Multidimensional deprivation index using small area estimation methods:

An application for the adult population in Colombia

Andrés Gutiérrez[1], Alejandra Arias-Salazar[2], Stalyn Guerrero-Gómez [1], Natalia Rojas-Perilla[3] , Xavier Mancero[1], Hanwen Zhang[4]

Economic Commission for Latin America and the Caribbean[1]
Freie Universität Berlin[2]
United Arab Emirates University[3]
Universidad Autónoma de Chile[4]

Table of Contents

Motivation

- Poverty and multidimensional poverty leading topics in national and international agendas: "End poverty in all its forms everywhere" (UN General Assembly, 2015).

- Necessity of quality data on poverty in its different expressions.

- Disaggregated information: geographically and based on relevant characteristics (e.g. sex, age, ethnicity)

The ECLAC multidimensional deprivation index (MDI)

- ECLAC is developing a regionally comparable MDI for 18 Latin American countries.
- The index considers the person (and not the household) as unit of analysis.
- The MDI is based on the Alkire & Foster (2007) methodology.
- Complements the Global Multidimensional Poverty Index using regionally specific indicators and thresholds.

The ECLAC multidimensional deprivation index (MDI)

- ECLAC is developing a regionally comparable MDI for 18 Latin American countries.
- The index considers the person (and not the household) as unit of analysis.
- The MDI is based on the Alkire & Foster (2007) methodology.
- Complements the Global Multidimensional Poverty Index using regionally specific indicators and thresholds.
- It considers 5 dimensions and 8 indicators:

Dimension	**Indicator**	**Weight**	**Available in census**	**Target population**
Housing	Poor housing materials	1/10	Yes	Adults and seniors
	Overcrowding	1/10	Yes	Adults and seniors
Water and sanitation	Lack of drinking water	1/10	Yes	Adults and seniors
	Lack of sanitation	1/10	Yes	Adults and seniors
Energy and connectivity	Lack of internet service	1/10	Yes	Adults and seniors
	Lack of electricity	1/10	Yes	Adults and seniors
Education	Unfinished education	2/10	**No***	Adults and Seniors
Employment and social protection	No or insufficient pension	2/10	**No**	Seniors
	Unemployment or insufficient employment-related income	2/10	**No**	Adults

The ECLAC multidimensional deprivation index (MDI)

The MDI is computed as:

$$\text{MDI}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} I(q_{dj} > z) \tag{1}$$

The indicator function $I(\cdot)$ equals 1 when the condition $q_{dj} > z$ is met, i.e. $q_{dj} > 0.4$.

q_{dj} is a weighted quantity considering the $k = 8$ indicators that comprise the index.

$$q_{dj} = 0.1 \sum_{k=1}^{6} y_{dj}^k + 0.2 \sum_{k=7}^{8} y_{dj}^k$$

y_{dj}^k indicates if the person j in domain d has a deprivation or not in each of the $k = 8$ indicators.

Case study: Multidimensional deprivation index in Colombia

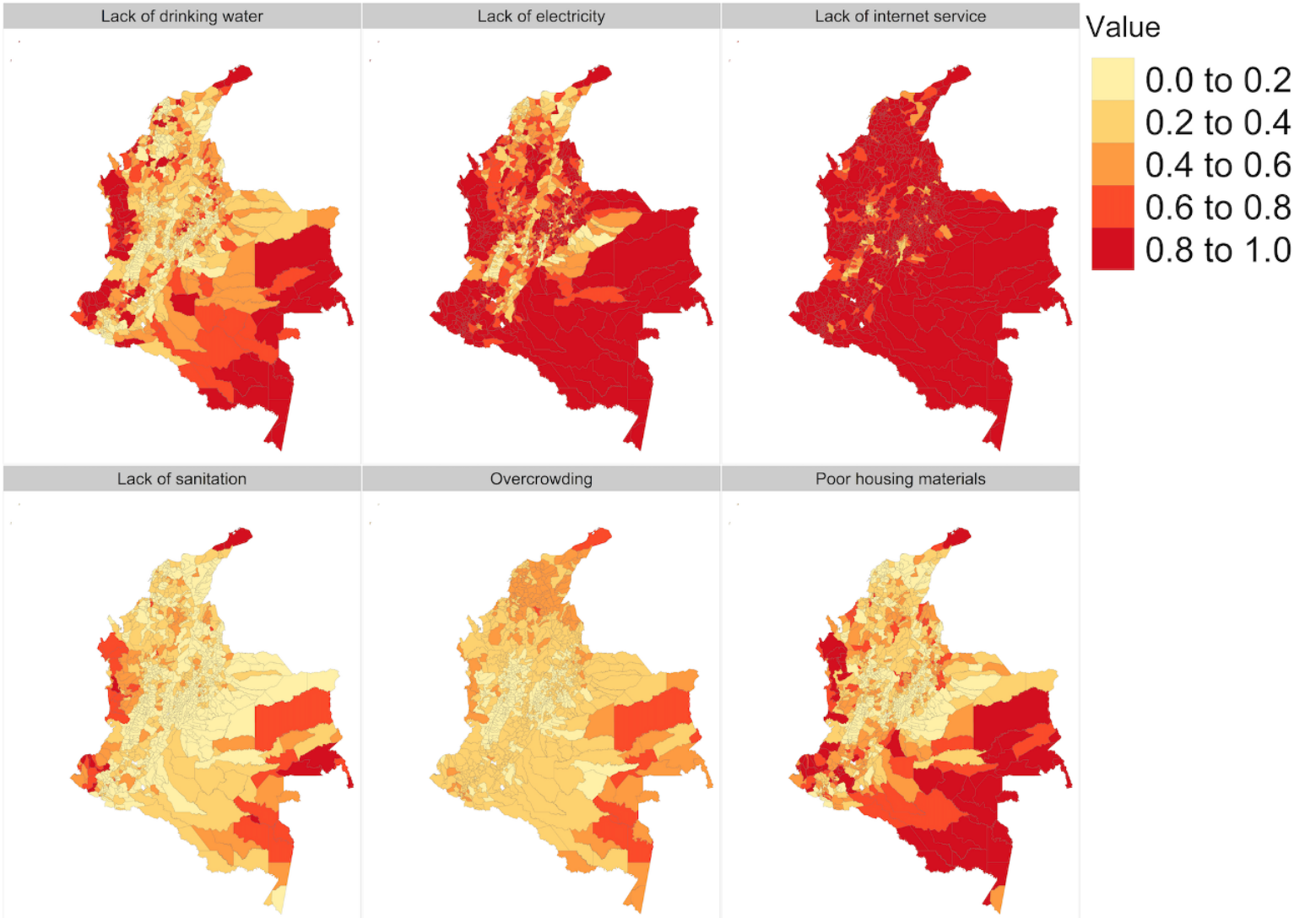


Figure: Direct estimates of MDI components for 6 available indicators at municipality level.

Case study: Multidimensional deprivation index in Colombia

- **Objective:** Producing reliable estimates of the multidimensional deprivation index and its components (indicators and dimensions) for the adult population of Colombia at first (departments) and second administrative division (municipalities).

Case study: Multidimensional deprivation index in Colombia

- **Objective:** Producing reliable estimates of the multidimensional deprivation index and its components (indicators and dimensions) for the adult population of Colombia at first (departments) and second administrative division (municipalities).

- **Challenge:** 2/8 indicators are missing so the final MDI cannot be computed.

Case study: Multidimensional deprivation index in Colombia

- **Objective:** Producing reliable estimates of the multidimensional deprivation index and its components (indicators and dimensions) for the adult population of Colombia at first (departments) and second administrative division (municipalities).
- **Challenge:** 2/8 indicators are missing so the final MDI cannot be computed.
- **Identified scenarios:**
  - **Only 1 missing indicator:** Using a unit-level Bernoulli logit mixed model.

Case study: Multidimensional deprivation index in Colombia

- **Objective:** Producing reliable estimates of the multidimensional deprivation index and its components (indicators and dimensions) for the adult population of Colombia at first (departments) and second administrative division (municipalities).
- **Challenge:** 2/8 indicators are missing so the final MDI cannot be computed.
- **Identified scenarios:**
  - **Only 1 missing indicator:** Using a unit-level Bernoulli logit mixed model.
  - **Two missing indicators:** Finding the the expected value of the linear combination of the them.

Case study: Multidimensional deprivation index in Colombia

- **Objective:** Producing reliable estimates of the multidimensional deprivation index and its components (indicators and dimensions) for the adult population of Colombia at first (departments) and second administrative division (municipalities).
- **Challenge:** 2/8 indicators are missing so the final MDI cannot be computed.
- **Identified scenarios:**
  - **Only 1 missing indicator:** Using a unit-level Bernoulli logit mixed model.
  - **Two missing indicators:** Finding the the expected value of the linear combination of the them.
  - **More than two missing indicators:** Using a Monte Carlo simulation approach to obtain "hard" estimates (0,1) for each individual in each missing indicator.

Case study: Multidimensional deprivation index in Colombia

- **Objective:** Producing reliable estimates of the multidimensional deprivation index and its components (indicators and dimensions) for the adult population of Colombia at first (departments) and second administrative division (municipalities).
- **Challenge:** 2/8 indicators are missing so the final MDI cannot be computed.
- **Identified scenarios:**
  - **Only 1 missing indicator:** Using a unit-level Bernoulli logit mixed model.
  - **Two missing indicators:** Finding the the expected value of the linear combination of the them.
  - **More than two missing indicators:** Using a Monte Carlo simulation approach to obtain "hard" estimates (0,1) for each individual in each missing indicator.
- **Assumptions:** There is no dependence when there are two or more missing indicators, and there are no causal relationships between the dependent and independent variables.

Unit-level Bernoulli logit mixed model

- The variable of interest is binary ($y_{dj} = 0$ or 1),

- The target estimation can be, the proportion per domain: $\bar{Y}_d = \pi_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$, with $j = 1, \ldots, N_d$, $d = 1, \ldots, D$, and π_{dj} the probability that a specific unit j in the domain d obtains the value 1.

Unit-level Bernoulli logit mixed model

- The variable of interest is binary ($y_{dj} = 0$ or 1),
- The target estimation can be, the proportion per domain: $\bar{Y}_d = \pi_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$, with $j = 1, \ldots, N_d$, $d = 1, \ldots, D$, and π_{dj} the probability that a specific unit j in the domain d obtains the value 1.
- The generalized linear mixed model (GLMM) with a logit link function is defined as:

$$\text{logit}(\pi_{dj}) = \log\left(\frac{\pi_{dj}}{1-\pi_{dj}}\right) = \eta_{dj} = x_{dj}^T \beta + u_d,$$

with β the vector of fixed-effects parameters, u_d the random area-specific effect for d and $u_d \sim N(0, \sigma_u^2)$.

- u_d are assumed independent with $y_{dj}|u_d \sim \text{Bernoulli}(\pi_{dj})$ with $E(y_{dj}|u_d) = \pi_{dj}$ and $Var(y_{dj}|u_d) = \sigma_{dj} = \pi_{dj}(1-\pi_{dj})$.

Point estimation - one missing indicator

The plug-in predictor of π_{dj} is:

$$\hat{\pi}_{dj}^{in} = \frac{\exp(\mathbf{x}_{dj}^T \hat{\beta} + \hat{u}_d)}{1 + \exp(\mathbf{x}_{dj}^T \hat{\beta} + \hat{u}_d)}, \tag{2}$$

which would allow obtaining the plug-in predictor of $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{in} = \frac{1}{N_d}\left(\sum_{j \in s_d} y_{dj} + \sum_{j \in r_d} \hat{\pi}_{dj}^{in}\right), \tag{3}$$

where s and r represent the in- and out-of-sample observations respectively.

Point estimation - one missing indicator

The plug-in predictor of π_{dj} is:

$$\hat{\pi}_{dj}^{in} = \frac{\exp(\mathbf{x}_{dj}^T \hat{\beta} + \hat{u}_d)}{1 + \exp(\mathbf{x}_{dj}^T \hat{\beta} + \hat{u}_d)}, \tag{2}$$

which would allow obtaining the plug-in predictor of $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{in} = \frac{1}{N_d}\left(\sum_{j \in s_d} y_{dj} + \sum_{j \in r_d} \hat{\pi}_{dj}^{in}\right), \tag{3}$$

where s and r represent the in- and out-of-sample observations respectively.

* In this case study, the two missing indicators Y_1 = *education* and Y_2= *employment* could be obtained with this procedure. **However, it does not allows to compute the final MDI.**

Point estimation - two missing indicators

Let's define the poverty status for each individual j, i.e. multidimensional poor (1) or not (0) based on a threshold δ:

$$Z_j = \begin{cases} 1 & \text{if } X_j \geq \delta, \\ 0 & \text{if } X_j < \delta \end{cases}$$

Point estimation - two missing indicators

Let's define the poverty status for each individual j, i.e. multidimensional poor (1) or not (0) based on a threshold δ:

$$Z_j = \begin{cases} 1 & \text{if } X_j \geq \delta, \\ 0 & \text{if } X_j < \delta \end{cases}$$

where

$$X_j = \underbrace{\alpha\left(Y_{1j} + Y_{2j}\right)}_{W_j \text{ unknown}} + \underbrace{k_j.}_{\text{known}}$$

Point estimation - two missing indicators

Let's define the poverty status for each individual j, i.e. multidimensional poor (1) or not (0) based on a threshold δ:

$$Z_j = \begin{cases} 1 & \text{if } X_j \geq \delta, \\ 0 & \text{if } X_j < \delta \end{cases}$$

where

$$X_j = \underbrace{\alpha\left(Y_{1j} + Y_{2j}\right)}_{W_j \text{ unknown}} + \underbrace{k_j.}_{\text{known}}$$

By finding the mass probability function of W_j, the domain proportion is:

$$\bar{Z}_d = \frac{1}{N_d} \sum_{j \in U_d} Z_j.$$

The target predictor is then the estimator of $\bar{Z}_d$, which is given by the expectation of $\bar{Z}_d$:

$$\widehat{\text{MDI}}_d = \hat{\bar{Z}}_d = E\left(\bar{Z}_d\right)$$

Details

Point estimation - more than two missing indicators

A Monte Carlo simulation approach can be implemented when more than three indicators are missing:

1. Use the sample data to fit a unit-level Bernoulli logit mixed model for each indicator and estimate $\hat{\beta}^k$, $\hat{u}_d^k$, and finally $\hat{\pi}_{dj}^{in,k}$.

2. for $l = 1, \ldots, L$ Monte Carlo simulations:

   - For each individual in the census, predict the probability of obtaining the value 1. i.e. $\hat{\pi}_{dj}^{in,k,(l)} \ \forall \ j \in U_d$.
   - Obtain Monte Carlo "hard" estimates $\tilde{y}_{dj}^{k,(l)}$ with $y_{dj}^k \sim \text{Bernoulli}(\hat{\pi}_{dj}^{in,k})$.
   - Compute the $\text{MDI}^{(l)}$, with 6/8 indicators already available in the census and the new indicators $\tilde{y}_{dj}^{k,(l)}$.

3. The final point estimate in each small area d is computed by taking the mean over each L simulation:

$$\widehat{\text{MDI}}_d = \frac{1}{L} \sum_{l=1}^{L} \text{MDI}_d^{(l)}$$

Uncertainty estimation: Following González-Manteiga et al (2007)

- Suitable when using a logistic mixed model for estimating any characteristic of the population.
- Small area robust wild bootstrap (SAWB) is a re-sampling procedure for the MSE estimation of an empirical predictor.

Uncertainty estimation: Following González-Manteiga et al (2007)

- Suitable when using a logistic mixed model for estimating any characteristic of the population.
- Small area robust wild bootstrap (SAWB) is a re-sampling procedure for the MSE estimation of an empirical predictor.

1. For $k = 1, 2$
2. For $b = 1, ..., B$
   - Using the already estimated $\hat{\beta}^k, \hat{\sigma}_u^{2,k}$, generate $u_d^{*,k}$ and simulate a bootstrap superpopulation $y_{dj}^{k,(b)} \sim \text{Bernoulli}(\pi_{dj}^{*,k})$ with $\pi_{dj}^{*,k} = \frac{\exp(\mathbf{x}_{dj}^T \hat{\beta} + u_d^*)}{1+\exp(\mathbf{x}_{dj}^T \hat{\beta} + u_d^*)}$
   - Calculate the $\text{MDI}_d^{(b)}$
   - Extract the bootstrap sample and obtain the estimated MDI following the point estimate - Monte Carlo approach $\widehat{\text{MDI}}_d^{(b)}$
3. $\text{MSE}\left[\widehat{\text{MDI}}_d\right] = 1/B \sum_{b=1}^{B} \left[\text{MDI}_d^{(b)} - \widehat{\text{MDI}}_d^{(b)}\right]^2$

Data sources and covariates

Data sources:

- National population and housing census, Colombia 2018.
- Employment and living conditions survey (Large Integrated Household Survey), Colombia 2018.
- Satellite imagery, Colombia 2016 from Earth Engine Data Catalog.

Data sources and covariates

Data sources:

- National population and housing census, Colombia 2018.
- Employment and living conditions survey (Large Integrated Household Survey), Colombia 2018.
- Satellite imagery, Colombia 2016 from Earth Engine Data Catalog.

Covariates:

Census and survey data

- Group of age
- Area (urban/rural)
- Department
- Sex
- 6 y_{dj}^{k} (k = water, energy, housing material, sanitation, overcrowding, internet)

Satellite imagery

- Intensity of nighttime lights
- Distance to cultivated areas (crops)
- Urbanization (human settlements)

Area level

- Unemployment rate

Results: Estimated MDI indicators

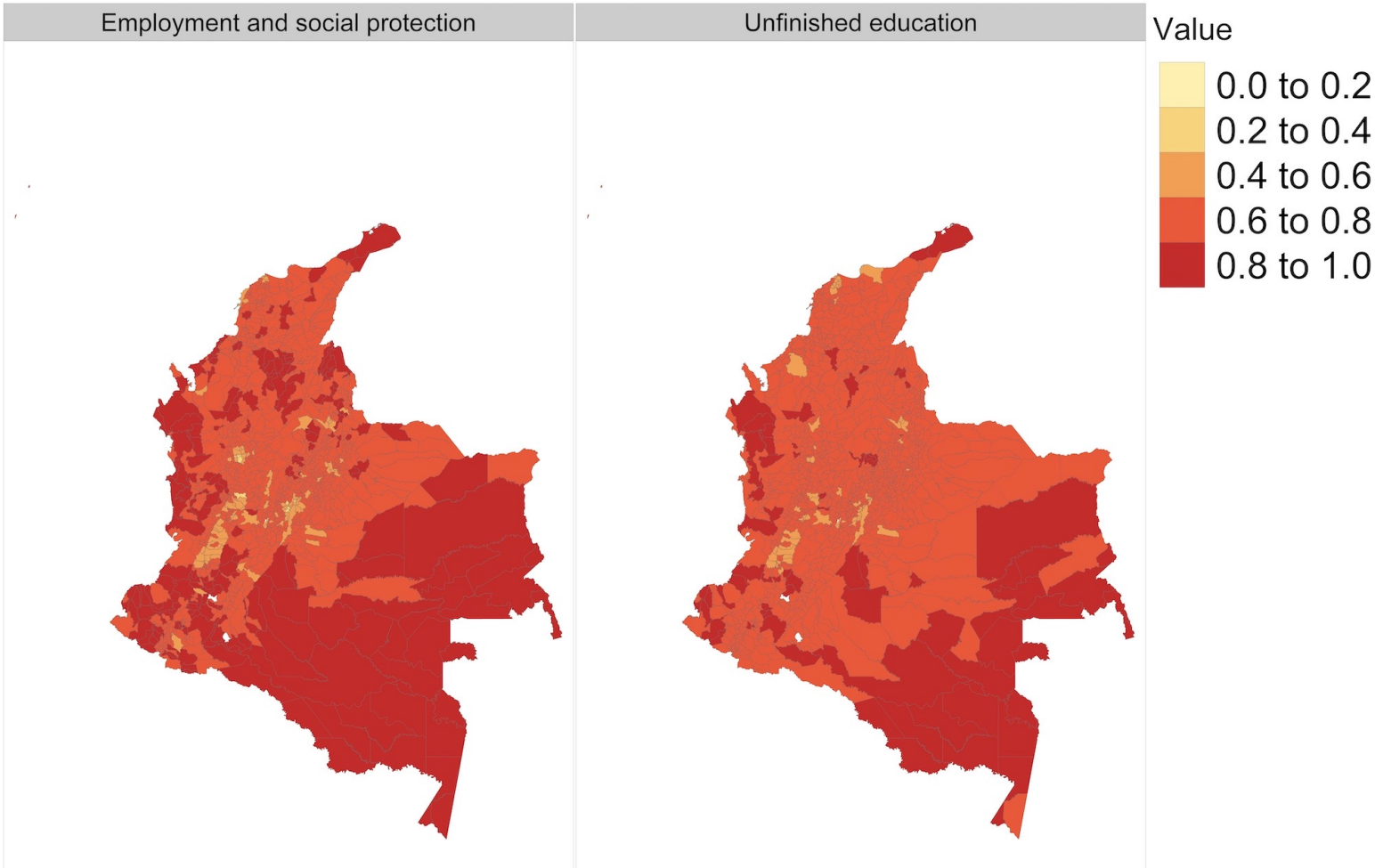
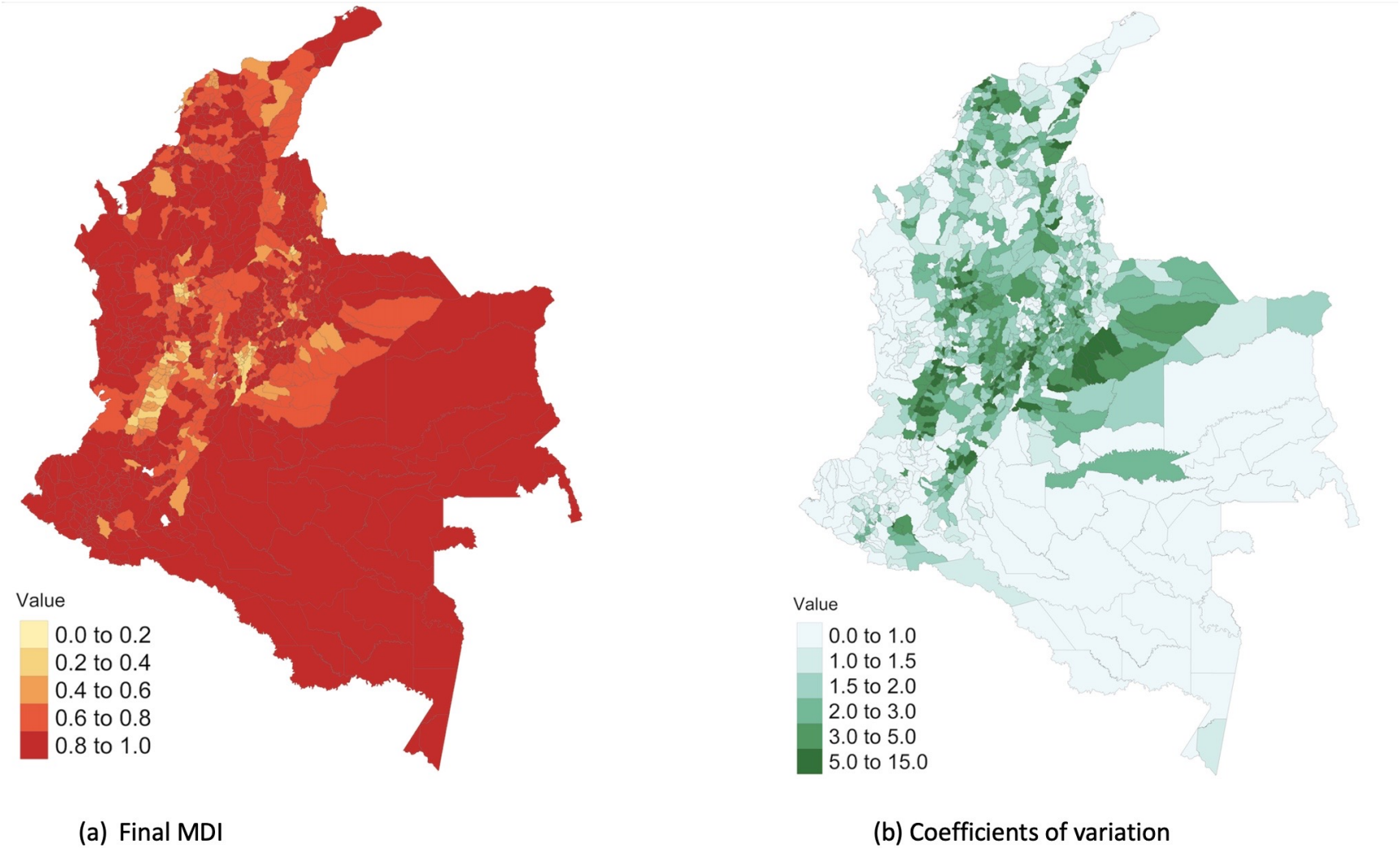


Figure: Model-based estimates for the indicators employment and social protection and unfinished education at municipality level.

Results: Estimated MDI



(a) Final MDI

(b) Coefficients of variation

Figure: Model-based estimates for final MDI (a) and coefficients of variation (b) for municipalities in Colombia.

Conclusion and further research

Unit-level Bernoulli logit mixed models can be used when one, two or more indicators of the MDI are not available in the census data.

Conclusion and further research

Unit-level Bernoulli logit mixed models can be used when one, two or more indicators of the MDI are not available in the census data.

Further research:

- Modeling (co-)relations: How to take into account dependencies and covariances between indicators? and dimensions?

Conclusion and further research

Unit-level Bernoulli logit mixed models can be used when one, two or more indicators of the MDI are not available in the census data.

Further research:

- Modeling (co-)relations: How to take into account dependencies and covariances between indicators? and dimensions?

- Time-gap between census and survey: Modeling all indicators to "update" them yearly.

Conclusion and further research

Unit-level Bernoulli logit mixed models can be used when one, two or more indicators of the MDI are not available in the census data.

Further research:

- Modeling (co-)relations: How to take into account dependencies and covariances between indicators? and dimensions?
- Time-gap between census and survey: Modeling all indicators to "update" them yearly.
- MSE estimation: Evaluate the performance of the point and variance estimators.

Conclusion and further research

Unit-level Bernoulli logit mixed models can be used when one, two or more indicators of the MDI are not available in the census data.

Further research:

- Modeling (co-)relations: How to take into account dependencies and covariances between indicators? and dimensions?
- Time-gap between census and survey: Modeling all indicators to "update" them yearly.
- MSE estimation: Evaluate the performance of the point and variance estimators.
- Benchmark: Possible alternatives and Bayesian approximations.

Conclusion and further research

Unit-level Bernoulli logit mixed models can be used when one, two or more indicators of the MDI are not available in the census data.

Further research:

- Modeling (co-)relations: How to take into account dependencies and covariances between indicators? and dimensions?
- Time-gap between census and survey: Modeling all indicators to "update" them yearly.
- MSE estimation: Evaluate the performance of the point and variance estimators.
- Benchmark: Possible alternatives and Bayesian approximations.
- Generalization: Compute all the indicators and dimensions of the MDI with SAE methods.

Main references

- Chandra H, Kumar S, Aditya K. (2018). Small area estimation of proportions with different levels of auxiliary data. *Biometrical Journal*. 60(2): 395-415.

- González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics data analysis*, 51(5): 2720–2733.

- Hobza, T., Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 32(3), 661.

- Jiang, J., Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53(2), 217-243.

- Morales, D., Esteban, M. D., Pérez, A., Hobza, T. (2021). *A course on small area estimation and mixed models. Methods, theory and applications in R.*

Thank you for your attention

Alejandra Arias-Salazar (alejandra.arias@fu-berlin.de)

The authors gratefully acknowledge support by UAEU Start-up Research Grant from the United Arab Emirates University

MDI predictor: two missing indicators

Point

Finding $W_j = \alpha\,(Y_{1j} + Y_{2j})$
The probability mass function of W_j is defined by:

$$f\,(W_j) = \begin{cases} (1-p_1)\,(1-p2) & \text{if } w_j = 0, \\ p_2\,(1-p_1) + p_1\,(1-p_2) & \text{if } w_j = \alpha, \\ p_1 p_2 & \text{if } w_j = 2\alpha, \\ 0 & \text{otherwise.} \end{cases}$$

1. Use the sample data to fit a unit-level Binomial logit mixed model for each indicator (Y_{1j}, Y_{2j}) and estimate $\hat{\beta}^{Y_1}$, $\hat{\beta}^{Y_2}$, $\hat{u}_d^{Y_1}$ and $\hat{u}_d^{Y_2}$.
2. For each individual j in the domain d of the census, predict the probability of obtaining the value 1. i.e. $\hat{\pi}_{dj}^{in,Y_1}$ and $\hat{\pi}_{dj}^{in,Y_2}$.
3. Define δ, as a fixed threshold that defines the number of deprivations.
4. $X_j = \alpha\,(Y_{1j} + Y_{2j}) + k_j$ a general linear combination of the two discrete random variables, with k_j a known parameters $\forall j \in U$
5. Using the additive condition proved above, calculate the expectation as the target estimator for each individual

$$\text{MDI}_d = E\left(\bar{Z}_d\right).$$