Model-based small area estimates of standardized test outcomes in Colombian schools in presence of missing data

a. Andrés Gutiérrez, b. Diego Lemus and c. Felipe Ortíz
a. Assessment Director, b. Subdirector of instrument design and c. Investigator
Instituto Colombiano para la Evaluación de la Educación - ICFES

August 2016

SAE 2016 Maastricht

MINEDUCACIÓN

icfes mejor saber

TODOS POR UN NUEVO PAÍS PAZ EQUIDAD EDUCACIÓN

Overview

Motivation

1. Saber 3° 5° 7° 9° tests.
    1. General aspects
    2. Sample Design
    3. Assessment
2. Problem
3. Solution
4. Results
    1. Application 5°
    2. Application 7°

Motivation

- High quality statistics at low price - SAE.
- Due to the standardized tests students answer just a sample of items - Multiple imputation.

Saber 3° 5° 7° 9° tests

External standardized assessments applied annually since 2012 to students that are in third, fifth, seventh or ninth grades in public or private, rural or urban schools in Colombia.

Sampling design

Rotating Panel, 50% keeps from the preceding year and the remaining 50% change.

- Multistage stratified sample design for the 50% of the panel that changed.
- It incluides implicit and explicit stratification.
  - Explicit: Combination of variables in the sampling frame, 21 strata.
  - Implicit: Controls the order of the schools in each explicit strata and reproduces the population proportions.
- Systematic design for the schools selection.
- Systematic design for students selection in each test (mathematics, sciences, etc).

Test scoring

The two parameter logistic model from the Item Response Theory (IRT) is used.

$$P(Y_{ij} = 1|\theta_i, a_j, b_j) = \frac{1}{1 + exp(b_j - a_j\theta_i)} \qquad (1)$$

Where,

- θ_i Represents the student's ability. i
- a_j Represents item discrimination. j
- b_j Represents item difficulty. j

Estimation using plausible values

- The plausible values methodology is based on the multiple imputation procedures Rubin (1987) applied to item response theory models.
- Currently is applied in international tests like TIMSS and PISA.

Expansion estimator

The assessment estimation for each school is obtained as the average of the generated K plausible values weighted by the sample design weight w_{di}.

$$\hat{\hat{Y}}_d^{(D)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\hat{Y}}_{dk}^{(D)}, \qquad \hat{\hat{Y}}_{dk}^{(D)} = \frac{\sum_{i=1}^{n_d} w_{di}\hat{\eta}_{ik}}{\sum_{i=1}^{n_d} w_{di}} \tag{2}$$

While the variance is obtained as:

$$\hat{V}ar\left(\hat{\hat{Y}}_d^{(D)}\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{V}ar_p\left(\hat{\hat{Y}}_d^{(D)}\right) + \left(1+\frac{1}{K}\right)\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\hat{Y}}_{dk}^{(D)} - \hat{\hat{Y}}_d^{(D)}\right)^2 \tag{3}$$

$\hat{V}ar_p\left(\hat{\hat{Y}}_d^{(D)}\right)$ comes from the use of resampling procedures as Jackknife.

$$\hat{V}ar_p\left(\hat{\hat{Y}}_d^{(D)}\right) = \frac{1}{n_d(n_d-1)}\sum_{i=1}^{n_d}\left(\hat{\hat{Y}}_{d(i)}^{(D)} - \hat{\hat{Y}}_d^{(D)}\right)^2 \qquad (4)$$

Therefore $\hat{\hat{Y}}_{d(i)}^{(D)}$, corresponds to the value of $\hat{\hat{Y}}_d^{(D)}$. But leaving aside the i-th student of the d school.

Indirect estimator

Indirect estimator $\hat{\bar{Y}}_d^{(l)}$ for school average is based on calibration restriction.

$$\hat{t}_{xd} = \sum_{i=1}^{n_d} g_{di} w_{di} x_{ik} = t_{xd} \tag{5}$$

In this case, the school performance considering plausible values is obtained as:

$$\hat{\bar{Y}}_d^{(l)} = \frac{1}{K} \sum_{k=1}^{K} \hat{\bar{Y}}_{dk}^{(l)}, \qquad \hat{\bar{Y}}_{dk}^{(l)} = \frac{\sum_{i=1}^{n_d} g_{di} w_{di} \hat{\eta}_{ik}}{\sum_{i=1}^{n_d} g_{di} w_{di}} \tag{6}$$

Indirect estimator

Variance for indirect estimator is determined by the residuals of the model:

$$\hat{\eta}_{ik} = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + \varepsilon_{ik} \tag{7}$$

and

$$\hat{\hat{\varepsilon}}_d = \frac{1}{K}\sum_{k=1}^{K}\hat{\hat{\varepsilon}}_{dk}, \qquad \hat{\hat{\varepsilon}}_{dk} = \frac{1}{n_d}\sum_{i=1}^{n_d}\hat{\varepsilon}_{ik} \tag{8}$$

The variance is given by:

$$\hat{V}ar\left(\hat{\hat{Y}}_d^{(I)}\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{V}ar_p\left(\hat{\hat{\varepsilon}}_d\right) + \left(1 + \frac{1}{K}\right)\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\hat{\varepsilon}}_{dk} - \hat{\hat{\varepsilon}}_d\right)^2 \tag{9}$$

$$\hat{V}ar_p\left(\hat{\hat{\varepsilon}}_d\right) = \frac{1}{n_d(n_d-1)}\sum_{i=1}^{n_d}\left(\hat{\hat{\varepsilon}}_{d(i)} - \hat{\hat{\varepsilon}}_d\right)^2 \tag{10}$$

Composite estimator:

- Composite estimator is an average of both direct and indirect
- Reduces the variance of the direct estimator.
- Reduces the bias of the indirect estimator.

$$\hat{\bar{Y}}_d^{(C)} = \phi_d \hat{\bar{Y}}_d^{(D)} + (1-\phi_d)\,\hat{\bar{Y}}_d^{(I)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\bar{Y}}_{dk}^{(C)} \tag{11}$$

$$\hat{\bar{Y}}_{dk}^{(C)} = \sum_{i=1}^{n_d} w_{di}\,(\phi_d + (1-\phi_d)g_{di})\,\hat{\eta}_{ik} \tag{12}$$

The constant ϕ_d for each school could be obtained minimizing the mean square error of the resulting estimator $MSE\left(\hat{\bar{Y}}_d^{(C)}\right)$, draw from a cross validation procedure on the selected schools.

Problem

- Estimation performance for schools not selected in the sample, using auxiliary information.
- It is possible to use SAE and take the schools as small areas.
- How is the multiple imputation incorporated in the small area estimation?

SAE

- Mix the estimation by sample and by models.
- Because of sample size associated to the sample is small, information is increased with auxiliary information and a linear mixed model.

Area and individual level models

- In SAE there are two types of models frequently used, area and individual level.
- In the first ones auxiliar information is required at area level that is easier to obtain than the information at individual level (this is required in the second ones).
- In the first ones the protection for microdata is ensured while in the second ones no.

Fay - Herriot Model

$$\hat{\hat{Y}}_d = \mathbf{z}_d^t \beta + v_d + e_d, \qquad d = 1, ..., m \tag{13}$$

- The previous estimates are included with the plausible values in terms of $\hat{\hat{Y}}_d$.
- The error term variance e_d could be estimated with variance of the previous estimators.

BLUP

In Fay - Herriot model, the best linear unbiased predictor is:

$$\hat{\hat{Y}}_d^{(BP)} = (1 - B_d)\,\hat{\hat{Y}}_d + B_d \mathbf{z}_d^t \beta \tag{14}$$

With $B_d = \dfrac{\psi_d}{\sigma_v^2 + \psi_d}$. If σ_v^2 is known, β can be estimated by weighted least squares estimator $\tilde{\beta}(\sigma^2)$

$$\tilde{\beta} = \tilde{\beta}(\sigma^2) = \frac{\sum_{d=1}^{m} Z_d \hat{\hat{Y}} / \left(\sigma_v^2 + \psi_d\right)}{\left[\sum_{d=1}^{m} Z_d Z_d^t / \left(\sigma_v^2 + \psi_d\right)\right]^{-1}} \tag{15}$$

to obtain the best linear unbiased predictor (BLUP) for $\hat{\hat{Y}}_d$.

$$\hat{\hat{Y}}_d^{(BLUP)} = (1 - B_d)\,\hat{\hat{Y}}_d + B_d \mathbf{z}_d^t \tilde{\beta} \tag{16}$$

EBLUP

If σ_v^2 is unknown it could be estimated with the mean square error of the model (13), Therefore the Empirical Best Linear Unbiased estimator is obtained for $\hat{\hat{Y}}_d$.

$$\hat{\hat{Y}}_d^{(EBLUP)} = \left(1 - \hat{B}_d\right) \hat{\hat{Y}}_d + \hat{B}_d \mathbf{z}_d^t \tilde{\beta} \tag{17}$$

Results

- In 2013 10.125 schools took Mathematics Saber 5° test.
- 576 were selected in a control sample.
- Is possible to obtain the estimated performances for the 9.549 non selected schools using SAE.
- The expansion, indirect and composite estimates are analyzed for the estimation.
- More than 15 auxiliary information variables. Performance for the schools in previous years.
- More than 20.000 different models are tested, the following model that minimizes the mean square error for the obtained forecasts is selected.

Selected Model

$$\hat{\hat{Y}}_d = \beta_0 + \beta_1 X_{1d} + \beta_2 X_{2d} + \beta_3 X_{3d} + \beta_4 X_{4d} + \beta_5 X_{5d} + v_d + e_d \qquad (18)$$

Where:

	Results in	Observed / Estimated	Year
$\hat{\hat{Y}}_d$	Mathematics 5°	Estimated	2013
X_{1d}	Language 3°	Observed	2013
X_{2d}	Mathematics 3°	Observed	2013
X_{3d}	Mathematics 5°	Observed	2012
X_{4d}	Mathematics 9°	Observed	2012
X_{5d}	Physics 11°	Observed	2012

Estimators comparison

Estimator	$\sqrt{\text{MSE}}$ Controlled application	$\sqrt{\text{MSE}}$ Not controlled application
Horvitz Thompson	8.1927	23.15
Calibration	8.32	23.1471
Composite $\phi = 2/3$	8.282	23.14

Table Square root of the mean square error for the forecasts at school level.

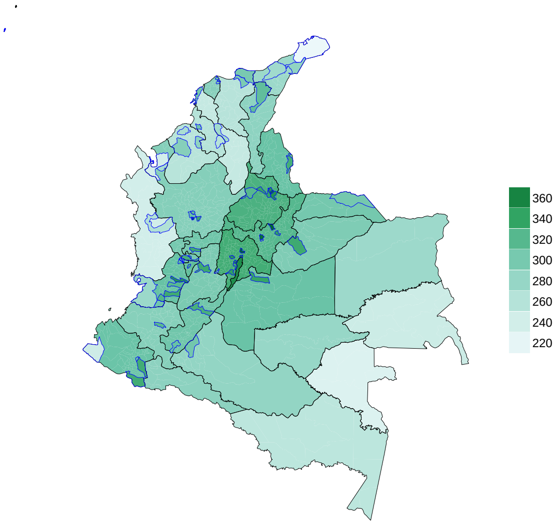

Figure Results observed in mathematics test.

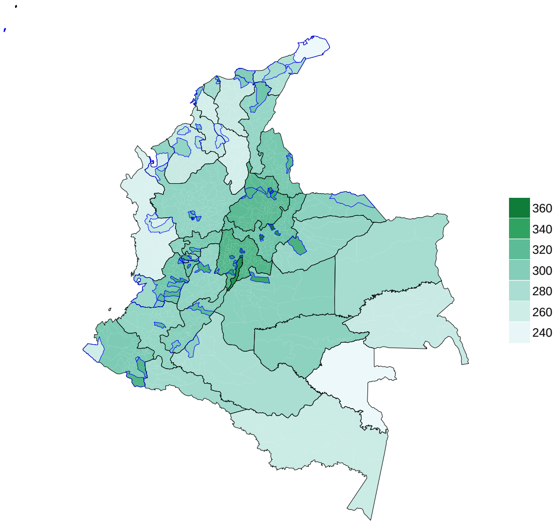

Figure Estimated results with the Horvitz-Thompson estimator.

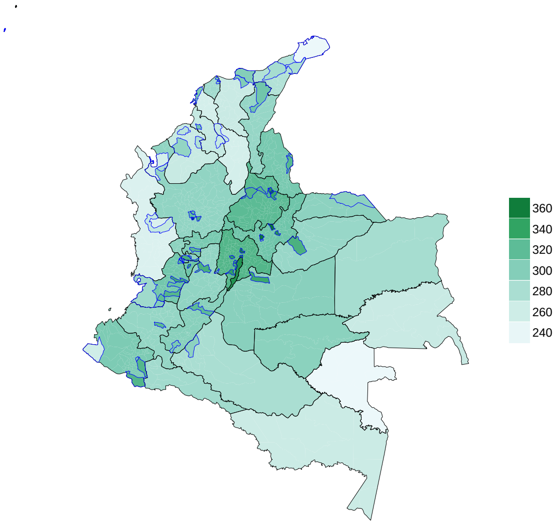

Figure Estimated results with the regression estimator.

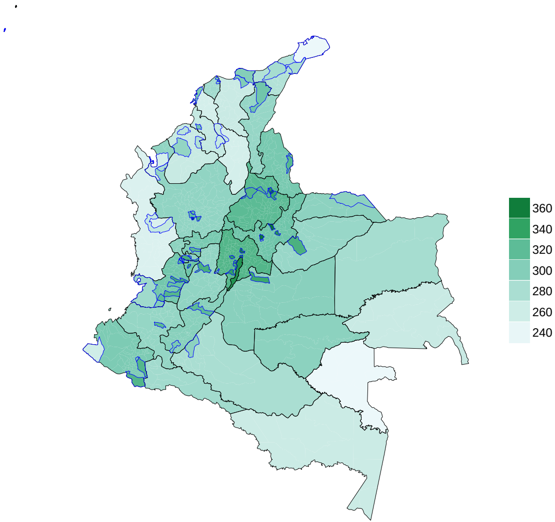

Figure Estimated results with the composite estimator.

SAE 7

In 2015, a random sample of 399 schools was selected and for the first time the performance of 7^{th} grade students was evaluated.

- Percentage of correct answers obtained for each area.
- Language, mathematics, sciences and citizenship skills.

Using cross validation, an approach of the $\sqrt{MSE}$ behavior is obtained for the estimation in each area.

Area	n	N	$\sqrt{MSE}$
Science	154	9846	0.4378%
Language	161	9837	0.4215%
Mathematics	155	9841	0.466%
Citizenship skills	161	9841	0.3685%

Table Square root of mean square error for forecasts at school level in the control sample.

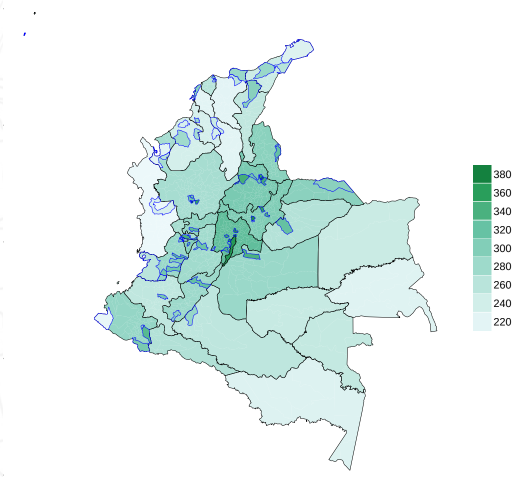

Figure Estimated results for the mathematics test with the proposed methodology.

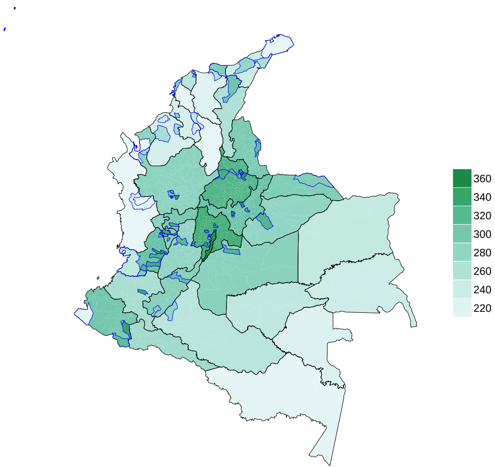

Figure Estimated results for the science test with the proposed methodology.

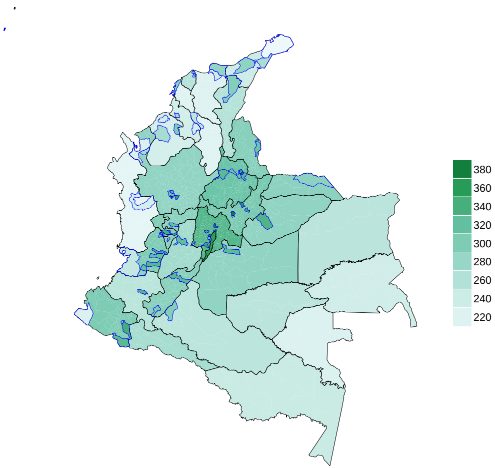

Figure Estimated results for the language test with the proposed methodology.

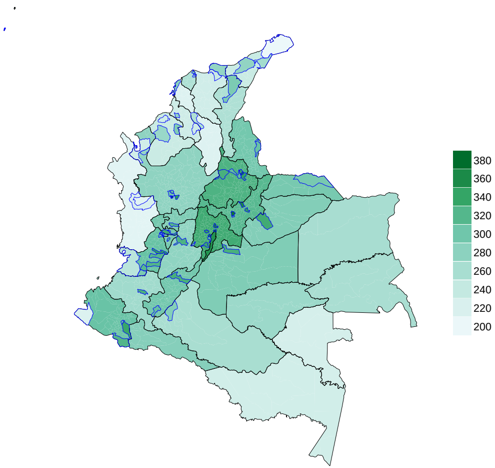

Figure Estimated results for the citizenship skills test with the proposed methodology.

References

- R.E. Fay y R.A Herriot. Estimation of income from small places: An application of james-stein procedures to census data. Journal of the American Statistical Association, (74): 269-277, 1979.
- D.B Rubin. Multiple imputation for nonresponse in surveys. John Wiley and Sons, Inc., 1987.
- ICFES. Pruebas SABER 3, 5 y 9. Lineamientos para las aplicaciones muestral y censal 2014. ICFES, 2014.
- ICFES. Manual de calificaci?n. prueba cognitiva saber 3, 5, 7 y 9. Informe t?cnico., ICFES, 2015.
- J.N Rao. Small Area Estimation. Wiley - Interscience, 2003.
- M. von Davier, E. Gonzalez, y R.J. Mislevy. What are plausible values and why are they useful? IERI Monograph series: issues and methodologies in large-scale assessments, 2009.
- K. Wolter. Introduction to variance estimation. Statistics for Social and Behavioral Sciences. Springer, 2ed, 2007.