UN-ECLAC

Governments use subnational poverty maps to design, implement, and monitor development policies more effectively by targeting the places or population groups that urgently need them. The Statistics Division of the United Nations Economic Commission for Latin America and the Caribbean (UN-ECLAC) has developed an SAE methodology that makes it possible to combine census information with household survey data to provide disaggregated estimates at the levels for which direct estimates from household surveys are generally too imprecise. 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 programs, among other uses.

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, the inference resulting from the surveys is not precise or accurate. In general, the higher the disaggregation, the less efficient the estimators become, and their reliability declines ostensibly. In the case of some complex indicators, this can even 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. Here we describe a series of steps for applying bayesian unit-level models to estimate indicators of interest relating to household income. These include the mean income, incidence, gap, and intensity of poverty and extreme poverty—at the state level in Latin American countries. 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.

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. 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 the 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, and include the traditional "headcount index" (the proportion of the population living below the poverty line), and the "poverty gap" (which measures the average distance between the income of the poor and the poverty line, weighted by the incidence of poverty).

Data sources

Implementation of the SAE models 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. For example:

- 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, 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.

Small area estimation models

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, as proposed by Guadarrama, Molina, and Rao (2018). The method assumes that the transformed income variables follow a nested error model including random effects for the subdivision of interest d. According to Molina (2019), for FGT indicators, the best linear predictor (the one that minimizes the mean squared error) 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. In household surveys of Latin America, the ratio of the number of units selected in the sample relative to the country’s population is very close to zero; so the census-EB predictor performs quite similarly to the pseudo-EBP.

Procedure for generating poverty maps

To produce geographically disaggregated poverty indicators and maps to visualize the resulting estimations helps afford policymakers a clear view of the incidence of poverty and extreme poverty in different geographic domains. The poverty mapping process involves the following stages:

- **Stage 1:** Standardization and harmonization of the databases, estimation of inequality, income, and poverty indicators.
- **Stage 2:** Estimation of the SAE model for income, inequality, or poverty indicators, the definition of interactions, and selection of auxiliary variables.
- **Stage 3:** Prediction in the subdomains of interest through the use of the EBP approach.
- **Stage 4:** Validation of model assumptions and benchmarking with survey estimates.
Stage 5: Parametric bootstrap simulation to estimate MSE
Stage 6: 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. Therefore, 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 D. 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 transformed variable is created to guarantee the normality assumptions of the model.

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 analyze 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 analyze the fit of the covariates.

The third stage uses the census data and predefined common covariates to predict the income on all of the units of the census. 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 to (i) eliminate the bias produced by a deficient 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 national published figures estimated by ECLAC from BADEHOG. The assumptions of the nested error model 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 fifth 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.

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 such as shown next. The map at the left shows the mean income estimates (in poverty lines), the middle one shows the poverty rate, and the map from the right shows the extreme poverty rate.
The ECLAC approach allows to focus on a single country and disaggregate the estimates not only at the geographical level but also incorporating other subgroups of interest such as sex, zone, age, disability, ethnicity, and education. For example, the following map shows how the extreme poverty rate is distributed in the Peruvian case by education (the first column represents no education, the second column represents 1 to 6 years of education, third column 7 to 12 years, and fourth column more that 12 years of education) and ethnicity (the first row represents indigenous people, the second row represents Afro-Peruvian people and the third row is devoted to the rest of the population).

References

- Economic Commission for Latin America and the Caribbean (ECLAC), “Medición de la pobreza por ingresos: actualización metodológica y resultados”, Metodologías de la CEPAL, No. 2 (LC/PUB.2018/22-P), Santiago, 2018. [AUTHOR: Please confirm this reference. OK]