Multiple deprivation index using small area estimation methods: an application for the adult population in Colombia

Andrés Gutiérrez-Rojas[1], Alejandra Arias-Salazar[2], Diego Lemus-Polanía[1], Natalia Rojas-Perilla[3] and Xavier Mancero[1]

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

Abstract

Based on the new “Multiple deprivation index" (MDI) for Latin American countries produced by the Economic Commission for Latin America and the Caribbean, this paper shows the case study of Colombia to obtain estimates of this multidimensional index in small domains. This country has a recent population census (2018) providing most of the information required to compute the MDI at municipality level. However, two of the indicators required to compute the index are not available in the census, namely: “Employment-related income" and “Lack of social security". Therefore, small area estimation (SAE) methods are implemented to obtain unit-level estimates for those specific indicators. Since two different SAE procedures are carried out, a parametric bootstrap algorithm is proposed to provide respective uncertainty measures.

Keywords: Small area estimation; Multidimensional poverty; Composite indicators; Empirical best predictor; Generalized linear mixed model

1 Introduction

Poverty is, and have been, one of the leading topics in national and international agendas for decades. A recent example is the first goal of the 2030 agenda for Sustainable Development (SDG): “End poverty in all its forms everywhere", as well as its 1.2.2 indicator which measures “proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions" (**?**). In this sense, obtaining quality data on poverty in its different expressions takes relevance as well as produce desegregated information (e.g. geographically or considering characteristics of the population) to identify and develop strategies to eradicate it.

Traditionally national and international organisms require the uni-dimensional poverty measure based on income and/or expenditure. The well known Foster-Greer-Thorbecke (FGT) poverty measures (**?**), provide information about the head count ratio, poverty gap index and the severity of poverty. However, several studies had revealed that poverty is a complex phenomenon that should be analyzed considering a group of factors and not only one, usually monetary (Lemmi, A. A., and Betti, G. (Eds.), 2006; Belhadj, B., Limam, M., 2012). Alkire Foster (2007), proposed a methodology based on previous research to measuring poverty, taking into account not achievements, but deprivations of individuals and households. Recently, this methodology has been widely used by many countries because there is no restriction on the number of dimensions and indicators, and furthermore, each nation can define the composite of their index. In the case of Latin America, it has been clarified that considering not only the monetary approach but also the multidimensional one, has helped to identify that poverty is a structural phenomenon in the region (**?**). Due to the complexity of the problem, alternatives have been sought in the region to have a standardized methodology that allows not only to study the phenomenon in an adequate way, but also to be able to compare between countries. In the early eighties the Economic Commission for Latin America and the Caribbean (ECLAC), presented the unsatisfied basic needs (UBNS) index as a method to represent the non-monetary capacities of the persons and households, but at the same time highly correlated with income (**?**)

More recently, in 2022 the multidimensional deprivation index (Cepal) has the purpose of ..... in comparison with the UBN.

As aforementioned, desegregation is also required. The first alternative are administrative records and population and housing censuses because they collects information at individual level. In Latin America countries lack of administrative systems and quality usually does not fulfil necessities to produce poverty information. On the other hand, census are produced every 5- 10 years, income information is not available and other relevant information is also not there. The other alternative, is to use survey data having the limitation of small sample sizes.

Small area estimation (SAE) methods have the goal of producing reliable estimates in smaller domains, i.e. with adequate precision, by combining two or more sources of information. Most of these methodologies, usually classified as unit-

or area-level models, provide efficiency gains if the correlation between existing auxiliary information and the survey data is sufficient (**????**). Since the idea is to estimate the MDI for small domains, but also without losing the information of each indicator and dimension, unit level models are chosen.

The objective of this paper is to tackle the multidimensional problem of poverty estimation in Colombia by providing an estimation of the multidimensional deprivation index (MDI). To provide estimates of the MDI in small domains in Colombia, model-based estimation methods are applied and combined. Those methods are based on SAE methodologies that are data-driven and take the theoretical information from the survey into account. For this, the pseudo empirical best predictor (EBP) is applied in parallel to an empirical plug-in methodology to provide different indicators dimensions and the global MPIs for small domains. As uncertainty measure, the Mean Squared Errors are estimated via parametric bootstrap. The structure of the paper is as follows: Data sources and relevant characteristics of this country are mentioned in Section 2. SAE methods included in this study are briefly explain in Section 3, as well as the procedure to obtain MSE estimates for the corresponding point estimates. In section 4 we present a simulation exercise to validate our proposal. Section 5 is dedicated to show main results. Conclusions and further research are presented in Section 6.

2 Case study: multidimensional deprivation index for adult population in Colombia

Colombia is chosen as the first country in Latin America where the MDI for small areas is implemented. Unlike other countries in the region, Colombia counts on a recent population and housing census with most all the variables required to produce the MDI.

2.1 Data sources

2.1.1 Censo Nacional de Población y Vivienda (CNPV) 2018

Implementation of the SAE requires at least two sources of data. For this case study the first one is the population and housing census Censo Nacional de Población y Vivienda (CNPV) 2018 of Colombia. It is conducted by the National Statistical Office of Colombia (DANE: Departamento Administrativo Nacional de Estadística). Although it is planned to be every 10 years, the last census was carried out three years later as planned due to administrative and economic reasons. For first time, the information was collected via electronic census and the traditional face-to-face interview. Data collection phase was conducted during the 2018 in 10 months (**?**).

2.1.2 Gran Encuesta Integrada de Hogares (GEIH) 2018

Comprehensive survey of households (in spanish: Gran Encuesta Integrada de Hogares - GEIH) collects information on employment, income and expenses of individuals and households. This survey is part of the Household Survey Data Bank (BADEHOG), a repository of household surveys from 18 Latin American countries maintained by the ECLAC Statistics Division. The GEIH provides results yearly, that are representative for the national, national urban, national rural, regional, departmental, and for the capitals of the country's departments.

2.2 MDI in Colombia

The multidimensional deprivation index proposed by ECLAC is consistent with an individual well-being approach. It is focus on adult and senior population, considering a gender perspective.

The MDI, similarly as the Global Multdimensional Poverty Index (**?**), is defined as the product of two measures: the incidence of multidimensional poverty and the intensity of poverty:

$$\text{MDI} = \text{H} \cdot \text{A}$$

where H is the headcount ratio or incidence of multidimentional poverty and A is the intensity of poverty. The Headcount ratio is defined as:

$$\text{H} = \frac{q}{n}$$

with q the number of people who are multidimensionally poor and n the total population. The intensity of poverty A is defined as:

$$\text{A} = \frac{\sum_1^q s_i}{q}$$

with s_i the deprivation score that the $i_t h$ multidimentionally poor person experiences.

For this case study, the indicators and dimensions considered are explained in table 1. Each dimension has same weight (1/4) and one individual is defined as poor if he or she obtains 40% of total deprivations. Notice that this index is based on individual information, i.e. quantity of deprivations by individual. In case one member of the household is defined multidimensionally poor, all members will received the same status.

As shown in Table 1. two of nine indicators are not available in the census and SAE methods are implemented as will be explain in next Section.

Table 1: Dimensions, indicators of the MDI and data availability in the census

Dimension	Indicator	Weight	Available in census	Target population
Housing	Poor housing materials	1/8	Yes	Adults and seniors
	Overcrowding	1/8	Yes	Adults and seniors
Basic services	Lack of information communication technologies (ICTs)	1/16	Yes	Adults and seniors
	Lack of drinking water	1/16	Yes	Adults and seniors
	Lack of sanitation	1/16	Yes	Adults and seniors
	Lack of energy (electricity)	1/16	Yes	Adults and seniors
Education	Unfinished education	1/4	Yes	Adults
	Illiteracy	1/4	Yes	Seniors
Employment, social security and income	No pension	1/4	Yes	Seniors
	Insufficient employment-related income	1/8	**No**	Adults
	Unemployment, precarious employment and out of the labor force due to housework	1/8	**No**	Adults

3 Methodology

The goal is to provide indicators, dimensions and the final MPI for small domains. To do that, it is necessary to use unit level models (persons).

Most of the indicators are binary variables (deprivation: yes/no) and to obtain them, the EPP will be applied. However, since income information is also part of these indicators, a Pseudo EBP is also implemented.

3.1 SAE Methods

A definir:

- Porqué decidimos usar el EBP de Guadarrama con transformación log-shift? Cómo justificamos que esta es la mejor alternativa? Se podría justificar comparando “clásico" EBP con EBP de Guadarrama, y/o distintas transformaciones
- Lo mismo para el EPP, porqué el Plug-in en lugar del EBP para variables binarias?

In order to performing a complete poverty mapping and obtaining a local picture of poverty in a region, the use of an adequate estimation method becomes necessary. Unfortunately national surveys often are not suitable to give reliable statistical information at local levels to cover all regions within a country due to the high costs. Small area procedures are estimation procedures for parameters under very small sample sizes. For sufficiently large sample sizes, traditional estimators, such as the mean estimators, produce very convincing results, but when

applied to small sample sizes, such estimators often only have very limited reliability. This is often the case if the subject requires the data to be split into many small categories, e.g. municipalities. Even large surveys in a country will contain administrative units from which only very few or even no households have been sampled. For such domains, in practice, the sampling error is often huge ([Rao03]). Luckily there is an alternative to the classical estimators, the model based methods, which have been developed further in recent years. These methods use model assumptions to reduce the sampling error. The small area estimation methods, in particular, based on generalized linear mixed models are part of this class of methods. Their basic principle is to improve the estimation by extending the original too small sample. Explanatory a survey data set might be extended by the census of the whole state, even though the variable of interest is missing in the census data. Under strict assumptions on the sample and its distribution, such model based procedures result in much better estimates than classical procedures.

3.1.1 Pseudo-EBP Guadarrama (2018)

The estimation of non-linear indicators at highly disaggregated levels is commonly carried out by the empirical best (EB) predictor introduced by Molina and Rao (2010). This method assumes a non-informative selection mechanism of individuals and can suffer from biased estimators. To address the bias problem due to a non-represntative sample under this context, the model-based pseudo empirical best (PEB) estimator, proposed by Guadarrama (2018), incorporates weights from an informative sampling design. This is a weighed EB method, which is based on the weighted sample mean and uses the inverses of the inclusion probabilities as weights.

NOTATION

Let U denote a finite population of size N partitioned into D areas or domains (representing the small areas) $U_1, U_2, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, where $i = 1, \ldots, D$ refers to the ith area. Let y_{ij} be the target variable defined for the jth individual belonging to the ith area, with $j = 1, \ldots, N_i$.

The FGT index of type α for a region $i = 1, \ldots, D$ and a fixed threshold t is given by

$$F_i(\alpha, t) = \frac{1}{N_i} \sum_{j=1}^{N_i} F_{ij}(\alpha, t), \quad \alpha = 0, 1, 2 \tag{1}$$

where

$$F_{ij}(\alpha, t) = \left(\frac{t - y_{ij}}{t} \right)^{\alpha} I(y_{ij} \leq t)$$

with $I(A)$ an indicator function which returns 1 if A is a true expression and 0 otherwise. From this definition the following poverty measures are derived:

Denote by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ the design matrix containing p explanatory variables and define by s as the set of sample units, with s_i the in-sample units in area i and by r be the set of non-sampled units, with r_i the out-of-sample units

in area i. Let n_i denote the sample size in area i with $n = \sum_{i=1}^{D} n_i$. Hence, we define by $\mathbf{y}_i$ a vector with population elements of the target outcome for area i partitioned as $\mathbf{y}_i^T = \left(\mathbf{y}_{is}^T, \mathbf{y}_{ir}^T\right)$, where $\mathbf{y}_{is}$ and $\mathbf{y}_{ir}$ denote the sample elements s and the out-of-sample elements r in area i respectively.

Consider a generic index or a target parameter $\delta_i = h(\mathbf{y}_i)$, as a function of the population variable $\mathbf{y}_i$, for $i = 1, \ldots, D$. Let $\hat{\delta}_i$ an estimator of δ_i depending only on $\mathbf{y}_{is}$, the in-sample units for region i. The mean squared error for δ_i is defined as:

$$\text{MSE}\left(\hat{\delta}_i\right) = \text{E}_{\mathbf{y_i}}\left\{\left(\hat{\delta}_i - \delta_i\right)^2\right\}, \tag{2}$$

where $\text{E}_{\mathbf{y}_i}$ indicates the expectation with respect to the joint distribution of $\mathbf{y}_i$. The best predictor (EB) of δ_i that minimizes (2) is $\delta_i^B = \text{E}_{\mathbf{y}_{ir}}\left(\delta_i|\mathbf{y}_{is}\right)$, a function of $\mathbf{y}_{is}$ proportioned by the conditional expectation of $\mathbf{y}_{ir}$ given $\mathbf{y}_{is}$. Subtracting and adding $\hat{\delta}_i$ the next expression for the mean squared error is obtained as follows:

$$\text{MSE}\left(\hat{\delta}_i\right) = \text{E}_{\mathbf{y_i}}\left\{\left(\hat{\delta}_i - \delta_i^B\right)^2\right\} + 2\text{E}_{\mathbf{y_i}}\left\{\left(\hat{\delta}_i - \delta_i^B\right)\left(\delta_i^B - \delta_i\right)\right\} + \text{E}_{\mathbf{y_i}}\left\{\left(\delta_i^B - \delta_i\right)^2\right\}.$$

In this equation the third term does not depend on $\hat{\delta}_i$ and the second one is equal to zero as follows:

$$\begin{aligned}
\text{E}_{\mathbf{y}_i}\left\{\left(\hat{\delta}_i - \delta_i^B\right)\left(\delta_i^B - \delta_i\right)\right\} &= \text{E}_{\mathbf{y}_{is}}\left[\text{E}_{\mathbf{y}_{ir}}\left\{\left(\hat{\delta}_i - \delta_i^B\right)\left(\delta_i^B - \delta_i\right)\right\}|\text{E}_{\mathbf{y}_{is}}\right] \\
&= \text{E}_{\mathbf{y}_{is}}\left[\left\{\hat{\delta}_i - \delta_i^B\right\}\left\{\delta_i^B - \text{E}_{\mathbf{y}_{ir}}\left(\delta_i|\mathbf{y}_{is}\right)\right\}\right] \\
&= 0.
\end{aligned}$$

Since $\delta_i^B = \text{E}_{\mathbf{y}_{ir}}\left(\delta_i|\mathbf{y}_{is}\right)$, it is non-negative with min-value equal to zero, the EB is:

$$\hat{\delta_i}^B = \delta_i^B = E_{\mathbf{y}_{ir}}(\delta_i|\mathbf{y}_{is}), \tag{3}$$

which is also an unbiased estimator:

$$\text{E}_{\mathbf{y}_{is}}\left(\hat{\delta_i}^B\right) = \text{E}_{\mathbf{y}_{is}}\left\{\text{E}_{\mathbf{y}_{ir}}\left(\delta_i|\mathbf{y}_{is}\right)\right\} = \text{E}_{\mathbf{y}_i}\left(\delta_i\right).$$

Usually the joint distribution of $\boldsymbol{y}$ depends on $\boldsymbol{\theta}$, a vector of unknown model parameters. In this context, the empirical best predictor (EBP) of δ can be obtained by evaluating the expectation in 3 by substituting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is a suitable estimator of $\boldsymbol{\theta}$.

The unit-level nested error regression model.

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2), \tag{4}$$

where u_i, the area-specific random effects, and e_{ij}, the unit-level error, are assumed to be independent. Assuming normality for the unit-level error and the area-specific

random effects, the conditional distribution of the out-of-sample data given the sample data are also normal.

A Monte Carlo approach is used to obtain a numerically efficient approximation to the expected value of this conditional distribution as follows:

1. Use the sample data to obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and the weighting factors $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

2. For $l = 1, \ldots, L$:

   2.1. Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and obtain a pseudo-population of the target variable by:

   $$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)},$$

   where the predicted random effect $\hat{u}_i$ is defined as $\hat{u}_i = E(u_i|\mathbf{y}_{is})$.

   2.2. Calculate the indicator of interest $I_i^{(l)}$ in each area.

3. Finally, take the mean over the L Monte Carlo runs in each area to obtain a point estimate of the indicator of interest:

$$\hat{I}_i^{EBP} = \frac{1}{L} \sum_{l=1}^{L} I_i^{(l)}.$$

3.1.2 Empirical Plug-in Predictor (EPP) for binary variables

In cases when the variable of interest in small areas is binary, e.g. $y_{ij} = 0$ or 1, the target estimation in each domain is the proportion $\bar{Y}_i = \pi_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$, and π_{ij} is the probability that a specific unit j in the domain i obtains the value 1. Traditionally, it is assumed that the π_{ij} follows a GLMM with a logistic link function defined as:

$$\text{logit}(\pi_{ij}) = \log\Big(\frac{\pi_{ij}}{1-\pi_{ij}}\Big) = \eta_{ij} = x_{ij}^T \beta + u_i \tag{5}$$

with $j = 1, \ldots, N_i$, $i = 1, \ldots, D$, β is a vector of fixed effect parameter, and u_i the random area-specific effect for the domain i with $u_i \sim N(0, \sigma^2)$. u_i are assumed independent and $y_{ij}|u_i \sim \text{Binomial}(1, \pi_{ij})$ with $E(y_{ij}|u_i) = \pi_{ij}$ and $Var(y_{ij}|u_i) = \sigma_{ij} = \pi_{ij}(1-\pi_{ij})$. Under model 5, $E(y_{ij}|u_i) = \pi_{ij} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}$.

As explained in Jiang (2003), for this kind of data the Empirical Best Predictor must be computed by numerical approximation means and for this reason, the Empirical Plug-in Predictor EPP is popularly chosen as a more feasible alternative (Chandra et al (2012)). The EPP for small area proportion in area i is:

$$\hat{P}_i = \frac{1}{N_i}\Big(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{\pi}_{ij}\Big), \tag{6}$$

where $\hat{\pi}_{ij} = \hat{E}(y_{ij}|u_i) = \frac{\exp(\mathbf{x}_{ij}^T\hat{\beta}+\hat{u}_i)}{1+\exp(\mathbf{x}_{ij}^T\hat{\beta}+\hat{u}_i))}$. Here $\hat{\beta}$ and $\hat{u}_i$ are the estimate of fixed effects parameter and prediction of random effects parameter, respectively under model 5.

The EPP estimator is based on the observed value of study variable for each unit in the sample and predicted values of non-sampled units under model 5. This estimator requires unit-level values of auxiliary variables for the population units.

3.2 Metodología empleada

-Describir cómo hacemos el proceso para unir las estimaciones al “censo".

-Describir cómo se computa el índice.

3.3 Measure of uncertainty

Decidir qué vamos a hacer con el MSE??

- Bootstrap para obtener cada MSE por separado?, reportamos MSE independientemente?
- Bootstrap para obtener cada MSE conjunto? “sumamos" simplemente o buscamos otra alternativa?

4 Validation: Design-based simulation study

Queremos validar el MSE?

- In this section we illustrate some of the aspects of composite indicators in a SAE context evaluation. In particular, combining the results of model prediction we described in sections bellow, we present results for the estimation of the MPI. We discuss how the design-based simulation results can guide the production of the final set of MPI estimates in a SAE context. Analysis with the original sam. And we ple.
- Evaluation of the performance of the proposed methodology in the context of a real population and realistic sampling methods.
- Illustrating aspects of SAE evaluation using the GEIH data
-

5 Results

5.1 Employment-related income estimation

Iniciar con descripción del modelo. No hay seleccion de variables, ambos modelos usan las mismas. Por esto podríamos decidir si al iniciar la sección de resultados,

queremos presentar las variables utilizadas, los tamaños de muestra, población, dominios en /fuera de muestra, etc.?

An EBP model to obtain employment-related income information was conducted. Income per capita of each individual as dependent variable and the socio-economic covariates described in Table 2 as predictors. .

Table 2: Covariables included in the EBP model to obtain employment-related income estimates

Category	Variable
Geographical	1. Department
Socio-demographic	2. Age
	3. Sex
Education	4. Highest degree of education completed
Employment	5. Labor condition
Household	6. Proportion of employees in the household
conditions	7. Equivalized size of the household
	8. Overcrowding
	9. At least one member without health insurance
	10. Quantity of economically dependent members
Housing	11. Poor condition of the floor or ceiling
	12. Use of internet
	13. No garbage disposal system
	14. No exclusive toilet for the household

To select the model, several transformations were considered to achieve normality in the error terms, but the final version applies a Box-Cox transformation (**?**) with an optimal lambda () using the Restricted maximum likelihood (REML) approach. For this example, normality for both, the unit level and the random effects, is assumed. Also, the marginal $R^2 =$ and the conditional $R^2 =$ were observed.

Table 3: Summary statistics for sample and population sizes

	Min	1st Q	Median	Mean	3rd Q	Max
Sample domains						
(In-sample: %)						
Population domains						

- Chequear diagnósticos de los residuos
- QQ plots
- Indicadores básicos: R2, ICC
- Comparamos estos resultados con distintas transformaciones o no?

5.2 Social security and pension estimation

- Decidir si describimos las variables del modelo otra vez aquí o una subsección inicial, al comienzo de la sección de resultados
- Chequear residuos para los efectos aleatorios: QQplot y Shapiro test
- análisis de clasificación con los datos de encuesta: sensibilidad, especificidad, exactitud

5.3 Multidimensional deprivation index

- General incidence of poverty by: sex, age group, area.
- Description by indicator /domains
- Provide information with coefficients of variation

6 Concluding remarks and further research

- Time-gap between census and surveys.
- Possible (co-)relations between indicators and dimensions.
- Provide a formal MSE.
- Generalise the methodology for a) all Latin American countries, and if possible b) composite indicators.

Acknowledgments