**Revista Colombiana de Estadística - Applied Statistics**
July 2024, volume 47, issue 2, pp. 407–422
http://doi.org/10.15446/rce.v47n2.112779

# A Review of the Use of Small Area Estimation in Colombia

## Un recorrido en el uso de estimación en áreas pequeñas en Colombia

Felipe Ortíz Rico[1,a], Cristian F. Tellez Piñerez[1,b], Nicolas Ramirez-Vargas[2,c]

[1]Universidad Santo Tomás, Bogotá, Colombia
[2]Departamento Administrativo Nacional de Estadística, DANE, Bogotá, Colombia

**Abstract**

This article provides a review of the work carried out in Colombia on small area estimation. It considers initiatives of an academic nature, mainly originating from universities, as well as initiatives focused on generating official statistics from public offices and private companies in the country. The objective of the work is to update the interested reader on the progress of this methodology in the country and to encourage the community to deepen into the analysis and publication of content on small area estimation. Additionally, a summary of the main models used in small area estimation is presented.

***Key words*:** Colombia; Small area estimation; Review.

**Resumen**

En este artículo se presenta una revisión de los trabajos realizados en Colombia sobre estimación en áreas pequeñas. Se considera iniciativas de carácter académico originadas principalmente en las universidades, así como iniciativas orientadas a la generación de estadísticas oficiales provenientes de oficinas públicas del país y de empresas privadas. El trabajo busca actualizar al lector interesado sobre los avances acerca de esta metodología en el país y motivar a la comunidad a profundizar en el análisis y la publicación de contenidos sobre estimación en áreas pequeñas. Adicionalmente, se presenta un resumen de los principales modelos utilizados en estimación en áreas pequeñas.

***Palabras clave*:** Colombia; Revisión; Estimación en áreas pequeñas.

[a]M.Sc. E-mail: andresortiz@usta.edu.co
[b]Ph.D. E-mail: cristiantellez@usta.edu.co
[c]M.Sc. E-mail: nicolas9294@gmail.com

# 1. Introduction

There is a growing demand for the use of statistical analysis in various fields of knowledge. One of the reasons for employing such analysis is to extract relevant information for deducing characteristics or attributes of the population and, in this way, optimize decision-making processes with scientific arguments. Currently, it is essential to have accurate and cost-effective statistics for decision-making. In the education sector, for example, it is useful to generate statistics on the quality of education across all grade levels at the lowest geographic disaggregation that a country has, in order to follow up and formulate targeted educational public policies to improve educational quality. Achieving this using conventional sampling techniques is currently impossible due to the very large sample size required, incurring high costs.

In the same meaning, in the case of household surveys, it would be very useful to measure the income level, poverty, or unemployment of households at the lowest levels of geographical disaggregation. However, for the same reason mentioned earlier, achieving this with commonly used sample designs is impossible. A solution to the previously mentioned problems lies in the use of Small Area Estimation (SAE) techniques, which, in general terms, do not alter the structure of traditional sample designs used in continuous surveys, instead, they focus on estimating parameters through mixed models, making inferences based on regression models rather than the sample design (Rao, 2003).

In the context of Latin America and specifically in Colombia, it is observed that the concept of small area estimation has not yet achieved significant dissemination or adoption among specialists in sampling from poll companies and government agencies, such as the *Departamento Administrativo Nacional de Estadística* (DANE). In this regard, it is crucial to provide the readers of this article with a precise contextual framework. This involves defining the concept of small area estimation in a concise and clear manner, in order to establish the necessary foundation for the proper understanding of this article.

In accordance with Pfeffermann (2002) and Rao (2005), small area estimation involves the use of statistical techniques to estimate parameters in small subpopulations of interest within a larger survey. The term 'small area' in this context typically refers to a geographic region, a county, a census district, a municipality, a department, or a school district where the sample size is not sufficient to obtain high-quality estimates using the proposed sampling design.

Ghosh (2020) mentions that a "small domain" refers to the intersection of demographic characteristics such as age, gender, ethnicity, among others. It is important to highlight that not only the size of the area at a geographic level but also the size of the target population within an area are determining factors for small area estimation. For instance, if one wishes to conduct a survey targeting a specific population with a specified precision, the sample size for a particular subpopulation may not be sufficient to achieve a similar precision. If a survey is conducted with a determined sample size to attain the desired precision, for example, at the regional level, there might not be sufficient resources to conduct a second survey with the same precision at the municipal level.

Taking into account the previous definition, the use of small area estimation methodologies in probabilistic sampling surveys has a significant impact worldwide due to the quality of the estimates obtained (Rao & Molina, 2015). For this reason, at present there are numerous articles, as Ghosh & Rao (1994), Pfeffermann et al. (2013), Ghosh (2020), Molina (2020) and Rao (2020), that highlight the importance of these methodologies and their evolution over time. It is important to note that the classic book on Rao (2003) establishes some of the concepts that emerged from the year 2004 and that are updated in the work of Rao & Molina (2015).

As mentioned Ghosh (2020) in it is discussion, Statistics obtained through small area estimation have been used for a long time, though not necessarily labeled as "small area". In reality, such statistics already existed in England in the 11th century and in Canada in the 17th century, based on censuses or administrative records. Demographers have long been using a variety of indirect methods for estimating population in small areas and other characteristics of interest during intercensal periods. Nowadays, administrative records are an important source of information for small area estimation. For more information on this topic, see Erciulescu et al. (2021).

In Colombia, there is a scarcity of academic works and high-impact surveys at the local or national level that have utilized small area estimation methodologies. For this reason, the purpose of this article is to review the works and applications that have been carried out in Colombia using these statistical techniques, as well as to summarize the main models used in the last 50 years. The goal is to encourage the use of these methodologies in Colombia and thereby improve the quality of estimates in the country.

This article is structured as follows. Section 2 presents the theoretical framework of some SAE models and includes references to other useful models in practice. Section 3 reviews different studies and evaluations where small area estimation methodologies have been employed in Colombia and the benefits this has had for public policy decision-making. Section 4 provides a practical application using the Fay-Herriot model to estimate the average household income at the municipal level in Cundinamarca, Colombia. Finally, Section 5 discusses the main conclusions derived from the practical case and the literature review.

## 2. Some Models Used in SAE

Without a doubt, small area estimation has gained considerable strength in recent years due to the good results it offers. Behind these results are different models used for estimation. The choice of the appropriate model depends on the parameter to be assessed and the availability of auxiliary information. In this regard, this section will highlight three very popular models in SAE and also mention some works that can be very useful in practice.

## 2.1. Fay Herriot Model

This model was introduced by Fay & Herriot (1979) and is characterized by being an area-level model. The birth of this model occurred to estimate per capita income in small areas of the U.S. This model links the indicators of interest for all areas $\delta_d$, $d = 1, \ldots, D$, with a vector of $p$ auxiliary variables $x_d$ following the next lineal regression model:

$$\delta_d \quad = \quad \boldsymbol{x}_d^T \boldsymbol{\beta} + u_d, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coeficients and $u_d$ is the the area random effect $d$. These effects are assumed to be independent and identically distributed, $u_d \overset{iid}{\sim} \left(0, \sigma_u^2\right)$. Since the true values of $\delta_d$ are not known, direct estimates $\left(\hat{\delta}_d^{Dir}\right)$ must be used for adjust the model. This estimation should be unbiased under the sampling design used. Using the direct estimator involves considering a sampling error $\epsilon_d$, which must be included in the model. Therefore, the model is rewritten as:

$$\hat{\delta}_d^{Dir} \quad = \quad \delta_d + \epsilon_d, \tag{2}$$

where, $\epsilon_d$ are the sampling errors in each area. It is assumed that these errors are independent of each other and independent of the random effects of the areas, $\epsilon_d \overset{ind}{\sim} \left(0, \sigma_{\epsilon,d}^2\right)$. In the practice, the variances of the sampling errors are not known and must be estimated from the variance of the direct estimate, $\hat{\sigma}_{\epsilon,d}^2 = var_\pi \left(\hat{\delta}_d^{Dir} \mid \delta_d\right)$, where $\pi(.)$ refers to the used sampling design. It is important to clarify that, for the purposes of estimation and MSE, it is not required for the errors or random effects to be normally distributed. However, normality can be assumed to simplify the estimation procedures.

When combining (1) and (2), the resulting mixed linear model, known as the Fay Herriot model, is given by:

$$\hat{\delta}_d^{Dir} \quad = \quad \boldsymbol{x}_d^T \boldsymbol{\beta} + u_d + \epsilon_d. \tag{3}$$

As shown (Rao & Molina, 2015), the best unbiased linear empirical predictor (EBLUP) under the Fay Herriot model is given by:

$$\tilde{\delta}_d^{FH} \quad = \hat{\gamma}_d \, \hat{\delta}_d^{Dir} + (1 - \hat{\gamma}_d) \, \boldsymbol{x}_d^T \hat{\boldsymbol{\beta}}, \tag{4}$$

where, $\hat{\gamma}_d = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{\epsilon,d}^2}$, the variance of the random effect $(\hat{\sigma}_u^2)$ can be estimated by some of the estimation methods such as Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML), among others. On the other hand, the vector of model parameters can be estimated as:

$$\hat{\boldsymbol{\beta}} = \left( \sum_{d=1}^{D} \hat{\gamma}_d x_d x_d^T \right)^{-1} \left( \sum_{d=1}^{D} \hat{\gamma}_d x_d \hat{\delta}_d^{Dir} \right). \tag{5}$$

Additionally, assuming that errors $u_d$ and $\epsilon_d$ follow normal distributions, the Mean Squared Error (MSE) of the Fay Herriot model predictor, according to Prasad & Rao (1990), is given by:

$$\hat{MSE}\left(\tilde{\delta}_d^{FH}\right) = g_{1d}\left(\hat{\sigma}_u^2\right) + g_{2d}\left(\hat{\sigma}_u^2\right) + 2g_{3d}\left(\hat{\sigma}_u^2\right), \tag{6}$$

with,

- $g_{1d}\left(\hat{\sigma}_u^2\right) = \hat{\gamma}_d\,\hat{\sigma}_{\epsilon,d}^2$
- $g_{2d}\left(\hat{\sigma}_u^2\right) = \left(1-\hat{\gamma}_d\right)^2 \boldsymbol{x}_d^T \left(\sum_{d=1}^{D}\left(\hat{\sigma}_u^2+\hat{\sigma}_{\epsilon,d}^2\right)^{-1}\boldsymbol{x}_d\boldsymbol{x}_d^T\right)^{-1}\boldsymbol{x}_d$
- $g_{3d}\left(\hat{\sigma}_u^2\right) = \left(1-\hat{\gamma}_d\right)^2\left(\hat{\sigma}_u^2+\hat{\sigma}_{\epsilon,d}^2\right)^{-1} v\bar{a}r\left(\hat{\sigma}_u^2\right)$, with $v\bar{a}r\left(\hat{\sigma}_u^2\right)$ is the asymptotic variance of the estimator $\hat{\sigma}_u^2$ which depends on the estimation method used.

## 2.2. Battese Harter Fuller Model

This model was proposed by Battese et al. (1988) to estimate corn and soybean production in U.S. counties. This model is at the unit level and allows linearly relating the observations of the variable of interest, $Y_{di}$ for the $i$th individual in area $d$, with the values of $p$ auxiliary variables for each area in the form:

$$Y_{di} = \boldsymbol{x}_{di}^T\boldsymbol{\beta} + u_d + \epsilon_{di},\ i = 1,\ldots,N_d, \tag{7}$$

where $\boldsymbol{\beta}$ is the vector of coefficients for the auxiliary variables, $u_d$ is the random effect of area $d$ and $\epsilon_{di}$ is the unit level error. The random effects are considered independent of the errors, with $u_d \overset{iid}{\sim}\left(0,\sigma_u^2\right)$ and $\epsilon_{di} \overset{ind}{\sim}\left(0,k_{di}^2\sigma_\epsilon^2\right)$, being $k_{di}$ known constants that represent possible heteroscedasticity. The EBLUP under the (7) model is given by:

$$\hat{Y}_{di}^{EBLUP} = \boldsymbol{x}_{di}\hat{\boldsymbol{\beta}} + \hat{u}_d, \tag{8}$$

where $\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\boldsymbol{X}_{ds}\boldsymbol{V}_{ds}^{-1}\boldsymbol{X}_{ds}^T\right)^{-1}\left(\sum_{d=1}^{D}\boldsymbol{X}_{ds}\boldsymbol{V}_{ds}^{-1}\boldsymbol{y}_{ds}\right)$, $\boldsymbol{X}_d = \begin{pmatrix}\boldsymbol{X}_{ds}\\ \boldsymbol{X}_{dr}\end{pmatrix}$, $\boldsymbol{V}_d = \hat{\sigma}_u^2\mathbf{1}_{N_d}\mathbf{1}_{N_d}^T + \hat{\sigma}_\epsilon^2\mathbf{I}_{N_d}$ and

$$\boldsymbol{V}_d = \begin{pmatrix}\boldsymbol{V}_{ds} & \boldsymbol{V}_{dsr}\\ \boldsymbol{V}_{drs} & \boldsymbol{V}_{dr}\end{pmatrix} \tag{9}$$

where, $X_{ds}$ are the covariates in the domains observed in the sample and $X_{dr}$ are the covariates in the domains not observed in the sample. $V_{ds}$ is the variance-covariance matrix in the domains observed in the sample. In general terms, the subscript $r$ represents what is not observed in the sample and $s$ represents what is observed in the sample.

The estimation of the random effect using restricted maximum likelihood is given by, $\hat{u}_d = \hat{\gamma}_d \left(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T \hat{\boldsymbol{\beta}}\right)$, $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\epsilon^2}{a_{d.}}}$, $\bar{y}_{da} = a_{d.}^{-1} \sum_{i \in s_d} a_{di} Y_{di}$, $\bar{\boldsymbol{x}}_{da} = a_{d.}^{-1} \sum_{i \in s_d} a_{di} \boldsymbol{x}_{di}$, $a_{di} = k_{di}^{-1}$ and $a_{d.} = \sum_{i \in s_d} a_{di}$.

On the other hand, for the estimation of the Mean Squared Error, is used the parametric bootstrap technique described in González-Manteiga et al. (2008).

## 2.3. Best Empirical Predictor Under the Nested Errors Model

The best predictor (best/Bayes predictor, BP) based on the nested errors model was proposed by Molina & Rao (2010) to estimate general nonlinear indicators. For the adjustment of the BP model, it is assumed that the response variable has a logarithmic relationship defined by the following expression:

$$Y_{di} = \log\left(E_{di} + c\right), \tag{10}$$

where $E_{di}$ refers to a bijective transformation of purchasing power with distribution approximately normal and $c > 0$ .In this way, the model has the following structure:

$$Y_{di} = \boldsymbol{x}_{di}^T \boldsymbol{\beta} + u_d + \epsilon_{di},\ i = 1, \ldots, N_d, d = 1, \ldots, D. \tag{11}$$

Unlike the unit-level model of Battese et al. (1988), the model shown in (11) assumes normality for the random effects $u_d$ and for the errors $\epsilon_{di}$. It is also assumed that the vectors of variables for each area are independent. For a general indicator defined as a function of $\boldsymbol{y}_d$ (vector of data observed in the domains), that is to say, $\delta_d = \delta_d\left(\boldsymbol{y}_d\right)$, the best predictor is the one that minimizes the mean squared error and is given by:

$$\tilde{\delta}_d^{\,B}\left(\boldsymbol{\theta}\right) = E_{y_{dr}}\left[\delta_d\left(\boldsymbol{y}_d\right) \mid \boldsymbol{y}_{ds}; \boldsymbol{\theta}\right]. \tag{12}$$

The expected value is taken with respect to the distribution of the vector of out of sample values $\boldsymbol{y}_{dr}$ in the domain $d$ given the values in the sample $\boldsymbol{y}_{ds}$. This conditioned distribution depends on the true values of the model parameters for $\boldsymbol{\theta}$. Replacing $\boldsymbol{\theta}$ with a consistent estimator $\hat{\boldsymbol{\theta}}$ in the best predictor (12), the best empirical predictor (empirical best/Bayes, EB) is obtained, $\hat{\delta}_d^{EB} = \tilde{\delta}_d^B\left(\hat{\boldsymbol{\theta}}\right)$. For the estimation of the mean squared error, Molina & Rao (2010) propose an approximation via bootstrap.

In addition to the previous models, as mentioned by Tellez Piñerez (2020) the use of classical small area estimation models, for example, those proposed by Fay & Herriot (1979) or Battese et al. (1988) are not efficient when estimating parameters such as the proportion because they cannot guarantee that the predictions found are within the interval $(0, 1)$ this is why other approaches, such as Bayesian ones (Bayes empirical estimator) are widely used in practice (Rao & Molina, 2015).

On the other hand, in the work done by Salvati et al. (2012), they propose an estimator for the distribution function in small areas, through which proportions,

medians, modes, etc., can be estimated with auxiliary information at the area level. The model they use in this article to relate auxiliary information to the target or of interest variable is non parametric.

## 3. Small Area Estimation in Colombia

The small area estimation tools allow for better quality estimates of the parameters of interest than those based solely on the sampling design, as long as auxiliary information associated with the parameter of interest is available (Molina et al., 2007). This philosophy has promoted in recent years the production of official statistics at disaggregated levels; for example, in the labor force survey in Canada and for the measurement of labor force and agricultural production in the National Agricultural Statistics Service in the United States (STATISTICS, 2006). In Latin American countries, small area estimation is being used for measuring poverty in the municipalities of Mexico in 2020 and the percentage of people in income poverty in communes in Chile; however, in Colombia, it has not yet been regularly used for the production of official statistics by entities conducting nationwide surveys, whose results are the fundamental basis for the generation of public policy, such as DANE.

As was mentioned in the introduction, small area estimation techniques have been used since the 17th century in England; however, in Colombia, the first recorded work took place at *Instituto Colombiano para la Evaluación de la Educación* (ICFES) in 2016. In this case, the Fay Herriot model was employed to enhance school-level estimates generated by the Saber 3°, 5° and 9° exams (ICFES, 2016). In the same line of work with *Pruebas Saber* data, there are studies by Herrera & Zea (2019), where a comparison between direct and indirect estimators is conducted; in Flórez Gutiérrez & Gutiérrez (2019), where a comparison of models at the area and unit levels is developed to estimate the average score of the mathematics test at the level of *Entidades Territoriales Certificadas* (ETC) of the Saber 11 test in 2017; the work of Gutiérrez Pérez (2019) incorporates space-time effects into the Fay Herriot model to estimate the results of the Mathematics module of the Saber 11 test for the year 2018, calendar A. Guerrero & Trujillo (2019) work proposes a method for estimating regression coefficients in a model in the presence of missing values and applies their results using information from the main associated factors study conducted in 2012 by ICFES.

Additionally, in the field of item response theory, in Colombia has been developed in Colombia a doctoral thesis by Tellez Piñerez (2020), proposing a methodology that incorporates item response theory in small area estimation in the presence of missing data. In this work, an unbiased estimator is suggested for the average ability of students and a Bayesian estimator based on the beta distribution, for the proportion of students with a particular characteristic. Theoretical developments are complemented by two applications of this methodology. The first one uses the mathematics results from the PISA test presented in 2015, and the second one is based on the results of the Saber 3°, 5° and 9° tests applied by ICFES in Colombia. Given the theoretical developments found in this thesis,

three additional works were derived; The first one was developed by Triviño et al. (2020), who applies the estimator presented in Tellez Piñerez (2020) to improve the quality of the estimates obtained with the sampling design used in the TIMSS 2015 tests. The second one was developed by Jiménez Coley (2023), in which he computationally explores the effect of including the spatial factor in the estimator proposed by Tellez Piñerez (2020), finding that, in some cases, it improves the estimates. Finally, Piñerez et al. (2021) made an application in the Saber 3°, 5° and 9° tests in the mathematics area.

In other line of work, small area estimation has been used for estimating indicators such as poverty lines, unemployment rates, and average incomes of municipalities or localities. These studies were conducted by Zea & Ortiz (2018), who employed the Fay Herriot model to estimate the unemployment rate and average income levels in the municipalities of Cundinamarca using the *Encuesta Multipropósito* 2014 and auxiliary socio-demographic and economic information. In the same line, Castañeda et al. (2020) proposes a procedure based on the Fay Herriot model to estimate the average income of households at the locality level. The *Encuesta Multipropósito* is considered, along with auxiliary economic and demographic variables such as the multidimensional poverty index, the valuation index and population projections. Additionally, Mendoza & Zea (2020) work explores the use of principal components as input to improve model fitting and estimate the unemployment rate and average income of municipalities in Cundinamarca. Lastly, Gómez Pinto (2023) uses small area estimation, information from the *Encuesta Multipropósito* of 2017, and compositional data to build a multivariate Fay-Harriot model along with parametric bootstrap MSE to estimate monetary poverty in the 116 municipalities of Cundinamarca.

In the field of health and wellfare, the work of Ramirez Vargas et al. (2018) stands out as they generate estimates of the total crack cocaine consumers in each Zonal Planning Unit (UPZ) using the spatial Fay-Herriot model to focus the attention and surveillance efforts of the District Health Secretary. Additionally, in Romero Romero (2018) work, the Fay Herriot model with spatial effects is used to disaggregate teenage pregnancy rates for all municipalities in the country. This involves using departmental rates produced by the 2015 *Encuesta Nacional de Demografía y Salud* (ENDS) of Colombia and an adaptation of the spatial Fay Herriot model that incorporates municipal-level pregnancy rate information and administrative data on the characteristics of the municipalities.

At the same time, universities have also promoted the use of small area estimation methodologies in various fields and have generated innovations to classical tools. Evidence of this is seen in the 8 mentioned works produced at the Faculty of Statistics of Santo Tomás University and the 4 works generated by the Department of Statistics at the National University of Colombia in Bogotá. Among them is the work of Bernal & González (2017), who proposed a method to estimate a proportion in small areas using the beta distribution, a doubly differentiable generalized mixed model and finite mixtures. Additionally, Velez & Polo (2019) proposed a modification to the bootstrap procedure to estimate mean squared error considering conditional Pearson residuals.

At the governmental level, there is also evidence of the development and implementation of theoretical-practical exercises that include the use of such estimators, considering that in many instances of decision-making and measuring the effectiveness or feasibility of public policies, there are not ample resources for conducting primary data collection at a detailed geographical level to ensure high levels of precision in estimations. This is evident in the study conducted by the *Secretaria Distrital de Desarrollo Económico* (SDDE) of Bogotá in 2024, where they applied the Fay Herriot model at the level of UPZ to estimate the total number of employees in formal enterprises in each UPZ. This estimation resulted in very low estimation errors and, furthermore, paved the way for future research in the district or nation.

In this way, the work carried out by the Departamento Nacional de Planeación (2019), as they were responsible for estimating the Multidimensional Poverty Index (MPI) at the municipal level using the Fay-Herriot model based on information from SISBEN IV, allowing to obtain an estimation of the magnitude of households and individuals in poverty. Additionally, is highlighted the joint analysis conducted by the *Departamento para la Prosperidad Social* (DPS) and the DANE in 2019 in collaboration with CEPAL, managing to build the map of monetary poverty at the municipal level using information reported by the *Gran Encuesta Integrada de Hogares* (GEIH) and the *Censo Nacional de Población y Vivienda* (CNPV). This exercise incorporates the model of the best empirical predictor with nested errors, enabling a deeper understanding of the poverty situation in the country and the existing geographic phenomena based on the social and economic impact that this generates through (Departamento para la Prosperidad Social. & CEPAL., 2019).

Moreover, various consulting firms in the social sector have also ventured into the use of small area estimation. An example of this is the extension carried out by the firm Inclusión S.A.S, who, based on the results of the previously mentioned poverty map, updated and expanded the exercise to obtain a more recent estimation of the same indicator in 2020 (Inclusión S.A.S, 2020).

## 4. Application

In this section, it is present the results of an application exercise of the Fay Herriot model. The exercise involves estimating the average household income in 38 municipalities of the Cundinamarca department in Colombia using data from the *Encuesta Multipropósito* conducted in 2017 in the region. This survey is designed to measure statistical information on the social, economic, and environmental conditions of households and residents of Bogotá and 37 municipalities in Cundinamarca. It captured information from 320 thousand people in 109 thousand households. In Bogotá, 77 thousand households were surveyed, representing approximately 222 thousand people, and in the 37 municipalities of Cundinamarca, 32 thousand households were surveyed, representing approximately 98 thousand people.

To start, it is necessary to obtain the directly estimations of average incomes in each of the 38 municipalities of Cundinamarca are outlined in this section. For this purpose, the Hájek estimator is used, taking the data collected from the *Encuesta Multipropósito*. This estimator is preferred over the Horvitz-Thompson estimator, considering its greater efficiency for average estimation in domains like those analyzed in this application. Direct estimates are generated using the `survey` library of the R software. These estimates, along with sample sizes and their coefficients of variation, are presented in the columns $n$, $\hat{\bar{Y}}^{Dir}$ and $Cv\left(\hat{\bar{Y}}^{Dir}\right)$ of the Table A1. As a result, it can be observed that the coefficients of variation obtained are small in the majority of cases, primarily due to the fact that the sample sizes in each municipality are large.

Given that the sampling design of the *Encuesta Multipropósito* involves complex sampling, it was necessary to resort to a resampling technique or variance approximation for the precise estimation of the variance of the direct estimate. In this context, the *last cluster* technique was chosen, which is popular among researchers as it directly incorporates the final sampling weights or expansion factors provided by the DANE. The application of this technique involves selecting clusters in the first stage of sampling, followed by the choice of all units within the last selected cluster. This approach, by simplifying the selection process and leveraging weighted information, is crucial for obtaining accurate variance estimates within the framework of complex sampling designs. For more details, refer to Lumley (2011).

Now, for the Fay Herriot model adjustment, 3 auxiliary variables obtained from 3 different sources are used: The first one is the amount of contributions to the Integrated Contribution Settlement Form (PILA) at the municipal level, this information is downloaded from the portal: `http://datlascolombia.com/#/downloads`; the second one is the percentage of households in the SISBEN system that have access to an oven for the year 2016, this information is downloaded from the open data portal of the Colombian government at the following link `https://www.datos.gov.co/dataset/Acceso-a-bienes-por-hogar-seg-n-municipio-2016-en-/ktkb-ymp2`; the third auxiliary variable is the distance from the municipality to the city of Bogotá, which is obtained for each municipality from the portal: `ttp://co.lasdistancias.net/`. These variables have been chosen because, on the one hand, they do not come from any survey and are not associated with measurement errors, and on the other hand, they have a moderate correlation with the results of the direct income estimation. The utilized model can by written in the following form:

$$\hat{\delta}_d^{Dir} = \beta_0 + \beta_1 x_{1d} + \beta_2 x_{2d} + \beta_3 x_{3d} + u_d + \epsilon_d, \tag{13}$$

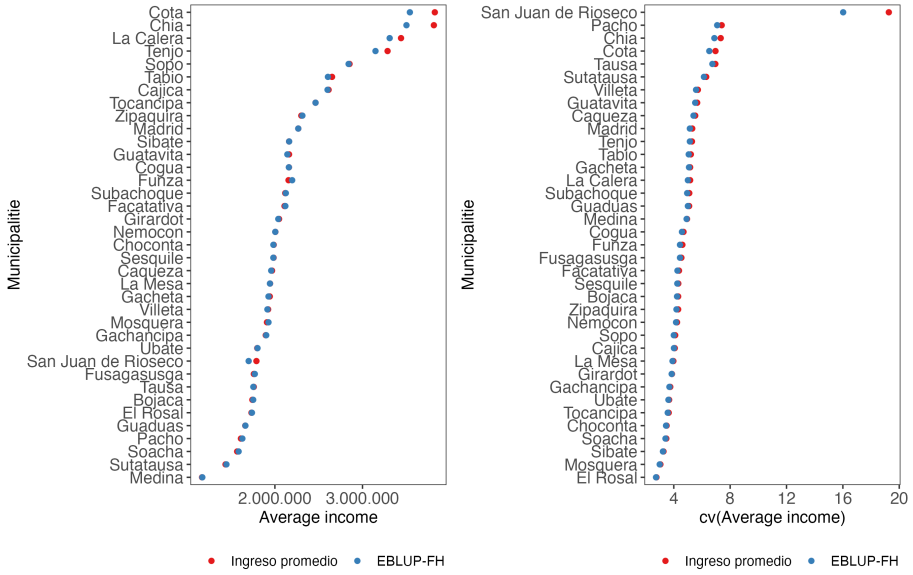
where $x_{1d}$ represents the quantity of parafiscal contributions made to the PILA, $x_{2d}$ represents the percentage of households per municipality that have access to a household oven, and $x_{3d}$ represents the distance of each municipality to Bogotá, the capital of the department.

Fay Harriot model is fitted using the function `mseFH` of the `sae` library of the statistical R software. The EBLUP results, obtained whit the expression (4),

are displayed in the column $\hat{\bar{Y}}^{FH}$ of the Table A1 along with the mean squared error, obtained with the expression (6), and it is estimated coefficient of variation $Cv\left(\hat{\bar{Y}}^{FH}\right) = \sqrt{\hat{MSE}\left(\hat{\bar{Y}}^{FH}\right)}/\hat{\bar{Y}}^{FH}$.

It can be observed in the Figure 1 (left) that the estimates generated by the Fay Herriot model are very similar to those generated by the direct estimator, and additionally, there is a decrease in the obtained sampling error (right).

FIGURE 1: Estimations (left) and sampling errors (right) for the average income for municipality.




## 5. Conclusions

As is well known, the precision of estimates in surveys decreases when making inferences in unplanned subgroups in the sampling design. However, the information collected in the survey can be used as initial input to produce estimates based on available auxiliary information from censuses or administrative records. This is directly related to the 2030 Agenda for Sustainable Development and its policy of leaving no one behind (left no one behind), and reaching the most vulnerable first presents many challenges and opportunities for the statistical community. Data disaggregation is essential to understand if the fruits of development are benefiting the entire spectrum of society, including the most vulnerable and the most disadvantaged. This is where the use of small area estimation becomes crucial to measure the characteristics of all required disaggregations.

As a result of this review, small area estimation techniques, while increasingly used in Colombia mainly in academic developments (recently), still lack broader dissemination and utilization in government entities that possess the auxiliary information enabling model adjustment and, consequently, estimations in these areas. In this regard, with the current developments in SAE, it is now possible to extend these techniques to the largest continuous surveys in the country, such as the GEIH, the *Encuesta de Calidad de Vida* (ECV), the *Encuesta Multipropósito*, among others, at the municipal level (without a significant increase in costs), allowing for a more detailed and precise understanding of the country's reality.

Now, it is important to note that, despite some limitations regarding the availability and quality of necessary data for the application of these techniques, various strategies have been developed to address these issues. These include the use of spatial models and the combination of different data sources. Furthermore, advancements in modeling in SAE allow for adjusting models using covariates with measurement errors. In other words, auxiliary variables taken from the results of other probabilistic surveys (Burgard et al., 2020). This expands the use of SAE models in achieving higher quality estimates for a greater number of disaggregation levels.

Regarding the models used for small area estimation, three main models were identified: the Fay Herriot model, the Battese Hater Fuller model and the best empirical predictor under the nested error model. Each of these models has its own advantages and limitations, and the selection depends on the type of parameter to be estimated and the characteristics of the available data. The Bayesian empirical estimator is also significantly emphasized, involving the use of Bayesian models for its implementation.

Finally, with the aim of promoting and supporting the use of small area estimation in future research in Colombia, it is imperative that government and state institutions adopt a policy of openness in information disclosure. This involves the comprehensive release of available data, considering variables inherent to the sampling design, such as Primary Sampling Units (PSUs) and strata. By doing so, the possibility of faithfully replicating the original sampling design is facilitated, a crucial step in generating the required inputs for model adjustment and, consequently, parameter estimation in areas where the sampling design was not able to reach. The release of design variables will not only strengthen the available databases but also enhance the quality and reliability of results obtained through SAE approaches, providing a solid foundation for evidence-based and informed decision-making.

[Received: February 2024 — Accepted: May 2024]

## References

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988), 'An error components model for prediction of county crop areas using survey and satellite data', *Journal of the American Statistical Association* **83**(401), 28–36.

Bernal, L. K. & González, L. M. (2017), Distribución beta para modelar proporciones en áreas pequeñas, Master's thesis, Universidad Nacional de Colombia-Sede Bogotá.

Burgard, J. P., Esteban, M. D., Morales, D. & Pérez, A. (2020), 'A fay-herriot model when auxiliary variables are measured with error', *Test* **29**, 166–195.

Castañeda, J., Tellez, C. & Fuquene, J. (2020), 'An alternative for the average income estimation using small area methods', *BEIO* .

Departamento Nacional de Planeación (2019), Estimación del índice de pobreza multidimensional mediante modelos de estimación en áreas pequeñas, Technical report, Departamento Nacional de Planeación.

Departamento para la Prosperidad Social. & CEPAL. (2019), Mapa de pobreza para los municipios en colombia 2018 -2019, Technical report, Departamento para la Prosperidad Social. and CEPAL.

Erciulescu, A. L., Franco, C. & Lahiri, P. (2021), *Use of administrative records in small area estimation*, John Wiley & Sons.

Fay, R. E. & Herriot, R. A. (1979), 'Estimates of income for small places: An application of James-Stein procedures to census data', *Journal of the American Statistical Association* **74**(366), 269–277.

Flórez Gutiérrez, J. A. & Gutiérrez, A. (2019), 'Comparación entre los modelos de unidad y área para la estimación en áreas pequeñas del puntaje de matemáticas en las entidades territoriales certificadas para la prueba saber 11'.

Ghosh, M. (2020), 'Small area estimation: its evolution in five decades', *STATISTICS* **41**.

Ghosh, M. & Rao, J. (1994), 'Small area estimation: An appraisal', *Statistical Science* **9**(1), 55–83.

Gómez Pinto, H. F. (2023), 'Estimación de la pobreza monetaria para los municipios de cundinamarca vía estimación en áreas pequeñas.'.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. & Santamaría, L. (2008), 'Bootstrap mean squared error of a small-area eblup', *Journal of Statistical Computation and Simulation* **78**(5), 443–462.

Guerrero, S. & Trujillo, L. (2019), Estimación de los coeficientes de regresión en áreas pequeñas utilizando valores plausibles en muestras probabilísticas, Master's thesis, Universidad Nacional de Colombia-Sede Bogotá.

Gutiérrez Pérez, M. E. (2019), 'Modelo fay-herriot espacio-temporal para la estimación de los resultados del módulo de matemáticas de la prueba saber-11 2018 calendario a desagregada por departamento'.

Herrera, A. & Zea, J. (2019), 'Comparación de estimadores directos e indirectos en estimación de áreas pequeñas'.

ICFES (2016), Estimación en áreas pequeñas del rendimiento cognitivo medio en el módulo de matemáticas de los estudiantes de quinto de primaria en las escuelas colombianas utilizando imputación múltiple, Technical report, ICFES.

Inclusión S.A.S (2020), Una actualización al mapa de pobreza para los municipios en colombia 2020, Technical report, Inclusión S.A.S.

Jiménez Coley, C. (2023), 'Estimación del rendimiento medio en las pruebas pisa 2018. un enfoque espacial desde la estimación en áreas pequeñas'.

Lumley, T. (2011), *Complex surveys: a guide to analysis using R*, John Wiley & Sons.

Mendoza, D. & Zea, J. (2020), 'Estimación de la tasa de desempleo e ingreso medio del departamento de cundinamarca utilizando SAE'.

Molina, I. (2020), 'Discussion of "small area estimation: Its evolution in five decades", by malay ghosh', *Statistics* **41**.

Molina, I. & Rao, J. (2010), 'Small area estimation of poverty indicators', *Canadian Journal of Statistics* **38**(3), 369–385.

Molina, I., Saei, A. & Lombardía, M. J. (2007), 'Small area estimates of labour force participation under a multinomial logit mixed model', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(4), 975–1000.

Pfeffermann, D. (2002), 'Small area estimation-New developments and directions', *International Statistical Review* **70**(1), 125–143.

Pfeffermann, D. et al. (2013), 'New important developments in small area estimation', *Statistical Science* **28**(1), 40–68.

Piñerez, C. T., Arias, I. R., Gómez, S. G. & Oyola, L. T. (2021), 'Estimation of educational establishments performance in saber 5° tests in colombia. an approach from small area estimation', *Boletın de Estadıstica e Investigación Operativa BEIO* p. 169.

Prasad, N. N. & Rao, J. N. (1990), 'The estimation of the mean squared error of small-area estimators', *Journal of the American Statistical Association* **85**(409), 163–171.

Ramirez Vargas, N. et al. (2018), 'Un modelo de áreas pequeñas para la estimación de la frecuencia de consumo de bazuco al interior de las unidades de planeamiento zonal de Bogotá'.

Rao, J. (2005), 'Inferential issues in small area estimation: some new developments', *Statistics in Transition* **7**(3), 513–526.

Rao, J. (2020), 'Discussion of "small area estimation: Its evolution in five decades", by malay ghosh', *Statistics* **41**.

Rao, J. N. (2003), *Small Area Estimation*, Wiley, New York.

Rao, J. N. & Molina, I. (2015), *Small Area Estimation*, 2 edn, John Wiley & Sons.

Romero Romero, J. J. (2018), 'Un modelo fay-herriot espacial para estimar la tasa de embarazo adolescente en colombia'.

Salvati, N., Chandra, H. & Chambers, R. (2012), 'Model-based direct estimation of small-area distributions', *Australian & New Zealand Journal of Statistics* **54**(1), 103–123.

STATISTICS, A. B. O. (2006), 'A guide to small area estimation-version 1.1. internal abs document', *http://www. nss. gov. au/nss/home. NSF/pages/Small+ Areas+ Estimates* .

Tellez Piñerez, C. F. (2020), Estimación de áreas pequeñas utilizando imputación múltiple en modelos logísticos de tres parámetros, PhD thesis, Doctorado en Ciencias-Estadística, Universidad Nacional, Bogotá, Colombia.

Triviño, A. F. P., Piñerez, C. F. T. & Oyola, L. T. (2020), 'Estimación de los resultados en matemáticas y ciencias de las pruebas timss 2015: un nuevo enfoque desde la metodología de áreas pequeñas', *Comunicaciones en Estadística* **13**(2), 62–77.

Velez, D. & Polo, M. (2019), Una adaptación del procedimiento bootstrap en la estimación del error cuadrático medio en áreas pequeñas con aplicación a datos colombianos., Master's thesis, Universidad Nacional de Colombia, Bogotá.

Zea, J. F. & Ortiz, F. (2018), 'Small area estimation methodology (SAE) applied on bogota multipurpose survey (EMB)', *Romanian Statistical Review* (1).

# Appendix

Table A1: Results for municipality

| Municipio | $n$ | $\hat{\bar{Y}}^{Dir}$ | $\hat{\bar{Y}}^{FH}$ | $Cv\left(\hat{\bar{Y}}^{Dir}\right)$ | $Cv\left(\hat{\bar{Y}}^{FH}\right)$ |
|---|---|---|---|---|---|
| Soacha | 1167 | 1569158.40 | 1585873.21 | 3.47 | 3.41 |
| Medina | 1104 | 1168317.92 | 1171757.66 | 4.94 | 4.91 |
| Fusagasusga | 1060 | 1759434.47 | 1770547.24 | 4.54 | 4.44 |
| Funza | 1052 | 2155588.61 | 2197191.73 | 4.61 | 4.44 |
| Mosquera | 1052 | 1909421.16 | 1926791.05 | 3.05 | 3.00 |
| Facatativa | 1046 | 2109775.25 | 2119535.12 | 4.37 | 4.26 |
| Cajica | 1037 | 2612933.24 | 2600837.92 | 4.07 | 4.01 |
| Ubate | 1009 | 1798447.87 | 1801684.02 | 3.67 | 3.62 |
| Girardot | 991 | 2047624.66 | 2038690.81 | 3.87 | 3.85 |
| Madrid | 983 | 2267032.79 | 2266636.63 | 5.31 | 5.14 |
| Zipaquira | 982 | 2301532.35 | 2313770.15 | 4.31 | 4.19 |
| Sibate | 981 | 2161960.00 | 2162548.21 | 3.27 | 3.23 |
| Sesquile | 954 | 1985047.82 | 1982041.04 | 4.32 | 4.25 |
| Sopo | 929 | 2851880.26 | 2841256.24 | 4.10 | 4.00 |
| El Rosal | 921 | 1732589.71 | 1737961.44 | 2.77 | 2.74 |
| Chia | 920 | 3814449.76 | 3501630.58 | 7.33 | 6.88 |
| Villeta | 903 | 1922691.90 | 1914414.27 | 5.71 | 5.58 |
| Cota | 900 | 3825328.53 | 3540593.30 | 6.96 | 6.52 |
| Choconta | 899 | 1986276.80 | 1982798.05 | 3.49 | 3.45 |
| San Juan de Rioseco | 889 | 1789287.90 | 1700510.57 | 19.26 | 16.02 |
| Tabio | 863 | 2652326.23 | 2605320.42 | 5.21 | 5.06 |
| Tocancipa | 854 | 2463669.54 | 2463921.03 | 3.64 | 3.57 |
| Gachancipa | 846 | 1895494.99 | 1901019.70 | 3.75 | 3.69 |
| La Calera | 836 | 3439726.65 | 3309951.00 | 5.16 | 5.00 |
| Pacho | 832 | 1612026.77 | 1628812.51 | 7.40 | 7.08 |
| Guaduas | 802 | 1661773.28 | 1663429.60 | 5.09 | 5.00 |
| Gacheta | 794 | 1941588.15 | 1926894.23 | 5.16 | 5.08 |
| La Mesa | 790 | 1944795.09 | 1943633.05 | 3.97 | 3.91 |
| Bojaca | 777 | 1742701.94 | 1752271.49 | 4.31 | 4.23 |
| Nemocon | 758 | 2005354.98 | 2004387.01 | 4.23 | 4.16 |
| Cogua | 747 | 2160422.25 | 2159495.09 | 4.69 | 4.58 |
| Subachoque | 675 | 2118053.65 | 2124639.55 | 5.10 | 4.95 |
| Tenjo | 670 | 3286038.97 | 3148129.41 | 5.29 | 5.14 |
| Guatavita | 667 | 2161448.16 | 2141973.87 | 5.66 | 5.52 |
| Caqueza | 642 | 1967425.31 | 1957913.22 | 5.52 | 5.40 |
| Sutatausa | 508 | 1436625.21 | 1446891.04 | 6.30 | 6.14 |
| Tausa | 246 | 1758951.15 | 1753847.56 | 6.95 | 6.74 |