ECLAC Approach for Poverty Mapping in Latin America: Small Area Estimation using Unit-level Models

Andrés Gutiérrez

Economic Commission for Latin America and the Caribbean

April 2022

Motivation

UNITED NATIONS

ECLAC

ECLAC experience

- Most Latin American and Caribbean countries regularly implement nationally representative household surveys to measure living conditions indicators (including poverty and income inequality).
  - These surveys can generally be disaggregated geographically by urban and rural areas and the level of the first administrative division.
  - Small Area Estimation (SAE) techniques allow obtaining such disaggregated estimates while improving inference quality.
- This talk presents the recent experience of the Statistics Division of the United Nations Economic Commission for Latin America and the Caribbean (ECLAC) in applying SAE techniques to estimate poverty indicators in seventeen (17) countries of Latin America.

UNITED NATIONS
ECLAC

Some history

- ECLAC regularly produces standardized national estimates of extreme poverty and poverty for Latin American countries using a methodology that aims to achieve regional comparability.
  - Even though countries in the region publish their own official poverty statistics, the diversity of procedures and assumptions used in these estimates prevent direct comparison.
  - ECLAC approach for measuring poverty classifies a person as poor when the per capita income of their household is lower than the poverty line proposed by ECLAC, based on the cost of meeting their food needs and other basic non-food needs.

UNITED NATIONS
ECLAC

ECLAC Poverty Lines

- The cost of food needs is estimated through the construction of basic food baskets, which provide the recommended amounts of energy and nutrients while reflecting the consumption habits of the population.
  - Consumption habits are captured through household income and expenditure surveys and correspond to those of a particular subset of the population (reference population) based on criteria established by ECLAC.
  - The monthly cost of the basic food basket is referred to as the extreme poverty line.
- The poverty line is obtained as the product of the extreme poverty line by the Orshansky coefficient of the same population of reference used to define the basic food basket.

UNITED NATIONS
ECLAC

Direct Estimators

- Household surveys are designed and implemented by national statistical offices (NSO) to generate representative statistics at a predefined level of aggregation, generally based on large geographic subdivisions, sex, or socioeconomic groups of the population.
  - 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.
  - 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.

Stages for Poverty Mapping

- Stage 1. Standardization and homologation of covariates in the databases (censuses and household surveys).
- Stage 2. Updating intercensal counts related to covariates preserving the census structures while updating marginals from the household survey.
- Stage 3. Definition of the models for indicators related to income and poverty, considering possible interactions, selection of auxiliary variables and estimation of model coefficients.
- Stage 4. Prediction of poverty on censal poststrata and small areas, estimation of the MSE based on Bootstrap replicas.
- Stage 5: Validation of model assumptions and benchmarking using ECLAC estimates of mean income and poverty at the national, urban, and rural levels.
- Stage 6: Generation of maps for 17 countries of Latin America.

UNITED NATIONS
ECLAC

SDG and Sources of Information

UNITED NATIONS

ECLAC

SDGs and 2030 Agenda

- The 2030 Agenda for Sustainable Development comprises 17 sustainable development goals (SDG) that integrate the different dimensions of development, such as the economic, social, and environmental.
  - It focuses on the most vulnerable subgroups of the population.
  - This is why the Leave No One Behind (LNOB) mandate claims to disaggregate SDG indicators by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics.
- For instance, SDG 1 (End poverty in all its forms everywhere) aims to eradicate extreme poverty for all people everywhere by 2030.

Household Surveys

- They are obtained from ECLAC's Household Survey Data Bank (BADEHOG).
  - This is a repository of household surveys from 18 Latin American countries maintained by the Statistics Division.
- For example, in the case of Chile, the 2017 National Social and Economic Survey (CASEN survey) corresponds to a representative sample at the national, regional, urban, and rural levels.
- For Colombia, the Great Integrated Household Survey of 2018, which is representative of the national, urban, rural, and regional levels, along with departments and their capitals, was used.
- In the case of Peru, the 2017 National Household Survey (ENAHO), which is representative of the national, urban, rural, and departmental levels, was considered. Table 1 shows a comprehensive summary of the household surveys used in this system.

Country	Survey	Year
ARG	Permanent Household Survey (EPH)	2019
BOL	National Household Survey	2020
BRA	National Survey by Continuous Household Sample	2020
CHL	National Socioeconomic Characterization Survey (CASEN)	2020
COL	Large Integrated Household Survey	2020
CRI	National Household Survey (ENAHO)	2020
DOM	National Continuous Labour Force Survey (ENCFT)	2020
ECU	National Survey on Employment, Unemployment and Underemployment (ENEMDU)	2020
GTM	National Survey on Living Conditions	2014
HND	Multipurpose Household Survey	2019
MEX	National Household Income and Expenditure Survey (ENIGH)	2020
NIC	National Household Survey on Living Standard Measurement	2014
PAN	Multipurpose Survey	2019
PER	National Household Survey - Living Conditions and Poverty	2020
PRY	Continuous Permanent Household Survey (EPHC)	2020
SLV	Multipurpose Household Survey	2020
URY	Continuous Household Survey	2020

National Population Censuses

- They have been accessed through the ECLAC's census data bank, maintained by the Population Division (CELADE), which has pursued its ongoing activity of disseminating the census information derived from its data bank.
- Most Latin American countries have transmitted their databases to CELADE, allowing the availability of microdata from the previous and current rounds of censuses.
  - We used the software Redatam for statistical processing specialized in microdata of population and housing censuses developed by CELADE.
  - This computational solution is a database management tool that administrates large volumes of census microdata with a hierarchical structure down to the smallest area of the census exercise (blocks).
  - Through this platform, we have all of the censuses in the region at our disposal.

UNITED NATIONS

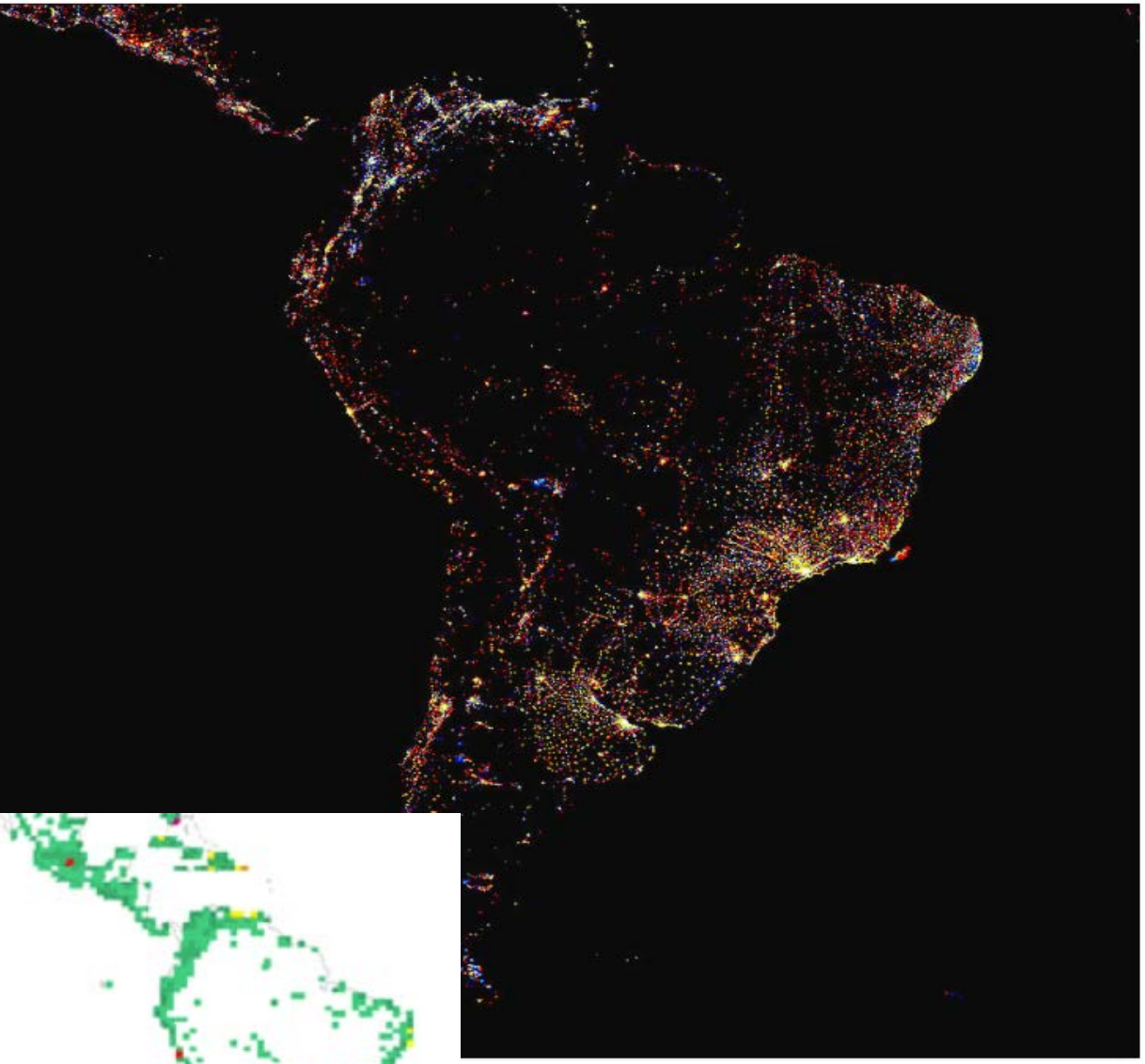
ECLAC

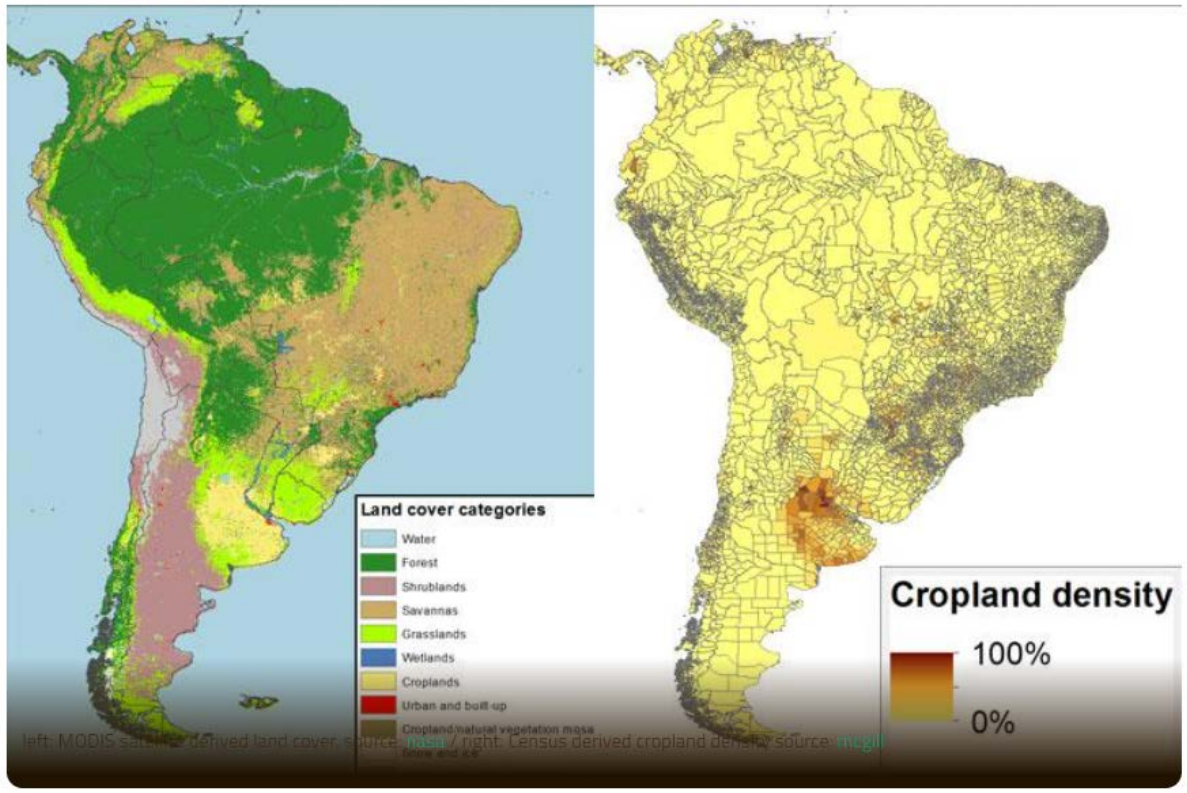
Country	Survey	Year
ARG	Censo Nacional de Población, Hogares y Viviendas 2010	2010
BOL	Censo de Población y Vivienda 2012	2012
BRA	Censo Demográfico 2010	2010
CHL	Censos de Población y Vivienda	2017
COL	Censo Nacional de Población y Vivienda - CNPV - 2018	2018
CRI	X cenSo nacIonal De PoBlacIón Y VI De VIVIenD	2011
DOM	IX Censo Nacional de Población y Vivienda 2010	2010
ECU	VII Censo de Población y VI de Vivienda	2011
GTM	XII Censo Nacional de Población y VII de Vivienda	2018
HND	XVII Censo de población y VI de Vivienda 2013	2013
MEX	Censo de Población y Vivienda 2020	2020
NIC	Censo de Población y Vivienda de Nicaragua de 2005	2005
PAN	Censo 2010	2010
PER	XII Censo de Población, VII de Vivienda y III de Comunidades Indígenas o Censo peruano de 2017	2017
PRY	II Censo Nacional Indígena de Población y Vivienda 2002. Pueblos Indígenas del Paraguay	2002
SLV	VI Censo de Población y V de Vivienda 2007	2007
URY	Censo de Población 2011	2011

Satellite Imagery

- We access this information trough Google Earth Engine, which provides facilities to analyze and obtain this data through the Javascript and Python programming languages, and recently since 2021 in R with the rgee package.
- Among the main advantages of information based on remote sensing is the ease of access to information with high geographic coverage that is impossible to obtain by traditional means such as surveys or administrative records.
- On the other hand, data panels can be built at a low marginal cost of variables as diverse as night lights, rainfall, wind speed, floods, topography, forest cover, types of crops, urban development, kind of road services, among many other variables that can be proxies for different economic aspects.

- Night lights are a notable example of the use of satellite images.
  - The intensity of the captured lights has been used as proxy for different socioeconomic variables such as the product gross domestic product.
  - However, there have been strong debates about the assumption that night lights can serve as a good proxy for development, especially in low and middle-income rural areas.
- The use of satellite images, particularly night lights, urban cover fraction and crop cover fraction give countries the option of compensating for the lack of population censuses and detailed surveys.





left: MODIS satellite derived land cover, source: nasa / right: Census derived cropland density source: mcgill

Updating intercensal counts

UNITED NATIONS

ECLAC

Census counts

- Population and housing censuses in Latin American countries are the main resources for obtaining detailed information on socio-demographic issues.
- In most developing countries, population censuses are carried out only every 10-15 years, which means the auxiliary information is outdated.
- We need to update census counts because not all of the countries in Latin America have recent censuses.
- This way, while preserving the structure of the old census, we can take advantage of the margins provided by the survey to update the census counts.

SPREE

- The need to work with up-to-date census tables arise from the field of demography to obtain post-census counts.
  - Rao (2003) reviews methods used to update census tables using demographic methods.
  - Table updating techniques have also been used to generate synthetic population data.
- One solution to update census counts is given by the Structure Preserving Estimator (SPREE) in one or more categorical variables of interest according to study domains for post-census years.

SPREE

- The SPREE method is popular in the context of estimates in small areas.
- The procedure to obtain updated SPREE tables is carried out employing the iterative proportional fitting (IPF) procedure (Deming, 1940).
- This process is also found in the literature as raking ration, or multiplicative raking and adjusts the counts of a contingency table based on a set of given margins.
- The SPREE method assumes the allocation structure obtained from an updated survey providing recent and reliable margins (rows and columns).

Unit-level models

UNITED NATIONS

ECLAC

Empirical Best Predictor

- ECLAC uses a unit-level model with adjustments to the complex sampling design to estimate average income. This model was first proposed by Guadarrama, Molina, and Rao (2018), and it induces an approximation of the best empirical predictor (Pseudo-EBP) based on the model with nested errors (Molina and Rao, 2010).
- This method assumes that the transformed income variable $y_{di}^{*} = log\,(y_{di} + c)$ follows the model described below (for simplicity, we will denominate the transformed variable as y_{di}).

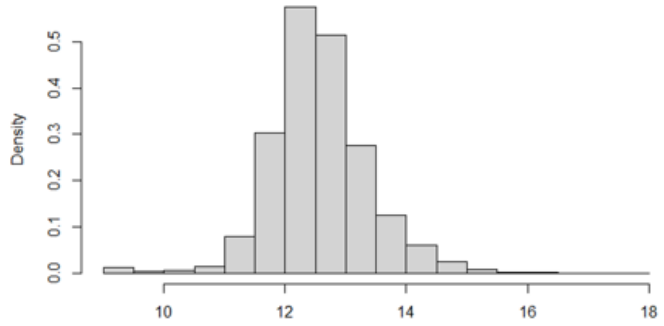
$$y_{di}^{*} = \mathrm{x}_{di}^{T}\boldsymbol{\beta} + u_d + e_{di}.\quad i = 1, \ldots, N_d,\ d = 1, \ldots, D,$$

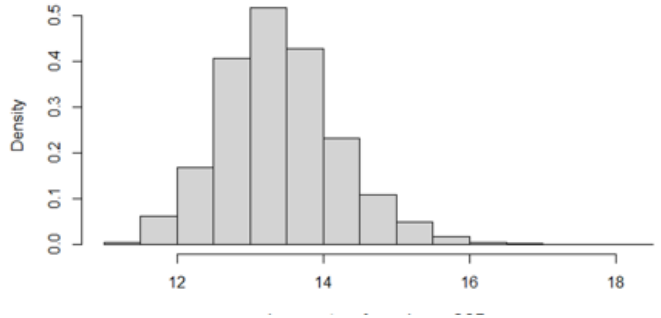
UNITED NATIONS

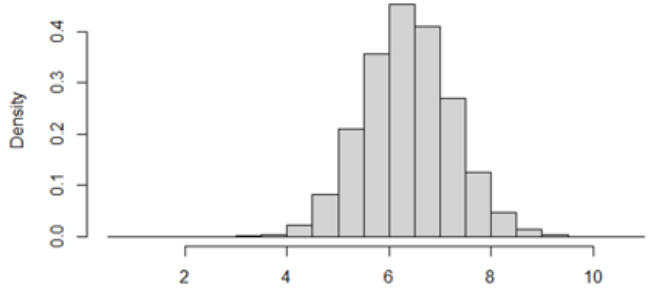
ECLAC

Transformation

- The model considers a transformation of the variable per-capita income that guarantees an approximately normal distribution. For this purpose, the Box-Cox and Log-Shift transformation families were explored. The latter was chosen to carry out the transformation of income in the models of the three countries, although the parameters associated with each country were different.







UNITED NATIONS
ECLAC

Best Linear Predictor

- Under this model, the vectors $\boldsymbol{y_d}$ are independent and follow a normal distribution with mean $\boldsymbol{\mu_d} = \boldsymbol{X_d\beta}$ and covariance matrix given by:

$$V_{\boldsymbol{d}} = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T + \sigma_e^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^T.$$

- According to Molina (2019), for those FGT indicators that can be defined as a function of y_d – that is, $\boldsymbol{\delta_d} = \boldsymbol{f}(\boldsymbol{y_d})$ – the best linear predictor is the one that minimizes the Mean Square Error (MSE) and is given by:

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\boldsymbol{y_{dr}}}[\delta_d(\boldsymbol{y_d})|\boldsymbol{y}_{ds}; \boldsymbol{\theta}]$$

UNITED NATIONS

ECLAC

Conditional expectation

- Since $\boldsymbol{y_d}$ follows a normal distribution, the conditional distribution $\boldsymbol{y_{dr}} \mid \boldsymbol{y_{ds}}$ will also be a normal distribution parameterized as follows:

$$\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds} \sim_{ind} N\left(\boldsymbol{\mu}_{dr|s}, \boldsymbol{V}_{dr|s}\right) \quad with \quad d = 1, \ldots, D$$

- To avoid the bias induced by ignoring the sampling design in the model, the parameters of the above distribution may consistently be estimated by including the sampling weights (w_{kd})

$$\hat{\boldsymbol{\mu}}_{dr|s} = \boldsymbol{X}_{dr}\hat{\boldsymbol{\beta}} + \hat{\gamma}_d(\bar{y}_{dw} - \bar{\boldsymbol{x}}_{dw}^T\hat{\boldsymbol{\beta}})\mathbf{1}_{N_d - n_d}$$
$$\hat{\boldsymbol{V}}_{dr|s} = (\hat{\sigma}_e^2 + \hat{\sigma}_u^2(1 - \hat{\gamma}_d))\mathbf{1}_{N_d - n_d}\mathbf{1}_{N_d - n_d}^T$$

UNITED NATIONS

ECLAC

Model selection

- We generate different linear models from several combinations of the covariates (with and without intercept) and compare them.
- We consider the number of significant variables and goodness of fit measures (AIC or BIC) in the analysis and comparability of the models.
- Furthermore, we use Ridge and adapted Lasso regressions to obtain a first impression of the feasibility of a set of covariates.
- We consider a Monte Carlo simulation procedure to estimate poverty indicators since the expectation that defines the best predictor often cannot be calculated analytically.

MSE on Small Areas and Benchmarking

UNITED NATIONS

ECLAC

Bootstrap

- We apply a parametric Bootstrap method to estimate the ECM of the Census-EB predictor.
- The unit-level models fitted to the survey data are replicated using census microdata. The mean square error estimator is given by:

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

UNITED NATIONS

ECLAC

Quality criterium

- With the estimated mean square error, the coefficient of variation (a measure that allows defining the quality of the estimates) has the following expression:

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

- We exclude from the map any province, commune, and municipality with a coefficient of variation ($\widehat{CV}$) greater than 30%, as they do not reach the desired precision.

UNITED NATIONS

ECLAC

Benchmarking on national figures

- Another stage of the procedure consists in benchmarking the results using the estimated FGT indicators from the survey at the national, national urban, national rural, and first administrative division levels.
- This guarantees consistency between the published figures as the aggregations of provinces, communes, and municipalities must be identical to those reported at different levels of disaggregation.
- This procedure also reduces the bias produced by a model miss-specification and improves the estimates' quality in the provinces, communes, and municipalities.

Benchmarking constraints

- The Monte Carlo simulation makes it possible to access the prediction vector for all the households in the pseudo-census. In addition, we have the poverty estimate from the survey. These quantities have the following ratio representation $R_q = \frac{t_{yq}}{t_{xq}}$.
- The purpose of the algorithm is to find a set of weights $d_k (k \in U)$, such that they meet the following restrictions:

$$R_q = \frac{\sum_{k \in S} d_k \, y_{qk}}{\sum_{k \in S} d_k x_{qk}}$$

UNITED NATIONS

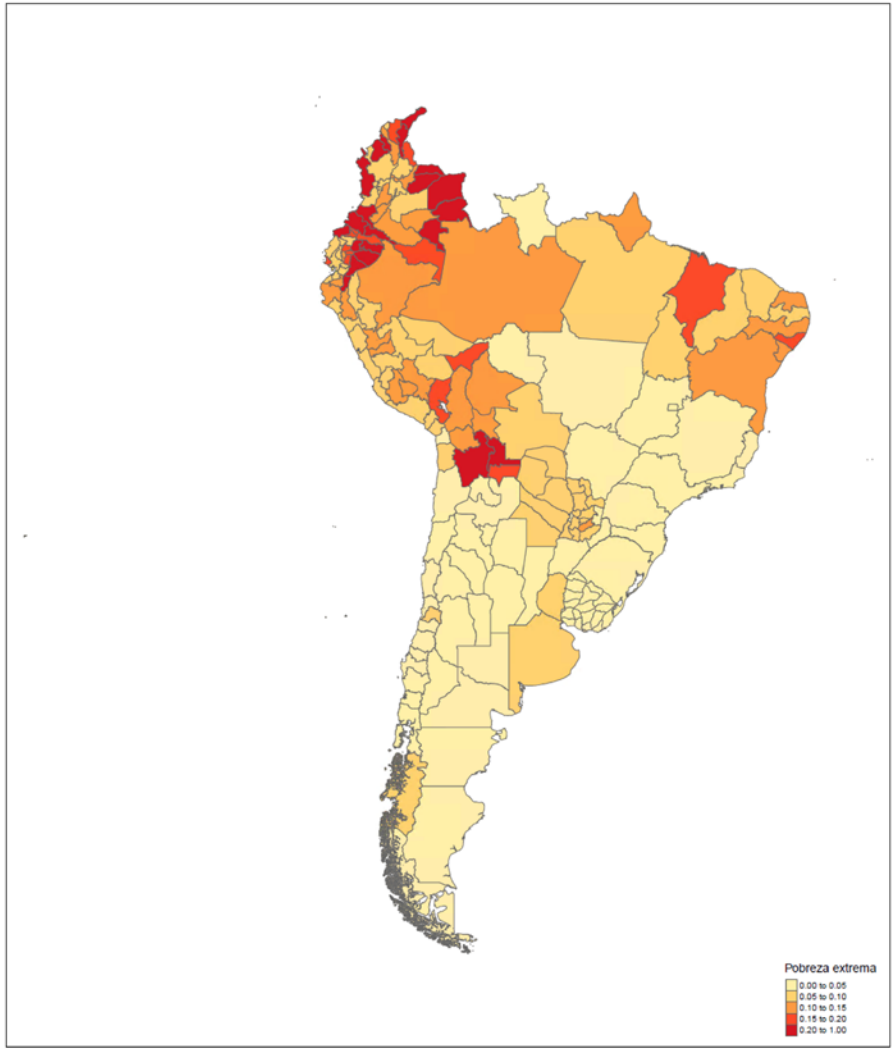
ECLAC

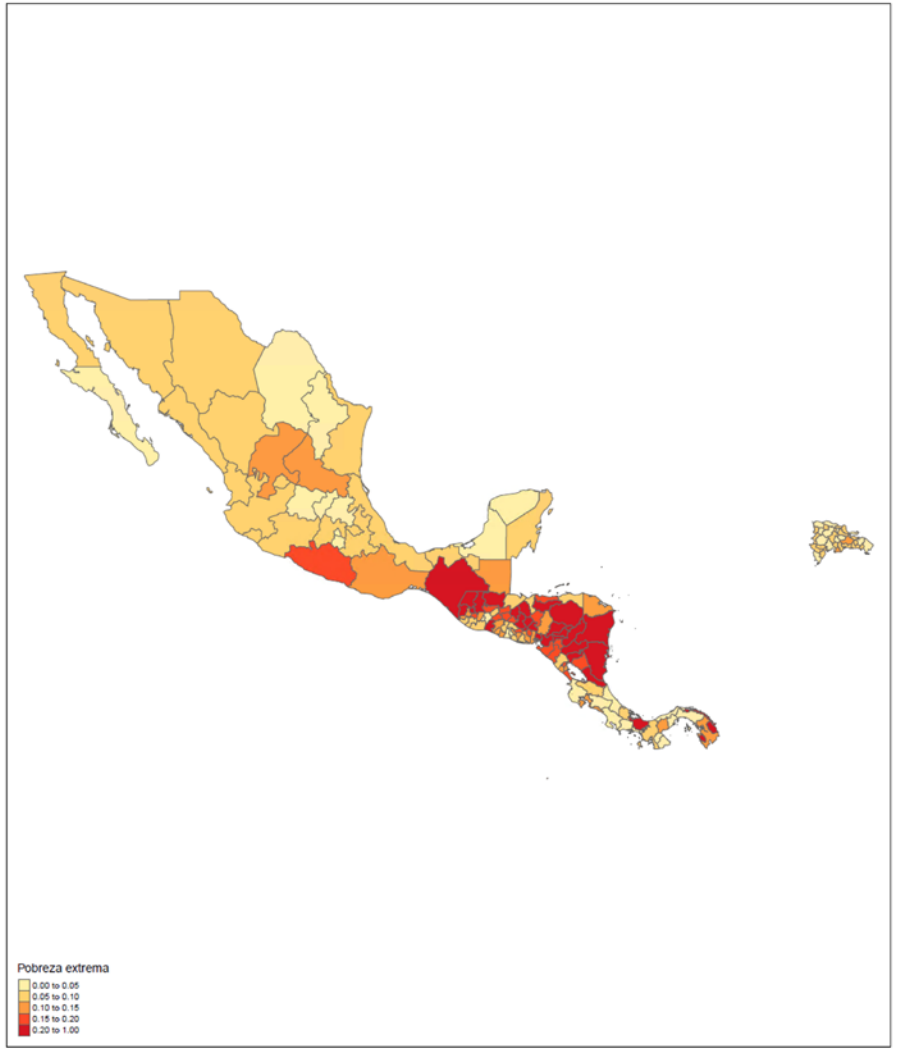
Maps

UNITED NATIONS

ECLAC

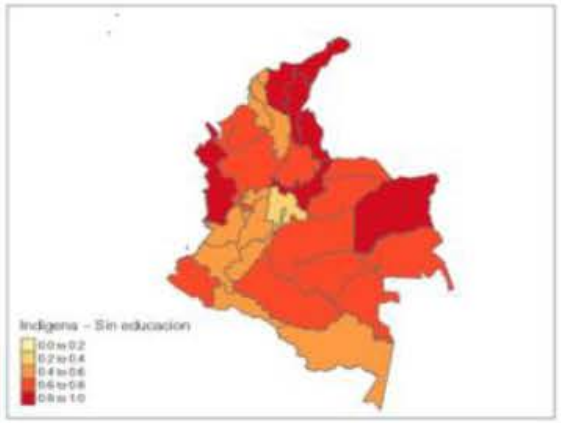
Extreme Poverty

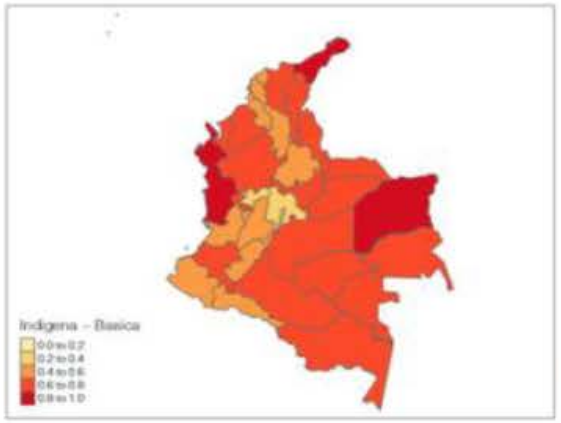


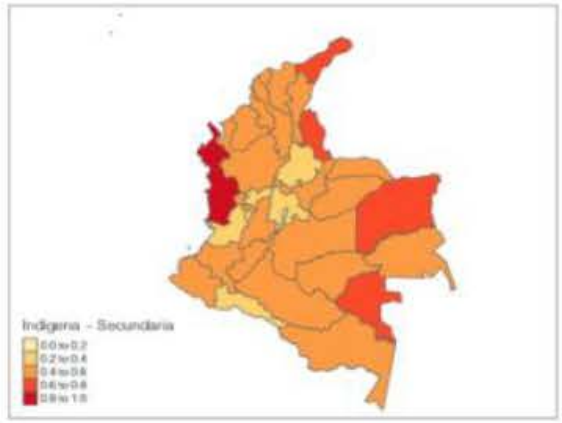


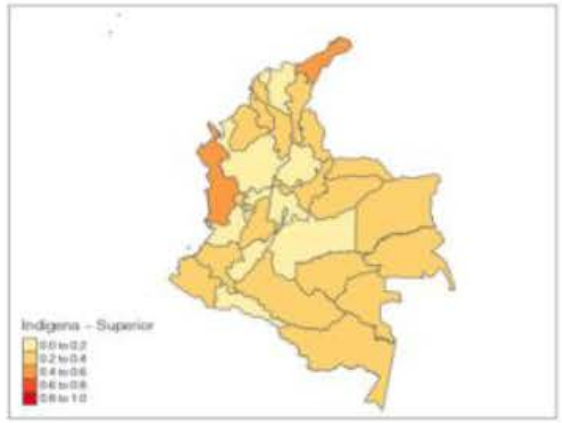
UNITED NATIONS

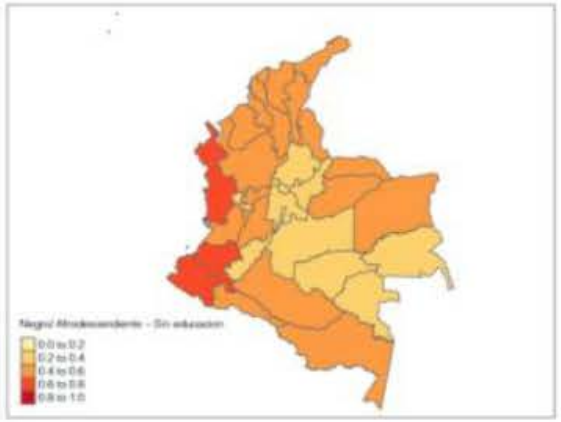
ECLAC

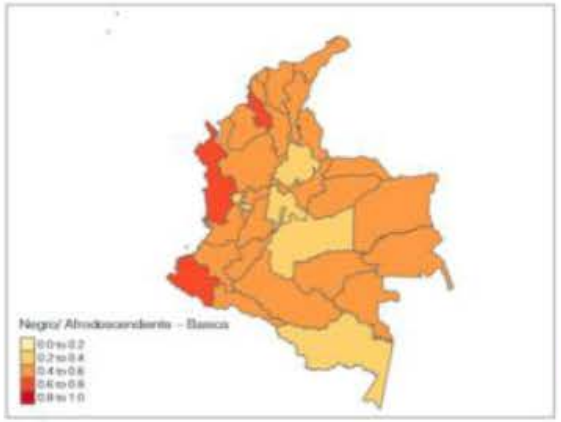


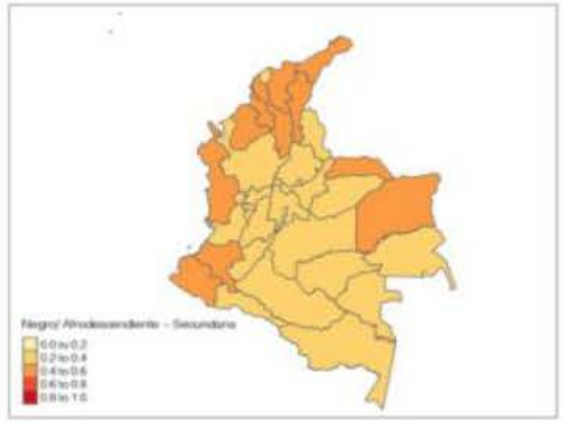


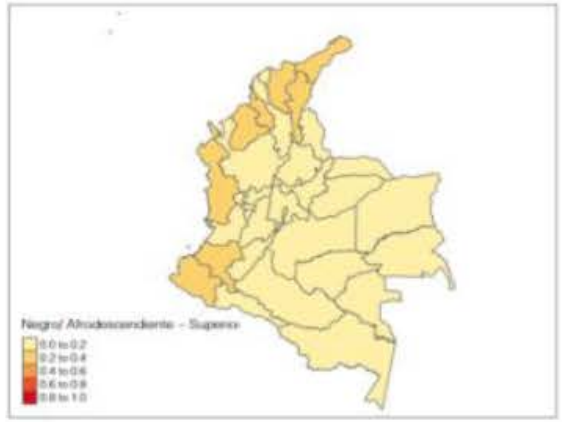


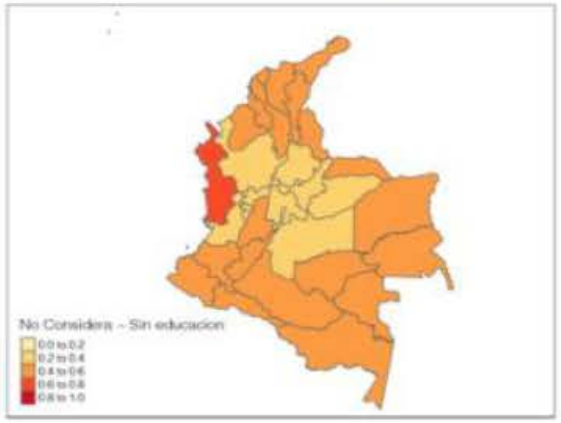


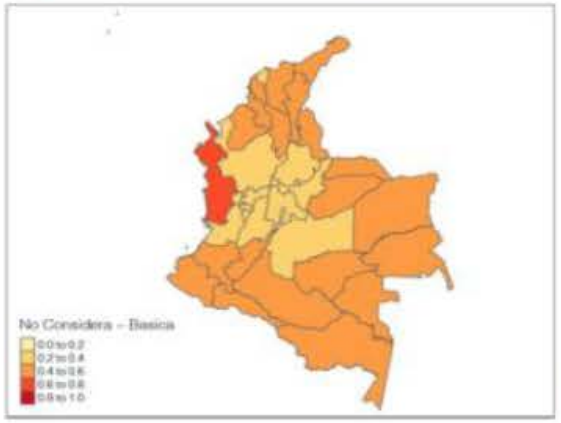


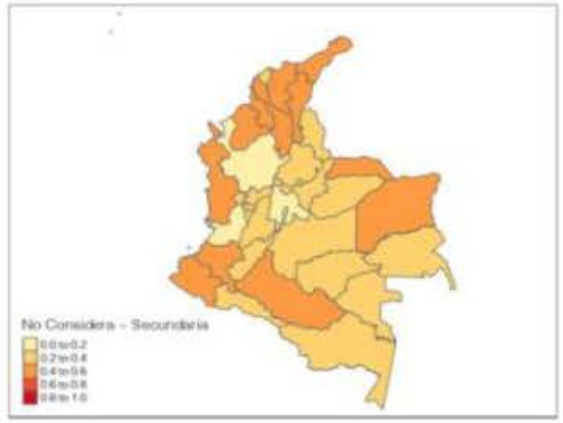


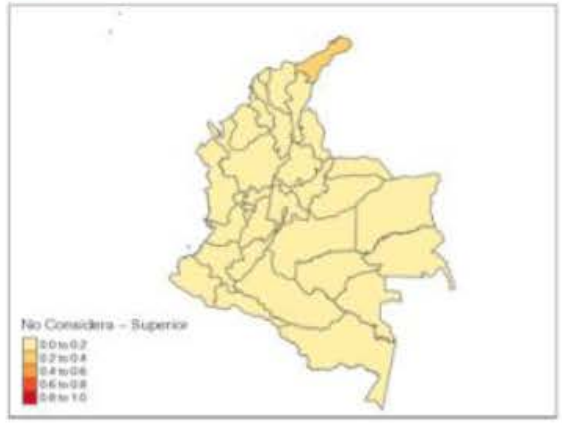






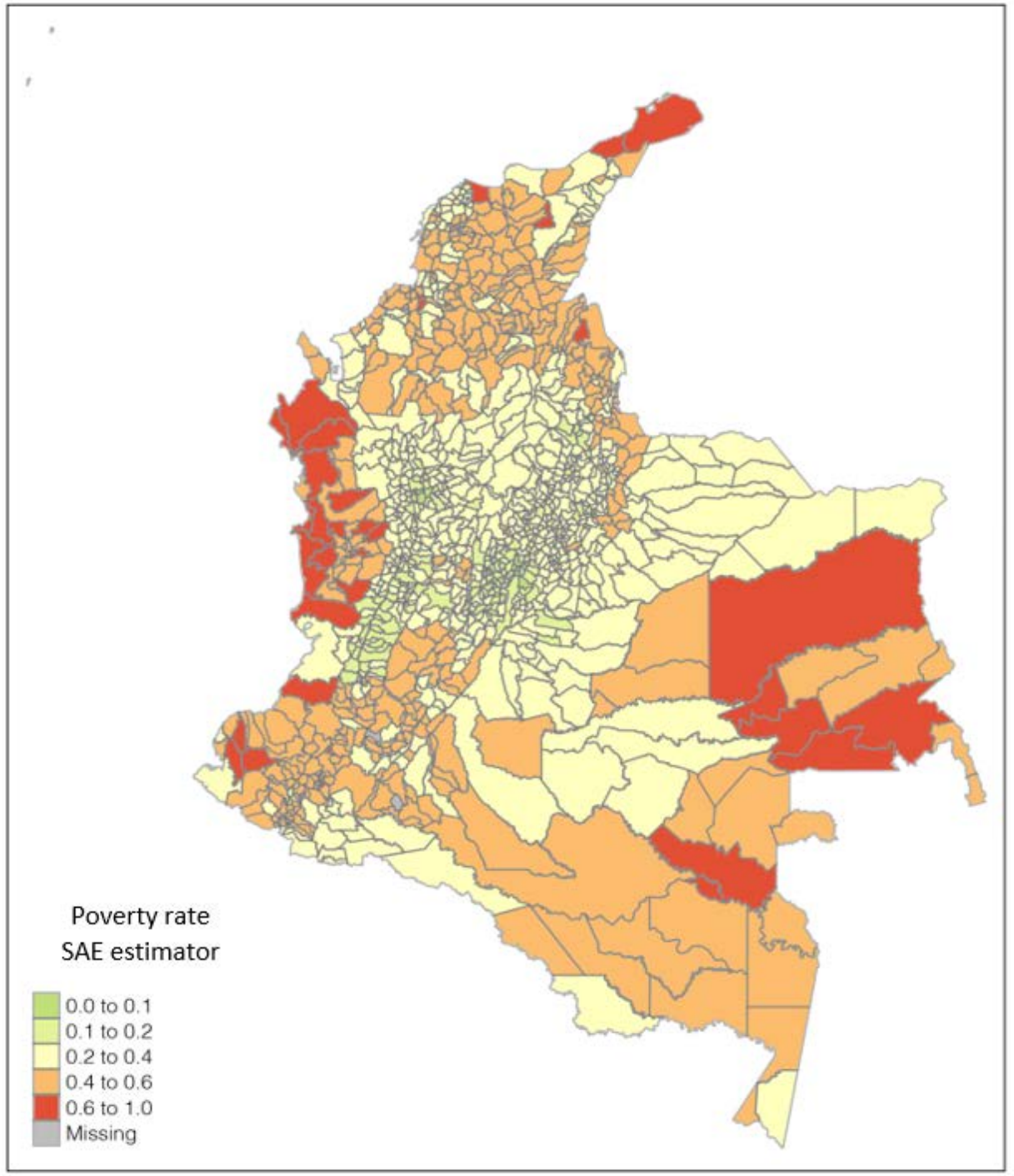






UNITED NATIONS

ECLAC



UNITED NATIONS

ECLAC

¡Thank you!

Andrés Gutiérrez

Regional Advisor on Social Statistics

Division of Statistics

UN-ECLAC

andres.gutierrez@un.org

UNITED NATIONS

ECLAC