uc3m | Universidad Carlos III de Madrid

Master Degree in Statistics for Data Science

2021-2022

Master Thesis

Small Area Estimation at two levels in the Colombian Saber 11 Exam

Andrés Mejía Rodríguez

Isabel Molina Peralta

Madrid, September 2022

SUMMARY

In this document we use compulsory Colombian Saber 11 exam data to explore how does a two level Nested Hierarchical Model based estimator for small area behave compared with other known estimators and the exact values of the parameters.

Keywords: Small Area Estimation, Nested Hierarchical Model, Fey-Herriot Model, Saber 11, Semi-parametric Bootstrap, Simulation

DEDICATION

To my brother for making this possible, to my my family for making this worthwhile, to my friends for making this fun.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

It is not uncommon for surveys to be built for measuring a population statistic with a certain degree of precision for example a presidential election or a poverty measure, also is not uncommon after all the data has been collected to try to estimate the same statistic in a domain or area of that sample. For example to find how a certain region behaves with respect to an election or a poverty measure. This is usually a problem since errors of those statistics increase as the sample size decreases. In this scenario, some of our areas will be fine, but most of them will have an unacceptable level of error due to their small sample size. It can even be the case that some areas have no individuals in the sample at all. Small area estimation attempts to solve this problem "borrowing strength" from auxiliary information about the population. It builds a model of the target variable and uses that information to "fill the gaps" and with these "data" improve the quality of our estimation. This work aims at understanding how well this works. We will take an existing census and generate a sample, using this sample we will generate both direct estimates and small area ones and their errors. We will then see how sensible are the methods to changes to the process. What will happen if we chose another sample? What will happen if the effect areas are different? How do they compare to the ground truth?

To do this we have the 2021 results of the second call of the Colombian Saber 11 exam [1]. Colombian law 1324 of 2009 [2] requires all schools to enroll all their students in the exam making it in practice an annual census of all graduating students. The exam includes scores of the individual subjects evaluated as well of a global score. Also with the inscription in the exam a socioeconomic survey is applied.

We will use a well-known and commonly used sampling design and will build our estimations based on that original sample. We will finally make Model Based and Design Based Validation Experiments validating not only our original sample and randomly generated samples as well.

2. DATA STRUCTURE AND SAMPLING

2.1. Data Description

The Saber 11 Exam is compulsory for all last year high school students in Colombia. There are two calls one in August aimed mostly to A calendar students, and one in March aimed mostly to B calendar students. The exam evaluates five areas: Critical Reading, Mathematics, Social Sciences and Education for Citizenship, Science, and English, We are interested in the global score that ranges from 0 to 500, 250 is the national mean score. Inscription to the exam includes a socioeconomic questionnaire collecting information about amenities in the household, income, stratum of the dwelling, education, and occupation of the parents among others.

Our dataset consists of the anonymized records of all 332810 students who took the exam in the second call of 2021. The dataset includes both the exam results as well as the socioeconomic survey [1]. The variables corresponding to the performance of the student on the test are described in table 2.1, We will focus on the total score of the students as our target variable.

The database also contains geographical information of both the school and the student. We will use information about the students home to build the sample, the relevant variables are the DIVIPOLA codes of department and municipalities [3]. Those can be seen in table 2.2

The variables corresponding to the results of the socioeconomic survey are described in table 2.3. Note that this list is not exhaustive and only includes the variables that we will use in this work. Some of the variables where transformed from their original values in order to reduce the number of levels in each variable (for example education level of the parents). Also INSE is an important numerical variable that is not easy to interpret. It is a numerical score of the socioeconomic situation of the student. To see more information

Variable Name	**Description**	**Range**
PUNT_GLOBAL	Total Score of the test	0-500
PUNT_"TEST"	Score of one of each of the five subjects MATEMATICAS-MATH LECTURA CRITICA-CRITICAL READING C_NATURALES-NATURAL SCIENCES SOCIALES_CIUDADANAS-EDUCATION FOR CITIZENSHIP INGLES-ENGLISH	0-100
PERCENTIL_"TEST"	Percentile of each of the five subjects	1-100
DESEMPEÑO_"TEST"	Levels bases on the Score of each of the five subjects	1,2,3,4

TABLE 2.1. PERFORMANCE VARIABLES

Variable Name	Description
ESTU_COD_RESIDE_DEPTO	DIVIPOLA Code of the department of the student's housesold
ESTU_COD_RESIDE_MCPIO	DIVIPOLA Code of the municipaliry of the student's housesold

TABLE 2.2. LOCATION VARIABLES

of this variable see the available ICFES documentation [4].

Missing data were handled assigning them to their own category (when categorical) or removing the record from the database if the INSE score was missing. There were no missing records for the Global Score.

2.2. Sampling Design and Description of Small Areas

The sampling frame is defined as all the records on the database. A stratified sample was then built using a variance constrained allocation model using Global Score (see [5]). The information used to estimate the prior variance was the total score of the exam in the 2020 second call. Each strata is one of the 1118 municipalities in Colombia, a minimum of 5 students per municipality were sampled whenever possible, even when the allocation called for less students. This was done to assure that all areas have elements in the sample.

The resulting sample consisted in 38479 students, with strata having sizes ranging from 3 to 5652. Note that even when imposing a minimum size of five we still got some municipalities with a smaller sample size. This happens because those areas whole population is less than five.

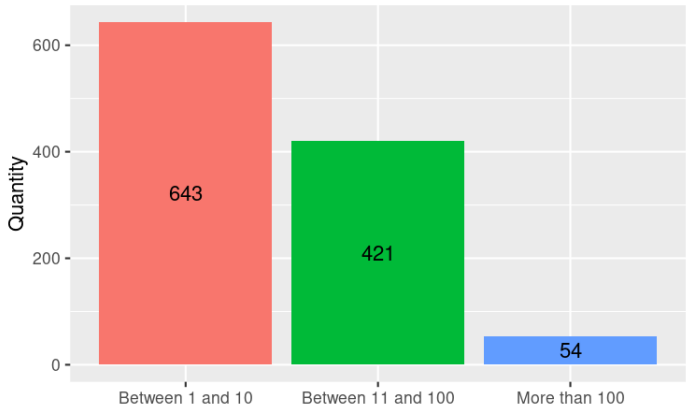

Fig. 2.1. Number of municipalities with a given sampled individuals

As we can see from figure 2.1 over half the municipalities have a sample size less than 10. Those areas also tend to concentrate most of the rural areas of the country (figure 2.2) and tend to be poorer as measured by the INSE (figure 2.3).

Variable Name	Description	Levels
FAMI_EDUCACIONMADRE	Highest Education Level Achieved by the mother	<elementary school Elementary school High School Trade School College Degree Postgraduate Degree
FAMI_EDUCACIONPADRE	Highest Education Level Achieved by the father	<elementary school Elementary school High School Trade School College Degree Posgraduate Degree
COLE_NATURALEZA	Is the school public?	Non Public Public
FAMI_ESTRATOVIVIENDA	Stratum of the householsd	No strata Stratum 1 Stratum 2 Stratum 3 Stratum 4 Stratum 5 Stratum 6
ESTU_GENERO	Student Gender	F M
COLE_AREA_UBICACION	Is the school in an urban or rural area?	Urban Rural
FAMI_NUMLIBROS	Number of books in the household	0-10 11-25 26-100 >100
ESTU_TIENEETNIA	Is the student from an etnic minority?	Yes No
FAMI_CUARTOSHOGAR	Number of rooms in the household	One, Two Three, Four Five, >=Six
FAMI_TIENEINTERNET	Does the household have internet?	Yes No
COLE_CALENDARIO	What calendar does the school follow?	A (school year starts in january) B (school year starts in August) Other
COLE_JORNADA	What time do the students go to school?	Morning, Afternoon Night, Saturdays Complete, Unique
ESTU_INSE_INDIVIDUAL	Students Individual SocioEconomic Score	See text

TABLE 2.3. SOCIO ECONOMIC VARIABLES

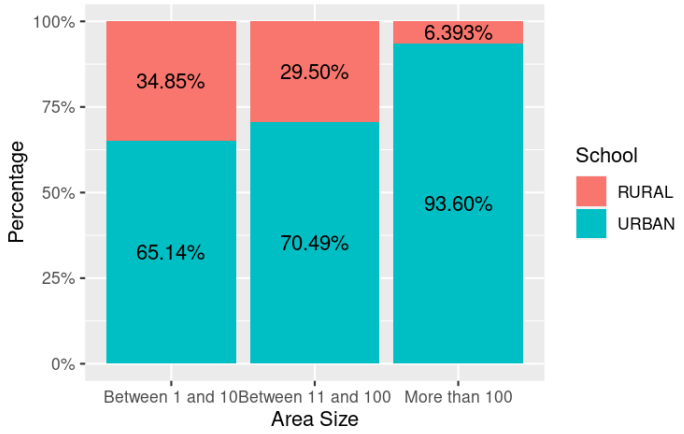

Fig. 2.2. Schools location Rural vs Urban and Municipality Sample Size

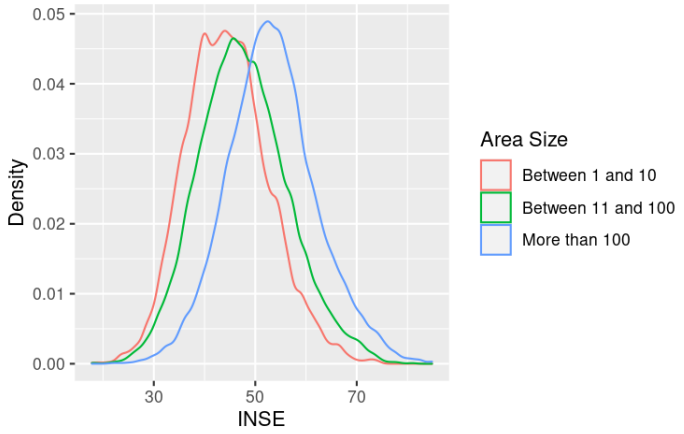

Fig. 2.3. INSE scores and Sample Size

3. ESTIMATION RESULTS

3.1. Direct Estimators

Consider the total score of a student i who lives in a department j and in a municpality k (S_{ijk}). Given that in each municipality we have taken a simple random sample, then, the direct estimation of the mean score (i.e the Horvitz-Thompson estimator) is [5]:

$$\hat{\mu}_k = \frac{\sum_{i \in k} S_{ijk}}{n_k}, \tag{3.1}$$

where n_k is the sample size of each area. When calculating the sample error in each area we must consider not only that they are small in the sample but also they are small in the frame, that is their whole population is small, so a correction for the finite sample size is necessary. Thus the expression for the sample error of the total score in each area is given by [5].

$$\text{se}_k = \frac{\hat{\sigma_k}}{\sqrt{n_k}} \sqrt{\frac{N_k - n_k}{N_k - 1}}, \tag{3.2}$$

where $\hat{\sigma_k}$ is the estimated standard deviation of each area and $N - k$ is the total population in each area. From this expression we can see that the bigger the sample size the smaller the error. Figure 3.1 shows the relationship between the estimated error and the logarithm of the area size. We can see that the bigger mean square errors are in small areas.

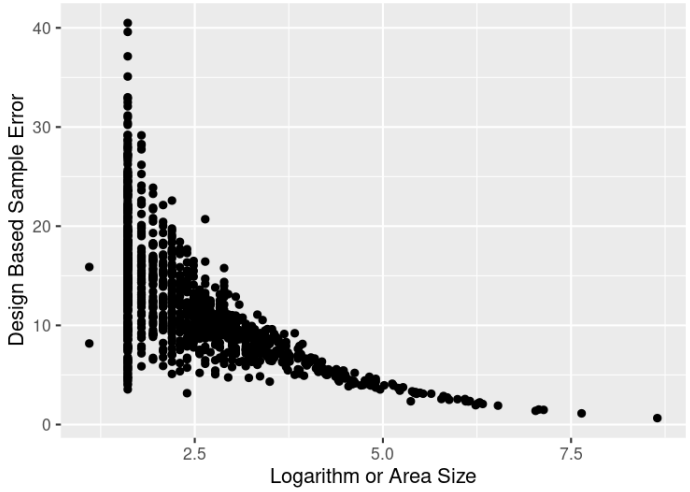

Fig. 3.1. Distribution of the Design Based Sample Error vs Logarithm of the Sample Size

3.2. Model Based Estimation Result

Direct Estimators only consider information about the target variable to estimate the population means, on the other hand Model Based Estimation "borrows strength" from the accompanying socioeconomic variables to improve results. For this we will use a modified Fey-Herriot Model [6]. In this approach the data is assumed to follow a linear mixed model with a single random effect accounting for the "area effect". We will use an extension of this idea used by Marhuenda, Molina, Morales and Rao [7] of using the two random effects.

3.2.1. Nested Hierarchical Model

The model we will consider is the score S_{ijk} for student i in department j and municipality k, based in the work or Marhuenda, Molina, Morales and Rao [7]:

$$S_{ijk} = \mathbf{x}_{ijk}^{'}\beta + u_{1,j} + u_{2,jk} + e_{ijk} \tag{3.3}$$

where $i \in \{1, \ldots, 32\}$ and $j \in \{1, \ldots, 1118\}$, $\mathbf{x}_{ijk}$ is a matrix with the values of socioeconomic information for the individuals and β os the vector of regression coefficients (this will be the same for all subdomains and domains). $u_{1,j}$ is the effect of each department in the quality of the students and $u_{2,jk}$ is the effect of each municipality. This and the error are assumed to be normal, mutually independent random variables satisfying.

$$\begin{aligned} u_{1,j} &\sim \mathcal{N}(0, \sigma_1^2) \\ u_{2,jk} &\sim \mathcal{N}(0, \sigma_2^2) \\ e_{ijk} &\sim \mathcal{N}(0, \sigma_3^2) \end{aligned} \tag{3.4}$$

There is ample literature showing how similar models use the socio-economic data from the Saber 11 to predict the students score. Timarán et al. [8] used regression trees, Gomez et al.[9] used logistic regression to predict performance, and Gutierres et al. [10] used those to predict the performance in the English component of the test.

The auxiliary variables $\mathbf{x}_{ijk}$ correspond to socioeconomic information, they can be found in table 2.3. To assess the quality of the model we will first check the normality of the random components and error as in equation 3.4. The distribution of residuals, municipality effect and department effect are in figures 3.2, 3.3 and 3.4 respectively.

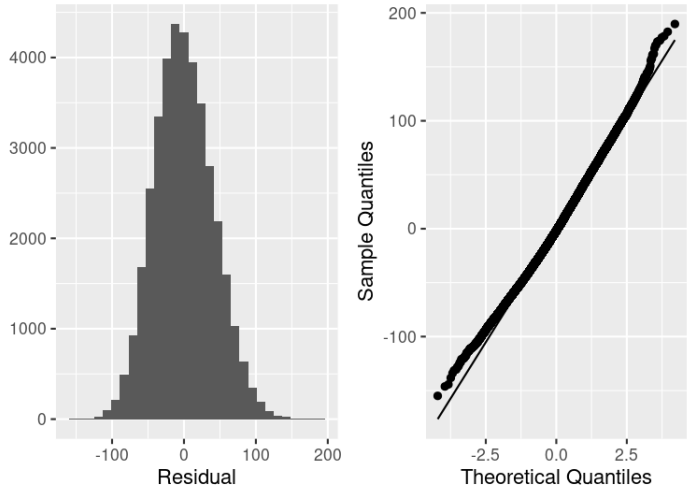

Fig. 3.2. Distributions of Residuals in Model 3.3

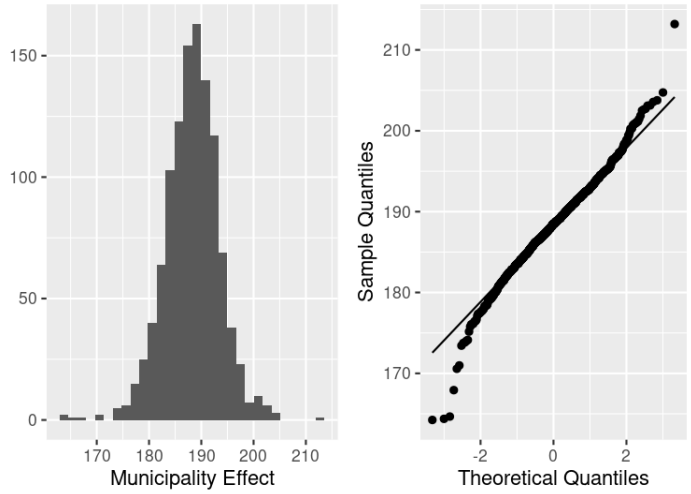

Fig. 3.3. Distributions of Municipality Effect in Model 3.3

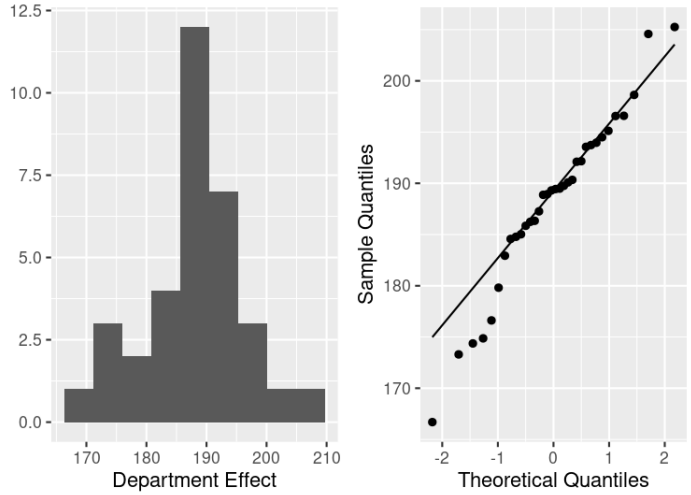

Fig. 3.4. Distributions of Department Effect in Model 3.3

Plots 3.2 and 3.4 corresponding to residuals and department effect seem to be consistent with a normal distribution, however figure 3.3 corresponding to municipality effect seems to be light tailed.

3.2.2. Empirical Best Predictors

We now move to using our model to build a prediction for the mean score in each municipality. An intuitive way to do this is to keep the sampled data and "fill" the non sampled data with predictions from our model. To formalize this concept let U_{jk} and s_{jk} be the population and sample sets of municipality k in department j. We then have that $|U_{jk}| = N_k$ and $|s_{jk}| = n_k$, We assume that all domains follow the model given by equations 3.3 and 3.4. Then following [7] we have that the Empirical Best estimator is:

$$\hat{\delta}_{jk}^{EB} = \frac{1}{N_k}\left\{\sum_{i\in s_{jk}} S_{ijk} + \sum_{i\in U_{jk}-s_{jk}} E[S_{ijk}|\mathbf{x}_{ijk} : \hat{\theta}]\right\} \tag{3.5}$$

where $\hat{\theta}$ is a consistant estimator $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_0^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$. Note that this estimator even works for domains with no elements on the sample. In that case the estimator is obtained predicting the scores for all the students in that domain.

3.2.3. Bootstrap error

One way to estimate the mean squared error of estimators is to use a bootstrap scheme. González-Manteiga et al [11] use the following semiparametic approach to estimate the MSE of the estimator.

Let U be a finite population and s be the sample found using the design described in section 2.2. A superpopulation is then defined as a not yet realized population following equations 3.3 and 3.4.The method used can be summarized as follows:

1. From s calculate consistent estimators $\hat{\beta}, \hat{\sigma}_0^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$ of $\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2$
2. For each department draw an independent value $u_{1.j}^*$ a normal with variance $\hat{\sigma}_1^2$
3. For each municipality draw an independent value $u_{2.j}^*$ a normal with variance $\hat{\sigma}_2^2$
4. For each individual in U draw an independent value e_{ijk}^* a normal with variance $\hat{\sigma}_3^2$
5. Generate a population U^* the superpopulation model given by

$$S_{ijk}^* = \mathbf{x}_{ijk}^{'}\beta + u_{1,j}^* + u_{2,jk}^* + e_{ijk}^*. \tag{3.6}$$

6. Calculate the values of the mean score for this population in each small area

$$\delta^*_{jk} = \sum_{i \in U_{jk}} S^*_{ijk}. \tag{3.7}$$

7. Build a sample s^* that contains the elements with the same indices as s.

8. Build estimator δ^{EB*}_{jk} as 3.5.

9. Calculate $W_m = (\delta^{EB*}_{jk} - \delta^*_{jk})^2$

10. Repeat steps 2-8 (index by m) B times, the estimator is given by

$$\text{mse}_*(\delta_{jk}) = B^{-1} \sum_{m=1}^{B} W_m. \tag{3.8}$$

We calculated then the MSE of the estimator for each municipality using 32000 bootstrap samples in each case. Comparing figures 3.1 and 3.5 we can see that the error is almost a smooth function of the sample size, we also see that in absolute scale the MSE tends to smaller than the design based error.

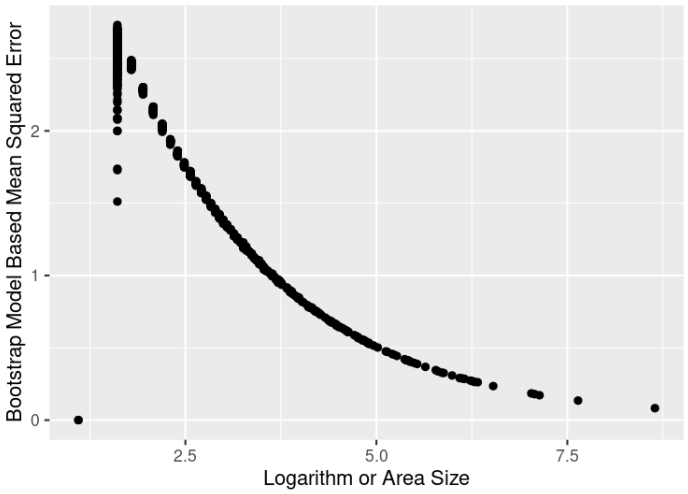

Fig. 3.5. Distribution of the Bootstrap Model Based Mean Squared Error vs Logarithm of the Sample Size

4. COMPARISON WITH GROUND TRUTH

In this chapter we compare the population mean Global Score of each municipality with the estimated mean score using the Direct Estimator and the two level Model Based estimator. We will also compare it with two simpler versions of the same model, where instead of using two nested effects we will consider models similar to 3.3. In 4.1 we consider that we have only one small area for each department and in 4.2 we only consider effect in each municipality and no overall effect for each department. Note that these models are classical Fey-Herriot Models [6].

$$S_{ijk} = \mathbf{x}_{ijk}^{'}\beta + u_{1,j} + e_{ijk} \tag{4.1}$$

$$S_{ijk} = \mathbf{x}_{ijk}^{'}\beta + u_{2,jk} + e_{ijk} \tag{4.2}$$

We are capable of comparing the results with the population value of the parameter given that we know the whole population values in advance. Let Δ_{jk} be the mean of the total score in municipality jk

4.1. Absolute difference

Let Δ_{jk} be the true value of the parameter then the absolute difference error is given by

$$|\delta_{jk} - \Delta_{jk}|. \tag{4.3}$$

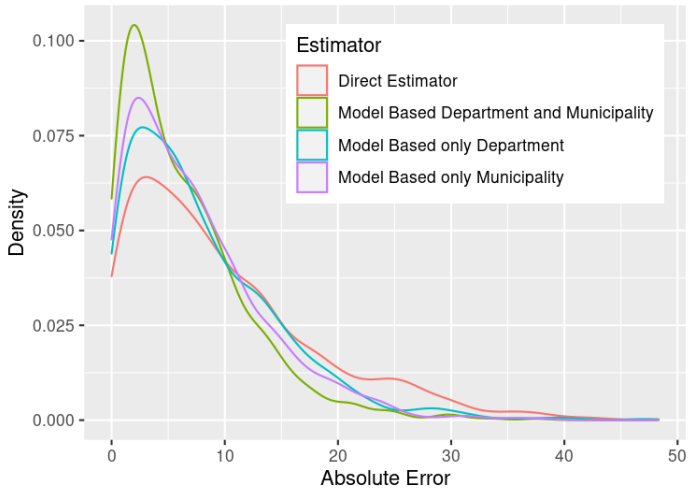

Fig. 4.1. Distribution of the Absolute Error in each Municipality

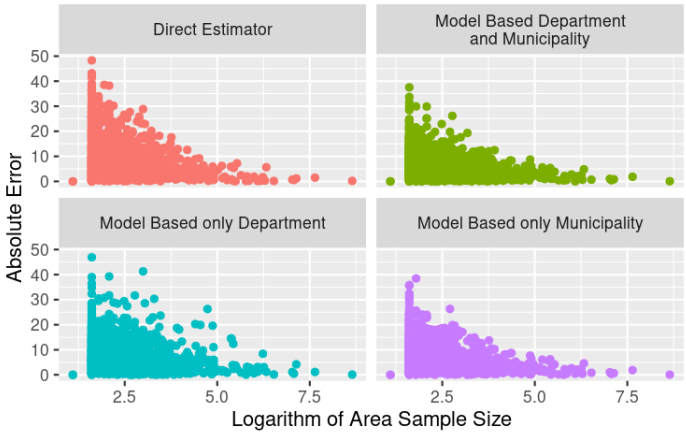

Fig. 4.2. Distribution of the Absolute Error in each Municipality vs Logarithm of the Sample Size

We can from figures 4.2 and 4.2 that model based approaches that consider municipality in general tend to give lower absolute error compared with the direct estimator. As expected all estimators reduce their error as the sample size gets bigger. In other words is only in the small areas that estimators give large errors.

We can also see that in general Model Based approaches that perform better that the Direct Estimator, and Model Based approaches that include municipality perform better than those without. We can also see that this performance tends to be consistent across area sizes.

4.2. Relative Differences

We now focus on the relative differences

$$\frac{|\delta_{jk} - \Delta_{jk}|}{\Delta_{jk}}. \tag{4.4}$$

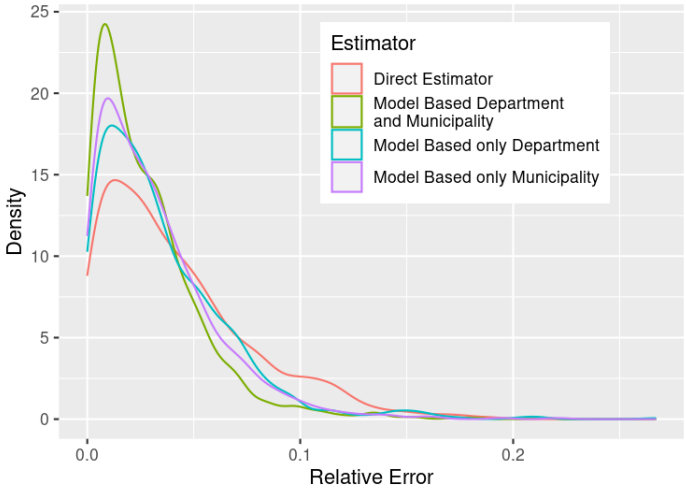

Fig. 4.3. Distribution of the Relative Error in each Municipality

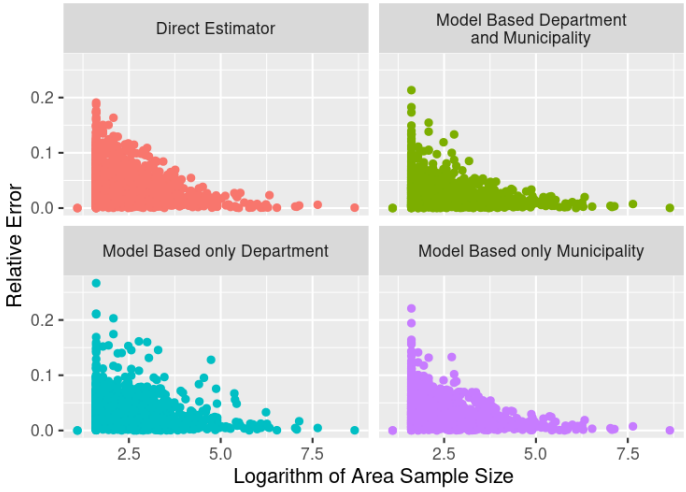

Fig. 4.4. Distribution of the Relative Error in each Municipality vs Logarithm of the Sample Size

Relative differences behave similarly than absolute differences with Model Based Estimators having lower relative error, here we can see that one result of misespecification in the model that only includes department is that it seems to have some municipalities with higher error than the other models.

5. SIMULATION EXPERIMENT

5.1. Design Based Validation Experiment

In this section we will explore a design based simulation study. As the data in the Saber 11 exam is a census of a population we can find out the exact value of the parameter we are estimating, this allows us to account for the effect of model error apart from sampling error.

From our population drew 3000 samples s_w, $w \in \{1, \ldots, 3000\}$ from the dataset using the same method described in section 2.2. Using each sample we will build the direct estimator (eq 3.1), nested two levels estimator (eq 3.3), one level department estimator (eq 4.1) and one level municipality estimator (eq 4.2) for that sample $(\delta_{jk})_w$ and Δ_{jk} the true value of the population mean in that municipality.

$$\hat{\text{Bias}}_{\delta_{jk}} = \frac{1}{3000} \sum_{w=1}^{3000} (\delta_{jk})_w - \Delta_{jk} \tag{5.1}$$

$$\hat{\text{MSE}}_{\delta_{jk}} = \frac{1}{3000} \sum_{w=1}^{3000} ((\delta_{jk})_w - \Delta_{jk})^2 \tag{5.2}$$

Figure 5.1 shows the distribution of the result across municipalities. We can see that the that for the direct estimator the bias in almost all municipalities is close to zero, the other estimators have more significant bias. Of the remaining estimators we see that the one that considers both department and municipality is tighter than the other two.

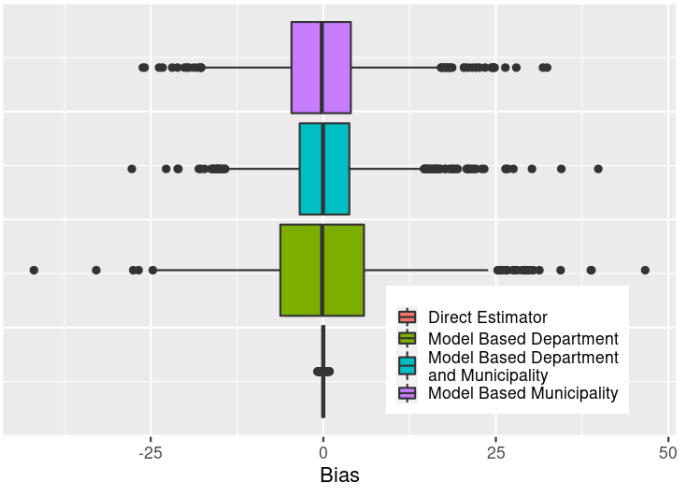

Fig. 5.1. Bias in Design Simulation

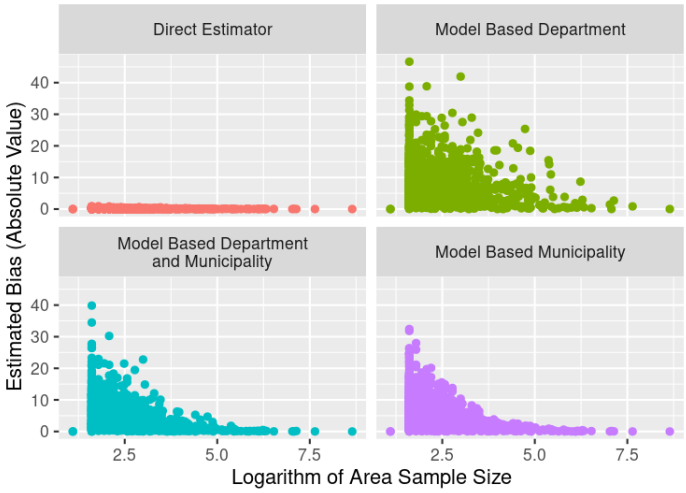

Fig. 5.2. Bias in Design Simulation vs Sample Size

In figure 5.2 We can see that the bias of the estimator decreases as the sample size increases. The bias is very low for the direct estimation and tends to be higher in small areas.

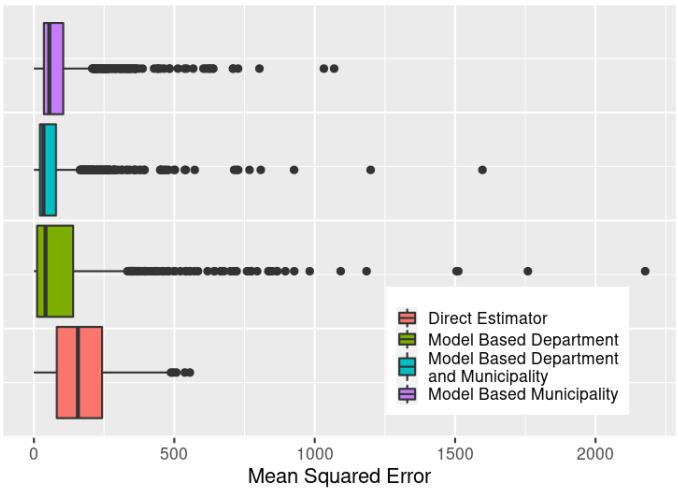

Fig. 5.3. MSE in Design Simulation

In contrast figure 5.3 shows us that the MSE of the Model based estimators tends to be lower than the direct estimator in most of the cases (mostly the interquartile range is lower in all model based estimators). However there are some outliers that have a MSE that is higher than the direct estimator.

Fig. 5.4. MSE in Design Simulation vs Sample Size

Again in figure 5.4 the size of the MSE the direct estimator tends to decrease as sample size increases. We can see that the MSE can spike in smaller sample sizes but overall the Model Based Estimators that include municipality tend to behave quite well.

5.2. Model Based Validation Experiment

For this simulation experiment the ground truth does not come from the data itself but from the model described in equations 3.3-3.4.

1. Set the values of $\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2$.
2. For each department draw an independent value $u_{1.j}$ a normal with variance σ_1^2.
3. For each municipality draw an independent value $u_{2.j}$ a normal with variance σ_2^2.
4. For each individual in U draw an independent value e_{ijk} a normal with variance σ_3^2.
5. Generate a population U^* the superpopulation model given in equations 3.3 and 3.4.
6. Calculate the values of the mean score for this population in each small area.
7. Build a sample s^* that contains the elements with the same indices as our original sample.

From this sample the estimators of the mean global score for each municipality are built. Note that the procedure is almost identical to the bootstrap described in section 3.2.3. The main difference being the starting point, in that case the model parameters are assumed to coincide to the first estimator build on the real sample. Here those values can be set arbitrarily. Whoever here we can check properties such as bias and the size of the MSE under those hypothesis.

Also note that the ground truth of this simulation refers to the true values of the parameters $\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2$ and not the value of the mean score in each municipality. And the mean score value in each municipality is a random variable.

The following plots show the comparison of the bias and MSE for 10000 simulations each individual point is a municipality. The values of $\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2$ were chosen to be consistent with section 3.2.3.

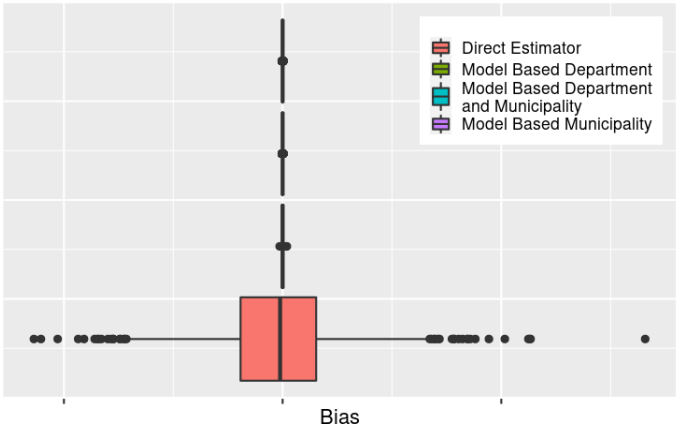

Fig. 5.5. Bias in Model Simulation

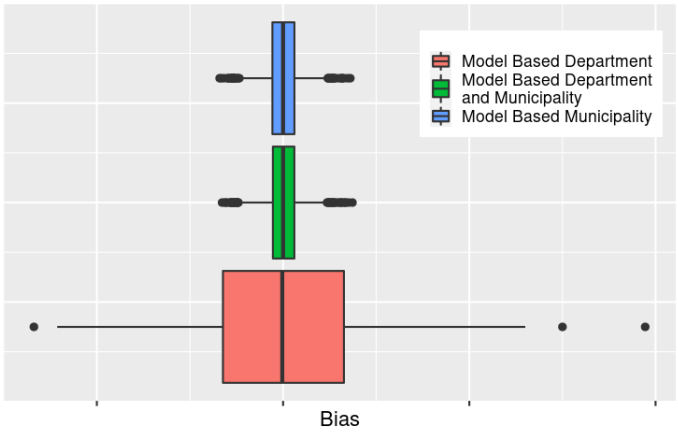

Fig. 5.6. Bias in Model Simulation without Direct Estimation

In figures 5.5 and 5.6 we can see that the direct estimator has a bias that is orders of magnitude bigger than any of the model based estimators when the ground truth comes form a model based framework. Of the design based models we can see that the model that uses only the broader department area has more bias than any of those models that consider municipality. These two models behave similarly one to another.

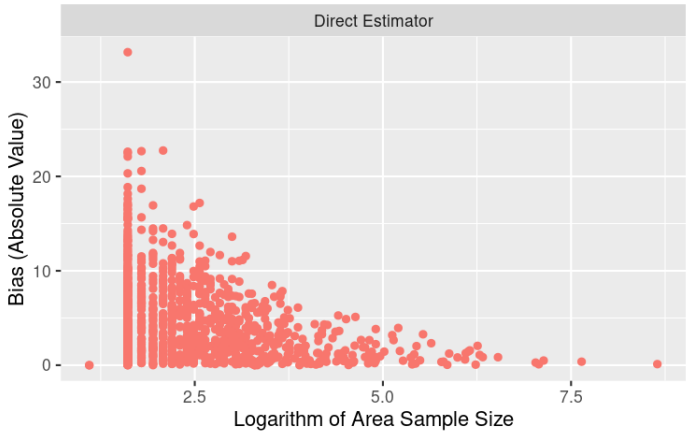

Fig. 5.7. Absolute value of Bias vs log sample size in model based experiment

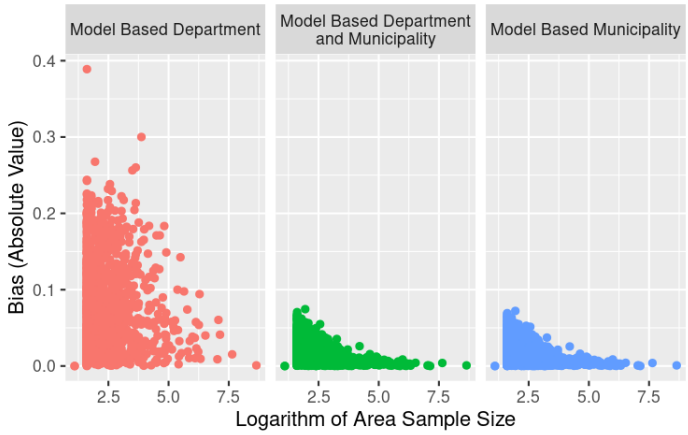

Fig. 5.8. Absolute value of Bias vs log sample size in model based experiment

If we analyse the effect that size has on the bias (figures 5.7 and 5.8)we can see that the direct estimators has to be plotted in its own plot due to it having a bias orders of magnitude bigger than the other estimators. It behaves as expected having a lower bias as the sample size increases. When we look at the model based estimators we see that the estimator that only has department as area effect has higher bias and although it decreases with sample size it does not seem to decrease as fast.

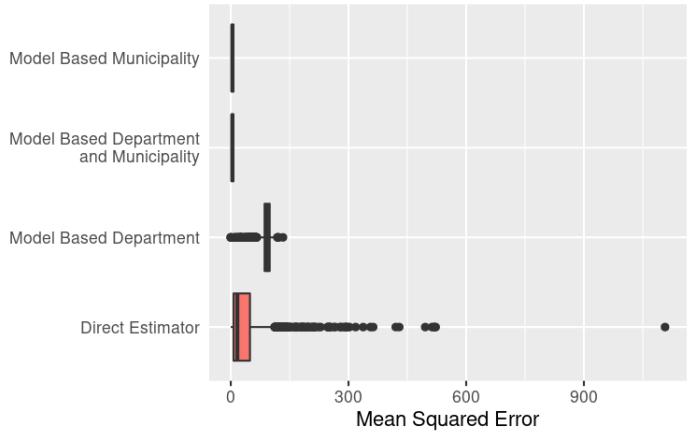

Fig. 5.9. MSE in Model Simulation

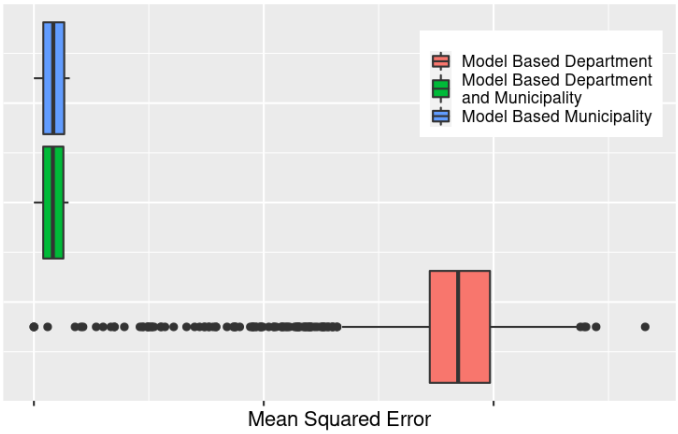

Fig. 5.10. MSE in Model Simulation without Direct Estimation

A similar situation occurs when analysing the MSE. In figures 5.9 and 5.10 we can see that, again the direct estimator has a MSE orders of magnitude bigger than any of the model based estimators. And of the design based models we can see that the model that uses the department areas has bigger MSE than those models that consider municipality.

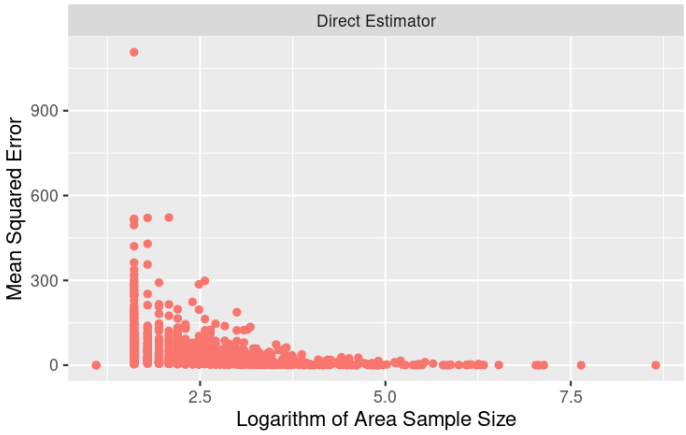

Fig. 5.11. MSE vs log sample size in model based experiment

Fig. 5.12. MSE vs log sample size in model based experiment

Figures 5.11 and 5.12 show that we have a similar situation than in bias. That is that direct estimator have a MSE orders of magnitude bigger than any model based one. There is however another effect worth mentioning. If the model was misspecified making the area wider than the true effect increasing the sample size does not decrease the MSE as much.

APPENDIX

The code used in the project can be found at:

https://github.com/andresmejiaro/SmallAreaTFM

Where it can be downloaded. The following scripts are of interest

- *PlotsCh2-3.R* Plots of Chapter 2 and 3.1.
- *BootstrapGen.R* Generate Bootstrap Errors chapter 3.
- *Compare Models.R* Plots in ch 4 and 3.2-3.4.
- *Design Simulation Gen.R* creates the design based simulation in ch 5.
- *design based images.R* creates the plots 5.1-5.4.
- *Model Based Gen* creates the model based simulation in ch 5.
- *model based experiment images.R* create images 5.5-5.12.

BIBLIOGRAPHY

[1] R. de datos abiertos - data Icfes. Accessed May. 20, 2022, Instituto Colombiano para el Fomento de la Educación Superior (ICFES. (), [Online]. Available: https://www2.icfes.gov.co/web/guest/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos?.

[2] Colombian Congress, *Law 1324*, https://www.mineducacion.gov.co/1621/article-210697.html, 2009.

[3] *Codificación de la división político administrativa de colombia - divipola*. [Online]. Available: https://www.dane.gov.co/index.php/sistema-estadistico-nacional-sen/normas-y-estandares/nomenclaturas-y-clasificaciones/nomenclaturas/codificacion-de-la-division-politica-administrativa-de-colombia-divipola.

[4] ". I. C. para el Fomento de la Educación Superior (ICFES ", *Saber al detalle 4:¿cómo se construye el índice de nivel socioeconómico (inse) en el contexto de las pruebas saber?* https://www2.icfes.gov.co/documents/39286/2231027/Edicion+4+-+boletin+saber+al+detalle+.pdf/f9a33ad6-7559-99a5-5f7f-16d2f9b16f76?version=1.0&t=1647958803251, 2019.

[5] W. G. Cochran, *Sampling Techniques, 3rd Edition.* John Wiley, 1977.

[6] J. Rao, *Small Area Estimation*, ser. Wiley Series in Survey Methodology. Wiley, 2005. [Online]. Available: https://books.google.es/books?id=f8NY6M-5EEwC.

[7] Y. Marhuenda, I. Molina, D. Morales, and J. Rao, "Poverty mapping in small areas under a twofold nested error regression model," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, Jul. 2017. DOI: 10.1111/rssa.12306.

[8] R. Timarán-Pereira, J. Zambrano, and A. Hidalgo Troya, "Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°," *REVISTA DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN*, vol. 9, Feb. 2019. DOI: 10.19053/20278306.v9.n2.2019.9184.

[9] S. Gómez, D. Gutiérrez, and A. Ramírez, "Determinantes del rendimiento académico en colombia. pruebas icfes - saber 11o, 2009," *Revista Universidad EAFIT*, vol. 46, pp. 48–72, Jan. 2010.

[10] D. M. Gutiérrez Duque and C. A. Mayora Pernía, "Variables predictoras del desempeño escolar en exámenes estandarizados de inglés: Evidencias desde el examen de estado en colombia," *Revista Virtual Universidad Católica del Norte*, no. 62, pp. 33–62, Dec. 2020. DOI: 10.35575/rvucn.n62a3. [Online]. Available: http://34.231.144.216/index.php/RevistaUCN/article/view/1246.

[11] W. González-Manteiga, M. J. Lombardía, I. Molina, D. Morales, and L. Santamaría, "Bootstrap mean squared error of a small-area eblup," *Journal of Statistical Computation and Simulation*, vol. 78, no. 5, pp. 443–462, 2008. DOI: 10.1080/00949650601141811. eprint: https://doi.org/10.1080/00949650601141811. [Online]. Available: https://doi.org/10.1080/00949650601141811.