ECLAC approach to poverty mapping

El enfoque de CEPAL en el mapeo de la pobreza

Andrés Gutiérrez

2021

SDG / ODS

No Poverty / Poner fin a la pobreza

- 1.1 By 2030, eradicate extreme poverty for all people everywhere.
- 1.2 By 2030, reduce at least by half the proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions.

- 1.1. De aquí a 2030, erradicar para todas las personas y en todo el mundo la pobreza extrema.
- 1.2. De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales.

Leave no one behind / No dejar a nadie atrás

Sustainable Development Goal indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics.

Global indicator framework for the Sustainable Development Goals (A/RES/71/313).

Desglosar los ODS por ingreso, sexo, edad, raza, etnicidad, estado migratorio, discapacidad y ubicación geográfica, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales.

Marco de indicadores globales para los Objetivos de Desarrollo Sostenible (A/RES/71/313).

Household surveys limitations and the use of auxiliary information

Limitaciones de las encuestas de hogares y el uso de información auxiliar

¿What is it all about? / ¿De qué se trata?

Surveys that depend on a large sample size and a proper sampling strategy rely on a robust inferential system that provides precise and exact estimation in planned domains.

When the sample size of the survey is not enough, it is necessary to resort to external auxiliary information (censuses, administrative records, satellite images) so that a precise and exact inferential system can be built.

Las encuestas que dependen de un buen tamaño de muestra y una estrategia de muestreo adecuada se basan en un sistema inferencial sólido que proporciona una estimación precisa y exacta en los dominios planificados.

Cuando el tamaño muestral de la encuesta no es suficiente para sustentar la inferencia estadística requerida para algunos subgrupos de interés, es necesario recurrir a información auxiliar externa para que en conjunto se puede construir un sistema inferencial preciso y exacto.

Solutions / Soluciones

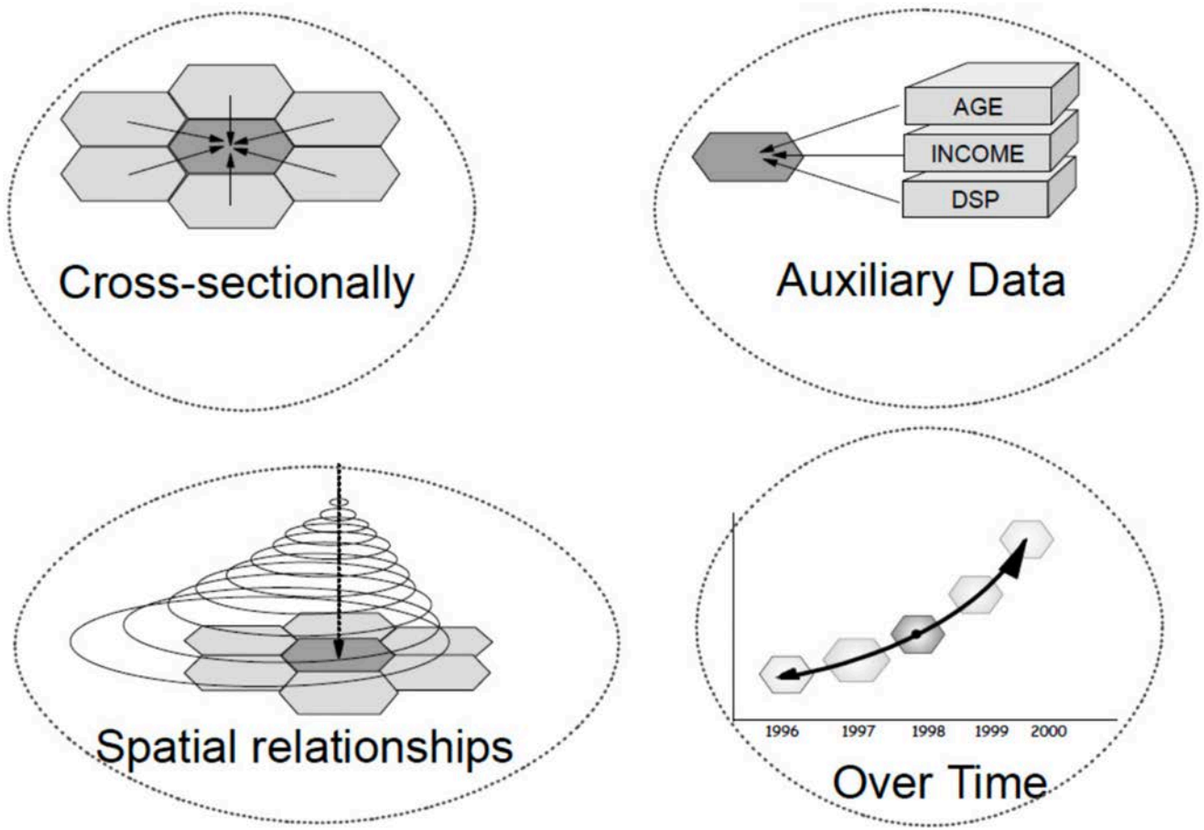
When the sample size does not allow obtaining reliable direct estimates for some domains of interest, the following options can be addressed:

1. Increase the sample size: this option raises costs, and it is unfeasible.
2. Use statistical methodologies that involve external auxiliary information to obtain reliable estimates (not direct) in the subgroups of interest, while keeping the survey sample size.

Cuando el tamaño de la muestra no permite obtener estimaciones directas confiables para algunos dominios de interés, se pueden abordar las siguientes opciones:

1. Incrementar el tamaño de la muestra: esta opción eleva los costos y es inviable.
2. Utilizar metodologías estadísticas que involucren información auxiliar externa para obtener estimaciones confiables (no directas) en los subgrupos de interés, mientras se mantiene el tamaño de la muestra de la encuesta.

Borrowing strength / Tomando fuerza prestada

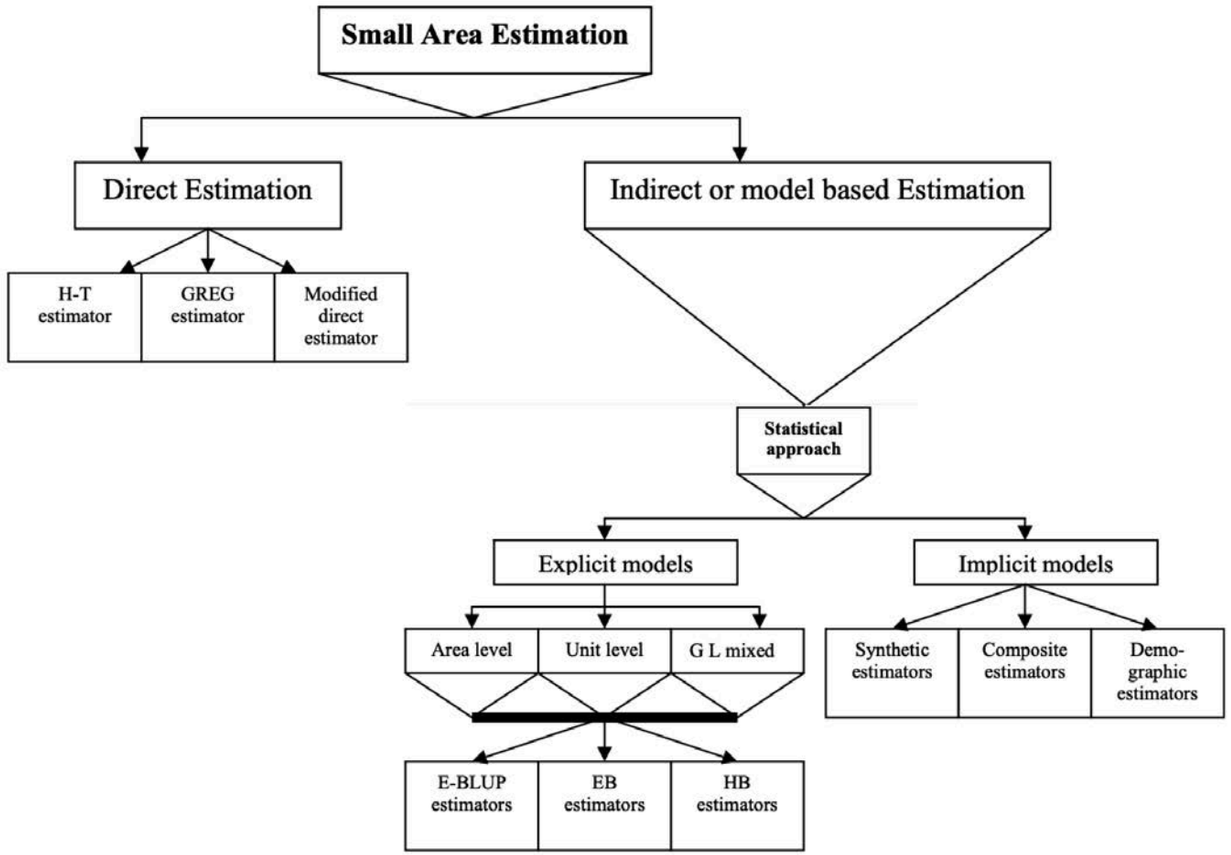


Source: Methodology of Modern Business Statistics (2014).

SAE methodologies in ECLAC

Metodologías SAE en la CEPAL

SAE methodologies / Metodologías SAE



Source: adaptation from Rahman (2008).

Two types of methods / Dos clases de métodos

SAE estimators used in ECLAC could be divided into two main types:

1. Estimators based on area models
2. Estimators based on unit models

The choice of the method that should be used in the estimation of the domains of interest is made depending on the level at which the auxiliary information is found (at the domain or aggregation level - at the household or person level)

Los estimadores de SAE se dividen en dos tipos principales:

1. Estimadores basados en modelos de área
2. Estimadores basados en modelos de unidad

La escogencia del método que se debe utilizar en la estimación de los dominios de interés se realiza dependiendo del nivel en el que se encuentre la información auxiliar (a nivel de dominio o agregación - a nivel de hogar o persona)

Area-level models

Modelos de áreas

Direct estimators / Estimadores directos

$$\hat{\theta}_{d}^{Dir} = \frac{\sum_{s_d} w_k y_k}{\sum_{s_d} w_k}$$

$$\widehat{AV}(\hat{\theta}_{d}^{Dir}) = \sum_{s} \sum_{s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}$$

- The sample size in each area is not planned in advance (since the sampling scheme follows a two-stage sampling).
- Any estimation of relative indicators (means and proportions) will have to make use of a ratio-type estimator: *random numerator and random denominator*.

When the sample size $n_d = \#(s_d)$ is not large enough, none of the above estimators will be precise neither consistent.

- El tamaño de muestra en cada área difícilmente es planificado de antemano (pues el esquema de muestreo es bietápico: UPM - Vivienda).
- Cualquier estimación de indicadores relativos (medias, proporciones) tendrá que usar un estimador de razón: *numerador y denominador aleatorios*.

Cuando el tamaño de muestra $n_d = \#(s_d)$ no es lo suficientemente grande, entonces ninguno de los anteriores estimadores será preciso, ni consistente.

Direct estimators / Estimadores directos

We will model the Direct Estimator so that in the areas where there is not enough sample, strength will be borrowed from the other areas.

Vamos a modelar el indicador directo para que en las áreas en las que no haya suficiente muestra se tome fuerza prestada de las otras áreas.

$$\hat{\theta}_{\mathrm{d}}^{\mathrm{Dir}} = \theta_d + \varepsilon_d \qquad \varepsilon_{\mathrm{d}} \sim \mathrm{N}(0, \mathrm{Var}(\hat{\theta}_{\mathrm{d}}^{\mathrm{Dir}}))$$
$$\theta_{\mathrm{d}} = \mathbf{x}'_d\boldsymbol{\beta} + u_d \qquad \mathrm{u_d} \sim \mathrm{N}(0, \sigma_{\mathrm{u}}^2)$$

$$\hat{\theta}_{\mathrm{d}}^{\mathrm{Dir}} = \mathbf{x}'_d\boldsymbol{\beta} + u_d + \varepsilon_d$$

Following (some) Bayes rule, we have that:

De la regla de Bayes, se tiene que:

$$\theta_{\mathrm{d}}|\hat{\theta}_{\mathrm{d}}^{\mathrm{Dir}} \sim \mathrm{N}\left(\theta_{\mathrm{d}}^{\mathrm{FH}}, \sigma^2_{\mathrm{d_{FH}}}\right)$$

$$\theta_{\mathrm{d}}^{\mathrm{FH}} = \mathrm{E}(\theta_{\mathrm{d}}|\hat{\theta}_{\mathrm{d}}^{\mathrm{Dir}}) = \gamma_d\theta_d^{Dir} + (1-\gamma_d)\mathbf{x}'_d\boldsymbol{\beta}$$
$$\sigma^2_{d_{FH}} = Var(\hat{\theta}_d^{Dir})\gamma_d$$

We like the Bayesian way / Nos gusta el enfoque bayesiano

Vol. 32, No. 1, pp. 97-103
Statistics Canada, Catalogue No. 12-001

Small Area Estimation Using Area Level Models and Estimated Sampling Variances

Yong You and Beatrice Chapman [1]

Abstract

In small area estimation, area level models such as the Fay–Herriot model (Fay and Herriot 1979) are widely used to obtain efficient model-based estimators for small areas. The sampling error variances are customarily assumed to be known in the model. In this paper we consider the situation where the sampling error variances are estimated individually by direct estimators. A full hierarchical Bayes (HB) model is constructed for the direct survey estimators and the sampling error variances estimators. The Gibbs sampling method is employed to obtain the small area HB estimators. The proposed HB approach automatically takes account of the extra uncertainty of estimating the sampling error variances, especially when the area-specific sample sizes are small. We compare the proposed HB model with the Fay–Herriot model through analysis of two survey data sets. Our results have shown that the proposed HB estimators perform quite well compared to the direct estimates. We also discussed the problem of priors on the variance components.

Key Words: Gibbs sampling; Hierarchical Bayes; Prior sensitivity; Sample size; Variance components.

We like the Bayesian way / Nos gusta el enfoque bayesiano

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 1-13
Statistics Canada, Catalogue No. 12-001-X

Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions

Benmei Liu, Partha Lahiri and Graham Kalton [1]

Abstract

The paper reports the results of a Monte Carlo simulation study that was conducted to compare the effectiveness of four different hierarchical Bayes small area models for producing state estimates of proportions based on data from stratified simple random samples from a fixed finite population. Two of the models adopted the commonly made assumptions that the survey weighted proportion for each sampled small area has a normal distribution and that the sampling variance of this proportion is known. One of these models used a linear linking model and the other used a logistic linking model. The other two models both employed logistic linking models and assumed that the sampling variance was unknown. One of these models assumed a normal distribution for the sampling model while the other assumed a beta distribution. The study found that for all four models the credible interval design-based coverage of the finite population state proportions deviated markedly from the 95 percent nominal level used in constructing the intervals.

Key Words: Weighted proportions; Hierarchical Bayes modeling; Beta distribution; credible interval.

Beta-logistic model for poverty / Modelo beta-logístico para la pobreza

Model 4: (beta-logistic model with unknown sampling variance)

Sampling model:

$$p_{iw} \mid P_i \overset{ind}{\sim} beta(a_i, b_i) \tag{3.8}$$

Linking model:

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2) \tag{3.9}$$

For both Model 3 and Model 4, the approximate variance function $\psi_i = [P_i(1-P_i)/n_i]deff_{iw}$ is used. The parameters a_i and b_i in Model 4 are given by:

$$a_i = P_i\left(\frac{n_i}{deff_{iw}} - 1\right), \text{ and } b_i = (1-P_i)\left(\frac{n_i}{deff_{iw}} - 1\right).$$

And, we like **STAN** / Nos gusta **STAN**

```
parameters {
  vector[p] beta;
  real<lower=0> sigma2_v;
  vector[N1] v;
}

transformed parameters{
  vector[N1] LP;
  real<lower=0> sigma_v;
  vector[N1] theta;
  LP = X * beta + v;
  sigma_v = sqrt(sigma2_v);
  for (i in 1:N1) {
     theta[i] = inv_logit(LP[i]);
  }
}
```

```
model {
  vector[N1] a;
  vector[N1] b;
  for (i in 1:N1) {
     a[i] = theta[i] * phi[i];
     b[i] = (1 - theta[i]) * phi[i];
  }
  // priors
   beta ~ normal(0, 100);
   sigma2_v ~ inv_gamma(0.0001, 0.0001);
  // likelihood
   y ~ beta(a, b);
   v ~ normal(0, sigma_v);
}

generated quantities {
  vector[N2] y_pred;
  vector[N2] thetapred;
  for (i in 1:N2) {
     y_pred[i] = normal_rng(Xs[i] * beta,
                            sigma_v);
     thetapred[i] = inv_logit(y_pred[i]);
  }
```

Unit-level models

Modelo de unidades

The GMR model / El modelo GMR

- ECLAC uses a unit-level model with adjustment to the complex sampling design for the estimation of 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).

- La CEPAL utiliza un modelo de nivel de unidad con ajuste al diseño complejo de muestreo para la estimación del ingreso promedio.
- Este modelo fue propuesto por Guadarrama, Molina y Rao (2018) e induce una aproximación del mejor predictor empírico (Pseudo-EBP) basado en el modelo con errores anidados (Molina y Rao, 2010).

The GMR model / El modelo GMR

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Small area estimation of general parameters under complex sampling designs☆

María Guadarrama [a,b,*], Isabel Molina [a], J.N.K. Rao [c]

[a] Department of Statistics, Universidad Carlos III de Madrid, Spain
[b] Luxembourg Institute of Socio-Economic Research (LISER), Luxembourg
[c] School of Mathematics and Statistics, Carleton University, Canada

ARTICLE INFO

Article history:
Received 9 February 2017
Received in revised form 23 November 2017
Accepted 25 November 2017
Available online 12 December 2017

Keywords:
Empirical best estimator
Informative sampling
Nested-error model
Poverty mapping
Unit level models

ABSTRACT

When the probabilities of selecting individuals (units) for the sample depend on the outcome values, the selection mechanism is said to be informative. Under informative selection, individuals with certain outcome values appear more often in the sample and, as a consequence, usual inference based on the actual sample without appropriate weighting might be strongly biased. An extension of the empirical best (EB) method for estimation of general non-linear parameters in small areas that handles informative selection by incorporating the sampling weights is proposed. Properties of this new method under complex sampling designs, including informative selection, are analyzed. Results confirm that the proposed weighted estimators significantly reduce the bias of unweighted EB estimators under informative sampling, and compare favorably under non-informative sampling. The proposed method is illustrated through an application to poverty mapping in a State from Mexico.

The nested-error model / El modelo de errores anidados

This method assumes that the transformed income variable $\mathrm{y}^{*}_{\mathrm{di}} = \log\,(\mathrm{y}_{\mathrm{di}} + \mathrm{c})$ follows the model:

Este método asume que la variable de ingresos transformados $\mathrm{y}^{*}_{\mathrm{di}} = \log\,(\mathrm{y}_{\mathrm{di}} + \mathrm{c})$ sigue el modelo:

$$\mathrm{y}^{*}_{\mathrm{di}} = \mathrm{x}'_{\mathrm{di}}\boldsymbol{\beta} + \mathrm{u}_{\mathrm{d}} + \mathrm{e}_{\mathrm{di}}; \qquad \mathrm{i} = 1, \ldots, \mathrm{N}_{\mathrm{d}},\ \mathrm{d} = 1, \ldots, \mathrm{D},$$

Where:

- $\boldsymbol{\beta}$ is the vector of regression coefficients,
- $\mathrm{u}_{\mathrm{d}} \overset{\mathrm{iid}}{\sim}\ \mathrm{N}\,(0,\ \sigma^{2}_{\mathrm{u}})$ is the area random effect, and
- $\mathrm{e}_{\mathrm{di}} \overset{\mathrm{iid}}{\sim}\ \mathrm{N}\,(0,\ \sigma^{2}_{\mathrm{e}})$ are the errors for individuals in the *d*-th area and are considered independent from the random effects.

En donde:

- $\boldsymbol{\beta}$ es el vector de coeficientes de regresión,
- $\mathrm{u}_{\mathrm{d}} \overset{\mathrm{iid}}{\sim}\ \mathrm{N}\,(0,\ \sigma^{2}_{\mathrm{u}})$ es el efecto de área, y
- $\mathrm{e}_{\mathrm{di}} \overset{\mathrm{iid}}{\sim}\ \mathrm{N}\,(0,\ \sigma^{2}_{\mathrm{e}})$ son los errores del modelo para los individuos del área *d*-ésima.

Weigthed estimation / Estimación ponderada

Since $\mathbf{y}_{\mathrm{d}}$ follows a normal distribution, the conditional distribution $\mathbf{y}_{\mathrm{dr}} \mid \mathbf{y}_{\mathrm{ds}}$ will also be a normal distribution parameterized as follows:

Dado que $\mathbf{y}_{\mathrm{d}}$ sigue una distribución normal, la distribución condicional $\mathbf{y}_{\mathrm{dr}} \mid \mathbf{y}_{\mathrm{ds}}$ también será normal parametrizada de la siguiente manera:

$$\mathbf{y}_{\mathrm{dr}} | \mathbf{y}_{\mathrm{ds}} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{dr|s}}, \mathbf{V}_{\mathrm{dr|s}}\right) \quad \text{con d} = 1, \ldots, \mathrm{D}$$

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}.

Para evitar el sesgo inducido al ignorar el diseño de muestreo en el modelo, los parámetros de la distribución anterior se pueden estimar constantemente incluyendo los pesos de muestreo w_{kd}.

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

$$\hat{\mathbf{V}}_{dr|s} = (\hat{\sigma}_e^2 + \hat{\sigma}_u^2(1 - \hat{\gamma}_d))\mathbf{1}_{N_d - n_d}\mathbf{1}_{N_d - n_d}^T$$

Monte Carlo estimation / Estimación de Monte Carlo

Since it is not possible to identify and link the units of the sample with those of the census, then the approach used is a Census-EB type, as follows:

Como no es posible identificar y vincular las unidades de la muestra con las del censo, el enfoque utilizado es de tipo Census-EB, de la siguiente manera:

$$\tilde{\theta}_{\mathrm{d}} = \frac{1}{\mathrm{N_d}} \sum_{\mathrm{i} \in \mathrm{r_d}} \mathrm{E}(\mathrm{I}_{\mathrm{y_{di}} < \mathrm{z}} | \mathbf{y}_{\mathrm{ds}})$$

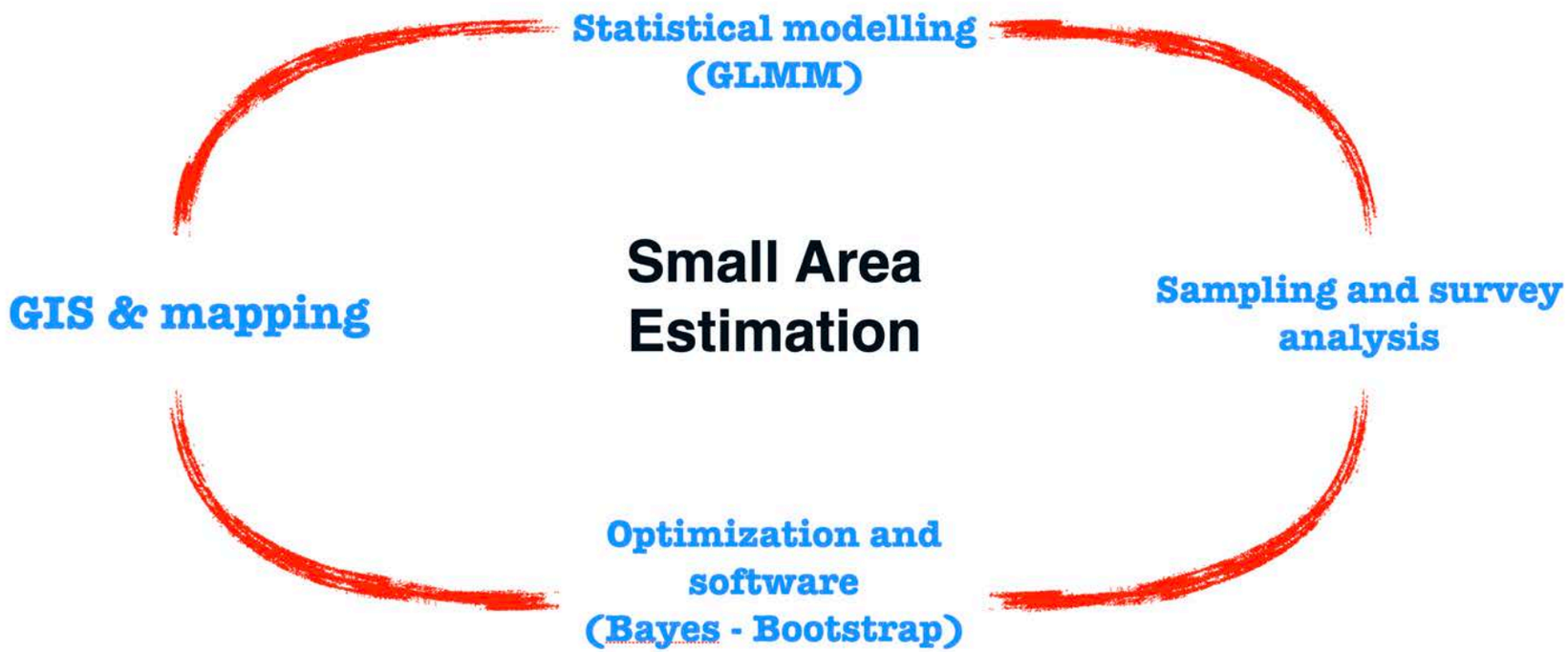
We consider a Monte Carlo simulation procedure to estimate the poverty indicators since often the expectation $\mathrm{E}(\mathrm{I}_{\mathrm{y_{di}} < \mathrm{z}} | \mathbf{y}_{\mathrm{ds}})$ that defines the best predictor cannot be calculated analytically.

Consideramos un procedimiento de simulación de Monte Carlo para estimar los indicadores de la pobreza, ya que a menudo la esperanza $\mathrm{E}(\mathrm{I}_{\mathrm{y_{di}} < \mathrm{z}} | \mathbf{y}_{\mathrm{ds}})$ que define el mejor predictor no se puede calcular analíticamente.

Some maps

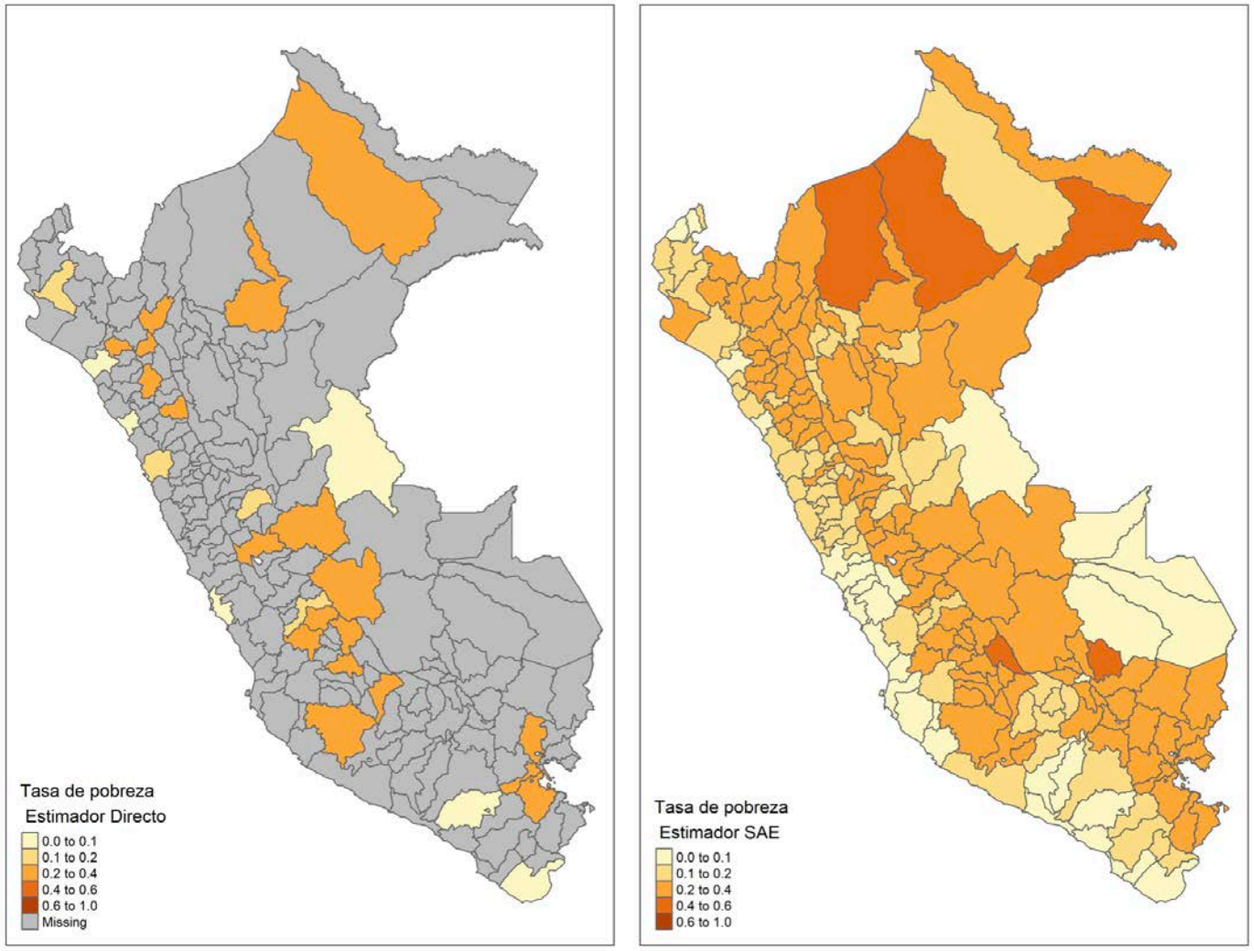
Algunos mapas

Processes in production of SAE / Procesos en la producción SAE

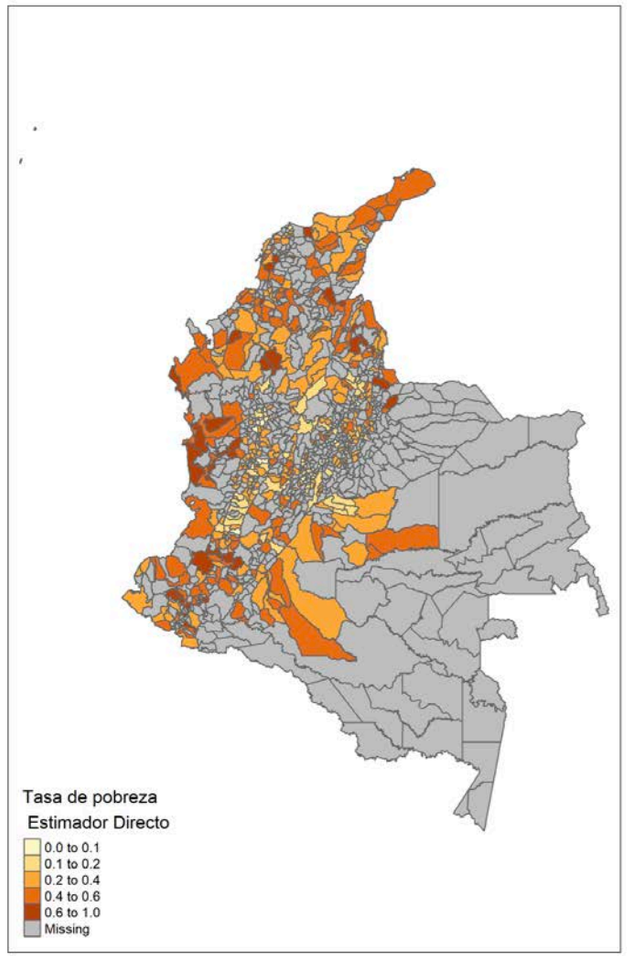


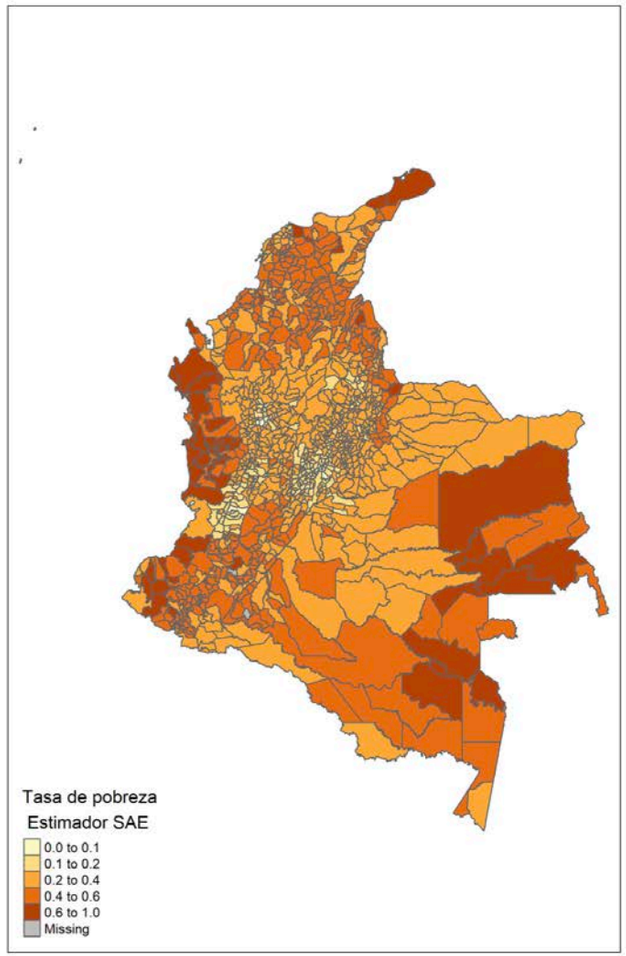
Source: adaptation from Kolenikov (2014).

Perú

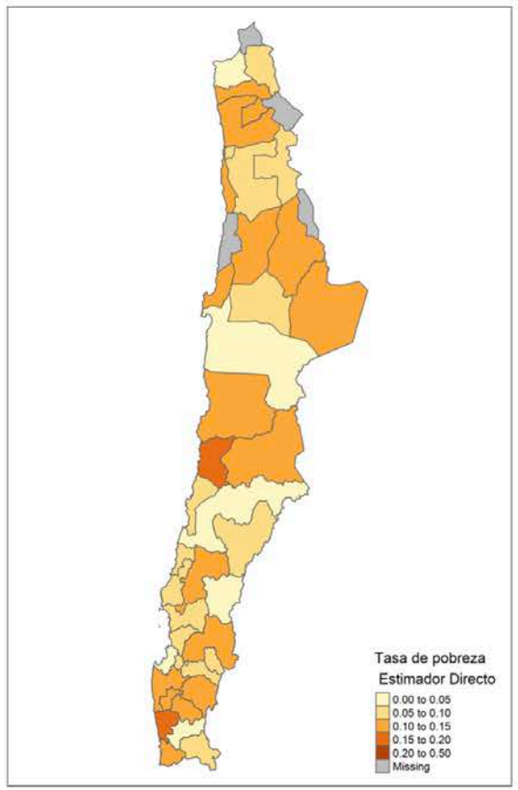

Tasa de pobreza
Estimador Directo
0.0 to 0.1
0.1 to 0.2
0.2 to 0.4
0.4 to 0.6
0.6 to 1.0
Missing
Tasa de pobreza
Estimador SAE
0.0 to 0.1
0.1 to 0.2
0.2 to 0.4
0.4 to 0.6
0.6 to 1.0

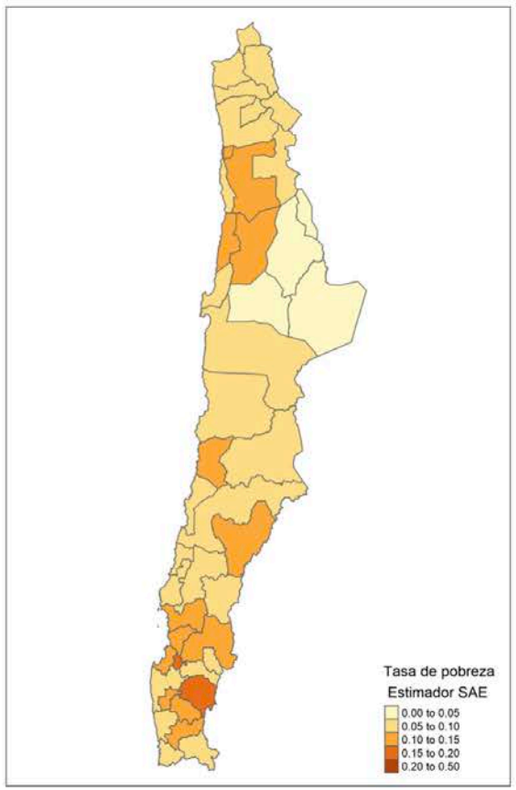
Colombia

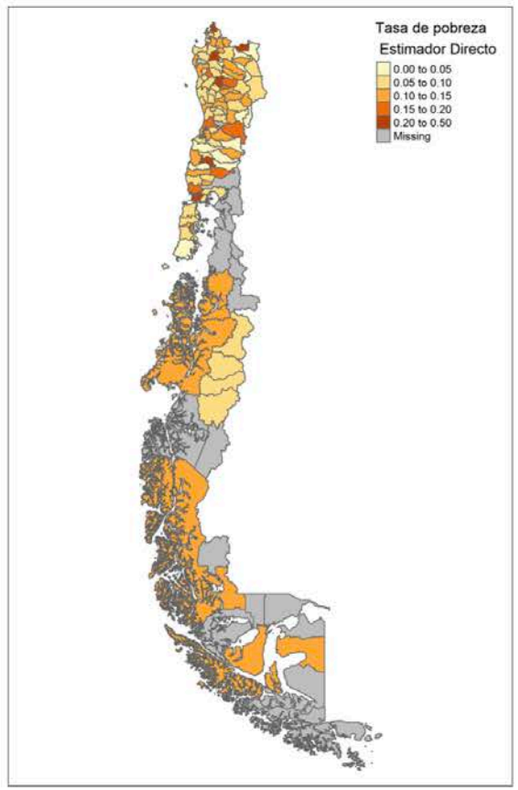

Tasa de pobreza
Estimador Directo
0.0 to 0.1
0.1 to 0.2
0.2 to 0.4
0.4 to 0.6
0.6 to 1.0
Missing

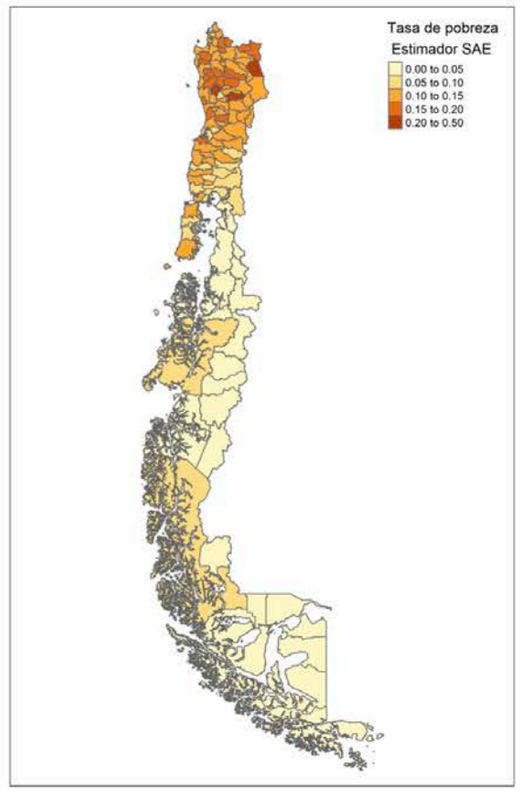

Tasa de pobreza
Estimador SAE
0.0 to 0.1
0.1 to 0.2
0.2 to 0.4
0.4 to 0.6
0.6 to 1.0
Missing

Chile


Tasa de pobreza
Estimador Directo
0.00 to 0.05
0.05 to 0.10
0.10 to 0.15
0.15 to 0.20
0.20 to 0.50
Missing


Tasa de pobreza
Estimador SAE
0.00 to 0.05
0.05 to 0.10
0.10 to 0.15
0.15 to 0.20
0.20 to 0.50

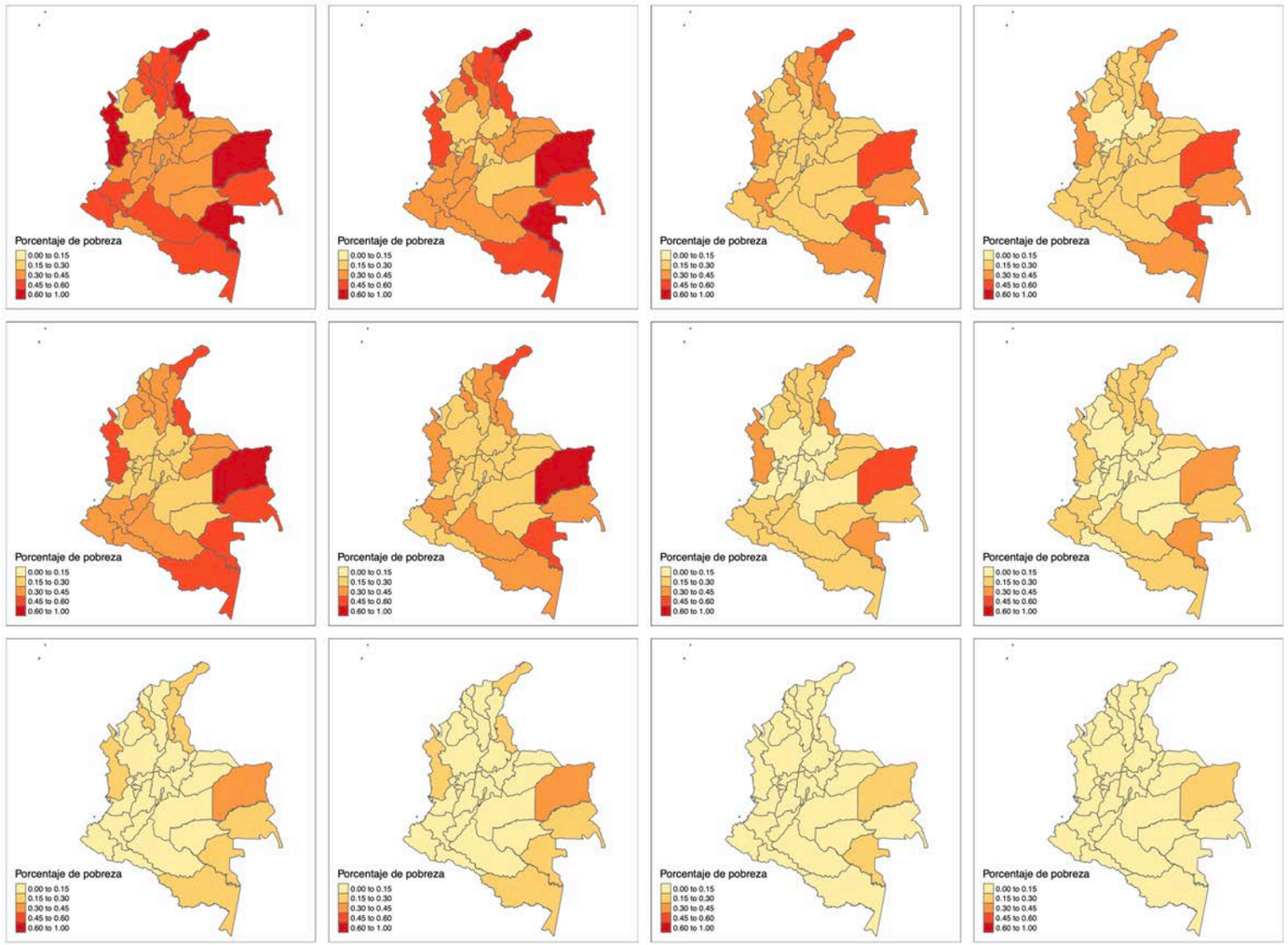

Tasa de pobreza
Estimador Directo
0.00 to 0.05
0.05 to 0.10
0.10 to 0.15
0.15 to 0.20
0.20 to 0.50
Missing


Tasa de pobreza
Estimador SAE
0.00 to 0.05
0.05 to 0.10
0.10 to 0.15
0.15 to 0.20
0.20 to 0.50

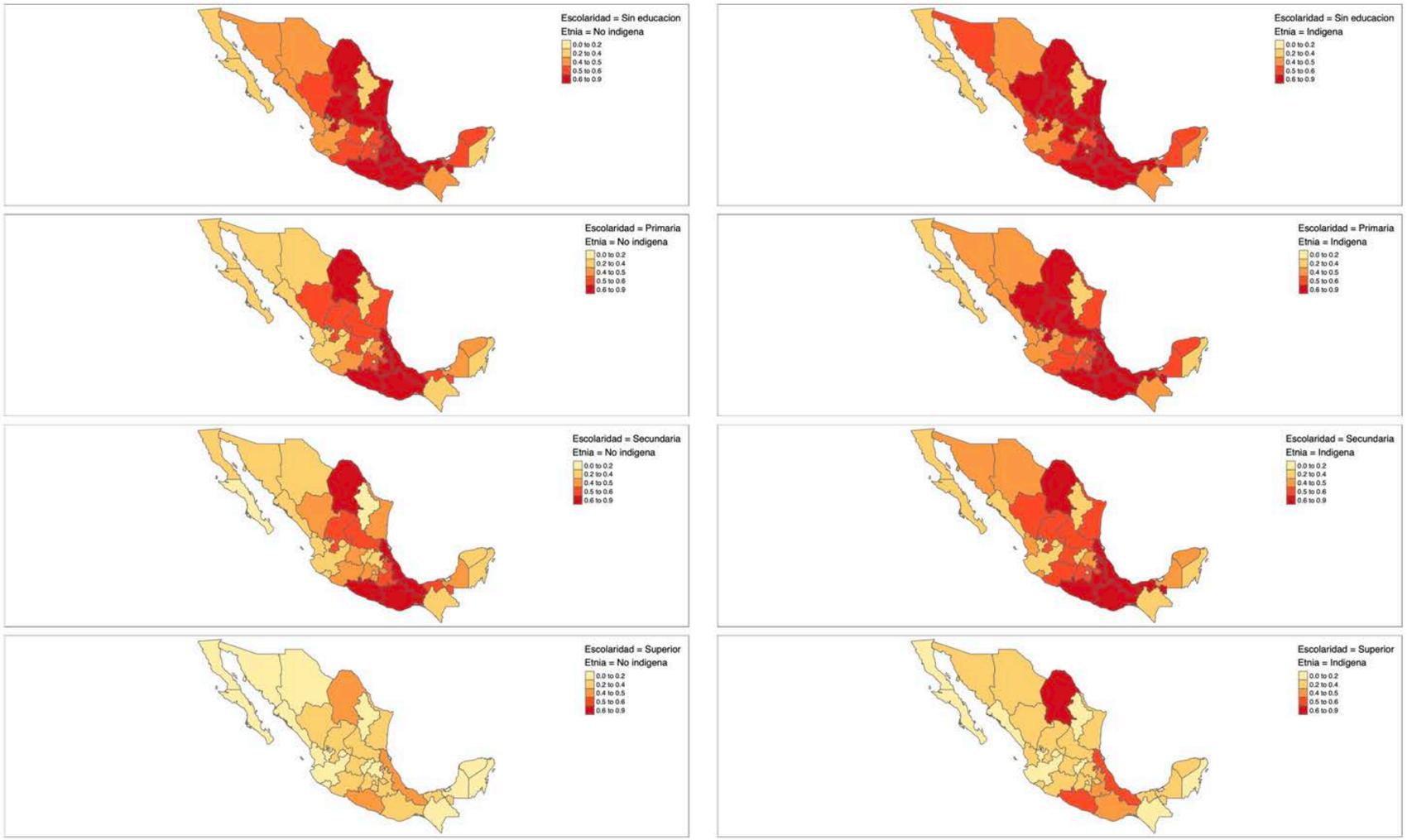
Leaving no one behind

No dejando a nadie detrás

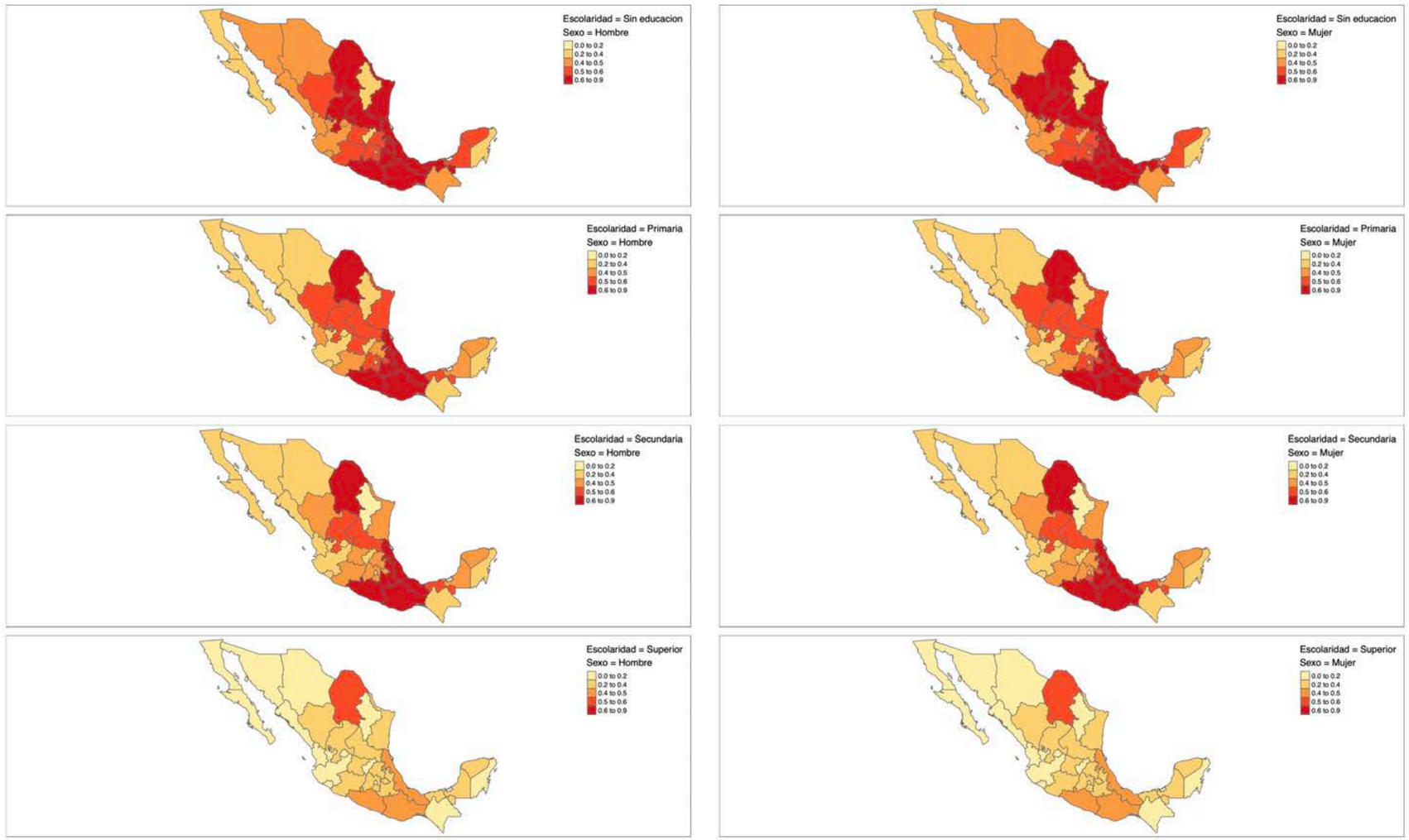
Colombia: age and education / edad y educación



México: ethnicity and education / étnia y educación



México: sex and education / sexo y educación



Thanks!

rolando.ocampo@un.org

xavier.mancero@un.org

andres.gutierrez@un.org

¡Gracias!