UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2018

Comparación de estimadores directos e indirectos en estimación de áreas pequeñas

Comparison of direct and indirect estimators in estimation of small areas

Edgar Andres Herrera Herrera[a]
edgarherrera@usantotomas.edu.co

José Fernando Zea Castro[b]
josezea@usantotomas.edu.co

Resumen

La Estimación en áreas pequeñas permite obtener resultados más precisos y confiables en dominios cuyos tamaños de muestra son nulos o cercanos a cero. Diferentes estimadores que dependen del diseño de muestreo como los estimadores directos, estimadores propuestos por la teoría para obtener mejores resultados en dominios donde el tamaño de muestra es pequeño, como los estimadores sintéticos y estimadores que usan modelos lineales mixtos para áreas pequeñas, son los tópicos que se presentan en este trabajo de grado, donde se observa el comportamiento de cada uno de ellos, para ser comparados a través del error cuadrático medio y de esta manera obtener conclusiones a cerca de su desempeño al momento de ser utilizados en dichas áreas donde el tamaño de muestra es pequeño.

***Palabras clave*:** Estimación en áreas pequeñas, dominios, tamaño de muestra, estimadores, estimadores directos, estimadores sintéticos, modelos lineales mixtos, error cuadrático medio.

Abstract

Estimating in small areas allows obtaining more accurate and reliable results in domains whose sample sizes are zero or close to zero. Different estimators that depend on the sampling design as the direct estimators, estimators proposed by the theory to obtain better results in domains where the sample size is small, such as synthetic estimators and estimators using mixed linear models for small areas, are the topics which are presented in this degree work, where the behavior of each one of them is observed, to be compared through the mean square error and in this way to obtain conclusions about their performance when they are used in those areas where the sample size is small.

***Keywords*:** Estimation in small areas, domains, sample size, estimators, direct estimators, synthetic estimators, mixed linear models, mean square error.

[a]Estudiante Estadística Universidad Santo Tomás
[b]Docente Estadística Universidad Santo Tomas

1. Antecedentes

Este apartado consta de la mensión de algunos estudios anteriores referidos al tema de estimación en áreas pequeñas y que son punto de referencia para el desarrollo del tema del presente trabajo de grado.

Estimación en áreas pequeñas es una rama de la estadística que trata el problema de estimar parámetros de subconjuntos llamados áreas pequeñas o dominios de la población a partir de muestras e información auxiliar. Debido a la falta de precisión de los estimadores directos al momento de estimar los parámetros correspondientes en dichos dominios o áreas pequeñas, autores como Ghosh, Rao, Jiang y Lahiri, han desarrollado nuevos procedimeintos de estimación basados en modelos lineales mixtos.

Los modelos de regresión lineal mixta incrementan la eficiencia de la información usada en el proceso de estimación estableciendo nexos o relaciones entre todas las observaciones de la muestra.

Los modelos de este estilo se han usado en Estados Unidos para estimar los ingresos per cápita en áreas pequeñas, en Canadá para estimar conteos no incluidos en el censo y para estudios de pobreza en población escolar. Pérez (2008)

En los últimos años, la demanda de estadísticas de áreas pequeñas ha aumentado considerablemente en todo el mundo. Esto se debe, entre otras cosas, a su creciente uso en la formulación de políticas y programas, en la asignación de fondos gubernamentales y en la planificación regional. La demanda del sector privado también ha aumentado porque las decisiones comerciales, en particular las relacionadas con las pequeñas empresas, dependen en gran medida de las condiciones socioeconómicas, ambientales y de otras índoles locales.

La estimación de áreas pequeñas es de particular interés para las economías en transición en los países de Europa central y oriental y en los países de la ex Unión Soviética. En la década de 1990, estos países se han alejado de la toma de decisiones centralizadas. Como resultado, las encuestas por muestreo ahora se utilizan para producir estimaciones para áreas grandes, así como para áreas pequeñas. Impulsadas por la demanda de estadísticas de áreas pequeñas, se celebró en Varsovia, Polonia, una Conferencia Científica Internacional sobre estadísticas de áreas pequeñas y diseños de encuestas en 1992, asi como también, se han llevado a cabo diferentes conferencias sobre SAE (Small Area Stimation) en diversos paises desde el año 2001 hasta el año 2017. J.N.K. Rao (2003)

El proyecto europeo SAMPLE (Small Area Methods for Poverty and Living Condition Estimate) es un proyecto de investigación financiado por la Comisión Europea en el marco del Seventh Framework Programme (FP7), que comenzó en marzo de 2008 y finalizó en marzo de 2011. Nueve socios de cuatro de Los diferentes países europeos, incluido España, participaron en el proyecto. Los socios españoles son la Universidad Carlos III de Madrid y la Universidad Miguel Hernández de Elche. El objetivo de este proyecto fue identificar y desarrollar nuevos indicadores y modelos de desigualdad y pobreza con atención a la exclusión social y la privación, así como el desarrollo y la implementación de modelos, medidas y procedimientos para la estimación de áreas pequeñas de indicadores tradicionales y nuevos. Este objetivo se logra utilizando datos de las encuestas europeas sobre ingresos y condiciones de vida con la ayuda de bases de datos administrativas locales. Molina & Morales (2009)

2. Problema

Son escasos Los estudios en donde se pueda evidenciar la implementación de los estimadores directos e indirectos (tanto basados en el diseño como en modelos) haciendo uso de los desarrollos de paquetes y

rutinas que han surgido en años recientes (survey, sae, entre otros paquetes de R enfocados al muestreo). Por otra parte, la comparación de estos estimadores en términos de error cuadrático medio no suelen detallarse para los diferentes estimadores disponibles en la literatura.

En este estudio se quiere responder a la pregunta de qué estimador y enfoque utilizar para estimación de áres pequeñas en un estudio de caso real haciendo uso de los desarrollos computacionales más recientes.

3. Introducción

La estimación en áreas pequeñas es una técnica estadística que pretende obtener estimaciones poblacionales de parámetros de interés en dominios cuyos tamaños de muestra sean nulos o cercanos a cero. Esta técnica SAE *(Small Area Estimation)*, que incorpora información auxiliar en la construcción de estimadores con el objetivo de obtener resultados precisos, no ha tenido un impulso notorio en cuanto a su utilización por entidades gubernamentales de estadística en colombia, con lo cual, se quiere dar un acercamiento ilustrando las ventajas de su utilización por medio de la comparación de estimadores directos e indirectos que enmarcan un desempeño en la estimación de áreas pequeñas.

Algunos ejemplos de aplicación de la herramienta SAE, se pueden encontrar en la Universidad Carlos III de Madrid, donde los investigadores Isabel Molina y Domingo Morales, realizan estimaciones de indicadores de probreza en Europa; El Instituto Nacional de Estadística de Madrid, estudia el comportamiento de la población española a través de aplicación en áreas pequeñas cuyo estudio fue estructurado para obaservar niveles de agregación nacional, regional y provincial.

Con el fin de construir los pasos y dar respuesta a los objetivos planteados en el presente trabajo de grado, se propone un diseño de muestreo en varias etapas estratificado MAS-MAS, aplicado sobre las pruebas saber 11 del año 2016 para estimar el promedio poblacional del puntaje global obtenido por los estudiantes que presentaron dichos exámenes. Con el fin de realizar la estimación planteada, se estructura un marco teórico donde se presentan los esimadores directos e indirectos y el uso de modelos lineales mixtos de unidad y de área, para dar un enfoque comparativo y de esta manera observar el comportamiento de dichos estimadores y así concluir cuáles de ellos tienen un mejor desempeño.

4. Objetivos

4.1. Objetivo General

Comparar los principales estimadores directos e indirectos en áreas pequeñas, para la estimación de parámetros como el total poblacional y la media poblacional, a través del error cuadrático medio, para observar el comportamiento de dichos estimadores a nivel de área y a nivel de unidad y de esta forma concluir qué estimadores presentan el mejor desempeño para la estimación de áreas pequeñas.

4.2. Objetivos Específicos

- Desarrollar un diseño de muestreo en varias etapas *estratificado-muetreo aleatorio simple* y *muestreo aletorio simple de conglomerados* sobre las pruebas saber 11 2016, tomando como variables de interés, las diferentes puntuaciones obtenidas en las asignaturas evaluadas a los estudiantes en el país y como información auxiliar, las proyecciones poblacionales por municipio.
- Realizar estimación del total y la media poblacional por medio de estimadores directos e indirectos de las puntuaciones de las asignaturas evaluadas en la prueba saber.
- Estudiar las propiedades de los estimadores de áreas pequeñas para el total y la media poblacional en términios de precisión, insesgamiento y capacidad predictiva sobre las áreas no observadas en

los estudios.

5. Marco teórico

Diferentes investigaciones gubernamentales o de índole empresarial, buscan acceder a la información de conjuntos poblacionales de interés, para observar los comportamientos de las unidades que hacen parte de estos conjuntos y de esta manera tomar decisiones. Dado que es difícil acceder a toda la información de las unidades (en cuyo caso se trataría de un censo), por costos o logística, la teoría del muestreo brinda estrategias para extraer una parte de estos conjuntos poblacionales denominadas muestra aleatoria y através de esta inferir a cerca de la población, por medio de la medición de estimaciones de parámetros como el total poblacional Y_k, la media poblacional $\bar{Y}_k$ o estimación de proporciones. Pero no solo surge la idea de estimar un total o una media para razonar a cerca de una población, sino que también puede ser de interés del investigador observar subconjuntos poblacionales o dominios llamados áras pequeñas, donde igualmente se desee realizar cualquier tipo de estimación.

De esta manera, se hará uso de los parámetros usuales, que se estimarán con base a estimadores directos e indirectos más comunes, con una descripción teórica y posteriormente su aplicación en las pruebas saber 11 del año 2016.

5.1. Estimadores Directos

5.1.1. Estimador de Horvitz - Thomson para el total

En términos generales, los estimadores directos, son aquellos que dependen del diseño. Con esto, un primer acercamiento al total poblacional y la media poblacional es el estimador Horvitz-Thomson.

$$\hat{Y}_d^{dir} = \sum_{j \in s_d} \frac{1}{\pi_j} y_j \tag{1}$$

La varianza estimada correpondiente a $\hat{Y}_d^{dir}$

$$\widehat{var}(\hat{Y}_d^{dir}) = \sum_{i \in s_d} \sum_{j \in s_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \tag{2}$$

Para el caso de la estimación de la media poblacional y teniendo en cuenta la definición del estimador directo del total $\hat{Y}_d^{dir}$ dato en (1), se tiene la estimación de la media en dominios como:

$$\hat{\bar{Y}}_d^{dir} = \frac{\hat{Y}_d^{dir}}{N_d} \tag{3}$$

El estimador de varianza para la media en dominios, y tomando como referencia Morales (2015) [3], se expresa como:

$$\widehat{var}(\hat{\bar{Y}}_d^{dir}) = \frac{1}{\hat{N}_d^2} \sum_{i \in sd} \sum_{j \in sd} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} (y_i - \hat{\bar{Y}}_d^{dir})(y_j - \hat{\bar{Y}}_d^{dir}) \tag{4}$$

Donde el error cuadrático medio y su respectiva estimación se obtienen como:

$$MSE(\hat{\bar{Y}}_d^{dir}) \approx var_\pi(\hat{\bar{Y}}_d^{dir}) \tag{5}$$

$$mse(\hat{\bar{Y}}_d^{dir}) = \widehat{var}_\pi(\hat{\bar{Y}}_d^{dir}) \tag{6}$$

5.1.2. Estimador General de Regresión

El estimador general de regresión tiene en cuenta el uso de información auxiliar, donde esta puede estar contenida en más de una variable y cuyo objetivo es incrementar la precisión en las estimaciones y de esta manera obtener mejores resultados al momento de hacer inferencias de la población.

Con lo anterior, se expresa el estimador general de resgresión como:

$$\hat{Y}^{greg} = \hat{Y}^{dir} + (\mathbf{t}_x - \hat{\mathbf{t}}_x)'\hat{\mathbf{B}} \tag{7}$$

Donde la estimación del parámetro de población $\mathbf{B}$ como $\hat{\mathbf{B}} = (\hat{B}_1, \hat{B}_2, \cdots, \hat{B}_J)'$ se obtiene a través de las siguientes expresiones:

$$\begin{aligned}\hat{\mathbf{B}} &= (\hat{B}_1, \hat{B}_2, \cdots, \hat{B}_J)' \\ &= \mathbf{T}^{-1}\hat{\mathbf{t}} \\ &= \left(\sum_s \frac{\mathbf{X_j}\mathbf{X_j'}}{\sigma_j^2\pi_j}\right)^{-1} \sum_s \frac{\mathbf{X_j}y_j}{\sigma_j^2\pi_j}\end{aligned}$$

Con lo anterior, $\hat{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{t}$, y a través de estas expresiones, es posible realizar algunos pasos algebraicos para reescribir el estimador general de regresión como sigue:

$$\hat{Y}^{greg} = \hat{Y}^{dir} + (\mathbf{t}_x - \hat{\mathbf{t}}_x)'\hat{\mathbf{T}}^{-1} \sum_s \frac{\mathbf{X}_j\check{y}_j}{\sigma_j^2}$$

$$= \sum_s \left[1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)'\mathbf{T}^{-1}\frac{\mathbf{X}_j}{\sigma_j^2}\right]\check{y}_j \tag{8}$$

En La expresión (4) se identifica los pesos dependientes de la muestra como:

$$g_{js} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)'\mathbf{T}^{-1}\frac{\mathbf{X}_j}{\sigma_j^2} \tag{9}$$

Con lo cual, el estimador de regresión se puede expresar como una función lineal de los valores π expandidos $\check{y}_j$ para $j \in s$. Särndal (1992) [2]

$$\hat{Y}^{greg} = \sum_s g_{js}\check{y}_j = \sum_s g_{js}\frac{y_j}{\pi_j} \tag{10}$$

5.1.3. Estimador General de Regresión en Dominios

Para el caso de dominios y tomando como referencia a Rao & Molina (2015) [6], el GREG estimador, puede adaptarse como:

$$\hat{Y}_d^{greg} = \sum_{j \in s_d} w_j^* y_j \tag{11}$$

Donde $\hat{Y}_d^{greg}$ Estimador general de regresión para el total Y en el dominio d

$\sum_{j \in s_d}$ Es la suma sobre los individuos j que pertenecen a la muestra s en el dominio d

w_j^* es el producto de los pesos del diseño $w_j(s)$ y los pesos estimados g_{js} Rao & Molina (2015) [6], con lo cual se expresa $w_j^* = w_j g_{js}$.

Una propiedades interesante de este estimador general de regresión para la estimación en dominios, consta de la propiedad aditiva y del insesgamiento aproximado del mismo.

i) Propiedad aditiva: $\hat{Y}_1^{greg} + \cdots + \hat{Y}_m^{greg} = \hat{Y}^{greg}$ Rao & Molina (2015) [6]

Donde la suma de la estimación de los totales en los dominio m es igual al total poblacional estimado $\hat{Y}^{greg}$

Para la estimación de la media, en Morales (2015) [3] pág.16, se tiene que el estimador GREG de la media en dominios es:

$$\hat{Y}_d^{greg} = \hat{\bar{Y}}_d^{dir} + (\bar{\mathbf{X}} - \hat{\bar{X}}_d^{dir})\hat{\beta} \tag{12}$$

Con lo cual, se tiene que el estimador GREG de la media en dominios se puede expresar como:

$$\hat{\bar{Y}}_d^{greg} = \sum_{j \in sd} h_{dj} w_{dj} y_{dj} \tag{13}$$

Con:

$$\bar{\mathbf{X}} = \sum_{d=1}^{D} \sum_{j \in sd} h_{dj} w_{dj} x_{dj}$$

y

$$h_{dj} = \frac{g_{dj}}{N_d} I_{Ud}(j) + (\mathbf{X}_d - \hat{\bar{X}}_d^{dir}) \left(\sum_{d=1}^{D} \sum_{i \in sd} w_{di} x'_{di} x_{di} \right)^{-1} x'_{dj}$$

con g_{dj} como:

$$g_{dj} = \frac{N_d}{\hat{N}_d} I_{Ud}(j) + N_d(\bar{X}_d - \hat{\bar{X}}_d^{dir}) \left(\sum_{d=1}^{D} \sum_{i \in sd} w_{di} x'_{di} x di \right)^{-1} x'_{dj}$$

La expresión respectiva del error cuadrático medio (MSE), dado que el estimador GREG es aproximadamente insesgado, se expresa como:

$$\begin{aligned} MSE(\hat{\bar{Y}}_d^{greg}) &\approx var_\pi(\hat{\bar{Y}}_d^{greg}) \\ mse(\hat{\bar{Y}}_d^{greg}) &= \widehat{var}_\pi(\hat{\bar{Y}}_d^{greg}) \end{aligned} \tag{14}$$

Estimador post-estratificado

Este estimador es un caso especial del estimador de regresión generalizado, que se caracteriza por tratar la subdivición de una población U en U_g grupos poblacionales llamados *post-estratos*, con tamaños poblacionales conocidos en cada uno de estos post-estratos como $N_g(g = 1, \cdots, G)$ y que son mutuamente excluyentes.

En la etapa de diseño, el tamaño del post-estrato N_g es conocido, se desconoce el número de individuos que pertenecerán al post-estrato n_g en la muestra realizada. Gutiérrez (2009). [5]

La teoría expuesta en Gutiérrez (2009) [5], sugiere la construcción del estimador mediante un modelo lineal de tipo ANOVA que supone la existencia de una sola característica de interés de tipo discreto con el supuesto de que ésta característica de interés esta altamente correlacionada con los post-estratos.

Con lo anterior, se propone que la carecterística de información auxiliar, toma la forma de un vector $\mathbf{X}_j = d_j = (0, 0, \cdots, 1, \cdots, 0, 0)'$ cuyos totales se exponen como $t_\mathbf{x} = (N_1, \cdots, N_g, \cdots, N_g)$.

La estimación de este vector de totales t_x conduce a la forma $\hat{t}_{\mathbf{x},\pi} = (\hat{N}_{1,\pi}, \cdots, \hat{N}_{g,\pi}, \cdots, \hat{N}_{g,\pi})'$

Con lo cual, el modelo propuesto es de la forma:

$$Y_j = d_j^{'}\beta + \epsilon_j = \beta_g + \varepsilon_j \tag{15}$$

Donde $\beta = (\beta_1, \cdots, \beta_g, \cdots, \beta_G)'$, cada uno de los ε_j con $j \in U$ son variables aleatorias independientes e identicamente distribuidas con media 0 y varianza σ_g^2. Nóte que $d_j = (d_{1j}, \cdots, d_{gj}, \cdots d_{Gj})'$ con:

$$d_{gj} = \begin{cases} 1 & si\ j \in U_g \\ 0 & en\ otro\ caso \end{cases} \tag{16}$$

Luego el modelo de superpoblación está dado por:

$$\begin{aligned} E_\varepsilon(Y_j) &= \mathbf{d}_j^{'}\beta = \beta_g \\ Var_\varepsilon(Y_j) &= \sigma_g^2 \end{aligned} \tag{17}$$

Para llevar a cabo la estimación de los betas β como $\hat{\mathbf{B}} = (\hat{B}_1, \hat{B}_2, \cdots, \hat{B}_G)'$ se tiene que:

$$\hat{B}_g = \left(\sum_{s_g} \frac{1}{\pi_j}\right)^{-1} \left(\sum_{s_g} \frac{y_j}{\pi_j}\right) = \frac{\hat{t}_{yU_g,\pi}}{\hat{N}_{U_g,\pi}} \tag{18}$$

A partir de la expresión (12) de la estimación de los betas para cada post-estrato $\hat{B}_g$, se construye el estimador del total poblacional de la característica de interés como:

$$\hat{Y}_d^{pst} = \sum_{\substack{g=1 \\ n_h \neq 0}}^{G} \sum_{j \in S_g} \frac{N_g}{\hat{N}_{g,\pi}} \frac{y_j}{\pi_j} \tag{19}$$

En cuanto a la estimación de la varianza del total estimador $\hat{Y}_d^{pst}$, toma la siguiente forma:

$$\widehat{Var}(\hat{Y}_d^{pst}) = \sum \sum_{S} \frac{\Delta_{ji}}{\pi_{ji}} \frac{e_j}{\pi_j} \frac{e_i}{\pi_i} \tag{20}$$

Con $e_j = y_j - \tilde{y}s_g$, para $g = 1, \cdots, G$

Una expresión amena para este estimador de post-estratificación, puede consultarse en Morales (2015) pp.13 [3] , donde solamente se realiza la suma de los tamaños N de los grupos g en cada dominio d multiplicados por el estimador directo de la media (dir) $\hat{\bar{Y}}$ del dominio d en el grupo g.

$$\hat{Y}_d^{pst} = \sum_{g=1}^{G} N_{dg} \hat{\bar{Y}}_{dg}^{dir} \tag{21}$$

Con una varianza estimada dada por:

$$\widehat{Var}(\hat{Y}_d^{pst}) = \sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_{dg}^2} \sum_{j \in s_{dg}} w_j(w_j - 1)(y_j - \hat{\bar{Y}}_{dg}^{dir})^2 \tag{22}$$

Para la estimación de la media a través de este estimador de post-estratificación, se tiene la expresión (21) pero diviendo por el tamaño del dominio N_d, con esto, la estimación de la media dada como $\hat{\bar{Y}}_d^{pst} = \frac{\hat{Y}_d^{pst}}{N_d}$ queda expresada de la siguiente forma:

$$\hat{\bar{Y}}_d^{pst} = \frac{1}{N_d} \sum_{g=1}^{D} N_{dg} \hat{\bar{Y}}_{dg}^{dir} \tag{23}$$

Con una varianza estiamada para la media estimada por post-estratificación en dominios como

$$\widehat{var}_{\pi}(\hat{\bar{Y}}_d^{pst}) = \frac{1}{N_d^2} \sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_{dg}^2} \sum_{j \in sdg} w_j(w_j - 1)(y_j - \hat{\bar{Y}}_{dg}^{dir})^2 \tag{24}$$

Tomando Morales(2015) [3], se tiene que el estimador de post-estratificación es aproximadamente insesgado, lo que conlleva al MSE error cuadrático medio a:

$$\begin{aligned} MSE(\hat{\bar{Y}}_d^{pst}) &\approx var_\pi(\hat{\bar{Y}}_d^{pst}) \\ mse(\hat{\bar{Y}}_d^{pst}) &= \widehat{var}_\pi(\hat{\bar{Y}}_d^{pst}) \end{aligned} \tag{25}$$

Algunas consideraciones:

- En cuanto a la información auxiliar que requiere el estimador, es información de tipo categórico [11]
- Dentro de cada uno de los post-estratos, la varianza de los residuales es constante. [11]
- El uso de información auxiliar reduce el sesgo, especialmente el sesgo causado por errores no muestrales tales como la ausencia de respuesta y la subcobertura del marco de muestreo. Zhang(2000) [5]
- El modelo de superpoblación ε es eficiente cuando la característica de interés es homogénea dentro de los post-estratos U_g para $(g = 1, \cdots, G)$, pero disímil y heterogénea entre cada uno de los post-estratos. Särndal, Swensson & Wretman (1992) [5].

  Cuando estas condiciones se cumplen, entonces el modelo ANOVA explicará una gran parte de la dispersión de la característica de interés. Gutierrez (2009) [5].

Estimador directo de Razón

Un caso especial del estimador general de regresión en dominios, se tiene si se considera una única variable auxiliar x en donde el modelo toma la forma como:

$$E_m(y_j) = \beta_d x_j$$

Y la varianza expresada como:

$$Var_m(y_j) = C_j = \lambda' x_j \qquad \forall \ j \in U_d$$

λ como un vector de constantes conocidas.

Bajo este modelo se obtienen el estimador de razón en dominios como:

$$\hat{Y}_d^{ratio} = \sum_{U_d} x_j \frac{\sum_{sd} a_j y_j}{\sum_{sd} a_j x_j} \tag{26}$$

Lo cual conlleva a la siguiente expresión:

$$\hat{Y}_d^{ratio} = \hat{Y}_d^{dir} \frac{X_d}{\hat{X}_d^{dir}} \tag{27}$$

Una observación se tiene si la viariable auxiliar x es la indicadora de pertenencia al dominio (δ_d), con lo cual el estimador $\hat{Y}_d^{ratio}$ coincide con el estimador de hayek $\tilde{Y}_d = N_d \tilde{Y}_{sd}$. Ferreira (2011) [8]

Por otra parte, bajo un diseño MAS de tamaño n de una población de N individuos, el estimador directo de razón $\hat{Y}_d^{ratio}$ queda expresado de la forma:

$$\hat{Y}_d^{ratio} = \sum_{U_d} x_j \hat{\beta}_d = \left(\sum_{U_d} x_j \right) \frac{\bar{Y}_{sd}}{\bar{x}_{sd}} \tag{28}$$

Donde $\bar{Y}_{sd}, \bar{x}_{sd}$ corresponde a las medias muestrales en U_d para la variable de interés y la variable auxiliar. La correspondiente estimación de la varianza se tiene como:

$$\widehat{Var}(\hat{Y}_d^{ratio}) = \left(\frac{\bar{x}_{U_d}}{\bar{x}_{sd}} \right)^2 N_d^2 \left(\frac{1}{n_{sd}} - \frac{1}{\hat{N}_d} \right) S_{e_{sd}}^2 \tag{29}$$

Donde $\hat{N} = \frac{N n_{sd}}{n}$. $\bar{x}_{Ud}$ y $\bar{x}_{sd}$ son las medias poblacional y muestral de la variable auxiliar del dominio U_d. Con $S_{e_{sd}}^2 = (n_{sd} - 1)^{-1} \sum_{sd} (y_j - \hat{\beta}_d x_j)^2$

5.2. Estimadores Indirectos

Debido a que los estimadores directos presentan errores estándar grandes en subconjuntos poblacionales cuyos tamaños de muestra son muy pequeños o casi nulos, la teoria ha dado un manejo para solventar este tipo de falencias a través de la generación de estimadores indirectos. Entre estos estimadores, pueden encontrarse, por ejemplo en Rao (2003) [7], *estimadores sintéticos*, *estimadores compuestos*, *estimadores James-Stein*.

Para el caso del presente trabajo, se tomará en cuenta el *estimador sintético* que tiene en cuenta un estimador directo para un área grande, donde dicha área, cubre áreas más pequeñas y finalmente deriva en un estimador indirecto para un área pequeña, suponiendo que esas áreas pequeñas tienen las mismas características o comportamientos del área que las cubre. Rao (2003) [7]

La expresión para este tipo de estimador, se toma de Morales (2015) pp. 11:

$$\hat{Y}_d^{synth} = \sum_{g=1}^{G} N_{dg} \hat{\bar{Y}}_g^{dir} \tag{30}$$

Donde también es posible expresar el estimardor en términos de la media como:

$$\hat{\bar{Y}}_d^{synth} = \frac{1}{N_d} \sum_{g=1}^{G} N_{dg} \hat{\bar{Y}}_g^{dir} \tag{31}$$

El respectivo estimador de la varianza para el estimador de la media en dominios $\hat{\bar{Y}}_d^{synth}$, se tiene como:

$$\widehat{var}(\hat{\bar{Y}}_d^{synth}) = \frac{1}{N_d^2} \sum_{g=1}^{G} \frac{N_{dg}^2}{\hat{N}_g} \sum_{j \in s_g} w_j (w_j - 1)(y_j - \hat{\bar{Y}}_d^{dir})^2 \tag{32}$$

El error cuadrático medio del estimador de la media $\hat{\bar{Y}}_d^{synth}$ toma la siguiente forma:

$$MSE\left(\hat{\bar{Y}}_d^{synth}\right) = var_\pi \left[\hat{\bar{Y}}_d^{synth}\right] + \left(B_\pi \hat{\bar{Y}}_d^{synth}\right]^2 \tag{33}$$

Y su respectiva estimación como:

$$mse\left(\hat{\bar{Y}}_d^{synth}\right) = \widehat{var}_\pi\left(\hat{\bar{Y}}_d^{synth}\right) + \left(\hat{\bar{Y}}_d^{synth} - \hat{\bar{Y}}_d^{dir}\right)^2 \tag{34}$$

5.3. Modelos Lineales Mixtos

Los modelos lineales clásicos, parten del hecho de que las observaciones extraídas son parte de la misma población y son independientes. Un modelo lineal mixto, tiene características más complejas debido a que tiene en cuenta factores específicos de las unidades de observación, dándoles un carácter de agrupamiento llamado *clúster*, considerando dependencia dentro de esos grupos o *clústers* formados pero independencia entre los mismos, identificando así, dos fuentes de variación: *entre clústers y dentro de los clústers.*

Debido a lo anterior, la construcción de un modelo para observaciones agrupadas, contempla en su estructura efectos fijos y efectos aleatorios que son parte inherente de la naturaleza de los datos y que conlleva a proponer una mirada diferente para buscar respuestas particulares a través de la elaboración de modelos mixtos que contienen la estructura de tales efectos para un análisis superior en comparación con un modelo líneal clásico.

En el contexto de la *estimación en áreas pequeñas*, los efectos aleatorios para la variación entre áreas no puede explicarse mediante la inclusión de variables auxiliares. Pushpal K Mukhopadhyay and Allen McDowell (2011) [9] De esta maneran, la teoría a establecido modelos que se definen como *modelos a nivel de área* y *modelos a nivel de unidad.*

5.3.1. Método de Fay Herriot - Modelos a Nivel de Área

Los modelos a nivel de área relacionan estimadores directos de área pequeña con datos auxiliares específicos de área. Si se tiene que $\hat{\bar{Y}}_d$ es la estimación de la media para el área d y que un vector de variables auxiliares $\bar{\mathbf{X}}_d = (\bar{x}_{d1}, \cdots, \bar{x}_{dp})$ es conocido, un modelo lineal mixto relaciona a $\bar{Y}_d$ con $\bar{X}_d$ de la siguiente forma:

$$\bar{Y}_d = \bar{\mathbf{X}}_d^T\beta + u_d + \hat{e}_d \tag{35}$$

Este modelo se estructura como un conjunto fijo de parámetros $\bar{\mathbf{X}}_d^T\beta$, con los efectos aleatorios de área u_d y $\hat{e}_d$ como los errores de muestreo. Se asume que $u_d \sim^{ind} (o, \sigma_u^2)$ y $\hat{e}_d \sim^{ind} (0, D_d)$

La media desconocida para el área d se expresa como:

$$\theta_d = \mathbf{X}_d^T\beta + u_d \tag{36}$$

Se hace uso del estimador predictivo BLUP que sólo se puede usar si σ_u^2 y D_d son conocidos. Con lo cual, Si el parámetro fijo β, σ_u^2 la varianza del factor aleatorio y D_d la varianza de los errores, son conocidos, la estimación del factor aleatorio u_d se muestra como:

$$\hat{u}_d = \gamma_d(u_d + \hat{e}_d) \tag{37}$$

donde $\gamma_d = (\sigma_u^2 + D_d)^{-1}\sigma_u^2$

Dato esto, se expresa el predictor BLUP para θ_i como:

$$\tilde{\theta} = \begin{cases} \mathbf{X}_d^T\beta + \gamma_d(\bar{Y}_d - \bar{\mathbf{X}}_d^T\beta) & Si \quad d \in A \\ \mathbf{X}_d^T\beta & Si \quad d \neq A \end{cases} \tag{38}$$

En la expresion del BLUP (28), A hace referencia a áreas pequeñas en las que se observa $\bar{Y}_d$.

Cuando estas dos cantidades son desconocidas, β y σ_u^2, se utiliza el estimador predictivo EBLUP haciendo el respectivo reemplazo en la expresión 28 por $\hat{\beta}$ y $\hat{\sigma}_u^2$, aunque también, puede ser escrito de la siguiente forma:

$$\hat{\theta}_d = \hat{\gamma}_d\bar{Y}_d + (1 - \hat{\gamma}_d)\bar{X}_d^T\hat{B} \tag{39}$$

Si σ_u^2 es un estimador insesgado de σ_u^2 entonces un estimador de MSEP es:

$$msep(\theta)_d = \begin{cases} \hat{\gamma}_d D_d + (1 - \hat{\gamma}_d)^2\bar{X}_d\hat{V}\{\hat{B}\}\bar{X}_d^T + 2(\sigma_u^2 + D_i)\hat{V}\{\hat{\gamma}_d\} & Si \quad d \in A \\ \sigma_u^2 + \bar{X}_d\hat{V}\{\hat{\beta}\}\bar{X}_d^T\beta & Si \quad d \neq A \end{cases} \tag{40}$$

Donde:

$$\hat{V}\{\hat{\gamma}_d\} = (\sigma_u^2 + D_d)^{-4}D_d^2\bar{V}(\sigma_u^2)$$

$\bar{V}(\sigma_u^2)$ es la varianza asintótica de σ_u^2

5.3.2. Método de Battisse - Modelos a Nivel de Unidad

Los Modelos mixtos a Nivel de Unidad, relacionan valores unitarios de una varibale de estudio con valores de variables auxiliares de cada unidad y toman en cuenta los efectos aleatorios específicos de la unidad para que las estimaciones sean más eficientes.

Si Y_{dj} es el valor de una variable de estudio en el área d y de la unidad j, para $d = 1, \cdots, m$ y $j = 1, \cdots, N_i$, donde m es el número de áreas pequeñas y N_i es el número de unidades de población en el área d.

Asumiendo que la información auxiliar está disponible para toda la población, por ende, es información específica cada unidad, se toma como un vector de la forma $\mathbf{X}_{dj} = (x_{dj1}, \cdots, x_{djq})^T$ con q el número de variables auxilares.

El modelo propuesto que relaciona Y_{dj} con $\mathbf{X}_{dj}$ toma la siguiente forma:

$$Y_{dj} = \mathbf{X}_{dj}^T\beta + u_d + \varepsilon_{dj} \tag{41}$$

Donde β es un conjunto fijo de parámetros de regresión, u_d son efectos aleatorios específicos de área y ε_{dj} son los errores de muestreo.

El parámetreo de media de área pequeña se define como:

$$\theta_d = \bar{X}_{d(p)}^T\beta + u_d \tag{42}$$

Donde $\bar{X}_{d(p)} = N_d^{-1} \sum_{j=1}^{N_d} x_{dj}$ es la media poblacional.

Se asume que $u_d \sim^{ind} (0, \sigma_u^2)$ y $\varepsilon_{dj} \sim^{ind} (0, \sigma_e^2)$

Si una muestra es seleccionada en el área d, es decir, n_d de N_d, estos valores muestrales también satisfacen el modelo poblacional (31).

Para las medias de área pequeña θ_d, si σ_u^2 y $sigma_e^2$ son conocidos, el mejor predictor lineal e insesgado es el predictor BLUP, que se estructura como:

$$\tilde{\theta}_d = \bar{X}_{d(p)}^T \tilde{\beta} + \gamma_d (\bar{Y}_{d\cdot} - \bar{X}_{d\cdot}^T \tilde{\beta}) \tag{43}$$

donde $\gamma_d = \left(\frac{\sigma_u^2 + \sigma_e^2}{n_d} \right)^{-1} \sigma_u^2$, $\tilde{\beta}$ es el predictor BLUP de β y $\bar{Y}_{d\cdot}$, $\bar{X}_{d\cdot}$, son las medias muestrales para el área d.

Si σ_u^2 y σ_e^2 son desconocidos y cuyas estimaciones $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$, entonces el predictor lineal empírico adeacuado es el EBLUP para la media θ_d dado por:

$$\hat{\theta}_d = \bar{X}_{d(p)}^T \hat{B} + \hat{\gamma}_d (\bar{Y}_{d\cdot} - \bar{X}_{d\cdot}^T \hat{\beta}) \tag{44}$$

donde $\hat{\gamma}_d = \left(\frac{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{n_d} \right)^{-1}$, $\hat{\beta}$ es un EBLUP para β

El error cuadrático medio (MSEP) respectivo, parte de la suposición de que los componenetes de error siguen una distribución normal, con lo cual el MSEP se expresa como:

$$MSEP(\hat{\theta}_d) \approx g_{1d}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) + \sigma_{3d}(\sigma_u^2, \sigma_e^2) \tag{45}$$

Donde:

$$g_{1d} = \frac{\gamma_d \sigma_e^2}{n_d}$$

$$g_{2d} = (\bar{X}_{d(p)} - \gamma_d \bar{X}_{d\cdot})^T \left(\sum_{d=1}^{m} A_d \right)^{-1} (\bar{X}_{d(p)} - \gamma_d \bar{X}_{d\cdot})$$

$$g_{3i}(\sigma_u^2, \sigma_e^2) = n_d^{-2} \left(\frac{\sigma_u^2 + \sigma_e^2}{n_i} \right)^{-3} h(\sigma_u^2, \sigma_e^2)$$

con:

$A_d = \sigma_e^{-2} \sum_{j=1}^{n_i} (X_{dj} X_{dj}^T - \gamma_d n_d \bar{X}_{d\cdot} \bar{X}_{d\cdot}^T)$
$h(\sigma_u^2, \sigma_e^2) = \sigma_e^4 \bar{V}_{uu}(\delta) + \sigma_u^4 \bar{V}_{ee}(\delta) - 2\sigma_e^2 \sigma_u^2 \bar{V}_{ue}(\delta)$
Con $\delta = (\sigma_u^2, \sigma_e^2)^T$, $\bar{V}_{uu}(\delta)$ y $\bar{V}_{ee}(\delta)$ son las varianzas asintóticas de los estimadores σ_u^2 y σ_e^2, y $\bar{V}_{ue}(\delta)$ como la covarianza asintótica de σ_u^2 y σ_e^2.

Con lo anterior, y asumiendo normalidad de los errores u_d y ε_{ij}, un estimador de MSEP se expresa como:

$$msep(\hat{\theta}_d) = g_{1d}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2d}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \tag{46}$$

Para llevar a cabo esta parte teórica de Modelos a Nivel de Área y a Nivel de Unidad, se referencia el artítculo de 2011 de Pushpal K Mukhopadhyay and Allen McDowell [9].

6. Metodología

Se han ilustrado los fundamentos teóricos de los estimadores que serán objeto de uso para lograr los objetivos de este trabajo, donde se destaca la necesidad de la busqueda de información auxiliar que será protagonista en aquellos estimadores que requieran de su incorporación para mejorar los resultados.

La aplicación y observación del comportamiento de los estimadores, serán analizados a través de los datos de las pruebas saber 11 del año 2016.

La información auxiliar tenida en cuenta, forma parte de la base de datos de las pruebas saber 11 2016, así como también de la base de datos de la medición de Desempeño Municipal que se encuentra publicada en el Departamento Nacional de Planeación DNP.

La estructura básica de esta metodología de aplicación, se basa en un diseño muestral en tres etapas EST-EST-MAS para la extracción de la muestra, a partir de la cual se implementan los estimadores directos, indirectos y por último los modelos de unidad de Battisse y de área de Fay-Herriot.

6.1. Diseño de Muestreo

El diseño de muestreo implementado en las pruebas saber 11-2016, consta de 3 etapas.

Primera Etapa

A través de un diseño de muestreo estratificado, se selcciona como unidad primaria los 1.111 municipios donde están ubicados los 12.214 instituciones educativas. En esta etapa, se agrupan en 8 estratos los municipios de ubicación de los colegios de los estudiantes, debido a que se busca homogeneidad en las observaciones, es decir, se busca agrupar municipios con cualidades similares.

En cada uno de estos 8 estratos, se utiliza un muestreo aleatorio simple, con la finalidad de seleccionar un total de 29 municipios.

Segunda Etapa

Con base en la primera etapa ya planteada, se planea un diseño de muestreo estratificado, pero teniendo encuenta que la estratifición se realiza sobre las instituciones educativas; con esto, se propone la creación de 7 estratos, donde cada uno de ellos, alberga colegios con características similares.

Tercera Etapa

Finalmente, apartir de la estratificación realizada en la segunta etapa, se hace uso de un diseño de muestreo aleatorio simple, con la finalidad de seleccionar un total de 358 colegios y, apartir de allí, comenzar los procesos de estimación en las pruebas saber 11 del año 2016.

En la figura 1, se ilustra las etapas realizadas con el fin de seleccionar la muestra como primera parte del trabajo a desarrolar.

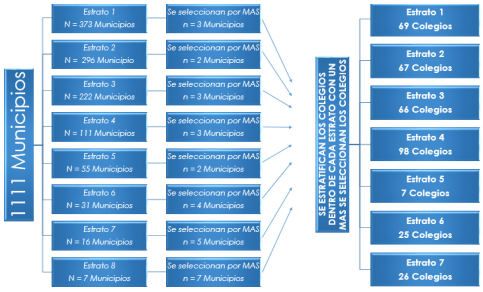

Figura 1: Esquema Diseño de Muestreo

6.2. Aplicación Estimadores Directos

Como se ha manifestado en el marco teórico, los estimadores directos son inherentes al diseño de muestreo, por lo tanto, un estimador insesgado para esta primera aplicación en las pruebas saber, será el uso del estimador Horvitz-Thomson.

Debido a que se trata como característica de interés el puntaje de matemáticas obtenido por los estudiantes que presentaron la prueba ICFES en el año 2016, se observará el promedio de este puntaje en los dominios especificados, que son los municipios seleccionados en la muestra.

Este estimador requiere de la construcción del total estimado en dominios $\hat{Y}_d$ donde se dividirá por el tamaño de los dominios N_d.

Con el uso de las funciones del paquete survey del lenguaje estadístico R, se obtienen las estimaciones por dominio del promedio del puntaje de matemáticas.

	MUNICIPIO	NOMBRE MUNICIPIO	ESTIMACIÓN DIRECTA	DESVIACIÓN ESTÁNDAR	COEFICIENTE DE VARIACIÓN
1	05001	Medellín	51.59017	1.2148748	2.3548570
2	08001	Barranquilla	51.62125	2.5265537	4.8944063
3	11001	Bogotá, D.C.	55.94195	1.0914459	1.9510331
4	13001	Cartagena	50.69562	2.1298737	4.2012971
5	13620	San Cristóbal	48.05345	0.1363276	0.2836999
6	13657	San Juan Nepomuceno	46.42323	0.3456440	0.7445498
7	15104	Boyacá	56.02443	0.1510454	0.2696064
8	15238	Duitama	59.86197	0.6882661	1.1497551
9	23001	Montería	52.25121	1.4774299	2.8275515
10	23417	Lorica	46.81286	0.5271954	1.1261766
11	23586	Purísima	44.75580	0.1369386	0.3059683
12	25154	Carmen de Carupa	51.26068	0.1303704	0.2543282
13	25290	Fusagasugá	53.54588	0.6519529	1.2175595
14	25430	Madrid	51.78581	0.4489841	0.8670022
15	27361	Istmina	47.55858	0.2658997	0.5590992
16	41359	Isnos	42.10097	0.2921520	0.6939318
17	47692	San Sebastián de Buenavista	44.27600	0.2426266	0.5479867
18	50001	Villavicencio	53.30836	0.8932474	1.6756234
19	50006	Acacías	52.40639	0.3530399	0.6736581
20	50313	Granada	50.85512	0.5313814	1.0448926
21	52378	La Cruz	52.34781	0.4874181	0.9311147
22	54001	Cúcuta	53.29712	1.9091202	3.5820326
23	63001	Armenia	55.96099	1.1722255	2.0947190
24	68547	Piedecuesta	58.27613	0.8269049	1.4189428
25	70001	Sincelejo	51.75501	0.8015236	1.5486880
26	73001	Ibagué	50.69695	0.8630438	1.7023585
27	76001	Cali	52.27600	0.9674050	1.8505718
28	76109	Buenaventura	45.58127	0.5760684	1.2638271
29	86865	Valle del Guamuez	43.32547	0.4197635	0.9688607

Figura 2: Estimaciones directas

En la figura 2, se plasman los resultados de las estimaciones por dominio, junto con sus respectivas desviaciones estándar y coeficientes de variación.

Estas primeras estimaciones muestran un buen comportamiento debido a que los coeficicentes de variación son menores al 5 %, lo cual quiere decir que son estimaciones eficientes.

Los mayores coeficientes de variación se presentan en los municipios de Barranquilla con el 4.89 %, Cartagena con 4.20 %, Cúcuta con el 3.58 %, Montería con el 2.82 % y Medellín con el 2.35 %.

Los menores coeficientes de variación, se presentan en los municipios de Carmen de Carupa con un 0.25 %, Boyacá con el 0.26 % y San Cristóbal con el 0.28 %. Para estos dominios, las etimaciones sugieren que, el promedio en el puntaje de matemáticas, para Carmen de Carupa fue de 51.26 puntos, en Boyacá el puntaje promedio fue de 56.02 puntos, y en San Cristóbal fue de 48.05 puntos.

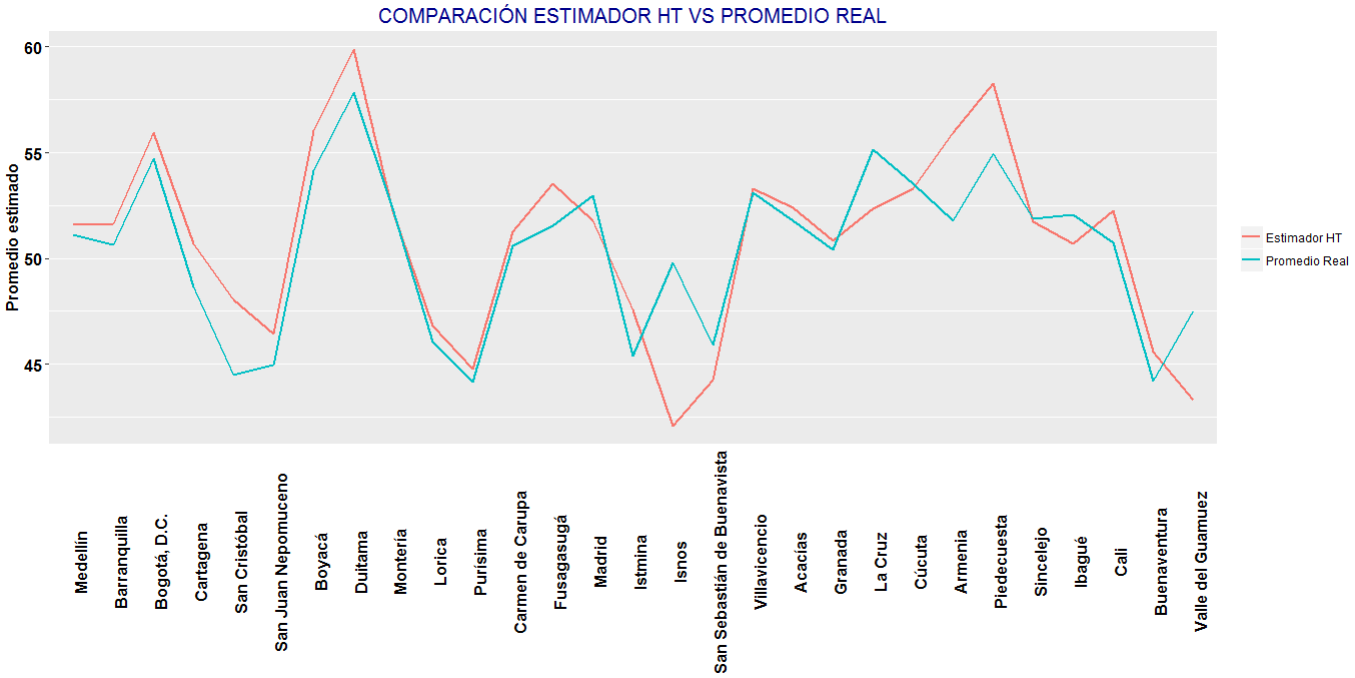

Figura 3: Gráfica de estimaciones directas Horvitz-Thomson vs Promedio Real

Pese a que los coeficientes de variación sugieren esimaciones eficientes, se evidencia en la figura 3, que las estimaciones realizadas por el estimador Horvitz-Thompson, en comparación con el promedio real, sobre-estiman los puntajes promedio de matemáticas en algunos municipios como San Cristóbal, Duitama, Fusagasugá y Piedescuesta. Para el municipio de Isnos, el estimador directo (Horvitz-Thompson), sub-estima la puntuación promedio de la asignatura de matemáticas.

6.3. Aplicación Estimador Post-estratificado

Para la construcción de este estimador, se utilizó la variable categórica *DESEMP_IGLES* que hace referencia al desempeño en inglés alcanzado por los estudiantes y personas que de manera individual presentaron la prueba.

Las categorías que contempla la variable *DESEMP_IGLES* son *A-, A1, A2, B+ y B1*, para las cuales se presenta la respectiva descripción en la siguiente tabla:

NIVEL		DESCRIPCIÓN
Nivel inferior	A-	No alcanza el nivel A1.
Usuario Básico	A1	Es capaz de comprender y utilizar expresiones cotidianas de uso muy frecuente así como frases sencillas destinadas a satisfacer necesidades de tipo inmediato
		Puede presentarse a sí mismo y a otros, pedir y dar información personal básica sobre su domicilio, sus pertenencias y las personas que conoce.
		Puede relacionarse de forma elemental siempre que su interlocutor hable despacio y con claridad y esté dispuesto a cooperar.
	A2	Es capaz de comprender frases y expresiones de uso frecuente relacionadas con áreas de experiencia que le son especialmente relevantes (información básica sobre sí mismo y su familia, compras, lugares de interés, ocupaciones, etc.)
		Sabe comunicarse a la hora de llevar a cabo tareas simples y cotidianas que no requieran más que intercambios sencillos y directos de información sobre cuestiones que le son conocidas o habituales
		Sabe describir en términos sencillos aspectos de su pasado y su entorno así como cuestiones relacionadas con sus necesidades inmediatas.
Usuario Independiente	B+	se aproxima a B2.
	B2	Es capaz de entender las ideas principales de textos complejos que traten de temas tanto complejos como abstractos, incluso si son de carácter técnico siempre que estén dentro de su campo de especialización.
		Puede relacionarse con hablantes nativos con un grado suficiente de fluidez y naturalidad de modo que la comunicación se realice sin esfuerzo por parte de ninguno de los interlocutores.
		Puede producir textos claros y detallados sobre temas diversos así como defender un punto de vista sobre temas generales indicando los pros y los contras de las distintas opciones.

Figura 4: Desempeño inglés - Fuente: ICFES [4]

Con lo anterior, se busca subdividir la población en g grupos que se caracterizan por las categorías del nivel de desempeño en inglés.

Dado que se trata de estimar dominios, un primer paso al pensar en este estimador de post-estratificación, trata de realizar un cruce de conteos entre los dominios con los grupos establecidos N_{dg}. En estos cruces, también se tiene que estimar el promedio a través del estimador directo $\hat{\bar{Y}}_{dg}^{dir}$, con lo cual, se obtendrá un estimador del total en cada cruce de dominio y grupo. Para estimar el promedio, se dividirá estas estimaciones de totales en cada celda (cruce de dominio y grupo), por el tamaño del dominio N_d. Una breve ilustración se puede observar en la tabla 1.

Dominio / Grupo(Post-estrato)	1	2	$\cdots$	g	$\cdots$	G
1	$N_{11} * \hat{\bar{Y}}_{11}^{dir}$	$N_{12} * \hat{\bar{Y}}_{12}^{dir}$	$\vdots$	$N_{1g} * \hat{\bar{Y}}_{1g}^{dir}$	$\vdots$	$N_{1G} * \hat{\bar{Y}}_{1G}^{dir}$
2	$\cdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\cdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d	$\cdots$	$\cdots$	$\cdots$	$N_{dg} * \hat{\bar{Y}}_{dg}^{dir}$	$\cdots$	$\vdots$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
D $\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$N_{DG} * \hat{\bar{Y}}_{DG}^{dir}$

Tabla 1: Ilustración: Estimaciones y conteos en cada cruce de grupo y dominio

Teniendo en cuenta la idea de estimar los promedios en cada celda a través de estimadores directos, se realizó este cálculo y se obtuvo la primera parte de la construccción del estimador de postestratificación, como se observa en la figura 5. Esta imagen (figura 5), presenta un breve resúmen de las 145 combinaciones que se realizaron. En la observación 24 y 25, no se presentó estimación, esto sucede debido a que, para el dominio 13620 perteneciente al municipio de San Cristóbal en el departamento de Bolívar, no se registraron desempeños en inglés que se clasificaran en las categorías de B+ y B1. Esto generará, que no se obtenga una estimación en dichos municipios en los cuales se presenten ausencias de registros.

	Domain	Grupo	Domin_Grupo	PUNT_MATEMATICAS	se
1	05001	A-	05001_A-	44.02181	1.06724849
2	05001	A1	05001_A1	52.86941	0.97308761
3	05001	A2	05001_A2	57.68670	1.19615739
4	05001	B+	05001_B+	67.31285	2.72872213
5	05001	B1	05001_B1	63.64754	1.19727608
6	08001	A-	08001_A-	41.84815	1.75025745
7	08001	A1	08001_A1	50.45015	1.64777039
8	08001	A2	08001_A2	59.58281	1.10329986
9	08001	B+	08001_B+	74.99702	1.43573647
10	08001	B1	08001_B1	63.95008	1.43574729
11	11001	A-	11001_A-	48.18325	0.93524659
12	11001	A1	11001_A1	52.52500	0.69644810
13	11001	A2	11001_A2	58.50163	0.74826018
14	11001	B+	11001_B+	71.97044	1.77777562
15	11001	B1	11001_B1	64.73831	0.92027243
16	13001	A-	13001_A-	42.53667	1.23218361
17	13001	A1	13001_A1	51.84088	1.64043137
18	13001	A2	13001_A2	60.70437	1.54698933
19	13001	B+	13001_B+	81.79393	3.15614039
20	13001	B1	13001_B1	65.07677	1.47098039
21	13620	A-	13620_A-	43.87795	0.20762468
22	13620	A1	13620_A1	51.55493	0.17664679
23	13620	A2	13620_A2	49.00000	0.00000000
24	13620	B+	<NA>	NA	NA
25	13620	B1	<NA>	NA	NA

Figura 5: Estimaciones de promedios en cada combinación de dominios y post-estratos

Los tamaños respectivos a cada combinación de dominios y post-estratos, son los conteos de los respectivos registros que se presentaron en los diferentes municipios de los desempeños en la asignatura de inglés. Se ilustra su calculo en la figura 6.

	A-	A1	A2	B+	B1
05001	540	398	220	34	159
08001	120	114	68	27	66
11001	489	523	335	61	248
13001	196	127	73	5	49
13620	8	10	1	0	0
13657	40	17	3	0	1
15104	7	6	5	0	1
15238	20	89	73	5	43
23001	60	46	22	13	26
23417	56	46	22	1	7
23586	23	6	4	0	0
25154	10	9	2	0	0
25290	53	51	45	2	18
25430	16	29	9	3	4
27361	23	17	3	0	0
41359	24	8	0	0	0
47692	17	4	0	0	0
50001	126	122	86	10	52
50006	30	25	15	1	5
50313	30	12	8	0	5
52378	23	15	3	2	5
54001	208	178	75	5	45
63001	100	142	83	16	48
68547	57	81	51	8	26
70001	56	52	40	2	15
73001	160	209	175	22	109
76001	327	416	262	35	196
76109	200	119	44	3	20
86865	18	6	0	0	0

Figura 6: Tamaños de los cruces de dominios y post-estratos

Con los calculos anteriores, se obtendrá una estimación de totales para cada municipio, pero dado que se trata de estimar los puntajes de matemáticas, estos totales no tendrán ningúna lógica, es por ello que se estima el promedio del puntaje, dividiendo los totales obtenidos por el tamaño de cada dominio N_d (figura 7).

CODIGO	05001	08001	11001	13001	13620	13657	15104	15238	23001	23417	23586	25154	25290	25430	27361	41359	47692	50001	50006	50313	52378	54001	63001	68547	70001	73001	76001	76109	86865
N_d	1351	395	1656	450	19	61	19	230	167	132	33	21	169	61	43	32	21	396	76	55	48	511	389	223	165	675	1236	385	24

Figura 7: Tamaños por dominio

Los resultados finales, se muestran en la tabla 2, donde se observan las estimaciones para el puntaje de matemáticas por municipios, junto con sus respectivos coeficientes de variación.

Para el municipio de Bogotá D.C., se muestra que la estimación del puntaje promedio de matemáticas fue de 55 puntos, con un coeficiente de variación del 0.0185 %. Esta estimación es eficiente debido a que el coeficiente de variación es pequeño o es menor al 5 %. De igual manera, para cada estimación realizada, el comportamiento es bueno en términos de eficiencia, por la poca variablidad que presentan las estimaciones de los puntajes de matemáticas

CÓDIGO	MUNICIPIO	ESTIMACIÓN	VARIANZA	CVE
05001	Medellín	51,7	0,3	1,1
08001	Barranquilla	53,3	0,6	1,5
11001	Bogotá, D.C.	55,0	0,2	0,8
13001	Cartagena	51,0	0,6	1,5
13620	San Cristóbal	*N/A*	*N/A*	*N/A*
13657	San Juan Nepomuceno	*N/A*	*N/A*	*N/A*
15104	Boyacá	*N/A*	*N/A*	*N/A*
15238	Duitama	59,7	0,1	0,6
23001	Montería	52,5	0,2	0,9
23417	Lorica	47,1	0,1	0,6
23586	Purísima	*N/A*	*N/A*	*N/A*
25154	Carmen de Carupa	*N/A*	*N/A*	*N/A*
25290	Fusagasugá	53,6	0,1	0,5
25430	Madrid	51,9	0,1	0,5
27361	Istmina	*N/A*	*N/A*	*N/A*
41359	Isnos	*N/A*	*N/A*	*N/A*
47692	San Sebastián de Buenavista	*N/A*	*N/A*	*N/A*
50001	Villavicencio	53,6	0,1	0,6
50006	Acacías	52,4	0,0	0,4
50313	Granada	*N/A*	*N/A*	*N/A*
52378	La Cruz	52,4	0,1	0,4
54001	Cúcuta	51,8	0,7	1,6
63001	Armenia	54,2	0,1	0,6
68547	Piedecuesta	55,5	0,1	0,6
70001	Sincelejo	52,3	0,1	0,7
73001	Ibagué	52,4	0,3	1,1
76001	Cali	52,7	0,3	1,0
76109	Buenaventura	46,2	0,1	0,7
86865	Valle del Guamuez	*N/A*	*N/A*	*N/A*

Tabla 2: Resultados Estimación Post-estratificado

Comparando las estimaciones realizadas con el promedio real del puntaje de matemáticas, se observa en la grafica 8, que para el municipio de montería, la estimación presenta un muy buen desempeño, debido a la fuerte aproximación con el promedio real. En algunos municipios se presentó subestimación, como por ejemplo en Armenia, Cali y Duitama.

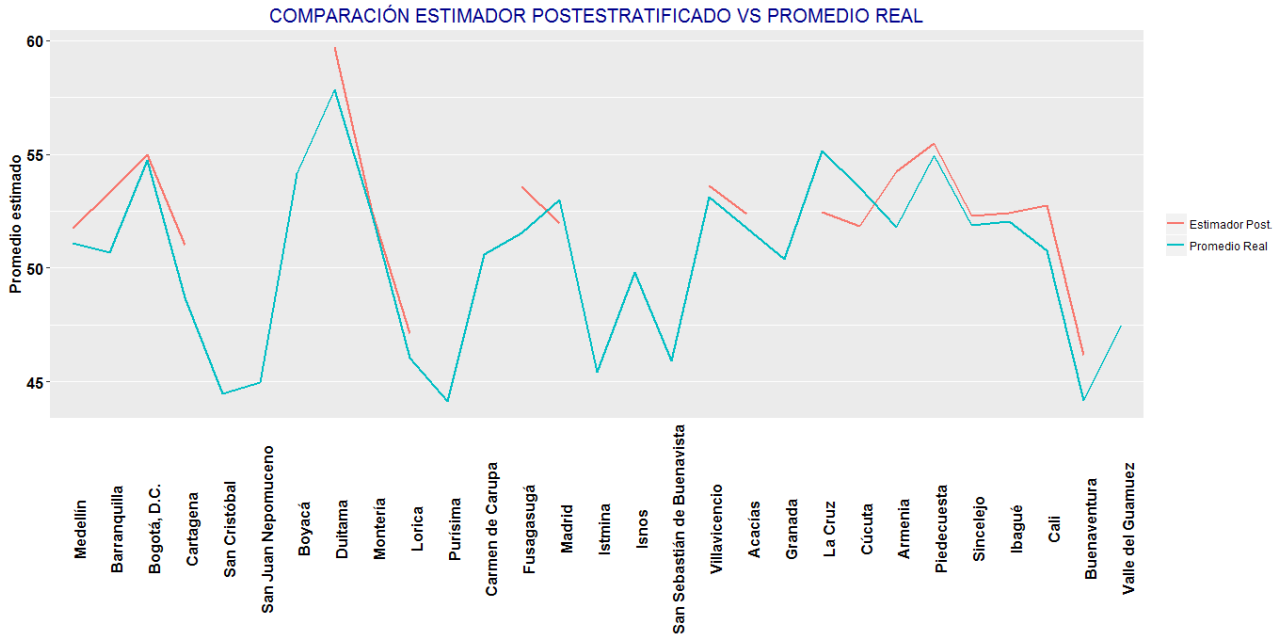

Figura 8: Gráfica de estimaciones del promedio puntaje de matemáticas. Estimador Postestratificado vs Promedio Real

6.4. Aplicación Estimador de Razón

La estructuración del estimador de razón, se basa en la utilización de una única variable de información auxiliar. Para ello, y dado que la teoría de muestreo sugiere una fuerte correlación entre la característica de interés Y_j y la variable de información auxiliar x, se realiza la busquedad de dicha variable en la base de datos de otras puntuaciones obtenidas en diferentes asignaturas, como lo son el puntaje de ciencias naturales, el puntaje de inglés, el puntaje de lectura crítica y el puntaje de ciencias sociales y ciudadanas.

Realizando los calculos de correlación, se identifica que la variable que contiene el puntaje de ciencias naturales, es la que mayormente está correlacionada con el puntaje de matemáticas con el 0.81 de correlación. Esto se puede observar en la figura 9, que relaciona la matriz de correlaciones.

Dado lo anterior, se toma como variable auxiliar PUNT_C_NATURALES que hace referencia a las puntuaciones obtenidas en esta asignatura.

	PUNT_C_NATURALES	PUNT_INGLES	PUNT_LECTURA_CRITICA	**PUNT_MATEMATICAS**	PUNT_SOCIALES_CIUDADANAS
PUNT_C_NATURALES	1	0,684617457	0,746811531	0,811920487	0,795449622
PUNT_INGLES	0,684617457	1	0,654796879	0,668237525	0,681441609
PUNT_LECTURA_CRITICA	0,746811531	0,654796879	1	0,746967726	0,793575358
PUNT_MATEMATICAS	0,811920487	0,668237525	0,746967726	1	0,771454474
PUNT_SOCIALES_CIUDADANAS	0,795449622	0,681441609	0,793575358	0,771454474	1

Figura 9: Matriz de correlación

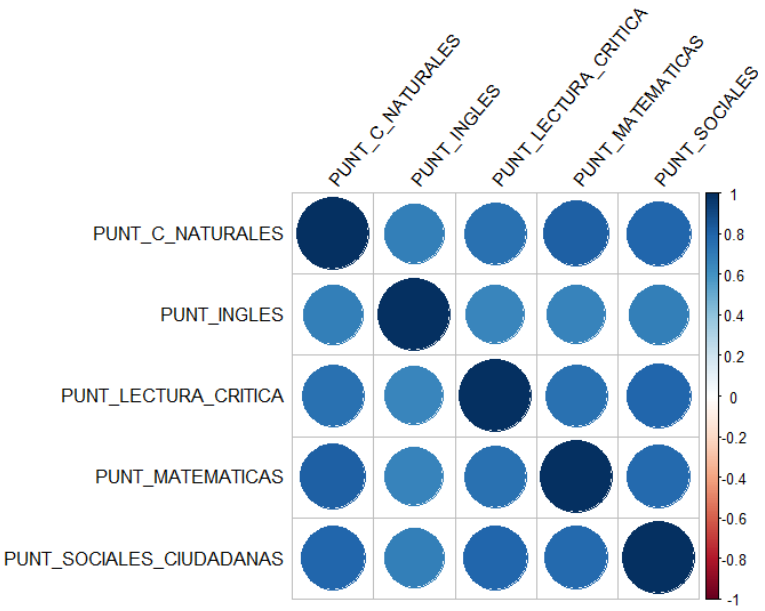

Figura 10: Diagrama de correlación

Se estima la razón como un cociente entre el total de la variable puntaje en ciencias naturales sobre su respectiva estimación en dominios, donde luego, se obtiene este estimador al multiplicar por el total estimado de la caracterísitca de interés para obtener así, una estimación de un total de Y. Como se pretende estimar el puntaje obtenido en la asignatura de matemáticas, este total estimado Y para cada dominio, se divide entre el tamaño de los dominios. Con lo cual, se tendría la estimación en las unidades correctas para cada municipio, como se evidencia en la tabla 3.

CÓDIGO	MUNICIPIO	ESTIMACIÓN	VARIANZA	CVE
05001	Medellín	51,9	0,1797	0,8160
08001	Barranquilla	50,4	0,4343	1,3088
11001	Bogotá, D.C.	55,3	0,0833	0,5222
13001	Cartagena	49,1	0,6360	1,6236
13620	San Cristóbal	46,1	0,0091	0,2074
13657	San Juan Nepomuceno	45,5	0,0106	0,2265
15104	Boyacá	54,2	0,0106	0,1905
15238	Duitama	57,5	0,0826	0,4999
23001	Montería	51,7	0,2029	0,8720
23417	Lorica	45,9	0,0530	0,5013
23586	Purísima	44,8	0,0073	0,1906
25154	Carmen de Carupa	50,5	0,0072	0,1681
25290	Fusagasugá	52,7	0,0458	0,4058
25430	Madrid	52,1	0,0388	0,3777
27361	Istmina	45,1	0,0189	0,3051
41359	Isnos	46,5	0,0664	0,5536
47692	San Sebastián de Buenavista	45,8	0,0394	0,4341
50001	Villavicencio	53,3	0,0714	0,5014
50006	Acacías	51,6	0,0217	0,2853
50313	Granada	50,7	0,0508	0,4441
52378	La Cruz	54,4	0,0537	0,4261
54001	Cúcuta	53,5	0,4753	1,2887
63001	Armenia	52,4	0,1163	0,6505
68547	Piedecuesta	54,6	0,0524	0,4196
70001	Sincelejo	50,3	0,0957	0,6148
73001	Ibagué	51,1	0,2348	0,9480
76001	Cali	50,7	0,1303	0,7119
76109	Buenaventura	44,9	0,0976	0,6960
86865	Valle del Guamuez	45,6	0,0605	0,5395

Tabla 3: Resultados Estimador de Razón

Los resultados de las estimaciones del puntaje de matemáticas se obtienen para cada municipio, asi, se observa por ejemplo, que la estimación del promedio de dicho puntaje para el municipio de Carmen de Carupa, es de 50.5 puntos. En general, las estimaciones para los municipios son eficientes, ya que los coeficientes de variación no son grandes, o mejor, se encuentran entre el 0.16 % y el 1.62 %.

En general, este estimador presenta buenos resultados. La figura 11 muestra las estimaciones del puntaje de matemáticas para cada municipio, en comparación con el promedio que realmente se presentó, identificando que el etimador de razón, tiene un buen ajuste en sus estimaciones en los municipios de Boyacá, Duitama, Monteria, Lorica, Carmen de Carupa, Villavicencio.

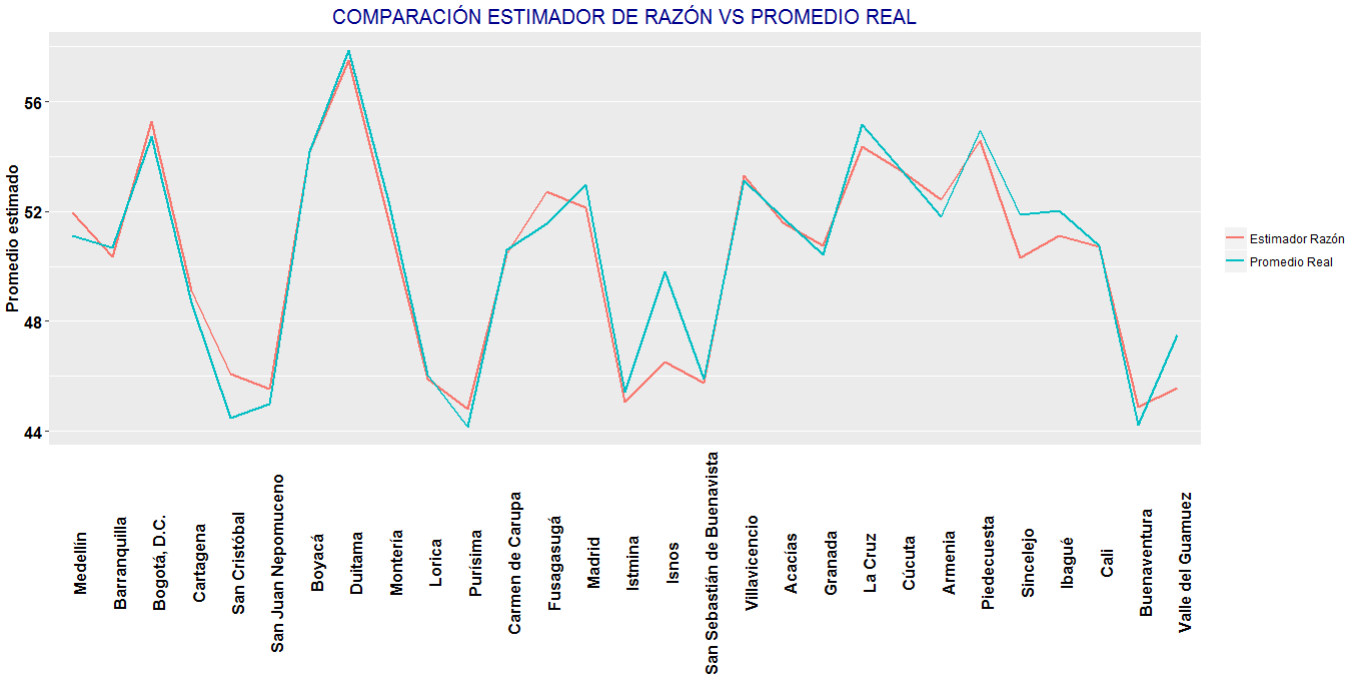

Figura 11: Gráfica de estimaciones promedio del puntaje de matemáticas. Estimador de Razón vs Promedio Real

6.5. Aplicación Estimador Sintético (estimador indirecto)

Este estimador, al igual que el estimador de post-estratificación, incorpora información de una variable de agrupación, donde se realizan cruces con la variable de interés; en el contexto del presente trabajo, dicha variable de agrupación es el desempeño de inglés, que tomará en cuenta la estimación directa del puntaje promedio de matemáticas en los promedios de cada categoría , debido a que, se trata de estimar grupos grandes, que derivaran en la estimación por dominios al ser multiplicadas por los tamaños del cruce entre los niveles del desempeño de inglés en los municipios. Lo anterior conforma la construcción del estimador sintético que genera los resultados de la tabla 4.

Se presentan los resultados de la estimación del promedio de las puntuaciones de matemáticas en cada municipio, donde se observa, que estas estimaciones no son eficientes en municipios como Medellín, Barranquilla, Bogotá y Cartagena. En municipios más pequeños, este estimador presenta un coeficiente de variación menor, donde el estimador tuvo un mejor desempeño, en municipios como San Cristóbal, Boyacá, Purísima, Carmen de Carupa.

CÓDIGO	MUNICIPIO	ESTIMACION	VARIANZA	CVE
05001	Medellín	51,9	84,0517	17,6502
08001	Barranquilla	54,4	24,0978	9,0320
11001	Bogotá, D.C.	53,7	95,8187	18,2284
13001	Cartagena	51,2	29,5114	10,6012
13620	San Cristóbal	48,8	1,5663	2,5635
13657	San Juan Nepomuceno	47,3	6,3479	5,3300
15104	Boyacá	51,4	1,1278	2,0658
15238	Duitama	56,3	15,4919	6,9968
23001	Montería	53,7	10,7389	6,1025
23417	Lorica	50,5	8,5251	5,7826
23586	Purísima	47,2	3,6571	4,0494
25154	Carmen de Carupa	48,7	1,6613	2,6467
25290	Fusagasugá	52,9	9,4712	5,8227
25430	Madrid	52,6	3,6589	3,6334
27361	Istmina	48,1	3,7195	4,0108
41359	Isnos	46,0	4,1296	4,4182
47692	San Sebastián de Buenavista	45,5	3,0343	3,8243
50001	Villavicencio	53,1	22,6364	8,9564
50006	Acacías	51,2	4,5993	4,1866
50313	Granada	49,8	4,3922	4,2073
52378	La Cruz	50,8	3,5366	3,7005
54001	Cúcuta	51,0	32,5078	11,1734
63001	Armenia	53,8	21,5100	8,6235
68547	Piedecuesta	53,7	12,2053	6,5071
70001	Sincelejo	52,3	9,3543	5,8499
73001	Ibagué	54,6	38,8999	11,4224
76001	Cali	53,9	72,1184	15,7442
76109	Buenaventura	49,4	29,4331	10,9790
86865	Valle del Guamuez	46,0	3,0972	3,8263

Tabla 4: Resultados Estimador Sintético

Gráficamente, observando la figura 12, el estimador sintético no tiene un buen desempeño, debido a la sub-estimación, o sobre-estimación que presenta en algunos municipios.

En el municipio de Medellín se observa una sobre-estimación del estimador, con respecto al promedio real registrado para este municipio. Al igual que en Medellín, en los municipios de San Cristóbal, Lorica, Purísima, Istmina, Ibagué, Cali y Buenaventura, se visualiza la sobre-estimación del estimador.

La sub-estimación relevante, se registra en el municipio de la cruz.

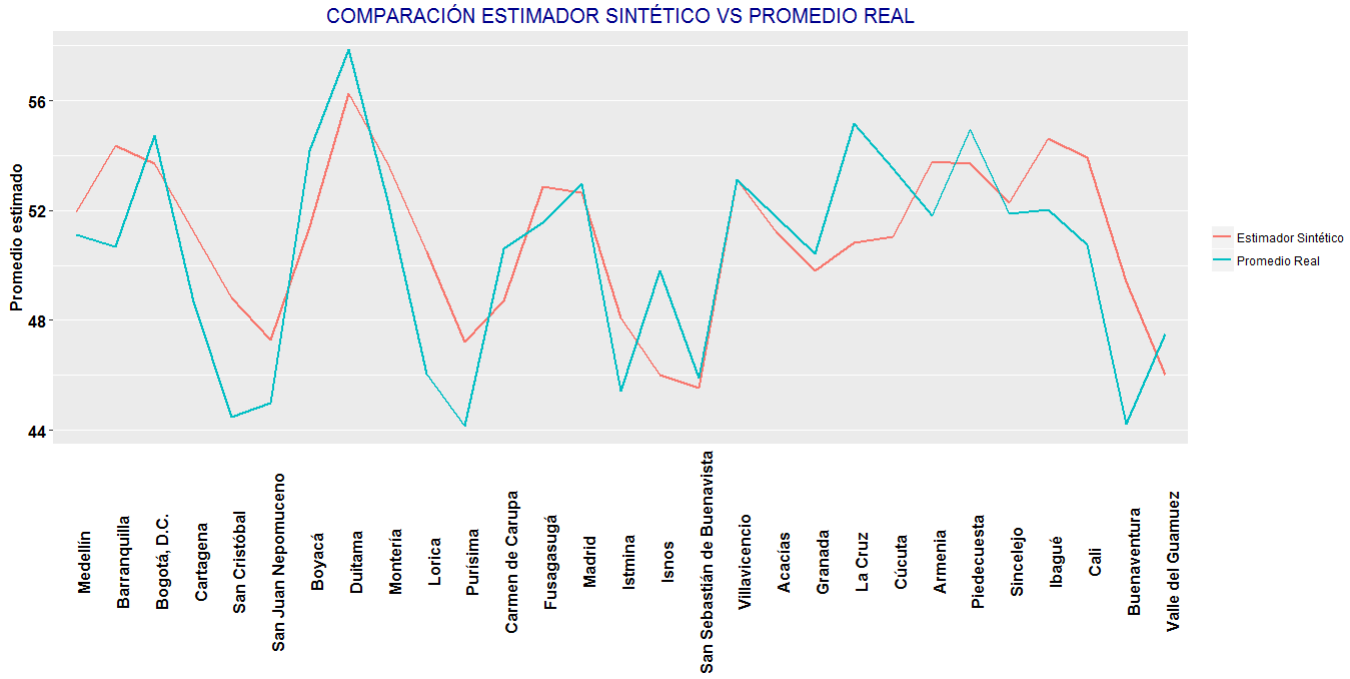

Figura 12: Gráfica de estimaciones promedio del puntaje de matemáticas. Estimador de Sintétitico vs Promedio Real

6.6. Aplicación Modelos Lineales Mixtos: Método de Fay Herriot - Modelos a nivel de área

6.6.1. Estimación directa del promedio de matemáticas por municipio

En los estimadores directos y en el estimador indirecto (estimador sintético), se identifican a los dominios como los municipios donde se realizan las estimaciones, en este contexto de modelos mixtos, dichos dominios son llamados áreas y es allí donde se requiere información auxiliar específica para cada área pequeña, que en este caso son los municipios seleccionados en la muestra, donde dicha información auxiliar requiere de la disponibilidad para cada área observada.

Los modelos a nivel de área, relacionan estimaciones directas de cada área pequeña con la información auxiliar específica disponible para cada área. De esta manera, y dado que el objetivo es estimar los promedios de los puntajes de matemáticas, se realizó la búsqueda de información auxiliar que estuviera especificada por cada dominio, donde se halló la Medición de desempeño Municipal que tiene como objetivo, *medir, comparar y ordenar los municipios según su desempeño integral entendiendo como capacidad de gestión y resultados de desarrollo teniendo en cuenta sus dotaciones iniciales, para incentivar una mejor gestión, calidad del gasto y la inversión orientada a resultados.* Medición de Desempeño Municipal DNP (2016) [10].

Como se evidenció en la aplicación del estimador directo para el promedio de matemáticas $\hat{\bar{Y}}_d^{dir}$, se obtienen estas estimaciones por municipio, las cuales buscan una asociación con la información auxiliar.

6.6.2. Modelo lineal Propuesto

Se ajusta el modelo lineal saturado como $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + u_d + \hat{e}_d$ que en términos del puntaje de matemáticas y las variables auxiliares se expresa como:

$Puntaje\ de\ matematicas = \beta_0 + \beta_1 * Movilizacion\ de\ recursos + \beta_2 * Ejecucion\ de\ Recursos + \beta_3 * Ordenamiento\ Territorial + \beta_4 * Gobierno\ Abierto\ y\ Transparencia + \beta_5 * Educacion\ 2016 + \beta_6 * Salud\ 2016 + \beta_7 * Servicios\ 2016 + \beta_8 * Seguridad 2016 + efectos\ aleatorios\ de\ area + errores\ de\ muestreo$

A través del método de Stepwise, se genera el modelo reducido:

$Puntaje\ de\ matematicas = \beta_0 + \beta_1 * Ordenamiento\ Territorial + \beta_2 * Gobierno\ Abierto\ y\ Transparencia + \beta_3 * Educacion\ 2016 + \beta_4 * Salud\ 2016 + \beta_5 * Seguridad\ 2016 + efectos\ aleatorios\ de\ area + errores\ de\ muestreo$

Este modelo reducido, tiene un coeficiente de determinación ajustado del 59 %, es decir que éste es el porcentaje de variabilidad ajustado por el modelo.

6.6.3. Método de Fay-Herriot

Una vez identificado el modelo reducido, se calcula el estimador de error cuadrático medio a través del EBLUP, bajo un modelo de Fay-Herriot, a través de la función mseFH de la librería sae del lenguaje estadísto R, con la cual, se llega a la estimación del puntaje promedio de matemáticas para cada municipio observado.

La tabla 5 presenta los resultados de la estimación, del método de Fey-Herriot donde se evidencian las estimaciones para cada municipio seleccionado, mostrando eficiencia en las estimaciones realizadas con coeficientes de variación

CÓDIGO	MUNICIPIO	ESTIMACION	MSE	CVE
05001	MEDELLIN	52,1	1,2815	2,1718
08001	BARRANQUILLA	52,8	3,7585	3,6690
11001	BOGOTA	55,5	1,0598	1,8540
13001	CARTAGENA	50,8	3,1053	3,4705
13620	SAN CRISTOBAL	48,1	0,0186	0,2835
13657	S.JUAN NEPOMUCENO	46,5	0,1180	0,7395
15104	BOYACA	56,0	0,0228	0,2694
15238	DUITAMA	59,8	0,4597	1,1346
23001	MONTERIA	52,4	1,8088	2,5668
23417	LORICA	46,9	0,2700	1,1086
23586	PURISIMA	44,8	0,0187	0,3058
25154	CARMEN DE CARUPA	51,3	0,0170	0,2542
25290	FUSAGASUGA	53,5	0,4108	1,1982
25430	MADRID	51,8	0,1981	0,8594
27361	ITSMINA	47,5	0,0703	0,5577
41359	ISNOS	42,2	0,0846	0,6901
47692	SAN SEBASTIAN	44,3	0,0586	0,5463
50001	VILLAVICENCIO	53,5	0,7518	1,6197
50006	ACACIAS	52,5	0,1232	0,6692
50313	GRANADA	50,9	0,2766	1,0333
52378	LA CRUZ	52,2	0,2326	0,9236
54001	CUCUTA	51,9	2,7718	3,2064
63001	ARMENIA	55,5	1,2108	1,9830
68547	PIEDECUESTA	58,0	0,6437	1,3837
70001	SINCELEJO	51,8	0,6105	1,5098
73001	IBAGUE	50,9	0,6928	1,6340
76001	CALI	52,3	0,8709	1,7835
76109	BUENAVENTURA	45,7	0,3258	1,2497
86865	VALLE GUAMUEZ	43,3	0,1744	0,9639

Tabla 5: Estimación del promedio de matemáticas por el método de Fey-Herriot

En la figura 13, se ilustra el comportamiento de este estimador, donde presenta un desempeño irregular debido a que, sobre-estima y sub-estima el puntaje de interés, una vez que se realiza la comparación con el promedio real del puntaje de matemáticas. Por ejemplo, en el municipio de Isnos, el método de Fay-Herriot sub-estima el promedio del puntaje, como también en el municipio del Valle del Guamuez.

La sobre-estimación importante, se presenta en Piedecuesta, Duitama, Fusagasugá, San Cristóbal.

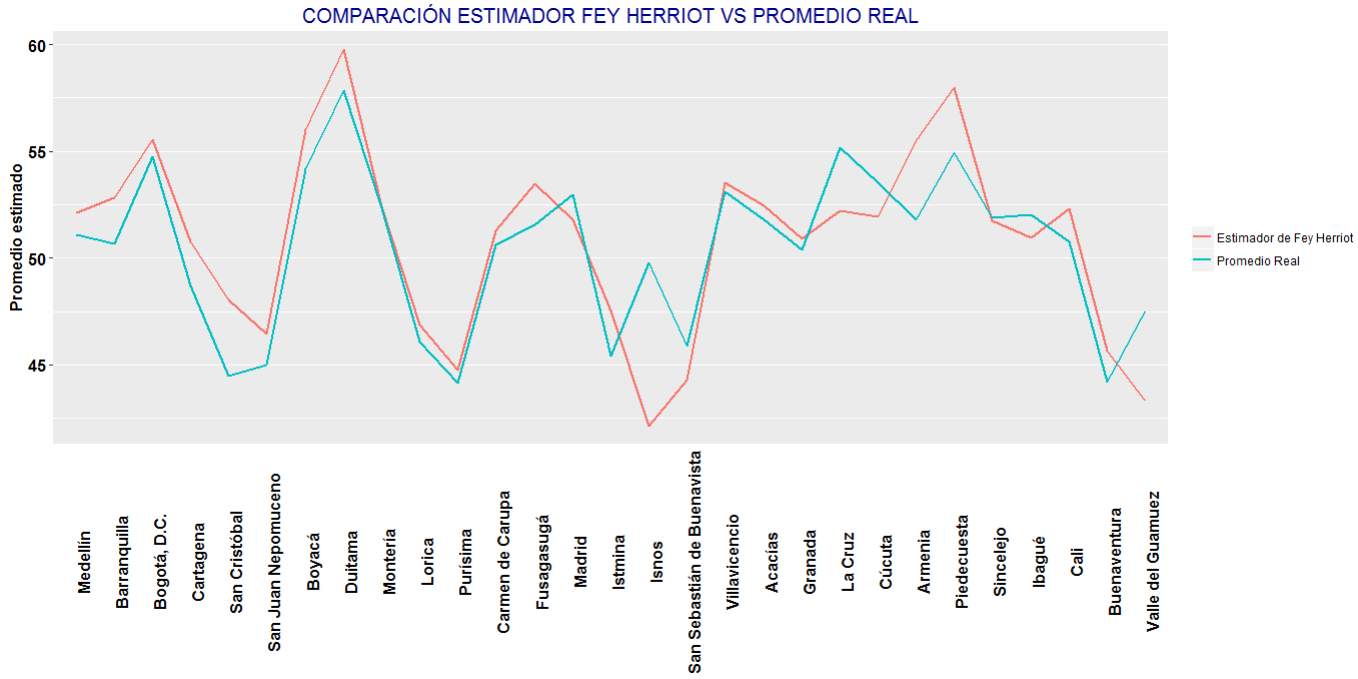

Figura 13: Gráfica de estimaciones promedio del puntaje de matemáticas. Estimador de Fay-Herriot vs Promedio Real

6.7. Aplicación Modelos Lineales Mixtos: Modelos a nivel de unidad

Para llevar a cabo la aplicación de los modelos mixtos a nivel de unidad, se utiliza el método de BATTISSE, donde se tomará como unidad observacional a los estudiantes. Se requiere que la información auxiliar esté dispobible para cada individuo $j - esimo$, con la cual se busca ajustar un modelo que relacione al puntaje de matemáticas con la matriz de variables auxiliares.

Como se ha envidenciado en la figura 9 de la matriz de correlación, la variable que contiene el puntaje de ciencias naturales, junto con la variable que mide el puntaje de ciencias sociales y ciudadanas, son las que presentan mayor correlación con el puntaje de matemáticas, de igual forma, también se hace uso de la variable categórica, que mide el desempeño de inglés.

Con estas tres variables auxiliares, se procede a calcular las estimaciones para las áreas observadas, y de esta forma, identificar que tan precisa es la estimación del puntaje de matemáticas.

La tabla 6 muestra los cálculos realizados por el método de Battisse, donde se evidencian estimaciones eficientes, debido a que registra, coeficientes de variación pequeños, que están entre 0.54 % y 1.70 %.

CÓDIGO	MUNICIPIO	ESTIMACIÓN	MSE	CVE
11001	Bogotá, D.C.	54,8	0,0220	0,2705
13001	Cartagena	48,5	0,0701	0,5464
13620	San Cristóbal	45,5	0,6026	1,7071
13657	San Juan Nepomuceno	45,3	0,3365	1,2792
15104	Boyacá	54,0	0,6999	1,5498
15238	Duitama	57,4	0,1428	0,6586
23001	Montería	51,8	0,1895	0,8408
23417	Lorica	45,9	0,1934	0,9571
23586	Purísima	45,5	0,4272	1,4367
25154	Carmen de Carupa	49,6	0,6666	1,6464
25290	Fusagasugá	52,3	0,1743	0,7981
25430	Madrid	52,4	0,2882	1,0248
27361	Istmina	45,3	0,3983	1,3935
41359	Isnos	48,8	0,4101	1,3125
47692	San Sebastián de Buenavista	46,1	0,5294	1,5795
50001	Villavicencio	53,3	0,0814	0,5353
50006	Acacías	51,2	0,2662	1,0071
05001	Medellín	51,2	0,0293	0,3345
50313	Granada	50,6	0,4165	1,2761
52378	La Cruz	54,0	0,3493	1,0942
54001	Cúcuta	52,8	0,0745	0,5167
63001	Armenia	52,5	0,0885	0,5669
68547	Piedecuesta	54,5	0,1286	0,6582
70001	Sincelejo	51,3	0,1781	0,8229
73001	Ibagué	52,1	0,0565	0,4566
76001	Cali	51,2	0,0308	0,3427
76109	Buenaventura	44,7	0,0906	0,6729
08001	Barranquilla	51,2	0,0991	0,6152
86865	Valle del Guamuez	47,6	0,5350	1,5380

Tabla 6: Estimación - Método de Battisse

Para observar el comportamiento del estimador de Battisse, la figura 14 relaciona el promedio real del puntaje de matemáticas, junto con las estimaciones realizadas, mostrando un desempeño importante, debido a la gran aproximación de estas esimaciones al puntaje promedio real, como ocurre en los Municipios de Medellín, Bogotá D.C., Cartagena, Boyacá, Lorica, Isnos, San Sebastián de Buenavista, Villavicencio, La cruz, Ibagué, y Valle del Guamez.

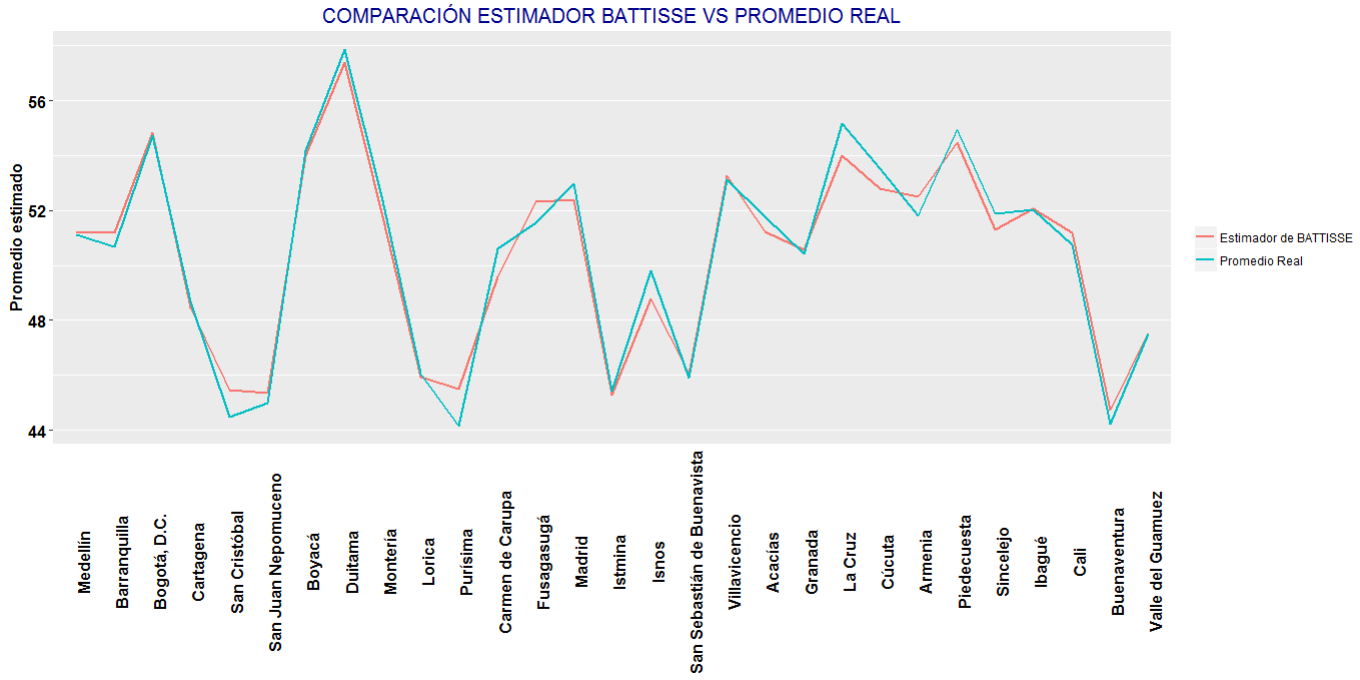

Figura 14: Gráfica de estimaciones promedio del puntaje de matemáticas. Estimador de Battisse vs Promedio Real

7. Conclusiones

Se ha observado el desempeño de cada método de estimación, generando resultados y observaciones interesantes. Los procedimientos y herramientas estadísticas utilizadas, pueden estar sujetas a mejoras y sugerencias de parte de algún lector interesado en observar el presente trabajo de grado.

Los hallazgos identificados en los resultados generados en cada procedimiento de los respectivos estimadores analizados, se expresan concluyendo que:

- El estimador basado en el método de Battisse, presenta estimaciones más precisas que el estimador directo, lo anterior se debe a que la información auxiliar disponible a nivel de unidad presenta una correlación aceptable con la variable de interés.
- El método de Fay-Herriot, en algunos municipios, registra un buen desempeño de estimación, pero, siendo su comportamiento muy similar al estimador de Horvitz-Thompson. Esto, debido a que la información auxiliar no muestra la mejor correlación con la característica de interés. El estimador sin embargo, presenta un buen desempeño para municipios que tuvieron en sus estimaciones directas coeficientes de variación muy altos.
- Por medio del estimador de razón, las estimaciones son precisas, ya que los resultados son aproximados a lo que pasa realmente con el promedio real del puntaje de matemáticas, con coeficientes de variación razonables, que acreditan su eficiencia.
- El estimador de Battisse presenta mejoras con respecto al estimador de Horvitz Thompson una vez que éste, no obtiene un buen desempeño en algunos municipios, de igual forma, el estimador de Horvitz-Thompson, en algunas áreas, registra buenas estimaciones presentando coeficientes de variación menores que el estimador de Battisse.
- Para obtener estimaciones precisas con el estimador sintético y el estimador postestratificado se requiere que en los cruces entre la variable de agrupación y el dominio exista un buen tamaño de muestra, además la variable de interés debe ser homógenea en cada una de las categorías de la variable de agrupupación.

Referencias

[1] Andrés Felipe Ortiz & José Fernando Zea (2017), *Estimación de áreas pequeñas para las condiciones de vida de los municipios de cundinamarca*, XXVII Simposio Internacional de Estadística - 5th International Workshop on Applied Statistics, Medellín, Antioquia, Colombia, 8 al 12 de agosto de 2017.

[2] Carl-Erick Särndal, Bengt Swensson, Jan Wretman(1992), *Model Assisted Survey Sampling*. Springer, Verlag New York, Inc.

[3] Domingo Morales González (2015), *ESTMACIÓN EN ÁREAS PEQUEÑAS: MÉTODOS BASADOS EN MODELOS*. Universidad Miguel Hernández de Elche.

[4] Guías, Descripción de los Niveles de Desempeño (2015), Instituto Colombiano para la Evaluación de la Educación ICFES.

[5] Hugo Andrés Gutiérrez (2009), *Estrategias de Muestreo, diseño de encuestas y estimación de parámetros*. Universidad Santo Tomás.

[6] J.N.K. RAO & Isabel Molina (2015), *Small Area Estimation (Second Edition)*. Published by John Wiley & Sons, Inc, Hoboken, New Yersey.

[7] J.N.K. RAO (2003), *Small Area Stimation*. ISBN=0-471-41374-7 (cloth), publisher=John Wiley & Sons, Inc., Hoboken, New Jersey.

[8] Juan Pablo Ferreira Neira, *Estimación en dominios*, Universidad de la república-Uruguay, Facultad de ciencias económicas y de Administración - Licenciatura en estadística, Tutor: Guillermo Zoppolo. Mayo de 2011.

[9] Pushpal K Mukhopadhyay and Allen McDowell (2011), *Small Area Estimation for Survey Data Analysis Using SAS Software*, SAS Institute Inc., Cary, NC.

[10] https://www.dnp.gov.co/programas/desarrollo-territorial/Estudios-Territoriales/Indicadores-y-Mediciones/Paginas/desempeno-integral.aspx, *Nueva Medición de Desempeño Municipal MDM*, Resultados y anexos de Medición de Desempeño Municiapal vigencia 2016. Departamento Nacional de Planeación DNP.

[11] https://www.youtube.com/watch?v=zTVuxf8f1a0, *Innovación Aprendizaje (2016) - DANE*. Videotutorial estimador de postestratificación módulo 2 MAM.