrate the distribution of poverty incidence across provinces, districts and municipalities in each country analysed. [6] To illustrate the effect of including auxiliary information in estimating poverty and extreme poverty in Colombia, map 1 displays the direct estimations, using household survey data only, and the estimations based on the SAE model, which combines census and survey data. Note that the municipalities for which precise estimates are not available, denoted in grey, predominate in the direct estimations map.

» Map 1. Colombia: incidence of extreme poverty, direct estimation and SAE estimation, 2019



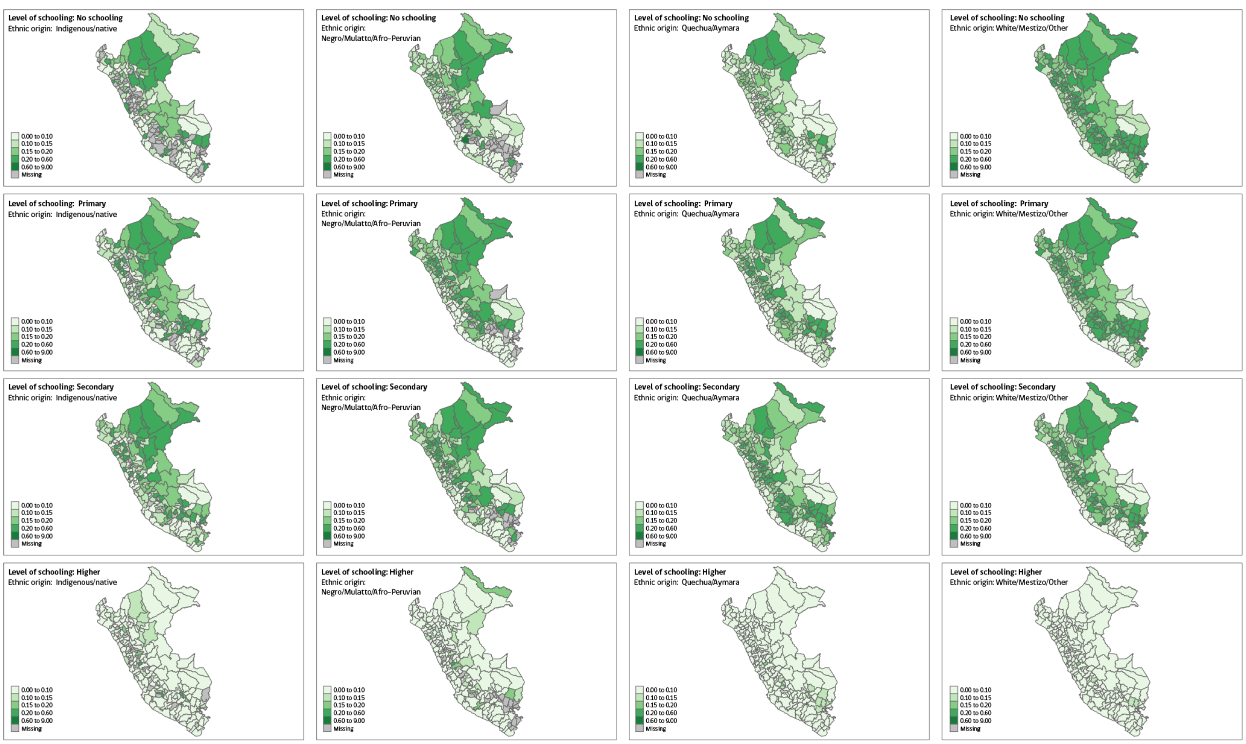
Source: Prepared by the authors.
Note: The boundaries and names shown on this map do not imply official endorsement or acceptance by the United Nations.

[6] As the incidence of poverty is lower in Chile than in Colombia and Peru, the cut-off points for the Chile map were adjusted to reflect the municipal dispersion of poverty more accurately.

In Colombia, the poorest municipalities are located in the outlying areas of the country; specifically, in the Pacific region (Chocó, Cauca and Nariño), the Caribbean region (La Guajira, Magdalena, Bolívar, Sucre, Córdoba and Cesar), the border departments (Norte de Santander and Arauca), and the Orinoco and Amazon region (Vichada, Guainía, Vaupés, Amazonas and Caquetá). In Peru, the highest poverty rates are found in the northern and southern border areas. Specifically, in the north, the highest incidence of poverty is concentrated in the regions of Loreto, Amazonas, the southern part of Cajamarca, and western La Libertad; similarly, near the border with the Plurinational State of Bolivia, there is a second focus of poverty in the regions of Puno and Cuzco.

The relevant disaggregations refer to geographic areas and population groups that are relevant for public policymaking and the follow-up and monitoring of the 2030 Agenda for Sustainable Development —by gender, age, employment status, ethnic group, disability status, and so forth. By way of illustration, map 2 displays the distribution of poverty in the Peruvian provinces by ethnic group and level of schooling. The poverty rate is higher among indigenous groups than the rate among Afrodescendants, which is higher than that of individuals who do not self-identify with any ethnic group. Similarly, as educational attainment increases, individuals are less likely to be living in poverty.

» Map 2. Peru: SAE estimation of the incidence of poverty by ethnic origin and level of schooling, 2019

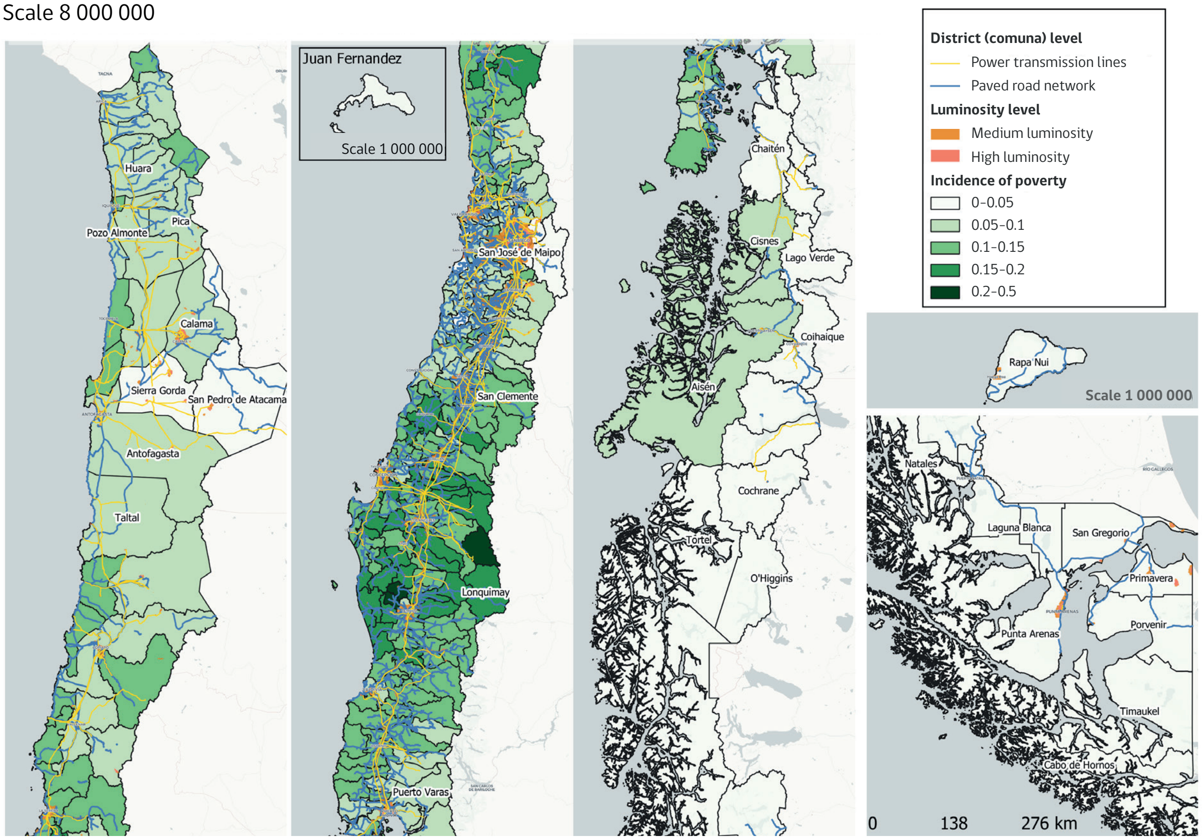


Source: Prepared by the authors.
Note: The boundaries and names shown on this map do not imply official endorsement or acceptance by the United Nations.

The results shown on the poverty maps can be contrasted with each country's infrastructure networks. For example, road coverage and access to electricity are key elements in integrating the economic system, facilitating transactions between firms and citizens. Map 3 shows that the areas with the lowest incidence of poverty are generally located in urban areas with denser road infrastructure and more efficient service provision.

» Map 3: Chile: Local infrastructure contrasted with the SAE estimation of the incidence of poverty

Incidence of poverty at the district level and access to infrastructure, Chile.
Scale 8 000 000



Source: Prepared by the authors.
Note: The boundaries and names shown on this map do not imply official endorsement or acceptance by the United Nations.

This publication, which is issued bimonthly, is prepared by the Statistics Division of the Economic Commission for Latin America and the Caribbean (ECLAC). This issue was prepared by Andrés Gutiérrez, Regional Expert on Social Statistics; Xavier Mancero, Chief of the Social Statistics Unit; and Felipe Molina and Diego Lemus, consultants of the Statistics Division.

https://www.cepal.org/en/work-areas/statistics

ECLAC

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