



UNIVERSIDAD
NACIONAL
DE COLOMBIA

UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

MAESTRÍA EN CIENCIAS - ESTADÍSTICA

**Estimación de Proporción en Áreas Pequeñas:
Enfoque basado en Aprendizaje Automático**

**Estimation of Proportions in Small Area Estimation:
Machine Learning Approach**

PRESENTADO POR:

Melanie Bernal Malpica

DIRECTOR:

Leonardo Trujillo Oyola

Bogotá, 2024

Agradecimientos

En primer lugar, agradezco a Dios por darme la fortaleza para completar este proyecto.

A mis padres, por su amor incondicional, por ser mi fuente de inspiración y por su apoyo constante en cada paso de mi formación académica.

También quiero expresar mi gratitud al profesor Leonardo Trujillo y a la profesora Natalia Rojas Perilla, por su valiosa orientación, apoyo, paciencia y por compartir su conocimiento a lo largo de este proceso, sus valiosos aportes hicieron posible este trabajo.

De igual manera, agradezco a mis amigos y compañeros de estudio, quienes con sus palabras de aliento, colaboración y compañía me ayudaron a culminar este proyecto.

Resumen

En los estudios de encuestas por muestreo, es común que los investigadores requieran estimaciones a nivel de dominios. Sin embargo, estos dominios suelen presentar una muestra reducida o incluso nula, lo que genera varianzas estimadas elevadas y, en consecuencia, estimaciones que no cumplen con los estándares de calidad requeridos. En los casos donde no hay muestra en un dominio específico, ni siquiera es posible calcular el estimador de interés utilizando el diseño muestral.

Para abordar esta problemática, surge la metodología de estimación en áreas pequeñas (SAE, por sus siglas en inglés), que permite obtener estimaciones confiables a partir del uso de información auxiliar disponible para toda la población. Esta metodología emplea modelos estadísticos que combinan los datos muestrales con predicciones sobre las unidades no observadas, permitiendo así obtener estimaciones precisas, incluso en dominios sin muestra. Generalmente, se utilizan modelos lineales mixtos para variables continuas y modelos lineales generalizados mixtos en el caso de proporciones.

Los modelos tradicionales requieren cumplir ciertos supuestos, como la relación lineal entre las variables auxiliares y la variable objetivo, así como la normalidad de los errores asociados. Además, presentan limitaciones como la multidimensionalidad y la sensibilidad a valores atípicos. Por esta razón, es necesario explorar enfoques más flexibles.

El propósito de este trabajo es presentar una metodología basada en modelos de aprendizaje automático con efectos mixtos, que permite calcular los estimadores en áreas pequeñas sin depender de los supuestos lineales. Esta estrategia ofrece ventajas como la robustez ante valores atípicos y una mejor selección de variables. Sustituyendo el modelo lineal por un modelo de aprendizaje automático, se siguen los mismos pasos de estimación del parámetro y su medida de error según la metodología SAE. Finalmente, se realizará un ejercicio de simulación basado en el modelo para comparar las estimaciones, el error cuadrático medio y el sesgo de cada metodología evaluada. Los resultados muestran que los modelos propuestos constituyen una alternativa viable, ya que logran estimaciones similares a las metodologías tradicionales, obteniendo una ganancia frente a los supuestos en la metodología tradicional.

Palabras clave: Estimación, área pequeña, proporción, modelos, semiparamétrico, machine learning.

Abstract

Sample surveys have been traditionally recognized as cost-effective means of obtaining information to provide estimates for different parameters, not only for the total population of interest but also for various subpopulations (domains) not large enough (even null) to support direct estimates of adequate precision and then not publishable. Small area estimation is a methodology that considers diverse methods to use available auxiliary information for the whole population to allow us to estimate the parameters in the domains (small areas). One possibility is to consider a linear mixed model or a mixed generalised model in the case of estimating a total population to estimate the variable of interest for the non-sampled units, allowing us to get an estimation for all the domains combining sampling units and non-sampling units. However, traditional models must fulfill some assumptions; for instance, the relationship between the auxiliary variables and the variable of interest must be linear, and the associated prediction errors must follow a particular probability distribution, raising problems of multicollinearity and outliers in some cases.

Therefore, we propose in this paper a strategy to substitute the traditional mixed generalised model for a more flexible one. In particular, we study a different approach using machine learning regression methods with mixed effects for estimating proportions in small areas without considering any assumptions and obtaining a gain in robustness for outliers and variable selection. Some approaches have already been proposed in the literature for small-area estimation of proportions. The idea is to substitute the linear model with a machine learning regression method following the same stages for estimating the parameter and its precision according to traditional small-area estimation methods. We present a simulation exercise considering model-based and design-based inferences (logistic mixed models, mixed effects random forest, and mixed effects tree boosting) to compare mean squared errors, biases, and computation times for all the methods considered. Also, an actual application for the evaluation of the National Program for the Substitution of Illicit Crops in Colombia is shown, considering these methods to estimate the proportion of families that have suffered forced eradication in the rural areas of the country.

Keywords: Estimation, small area, proportion, models, semiparametric.

Índice general

Índice de Figuras	9
Índice de Tablas	11
1 Introducción	1
1.1 Notación matemática	4
2 Marco Teórico	5
2.1 Estimación basada en el diseño y en el modelo	6
2.1.1 Estimación basada en el diseño	7
2.1.2 Criterios para la publicación de las estimaciones	8
2.1.3 Estimación basada en el modelo	11
2.2 Metodologías de estimación en áreas pequeñas	12
2.2.1 Estimador EBLUP basado en el modelo Fay-Herriot	12
2.2.2 Estimador EBLUP basado en el modelo con errores anidados	14
2.2.3 Mejor predictor empírico bajo el modelo con errores anidados (EB)	16
2.2.4 Estimación basada en modelos lineales generalizados mixtos	18
2.3 Exploración de métodos basados en aprendizaje automático en SAE	20
2.3.1 Modelo MERF para la estimación de medias en SAE	20
2.4 Árbol de decisión con efectos mixtos para la estimación de proporciones	25
3 Metodología propuesta	29
3.1 Estimador Plug-in	30
3.2 Estimación de la incertidumbre	35
3.3 Escenario de simulación	37
3.4 Aplicación en datos reales	47
4 Conclusiones y futuros trabajos	57

Índice de Figuras

1.1	Notación matemática.	4
2.1	Criterios para la publicación.	10
2.2	Comparación de técnicas	26
2.3	Bagging y Boosting	27
3.1	Función de enlace logístico	31
3.2	Ejemplo partición recursiva	32
3.3	Comparación del sesgo en la simulación	40
3.4	Comportamiento del sesgo según el tamaño muestral	40
3.5	Comparación del error cuadrático medio en la simulación	42
3.6	Comportamiento del ECM según el tamaño muestral	44
3.7	Comparación de la varianza en la simulación	45
3.8	Esquema del diseño	48
3.9	Tamaño muestral y poblacional por departamento	49
3.10	QQ-plot de los efectos aleatorio del modelo GLMM	51
3.11	Comparación medidas de ajuste	52
3.12	Comparación de las estimaciones	53
3.13	Comparación error cuadrático medio	55

Índice de Tablas

3.1	Comparación del sesgo en la simulación	41
3.2	Comparación del error cuadrático medio en la simulación	43
3.3	Comparación de las estimaciones en la simulación	46
3.4	Descripción de la muestra	48
3.5	Estimaciones directas por departamento	50
3.6	Comparación estimaciones	54

1. Introducción

Las encuestas por muestreo son una herramienta importante para para los gobiernos, centros de investigación, empresas, entidades públicas y demás organizaciones que deseen conocer alguna característica sobre una población de interés. Un censo no siempre resulta ser la mejor opción; su alto costo, complejidad de implementación, tiempo que lleva la recolección de toda la información, entre otras problemáticas hacen que evaluar la característica de interés sobre toda la población no sea lo mas eficiente. Es por esto que es necesario implementar una estrategia muestral probabilística donde, por medio de un método de selección se extrae de la población una muestra sobre la cual se mide la variable objetivo y haciendo uso del diseño muestral y los estimadores de los parámetros que se desean conocer, se puede inferir sobre esa muestra a toda la población.

Al plantear el diseño muestral, se calcula el tamaño de muestra óptimo que permite hacer estimaciones sobre la variable de interés con una precisión confiable. Cuando se hace el cálculo del tamaño muestral, normalmente se realiza para el total de la población o para los dominios en los que se espera tener estimación confiable, con el fin de obtener estimaciones de la variante del estimador con una magnitud pequeña. Sin embargo, en la implementación del estudio cuando ya se obtuvo la muestra puede surgir el interés de calcular estimaciones sobre algún dominio en la población no planificado. Como la estrategia muestral fue pensada para la estimación del total poblacional cuando se evalúan las estimaciones sobre los dominios se puede dar el caso de tener coeficientes de variación estimados altos que no permiten hacer inferencia válida a la población. También es posible que para ciertos dominios ninguno de sus elementos resulte seleccionado en la muestra por lo que ni siquiera se puede pensar en obtener una estimación haciendo uso del diseño muestral.

En la literatura se encuentran diferentes propuestas para dar solución a los problemas mencionados, la metodología de estimación de áreas pequeñas o modelos SAE utiliza información auxiliar disponible para toda la población con el fin de obtener estimaciones de

la característica de interés sobre los dominios con o sin muestra y con una precisión confiable que permita hacer válida la inferencia en estos dominios. En los métodos basados en el modelo para SAE se plantea que haciendo uso de los datos muestrales y la información auxiliar que se dispone, se construye un modelo lineal mixto donde la estructura jerárquica corresponde a las áreas o dominios de interés sobre los cuales se desea hacer estimación. A partir de las estimaciones de este modelo, se construye un estimador compuesto, el cual permite estimar sobre las áreas no muestreadas o las que poseen poca muestra con una mayor precisión en comparación a las estimaciones directas obtenidas a partir del diseño muestral. El problema de estimación en áreas pequeñas se ha estudiado incluso para la aplicación de pruebas estandarizadas en muestras de estudiantes e instituciones educativas como lo presentado en Tellez et al. (2020, 2021, 2024).

Los métodos tradicionales de estimación en áreas pequeñas, son usados para estimar parámetros lineales como totales, medias o proporciones. Sin embargo, para el caso de una proporción, haciendo uso del modelo a nivel de unidad, el costo computacional es muy alto, dado que para la estimación del parámetro se hace uso del método de simulación Monte Carlo y para la estimación del error cuadrático medio se usan técnicas de remuestreo, lo que hace que el proceso iterativo se vuelva extremadamente largo. En este trabajo se propone una metodología para la estimación de proporciones en áreas pequeñas, basada en la inferencia por modelo, en la cual el modelo lineal mixto clásico es sustituido por un modelo de aprendizaje supervisado.

El modelo lineal logístico mixto debe cumplir ciertos supuestos; entre estos, se debe asumir una relación lineal entre las variables auxiliares con la variable objetivo; se asume una distribución normal de los errores asociados al modelo. También es necesario considerar el problema de multicolinealidad, la multidimensionalidad y la detección de datos atípicos. Estos supuestos dificultan la implementación de los modelos lineales pues no siempre se pueden cumplir, es por esto que se hace necesario la implementación de una estrategia que nos permita sustituir el modelo lineal logístico mixto por un modelo más flexible, en donde no se deban cumplir estos supuestos y se pueda tratar los problemas mencionados. El propósito de este trabajo es presentar una metodología en la cual se sustituye el modelo lineal logístico mixto por un modelo de aprendizaje automático, más específicamente métodos de clasificación mediante árboles aleatorios con efectos mixtos, se evaluarán los métodos de bagging y boosting para reducir la varianza y el sesgo en las estimaciones. Bajo esta metodología no se deben cumplir los supuestos mencionados anteriormente, lo cual facilita su implementación, por lo tanto, el siguiente paso es evaluar las estimaciones por medio de las medidas de precisión y compararlas con las estimaciones directas y los métodos tradicionales de áreas pequeñas.

La metodología propuesta será evaluada por medio de un ejercicio de simulación, en donde se probará cada uno de los modelos en diferentes escenarios de muestra y supuestos, en donde se determina que el modelo GPBOOST el cual es un proceso gaussiano que usa la

técnica de ensamblaje boosting resulta ser una excelente alternativa no paramétrica para la estimación de la proporción en áreas pequeñas, mostrando resultados similares en cuanto a las estimaciones con el modelo mixto tradicional y una mejora en las estimaciones de las medidas de precisión. Finalmente se emplea esta técnica en datos reales, para obtener la estimación de una proporción en la encuesta de evaluación del programa PNIS (programa de sustitución de cultivos ilícitos en Colombia), obteniendo información no solo a nivel nacional si no a nivel de todos los departamentos participantes en el programa.

1.1. Notación matemática

Supongamos que se tiene una población finita U de individuos de tamaño finito N que consta de D subconjuntos denominados dominios, los cuales se denotan como U_1, U_2, \dots, U_D con tamaños poblacionales $N_1, N_2, \dots, N_i, \dots, N_D$, donde el índice $i = 1, \dots, D$ indica cada dominio o área. Para cada individuo dentro de la población se tiene un vector de información auxiliar denotado por \mathbf{x}_{ij} , el cual está compuesto por p variables.

Para cada individuo se define la variable respuesta y_{ij} la cual corresponde a la observación para el individuo j ($j = 1, \dots, N_i$) en el área i . El objetivo de este trabajo es estimar una proporción poblacional por lo que la variable objetivo es de tipo dicotómica $Y_{ij} \in \{0; 1\}$, el valor 1 indica la tenencia de la característica de interés. Para estimar esta proporción poblacional, se extrae una muestra s del universo U por medio de un diseño muestral probabilístico $p(s)$. Esta muestra consta de n unidades divididas en tamaños de muestra n_1, n_2, \dots, n_D para todas las áreas D . Denotamos a s_i como la submuestra del área i ; cabe resaltar que es posible tener tamaños de muestra $n_i = 0$. La variable objetivo Y_{ij} está disponible para cada unidad dentro de la muestra.

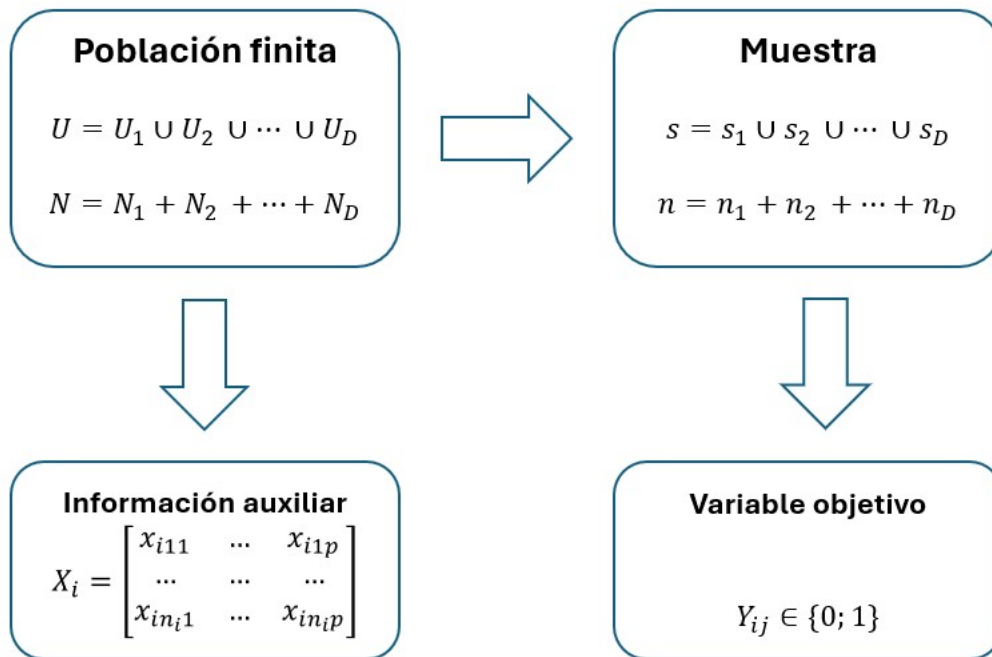


Figura 1.1: Notación matemática.

Fuente: Elaboración propia.

2. Marco Teórico

Cuando se requiere hacer estimación en dominios no planificados en la estrategia muestral en donde las estimaciones directas no proporcionan estimaciones confiables se tienen varias alternativas para su tratamiento. En primer lugar, podemos pensar en aumentar el tamaño de muestra en estos dominios; sin embargo, esta solución resulta ser muy costosa y poco eficiente. Es por esto que surge la metodología denominada estimación de áreas pequeñas, la cual consiste en estimar parámetros en subconjuntos de la población (llamados áreas pequeñas o dominios) haciendo uso de información auxiliar disponible para todos los individuos de la población o para todas las particiones conocidas como áreas o dominios. En esta metodología se usan las estimaciones directas construidas a partir del diseño muestral y con la ayuda de los datos auxiliares se construyen modelos que permitan hacer la estimación de la variable objetivo en las áreas con muestra pequeña o incluso nula.

Como se ha mencionado anteriormente esta metodología es posible gracias al uso de información auxiliar por lo que se define el vector $\mathbf{X}_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijp})$ el cual contiene p variables explicativas y está disponible para cada unidad j en el área i dentro del universo U .

2.1. Estimación basada en el diseño y en el modelo

En primer lugar se definen los principales parámetros a estimar, los cuales son:

- **Total poblacional:** se define como la suma de la variable de interés Y en el universo, se calcula mediante la siguiente expresión:

$$t_y = \sum_U y_j$$

El total poblacional de la variable Y en el área i es:

$$t_{yi} = \sum_{U_i} y_{ij}$$

- **Promedio o proporción poblacional:** se define como la suma de la variable de interés Y en el universo sobre el tamaño poblacional N , se calcula mediante la siguiente expresión:

$$\bar{t}_y = \frac{1}{N} \sum_U y_j$$

La proporción poblacional corresponde al caso particular de un promedio cuando y_j se asume como una variable dicotómica que toma el valor de 1 si el j -ésimo individuo tiene el atributo de interés y de 0 en otro caso.

El promedio poblacional de la variable Y en el área i es:

$$\bar{t}_{yi} = \frac{1}{N_i} \sum_{U_i} y_{ij}$$

- **Razón poblacional:** se calcula como el cociente de totales, el primer total asociado a una variable de interés Y , el segundo total asociado a una variable de interés Z , se calcula mediante la siguiente expresión:

$$R = \frac{t_y}{t_z} = \frac{\sum_U y_j}{\sum_U z_j}$$

La razón poblacional entre las variables Y y Z en el área i es:

$$R_i = \frac{\sum_{U_i} y_{ij}}{\sum_{U_i} z_{ij}}$$

Existen dos tipos de estimadores en muestreo, los que provienen de la inferencia basada en el **diseño** y la inferencia basada en el **modelo**, estos se presentan a continuación.

2.1.1. Estimación basada en el diseño

Bajo este tipo específico de estimación, se asume que el parámetro a estimar es una constante desconocida haciendo uso del diseño muestral para la estimación. La muestra se obtiene de un mecanismo de selección aleatorio por lo que para cada individuo se conoce la probabilidad de inclusión π_j , a partir de esta se construyen los pesos muestrales w_j que corresponden al inverso de la probabilidad de inclusión. El π -estimador, que definiremos más adelante, es el principal estimador basado en el diseño el cual es insesgado, sin embargo, a medida que el tamaño muestral disminuye la varianza estimada de éste aumenta.

- **Estimador del total:** para estimar el total se usa el estimador de Horvitz-Thompson (Särndal, 1992):

$$\hat{t}_{y_i} = \sum_{j \in s_i} w_{ij} y_{ij}$$

$w_{ij} = \frac{1}{\pi_{ij}}$ es el peso muestral o factor de expansión del individuo j en el área i

$$\hat{t}_y = \sum_{i=1}^D \hat{t}_{y_i}$$

donde, \hat{t}_{y_i} es la estimación de la variable objetivo Y en el área i para $i = 1, \dots, D$

La estimación de la varianza de \hat{t}_{y_i} y \hat{t}_y es Särndal (1992):

$$\hat{V}(\hat{t}_{y_i}) = \sum_{j \in s_i} \sum_{l \in s_i} \frac{\Delta_{jl}}{\pi_{jl}} \frac{y_j}{\pi_j} \frac{y_l}{\pi_l}$$

donde π_{jl} es la probabilidad de inclusión de segundo orden de los individuos i, j y $\Delta_{jl} = \pi_{jl} - \pi_j \pi_l$

$$\hat{V}(\hat{t}_y) = \sum_{j \in s} \sum_{l \in s} \frac{\Delta_{jl}}{\pi_{jl}} \frac{y_j}{\pi_j} \frac{y_l}{\pi_l}$$

- **Estimador del promedio o proporción:** para estimar el promedio se usa el estimador del total de Horvitz-Thompson (Särndal, 1992):

$$\hat{Y}_i = \frac{1}{N_i} \sum_{j \in s_i} w_{ij} y_{ij}$$

$$\hat{Y} = \frac{1}{N} \sum_{j \in s} w_j y_j$$

La estimación de la varianza de \hat{Y}_i y \hat{Y} es (Särndal, 1992):

$$\hat{V}(\hat{Y}_i) = \frac{1}{N_i^2} \sum_{j \in s_i} \sum_{l \in s_i} \frac{\Delta_{jl} y_j y_l}{\pi_{jl} \pi_j \pi_l}$$

$$\hat{V}(\hat{Y}) = \frac{1}{N^2} \sum_{j \in s} \sum_{l \in s} \frac{\Delta_{jl} y_j y_l}{\pi_{jl} \pi_j \pi_l}$$

A lo largo del documento se denota $\hat{\theta}_i^{Directo}$ como el estimador de la proporción de la variable objetivo Y en el dominio i , usando el diseño muestral, es decir a partir del estimador Horvitz-Thompson:

$$\hat{\theta}_i^{Directo} = \frac{1}{N_i} \sum_{j \in s_i} w_{ij} y_{ij}$$

2.1.2. Criterios para la publicación de las estimaciones

Determinar si la estimación obtenida a partir de una encuesta probabilística es o no publicable es un problema en el cual han trabajado diferentes oficinas de estadística. Se debe cumplir con ciertos criterios de calidad, los cuales aseguran que la cifra publicada es confiable, (CEPAL, 2023) establece un esquema, figura 2.1, con diferentes criterios para determinar si es o no publicable una estimación, las cuales tomamos como referencia para este trabajo.

Es necesario definir los siguientes términos los cuales hacen parte de los criterios de calidad:

- UPM: unidad primaria de muestreo, corresponde al primer nivel de selección en un diseño de muestreo por etapas. Por ejemplo, en encuestas de hogares, las UPM suelen ser los municipios, dentro de los cuales se seleccionan manzanas, viviendas o individuos en etapas sucesivas.
- Tamaño de muestra efectivo: cuando se obtiene una muestra a partir de un diseño muestral complejo, no se cumple que las respuestas y_1, y_2, \dots, y_n sean independientes e igualmente distribuidas debido a la forma jerárquica de la selección de los hogares y a la interrelación de la variable de interés con las UPMs, por esto se establece el tamaño de muestra efectivo el cual es una aproximación al valor del número de unidades en la muestra que cumple el principio de independencia e igualdad en la distribución.

$$n_{eff} = \frac{n}{DEFF}$$

$DEFF = \frac{\hat{V}(\hat{\theta})}{\hat{V}_{MAS}(\hat{\theta})}$, hace referencia al efecto del diseño que corresponde al cociente de la varianza estimada del estimador $\hat{\theta}$ bajo el diseño muestral propuesto sobre la varianza estimada bajo un muestreo aleatorio simple.

- Grados de libertad: Los grados de libertad cuantifican el número de unidades independientes, estos hacen parte del cálculo de los intervalos de confianza por lo que son uno de los criterios a tener en cuenta en la publicación. Bajo diseños muestrales complejos, estos se obtienen por la siguiente expresión:

$$gl = n_I - H$$

Donde n_I es el número de UPMs, las UPMs son las unidades muestrales de la primera etapa, y H es el número de estratos en la muestra.

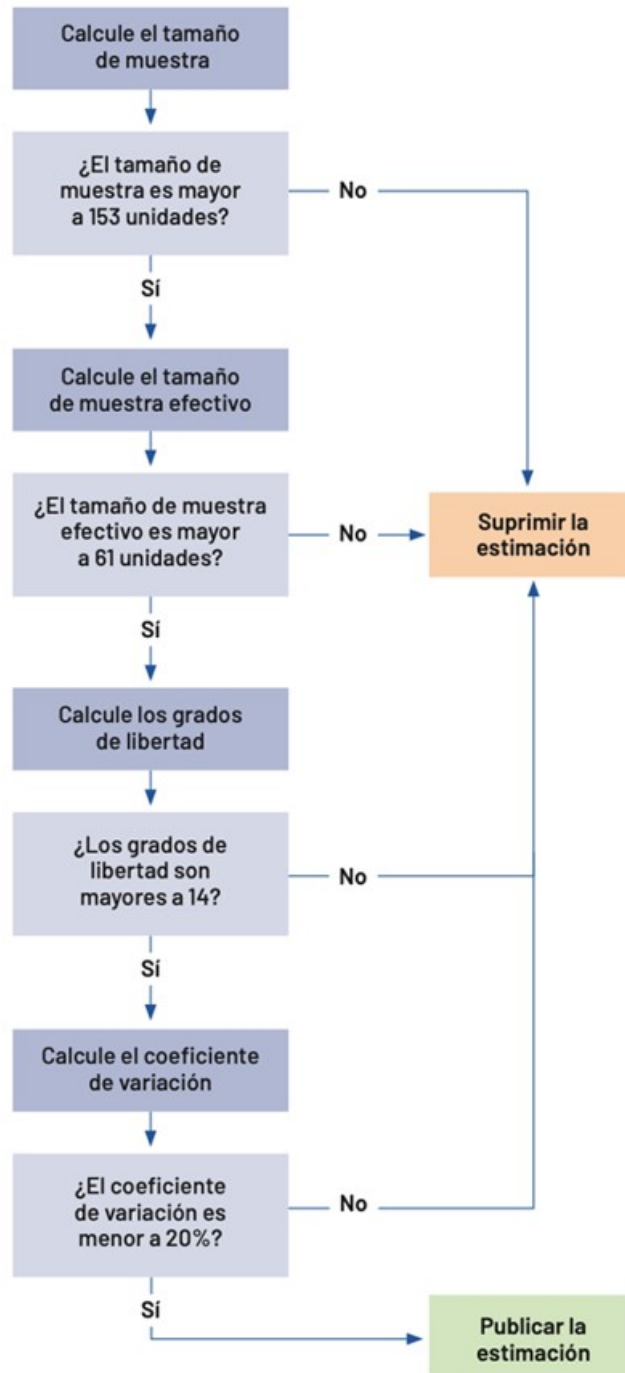


Figura 2.1: Criterios para la publicación.
Fuente: (CEPAL, 2023).

2.1.3. Estimación basada en el modelo

Bajo este tipo específico de estimación, se asume que el parámetro a estimar es una variable aleatoria. La población es considerada como la realización de un proceso aleatorio; es decir, se asume que las observaciones de la variable objetivo $Y = (y_1, y_2, \dots, y_N)'$ son realizaciones de las variables aleatorias Y_1, Y_2, \dots, Y_N , se asume que estas variables siguen una distribución de probabilidad y se hace uso de esta distribución para la estimación del parámetro de interés.

Es fundamental contar con información auxiliar fuertemente correlacionada con la variable objetivo para poder emplear esta metodología, normalmente se usan los microdatos de los censos y registros administrativos. Sin embargo esta información es difícil de conseguir debido a las políticas de confidencialidad y puede que no estén disponibles con toda la información (variables u observaciones en algunas variables) para uso público. Dependiendo de la variable objetivo y el tipo de estimador que se quiera emplear (total, media, proporción, razón, cuantiles, entre otros) es necesario contar con información a nivel de área o individuo. Por ejemplo supongamos que tenemos una variable objetivo Y de tipo continuo y queremos estimar la media de esta variable para el área i en el escenario de la estimación basada en el modelo es:

$$\hat{\theta}_i = \frac{1}{N_i} \left(\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} \mathbf{X}_{ij} \hat{\beta} \right)$$

La primera parte de la sumatoria corresponde a la suma de los datos muestrales en el dominio i , la segunda es la estimación por medio de un modelo lineal de los datos no muestrados en el dominio i . cuando estamos estimando parámetros lineales no es necesario tener la información a nivel individuo, por la forma como se define la sumatoria es suficiente conocer el total de las variables auxiliares en el dominio i , $t_{X_1} = \sum_{j=1}^{N_i} x_{ij1}$. Esto es de gran ayuda pues como mencionamos anteriormente tener acceso a la información a nivel de microdatos en ocasiones puede ser muy difícil.

Cuando nos encontramos en la situación donde se dispone de datos auxiliares altamente correlacionados con la variable respuesta se puede hacer uso de estimadores sintéticos o compuestos, los cuales son basados en el modelo:

- **Estimador sintético:** se basa en un modelo de regresión lineal el cual se puede plantear a nivel de individuo o a nivel de área, dependiendo de la información auxiliar que se disponga, está dado por la expresión Lahiri & Pramanik (2019) :

$$\hat{\theta}_i^{\text{Sintetico}} = \bar{\mathbf{x}}_i' \hat{\beta}$$

donde $\bar{\mathbf{x}}_i' = (\bar{x}_{i1}, \bar{x}_{i2}, \dots, \bar{x}_{ip})$ con $\bar{x}_{ip} = \frac{1}{N_i} \sum_{U_i} x_{ijp}$

$\hat{\beta}$ es el estimador de mínimos cuadrados ordinarios de β

-
- **Estimador compuesto:** El estimador directo es insesgado pero con varianza alta cuando se tiene poca muestra, mientras que el estimador sintético es sesgado pero con menor varianza, por lo que se considera una combinación lineal de estos denotado como el estimador compuesto Molina (2019):

$$\hat{\theta}_i^{\text{Comp}} = \alpha_i \hat{\theta}_i^{\text{Directo}} + (1 - \alpha_i) \hat{\theta}_i^{\text{Sintético}},$$

Para un $\alpha_i \in [0, 1]$. Cabe resaltar que cuando $\alpha_i = 0$, estamos en el escenario de $n_i = 0$, por lo que la estimación queda dada por la estimación del modelo. Existen diferentes métodos para la elección de α_i : una primera opción es elegir el α_i tal que minimice el error cuadrático medio; una segunda opción es seleccionarlo según el tamaño muestral del dominio, cuando se tiene una muestra grande el α_i debe ser mayor, de tal forma que se le de un mayor peso al estimador directo garantizando la disminución del sesgo y al tener buena muestra la varianza no será tan grande; en el caso contrario es mejor proporcionar un mayor peso al estimador sintético.

2.2. Metodologías de estimación en áreas pequeñas

Los métodos basados en modelos asumen un modelo para la población es decir la variable objetivo Y sigue una distribución y los valores y_i son valores aleatorios de esta distribución, así que los valores muestrales y la construcción del estimador se basa en el modelo o distribución de la población. Existen dos tipos de modelos en esta metodología:

- **Modelos a nivel de área:** la información auxiliar esta al nivel de área o el dominio a estimar, por lo que el modelo se establece a nivel de área, por ejemplo información auxiliar a nivel de municipio como la tasa de ocupación.
- **Modelos a nivel de unidad:** La información está a nivel de los individuos, es decir se tiene acceso a microdatos de la información auxiliar, por ejemplo registros administrativos o censales por persona, en este escenario se plantea el modelo a nivel de individuo o unidad.

(Molina, 2019) presenta una recopilación de las metodologías SAE desarrolladas en los últimos años:

2.2.1. Estimador EBLUP basado en el modelo Fay-Herriot

El modelo Fay-Herriot (FH) es un modelo lineal a nivel de área muy usado en las estimaciones SAE, fue propuesto por (Fay y Herriot, 1979), este modelo relaciona los parámetros de interés θ_i para todas las áreas con $i = 1, \dots, D$, asumiendo que éstos valores varían respecto de un vector de p variables auxiliares $\mathbf{x}_i' = (x_{i1}, x_{i2}, \dots, x_{ip})$ de forma constante para todas las áreas, siguiendo un modelo de regresión lineal (Molina, 2019):

$$\theta_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i, \quad i = 1, \dots, D \quad (2.1)$$

donde $\boldsymbol{\beta}$ es el vector que contiene los p coeficientes de la regresión y μ_i es el término de error de la regresión asociado al área i , estos efectos aleatorios representan la heterogeneidad de los parámetros θ_i , no explicada por las p variables auxiliares consideradas. En la expresión más simple, se asume que los u_i son independientes e idénticamente distribuidos, con varianza común σ_μ^2 desconocida: $\mu_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\mu^2)$.

Dado que los θ_i son desconocidos y se estiman con un error asociado al diseño muestral e_i (error muestral del área i), denotamos el estimador directo como:

$$\hat{\theta}_i^{Directo} = \theta_i + e_i$$

Se asume que los errores muestrales son independientes entre sí y también son independientes de los efectos aleatorios de las áreas, μ_i , con media cero y varianzas conocidas ψ_i ; así, $e_i \stackrel{\text{iid}}{\sim} (0, \psi_i)$. $\psi_i = \text{var}_\pi \left(\hat{\theta}_i^{Directo} \mid \theta_i \right)$, $i = 1, \dots, D$, se estiman con los microdatos de la encuesta.

Con estas expresiones se obtiene el modelo a nivel de área:

$$\hat{\theta}_i^{Directo} = \mathbf{x}'_i \boldsymbol{\beta} + \mu_i + e_i, \quad i = 1, \dots, D \quad (2.2)$$

Siguiendo la metodología de (Molina, 2019) el mejor predictor lineal insesgado (best linear unbiased predictor, BLUP) bajo el modelo FH es:

$$\tilde{\theta}_i^{FH} = \mathbf{x}'_i \tilde{\boldsymbol{\beta}} + \tilde{\mu}_i \quad (2.3)$$

donde $\tilde{\mu}_i = \gamma_i \left(\hat{\theta}_i^{Directo} - \mathbf{x}'_i \tilde{\boldsymbol{\beta}} \right)$ es el BLUP de μ_i , siendo $\gamma_i = \sigma_\mu^2 / (\sigma_\mu^2 + \psi_i)$ y donde $\tilde{\boldsymbol{\beta}}$ es el estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$ bajo el modelo (2.1), dado por

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^D \gamma_i \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \sum_{i=1}^D \gamma_i \mathbf{x}_i \hat{\theta}_i^{Directo}.$$

Sustituyendo $\tilde{\mu}_i = \gamma_i \left(\hat{\theta}_i^{Directo} - \mathbf{x}'_i \tilde{\boldsymbol{\beta}} \right)$ en el BLUP bajo el modelo FH dado en (2.3), podemos expresar el BLUP como una combinación lineal convexa del estimador directo y del estimador sintético de regresión, es decir,

$$\tilde{\theta}_i^{FH} = \gamma_i \hat{\theta}_i^{Directo} + (1 - \gamma_i) \mathbf{x}'_i \tilde{\boldsymbol{\beta}},$$

$\gamma_i = \sigma_\mu^2 / (\sigma_\mu^2 + \psi_i) \in (0, 1)$, vemos que este peso depende del tamaño muestral del área i a través de la varianza ψ_i del estimador directo y de la bondad de ajuste del modelo

sintético medido por σ_μ^2 ; así que para un área i en la que el estimador directo $\hat{\theta}_i^{Directo}$ sea eficiente por el tamaño muestral en el cual se tenga una varianza muestral ψ_i pequeña comparada con la heterogeneidad no explicada σ_μ^2 , $\gamma_i = \sigma_\mu^2 / (\sigma_\mu^2 + \psi_i)$ es cercano a uno y por lo tanto $\tilde{\theta}_i^{FH}$ le da un mayor peso al estimador directo. En el caso contrario donde las áreas i en las que el estimador directo tiene una varianza muestral ψ_i sea mayor que la heterogeneidad no explicada σ_μ^2 , entonces γ_i se acerca a cero y por lo tanto se le da más peso al estimador sintético de regresión $\mathbf{x}_i' \tilde{\boldsymbol{\beta}}$. Otros autores que han estudiado el modelo de Fay-Herriot son Li & Lahiri (2010), Marcheti et al. (2015), Casas-Cordero et al. (2016) y Avila et al. (2020).

2.2.2. Estimador EBLUP basado en el modelo con errores anidados

Este modelo (BHF) fue propuesto por Battese, Harter y Fuller (1988) en donde se modela la variable de interés en cada individuo j en un área i por medio de un modelo de regresión lineal:

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i + e_{ij}, \quad i = 1, \dots, D \quad (2.4)$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes de las variables auxiliares, común para todas las áreas, u_i es el efecto aleatorio del área i y e_{ij} es el error a nivel de individuo. Al igual que en el modelo (2.1) los efectos aleatorios se consideran independientes de los errores, con $\mu_i \stackrel{\text{iid}}{\sim} (0, \sigma_\mu^2)$ y $e_i \stackrel{\text{ind}}{\sim} (0, \sigma_e^2 k_{ij}^2)$, siendo k_{ij} constantes conocidas que representan la posible heterocedasticidad.

La media poblacional se puede descomponer en la suma de los valores observados en la muestra y los no muestreados en cada dominio:

$$\bar{Y}_i = N_i^{-1} \left(\sum_{j \in s_i} Y_{ij} + \sum_{j \in \bar{s}_i} Y_{ij} \right)$$

De esta misma forma se puede construir el estimador (Molina, 2019):

$$\tilde{Y}_i^{BLUP} = N_i^{-1} \left(\sum_{j \in s_i} Y_{ij} + \sum_{j \in \bar{s}_i} \tilde{Y}_{ij}^{BLUP} \right)$$

Donde $\tilde{Y}_{ij}^{BLUP} = \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}} + \tilde{u}_i$ es la estimación de la variable objetivo Y a nivel de individuo en los datos no muestreados. Se denota $\tilde{\boldsymbol{\beta}}$ al estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$ bajo el modelo (2.4).

Al igual que en el modelo (2.1) \tilde{u}_i es el estimador BLUP de u_i y los valores predichos \tilde{Y}_{ij}^{BLUP} son los BLUPs de las variables Y_{ij} , $j \in \bar{s}_i$ (Molina, 2019):

$$\begin{aligned}\tilde{Y}_{ij}^{BLUP} &= \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}} + \tilde{u}_i, \\ \tilde{u}_i &= \gamma_i \left(\bar{y}_{ia} - \bar{\mathbf{x}}'_{ia}\tilde{\boldsymbol{\beta}} \right), \gamma_i = \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_e^2/a_i)\end{aligned}$$

Donde $\bar{y}_{ia} = a_i^{-1} \sum_{j \in s_i} a_{ij} Y_{ij}$ y $\bar{\mathbf{x}}_{ia} = a_i^{-1} \sum_{j \in s_i} a_{ij} \mathbf{x}_{ij}$ las medias muestrales ponderadas de la variable respuesta y las variables auxiliares, respectivamente, con pesos $a_{ij} = k_{ij}^{-2}$, y $a_i = \sum_{j \in s_i} a_{ij}$.

Para dominios con tamaño de muestra pequeño ($n_i/N_i \approx 0$) el estimador BLUP toma la forma (Molina 2019):

$$\tilde{Y}_i^{BLUP} \approx \gamma_i \left\{ \bar{y}_{ia} + (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ia})' \tilde{\boldsymbol{\beta}} \right\} + (1 - \gamma_i) \bar{\mathbf{x}}'_i \tilde{\boldsymbol{\beta}}$$

Dado que $\gamma_i \in (0, 1)$, el estimador BLUP es un promedio ponderado entre el estimador $\bar{y}_{ia} + (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ia})' \tilde{\boldsymbol{\beta}}$ conocido como estimador de regresión, en el cual se toman los efectos de las áreas u_i como fijos en lugar de aleatorios y el estimador sintético de regresión, $\bar{\mathbf{x}}'_i \tilde{\boldsymbol{\beta}}$.

γ_i depende de los verdaderos valores de las componentes de la varianza del modelo (2.4), $\boldsymbol{\theta} = (\sigma_\mu^2, \sigma_e^2)'$. Sustituyendo $\boldsymbol{\theta}$ por un estimador consistente $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_\mu^2, \hat{\sigma}_e^2)'$ en el estimador BLUP, se obtiene el estimador EBLUP (Molina, 2019):

$$\hat{Y}_i^{EBLUP} = N_i^{-1} \left(\sum_{j \in s_i} Y_{ij} + \sum_{j \in r_i} \hat{Y}_{ij}^{EBLUP} \right)$$

ahora $\hat{\boldsymbol{\beta}}$ es el resultado de sustituir $\boldsymbol{\theta}$ por el estimador $\hat{\boldsymbol{\theta}}$ en $\tilde{\boldsymbol{\beta}}$, los valores predichos ahora son

$$\begin{aligned}\hat{Y}_{ij}^{EBLUP} &= \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \hat{u}_i \\ \hat{u}_i &= \hat{\gamma}_i \left(\bar{y}_{ia} - \bar{\mathbf{x}}'_{ia}\hat{\boldsymbol{\beta}} \right), \hat{\gamma}_i = \hat{\sigma}_\mu^2 / (\hat{\sigma}_\mu^2 + \hat{\sigma}_e^2/a_i)\end{aligned}$$

El estimador BLUP es insesgado bajo el modelo (2.4), el estimador EBLUP sigue siendo insesgado bajo el modelo (2.4), bajo ciertas condiciones sobre el estimador $\hat{\boldsymbol{\theta}}$. Sin embargo, ni el BLUP ni el EBLUP son insesgados bajo el diseño muestral, dado que no se considera el diseño muestral, así que bajo diseños muestrales con probabilidades desiguales, se puede tener un sesgo bajo el diseño no despreciable.

Estimación del Error cuadrático medio para el estimador EBLUP

(González-Manteiga et al., 2008) proponen estimar el error cuadrático medio (ECM) por remuestreo y (Molina, 2019) propone un procedimiento basado en Bootstrap paramétrico el cual consiste en:

1. Con los datos de la muestra ajustar el modelo $Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mu_i + e_{ij}$ para obtener los estimadores de los parámetros del modelo $\hat{\boldsymbol{\beta}}, \hat{\sigma}_\mu^2$ y $\hat{\sigma}_e^2$.

-
2. Simular los efectos aleatorios de las áreas siguiendo $\mu_i^{*(b)} \sim^{iid} N(0, \hat{\sigma}_\mu^2)$, $i = 1, \dots, D$.
 3. Simular, independientemente de los efectos de las áreas $\mu_i^{*(b)}$, los errores Bootstrap para las unidades de la muestra en el área, $e_{ij}^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2)$, $j \in s_i$.
 4. Simular las medias poblacionales de los errores en las áreas, $\bar{E}_i^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2/N_i)$, $i = 1, \dots, D$.
 5. Calcular las verdaderas medias Bootstrap de las áreas,

$$\bar{Y}_i^{(b)} = \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}} + \mu_i^{*(b)} + \bar{E}_i^{*(b)}, \quad i = 1, \dots, D$$

6. Usando la información auxiliar \mathbf{x}_{ij} , $j \in s_i$, estimar las variables respuesta para las unidades de la muestra a partir del modelo

$$Y_{ij}^{*(b)} = \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}} + \mu_i^{*(b)} + e_{ij}^{*(b)}, \quad j \in s_i, \quad i = 1, \dots, D.$$

7. Para la muestra original $s = s_1 \cup \dots \cup s_D$, sea $\mathbf{y}_s^{*(b)} = \left(\left(\mathbf{y}_{1s}^{*(b)} \right)', \dots, \left(\mathbf{y}_{Ds}^{*(b)} \right)' \right)'$ el vector Bootstrap de valores en la muestra.

Ajustar el modelo (2.4) con los datos Bootstrap $\mathbf{y}_s^{*(b)}$ y calcular los estimadores EBLUPs Bootstrap $\hat{Y}_i^{EBLUP*(b)}$, $i = 1, \dots, D$.

8. Repetir los pasos 2-7, para $b = 1, \dots, B$ obteniendo las medias $\bar{Y}_i^{*(b)}$ y los correspondientes EBLUPs $\hat{Y}_i^{EBLUP*(b)}$ para cada muestra Bootstrap b .

La estimación del ECM de los EBLUPs \hat{Y}_i^{EBLUP} , se calcula como:

$$ECM \left(\hat{Y}_i^{EBLUP} \right) = \frac{1}{B} \sum_{b=1}^B \left(\hat{Y}_i^{EBLUP*(b)} - \bar{Y}_i^{*(b)} \right)^2, \\ i = 1, \dots, D.$$

2.2.3. Mejor predictor empírico bajo el modelo con errores anidados (EB)

El mejor predictor empírico bayesiano (EB, empirical Bayes) es una alternativa cuando se desean estimar parámetros no lineales, este fue propuesto por Molina y Rao en el año 2010, aquí se asume que las variables Y_{ij} , siguen el modelo $Y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + \mu_i + e_{ij}$ donde los efectos aleatorios de las áreas μ_i y los errores e_{ij} se distribuyen normal.

Bajo este modelo, el vector de la variable de interés para cada área, $\mathbf{y}_i = (y_{i1}, \dots, y_{iN_i})'$, $i = 1, \dots, D$, son independientes y además $\mathbf{y}_i \sim^{ind} N(\boldsymbol{\mu}_i, \mathbf{V}_i)$, con un vector de medias

$\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$, donde \mathbf{X}_i es la matriz de información auxiliar a nivel de individuo y \mathbf{V}_i es la matriz de covarianzas dada por $\mathbf{V}_i = \sigma_\mu^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}' + \sigma_e^2 \mathbf{A}_i$, donde $\mathbf{A}_i = \text{diag}(k_{ij}^2; j = 1, \dots, N_i)$.

Para estimar un indicador que es función de \mathbf{y}_i ($\delta_i = \delta_i(\mathbf{y}_i)$), se establece un estimador insesgado y además que minimiza el error cuadrático medio:

$$\tilde{\delta}_i^B(\boldsymbol{\alpha}) = E_{\mathbf{y}_{i\bar{s}}} [\delta_i(\mathbf{y}_i) \mid \mathbf{y}_{i\bar{s}}; \boldsymbol{\alpha}] \quad (2.5)$$

$E_{\mathbf{y}_{i\bar{s}}}$ es la esperanza bajo la distribución del vector de valores fuera de la muestra $\mathbf{y}_{i\bar{s}}$ del dominio i dados los valores en la muestra $\mathbf{y}_{i\bar{s}}$. Esta distribución condicionada depende del verdadero valor de los parámetros del modelo $\boldsymbol{\alpha}$. Al remplazar $\boldsymbol{\alpha}$ por un estimador consistente $\hat{\boldsymbol{\alpha}}$ en el valor esperado (2.5), se obtiene el mejor predictor empírico, $\hat{\delta}_i^{EB} = \tilde{\delta}_i^B(\hat{\boldsymbol{\alpha}})$.

Para obtener la distribución de $\mathbf{y}_{i\bar{s}} \mid \mathbf{y}_{i\bar{s}}$, bajo el modelo $Y_{ij} = \mathbf{x}_{ij} \boldsymbol{\beta} + \mu_i + e_{ij}$ se debe descomponer las matrices \mathbf{X}_i y \mathbf{V}_i dividiendo la parte muestral y fuera de la muestra de cada dominio i :

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i\bar{s}} \\ \mathbf{y}_{i\bar{s}} \end{pmatrix}, \quad \mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i\bar{s}} \\ \mathbf{X}_{i\bar{s}} \end{pmatrix}, \quad \mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{i\bar{s}\bar{s}} & \mathbf{V}_{i\bar{s}s} \\ \mathbf{V}_{i\bar{s}s} & \mathbf{V}_{i\bar{s}s} \end{pmatrix}.$$

Dado que \mathbf{y}_i sigue una distribución normal, entonces la distribución condicional también sigue distribución normal,

$$\mathbf{y}_{i\bar{s}} \mid \mathbf{y}_{i\bar{s}} \stackrel{ind}{\sim} N(\boldsymbol{\mu}_{i\bar{s}|s}, \mathbf{V}_{i\bar{s}|s}), \quad i = 1, \dots, D,$$

donde el vector de medias condicionadas y la correspondiente matriz de covarianzas son:

$$\boldsymbol{\mu}_{i\bar{s}|s} = \mathbf{X}_{i\bar{s}} \boldsymbol{\beta} + \gamma_i \left(\bar{y}_{ia} - \bar{\mathbf{x}}_{ia}' \boldsymbol{\beta} \right) \mathbf{1}_{N_i - n_i},$$

$$\mathbf{V}_{i\bar{s}|s} = \sigma_\mu^2 (1 - \gamma_i) \mathbf{1}_{N_i - n_i} \mathbf{1}_{N_i - n_i}' + \sigma_e^2 \text{diag}_{i \in \bar{s}_i}(k_{ij}^2),$$

siendo $\mathbf{1}_k$ un vector de unos de tamaño k . Para un individuo $j \in \bar{s}_i$, se tiene

$$Y_{ij} \mid \mathbf{y}_{i\bar{s}} \sim N(\mu_{ij|s}, \sigma_{ij|s}^2),$$

donde la media y la varianza condicionadas vienen dadas por

$$\mu_{ij|s} = \mathbf{x}_{ij} \boldsymbol{\beta} + \gamma_i \left(\bar{y}_{ia} - \bar{\mathbf{x}}_{ia}' \boldsymbol{\beta} \right),$$

$$\sigma_{ij|s}^2 = \sigma_\mu^2 (1 - \gamma_i) + \sigma_e^2 k_{ij}^2.$$

Estimación del Error cuadrático medio para el mejor predictor empírico bayesiano

Para la estimación del ECM del estimador EB seguiremos la metodología planteada por (Molina y Rao, 2010):

-
1. Con los datos de la muestra ajustar el modelo $Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mu_i + e_{ij}$ para obtener los estimadores de los parámetros del modelo $\widehat{\boldsymbol{\beta}}, \hat{\sigma}_\mu^2$ y $\hat{\sigma}_e^2$.
 2. Simular los efectos aleatorios de las áreas siguiendo $\mu_i^{*(b)} \sim^{iid} N(0, \hat{\sigma}_\mu^2), i = 1, \dots, D$.
 3. Simular, independientemente de los efectos de las áreas $\mu_i^{*(b)}$ y los errores Bootstrap para las unidades de la muestra en el área, $e_{ij}^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2), i \in s_i$.
 4. Generar una población (o censo) Bootstrap de valores de la variable respuesta a través del modelo, $Y_{ij}^{*(b)} = \mathbf{x}_{ij}\widehat{\boldsymbol{\beta}} + \mu_i^{*(b)} + e_{ij}^{*(b)}, j = 1, \dots, N_i, i = 1, \dots, D$
 5. Definimos el vector censal de variables respuesta del área i , dado por $\mathbf{y}_i^{*(b)} = (Y_{i1}^{*(b)}, \dots, Y_{iN_i}^{*(b)})'$. Calcular los indicadores de interés a partir del censo Bootstrap $\delta_i^{*(b)} = \delta_i(\mathbf{y}_i^{*(b)}), i = 1, \dots, D$.
 6. Para la muestra original $s = s_1 \cup \dots \cup s_D$, sea $\mathbf{y}_s^{*(b)} = \left((\mathbf{y}_{1s}^{*(b)})', \dots, (\mathbf{y}_{Ds}^{*(b)})' \right)'$ el vector que contiene las observaciones Bootstrap cuyos índices están en la muestra, es decir, que contiene a las variables $Y_{ij}^{*(b)}, i \in s_i, i = 1, \dots, D$. Ajustar de nuevo el modelo $Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mu_i + e_{ij}$ a los datos Bootstrap $\mathbf{y}_s^{*(b)}$ y obtener los predictores EB Bootstrap de los indicadores de interés, $\hat{\delta}_i^{EB*(b)}, i = 1, \dots, D$.
 7. Repetir los pasos 2-6 para $b = 1, \dots, B$, y obtener los verdaderos valores, $\delta_i^{*(b)}$, y los correspondientes predictores EB, $\hat{\delta}_i^{EB*(b)}$, para cada área $i = 1, \dots, D$, y para cada réplica Bootstrap, $b = 1, \dots, B$.
 8. La estimación del ECM de los predictores EB, $\hat{\delta}_i^{EB}$, se calcula como:

$$\text{ECM}(\hat{\delta}_i^{EB}) = B^{-1} \sum_{b=1}^B \left(\hat{\delta}_i^{EB*(b)} - \delta_i^{*(b)} \right)^2, \quad i = 1, \dots, D$$

2.2.4. Estimación basada en modelos lineales generalizados mixtos

Los modelos expuestos anteriormente son para variables dependientes continuas, el objetivo de este trabajo es estimar una proporción, en la cual la variable de interés Y es cualitativa, en este escenario es habitual considerar modelos lineales generalizados mixtos (GLMM).

Sea $Y_{ij} \in \{0; 1\}$ la variable binaria que mide la característica de interés, el modelo de estimación en áreas pequeñas tradicional es el GLMM con efectos aleatorios de las áreas:

$$Y_{ij} \mid \mu_i \sim \text{Bern}(p_{ij}), g(p_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta} + \mu_i, \mu_i \sim^{iid} N(0, \sigma_\mu^2), j = 1, \dots, N_i, i = 1, \dots, D$$

donde μ_i es el efecto aleatorio del área i , β es el vector de coeficientes de la regresión y $g : (0, 1) \rightarrow R$ es la función enlace, consideraremos la función logística dada por $g(p) = \log(p/(1-p))$.

Siguiendo la idea de la sección 2.2.3, el mejor predictor bajo el modelo para la proporción \bar{Y}_i , es

$$\tilde{Y}_i^{EB}(\alpha) = E(Y_i | \mathbf{y}_{is}; \alpha) = \frac{1}{N_i} \left\{ \sum_{j \in s_i} Y_{ij} + \sum_{j \in \bar{s}_i} E(Y_{ij} | \mathbf{y}_{is}; \alpha) \right\}$$

La distribución de $Y_{ij} | \mathbf{y}_{is}$ depende del vector $\alpha = (\beta', \sigma_\mu^2)'$ de parámetros del modelo. En la práctica, obtenemos el predictor EB reemplazando α por un estimador consistente $\hat{\alpha}$, el cual se obtiene ajustando el modelo GLMM, en el mejor predictor, $\hat{Y}_i^{EB} = \tilde{Y}_i^B(\hat{\alpha})$.

Con el modelo ajustado el siguiente paso es calcular las esperanzas $E(Y_{ij} | \mathbf{y}_{is}; \hat{\alpha})$ para calcular el predictor EB, utilizando el Teorema de Bayes y dado μ_i y la independencia de las variables $\{Y_{ij}; j = 1, \dots, N_i\}$ la esperanza se puede expresar de la forma

$$E(Y_{ij} | \mathbf{y}_{is}; \hat{\alpha}) = \frac{E\{h(\mathbf{x}_{ij}\beta + \mu_i) f(\mathbf{y}_{is} | \mu_i); \hat{\alpha}\}}{E\{f(\mathbf{y}_{is} | \mu_i); \hat{\alpha}\}}, \quad j \in \bar{s}_d, \quad (2.6)$$

donde $h = g^{-1}$ y

$$\begin{aligned} f(\mathbf{y}_{is} | v_i) &= \prod_{j \in s_i} p_{ij}^{Y_{ij}} (1 - p_{ij})^{(1-Y_{ij})} \\ &= \prod_{j \in s_i} h(\mathbf{x}_{ij}\beta + \mu_i)^{Y_{ij}} \{1 - h(\mathbf{x}_{ij}\beta + \mu_i)\}^{(1-Y_{ij})}. \end{aligned}$$

La función inversa de g es $h(\mathbf{x}_{ij}\beta + \mu_i) = \exp(\mathbf{x}_{ij}\beta + \mu_i) / \{1 + \exp(\mathbf{x}_{ij}\beta + \mu_i)\}$, utilizando estos resultados, se puede aproximar las dos esperanzas de la expresión (2.6) mediante simulación Monte Carlo, generando $\mu_i^{(r)} \sim N(0, \hat{\sigma}_v^2)$, $r = 1, \dots, R$, con R el número de iteraciones y después calculando

$$E(Y_{ij} | \mathbf{y}_{is}; \hat{\beta}) \approx \frac{R^{-1} \sum_{r=1}^R h(\mathbf{x}_{ij}\hat{\beta} + \mu_i^{(r)}) \hat{f}(\mathbf{y}_{is} | \mu_i^{(r)})}{R^{-1} \sum_{r=1}^R \hat{f}(\mathbf{y}_{is} | \mu_i^{(r)})}, \quad j \in \bar{s}_i,$$

Posteriormente se hace uso del mejor predictor empírico bajo el modelo con errores anidados (EB) para obtener la estimación de la proporción a nivel de área.

2.3. Exploración de métodos basados en aprendizaje automático en SAE

Con los modelos tradicionales de áreas pequeñas, bajo inferencia basada en el modelo, se deben cumplir algunos supuestos como la relación lineal de las variables auxiliares con la variable de interés o que los errores asociados al modelo deben seguir una determinada distribución de probabilidad. Adicionalmente, surgen algunas problemáticas con estos modelos como la multidimensionalidad y la sensibilidad ante datos atípicos. Diferentes autores han trabajado en la exploración de técnicas semi-paramétricas y basadas en aprendizaje automático. Algunos de estos métodos son flexibles frente al supuesto de distribución preservando la linealidad, mientras que otros preservan el supuesto distribucional, pero se consideran especificaciones del modelo más generales que el modelo de regresión lineal.

Dentro de la revisión bibliográfica del uso de diferentes metodologías haciendo uso de modelos no paramétricos y modelos de aprendizaje automático en donde se tenga en cuenta la estructura jerárquica de los datos se resalta el trabajo de (Sela y Simonoff, 2012) que proponen un modelo mixto semiparamétrico que consiste en una parte de efectos aleatorios y un modelo de árbol no paramétrico de efectos fijos en el contexto de métodos basados en árboles. (Hajjem et al., 2011) proponen un enfoque similar bajo árboles de regresión de efectos mixtos (MERT).

La idea general de aplicar métodos basados en árboles en áreas pequeñas no es completamente reciente (Anderson, 2014; Bilton, 2017; De Moliner y Goga, 2018; Capitaine et al., 2021). Recientemente, (Dagdoug, 2021) analizó las propiedades teóricas de los bosques aleatorios en el contexto de datos de encuestas complejos para la estimación asistida por modelos. Como el desempeño superior de los bosques aleatorios sobre los árboles de regresión también se evidencia en los datos dependientes, (Hajjem, 2014) reemplaza la parte de efectos fijos en MERT con un bosque aleatorio o random forest, lo que lleva al modelo random forest de efectos mixtos (MERF). Recientemente Krennmair, P., Würz, N., & Schmid, T. proponen un modelo mixto a nivel de unidad semiparamétrico que combina la flexibilidad de los modelos basados en árboles con las ventajas estructurales de los modelos mixtos lineales para las medias a nivel de dominio.

2.3.1. Modelo MERF para la estimación de medias en SAE

(Krennmair, Würz, y Schmid, 2022) presentan uno de los primeros trabajos en áreas pequeñas en donde se aplican técnicas de aprendizaje automático, en este se hace uso del modelo MERF (Mixed Effects Random Forest) expuesto por (Hajjem et al., 2014) para estimar medias en áreas pequeñas. La metodología propuesta es:

Dado el vector de variables auxiliares \mathbf{x}_{ij} y la variable objetivo y_{ij} (continua) se establece la relación entre estos valores por medio de un modelo general de regresión de efectos mixtos:

$$y_{ij} = f(\mathbf{x}_{ij}) + \mu_i + e_{ij} \quad (2.7)$$

donde $\mu_i \sim N(0, \sigma_\mu^2)$ corresponde al efecto aleatorio asociado al i -ésimo dominio y $e_{ij} \sim N(0, \sigma_\epsilon^2)$ corresponde al error aleatorio para cada individuo j en el dominio i .

La función $f(\mathbf{X}_{ij})$ modela la media condicional de la variable objetivo y_{ij} dado el vector de variables auxiliares \mathbf{X}_{ij} . Los interceptos aleatorios a nivel de área, denotados por μ_i , representan la estructura de dependencia jerárquica de las observaciones de Y dentro de cada área i . Finalmente, se asume la presencia de errores a nivel unitario, e_{ij} , y se considera que los efectos aleatorios μ_i son independientes entre sí.

Si se define:

$$f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta} \quad (2.8)$$

donde $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ son los coeficientes de regresión, se obtiene un modelo de regresión lineal, la estimación y_{ji} queda dada por un modelo lineal mixto el cual es ampliamente utilizado en los modelos de estimación de áreas pequeñas a nivel de unidad:

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mu_i + e_{ij} \quad (2.9)$$

Assumiendo normalidad para el error de nivel unitario e_{ij} y los efectos aleatorios de dominio μ_i , se obtiene que la distribución condicional de los datos fuera de la muestra dados los datos de la muestra también siguen una distribución normal; por lo que los valores sintéticos (valores predichos) de la variable objetivo para toda la población del área i de tamaño N_i se generan a partir del siguiente modelo (Krennmair et al., 2022):

$$y_{ii}^* = \mathbf{x}_{ij}\boldsymbol{\beta} + \tilde{u}_i + \mu_i^* + \epsilon_{ii}^*, \quad \mu_i^* \sim N\{0, \sigma_\mu^2(1 - \gamma_i)\}$$

$$\epsilon_{ii}^* \sim N(0, \sigma_\epsilon^2), \quad \gamma_i = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\epsilon^2/n_i},$$

donde $\tilde{u}_i = E(\mu_i | y_s)$ es la media condicional de μ_i dados los valores de la variable objetivo en la muestra; $\mathbf{x}_{ij}\boldsymbol{\beta} + \tilde{u}_i$ es la media condicional de y_{ij} en la población dados los valores de la muestra, mientras que $\mu_i^* + \epsilon_{ij}^*$ es la media condicional de los valores predichos fuera de la muestra dados los datos muestrales.

Si se asume que el modelo $f()$ en la expresión (2.7) es un Random forest, se obtiene una metodología semi-paramétrica para la estimación de la media, en donde se combinan las ventajas predictivas de los Random Forest eliminando los supuestos de los modelos lineales con la capacidad de modelar estructuras jerárquicas de datos de encuestas utilizando efectos aleatorios.

Este método obtiene estimaciones óptimas de los componentes del modelo $\hat{\mu}_i$, $\hat{\sigma}_\mu^2$ y $\hat{\sigma}_e^2$, la propuesta que se presenta por medio del algoritmo MERF se ajusta a los parámetros óptimos para el Modelo (2.7) al estimar iterativamente:

- a. La función del random forest, asumiendo que el término de efectos aleatorios es correcto.
- b. Los efectos aleatorios parten de la suposición que las predicciones en de los datos de validación (predicciones *OOB* Out-of-Bag) del modelo son correctas.

Las predicciones *OOB* se basan en las observaciones no utilizadas en la construcción del subárbol de cada bosque, es decir corresponden a los datos de validación. La estimación de los componentes de la varianza $\hat{\mu}_i$, $\hat{\sigma}_\mu^2$ se obtienen implícitamente de los estimadores maximoverosimil dados los datos muestrales con los cuales se entrena el modelo. El estimador resultante para las predicciones basadas en el modelo MERF se presenta a continuación (Krennmair et al., 2022):

$$\hat{\mu}_{ij}^{MERF} = \hat{f}(\mathbf{x}_{ij}) + \hat{\mu}_i = \hat{f}(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + \hat{\sigma}_e^2/n_i} * \frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{OOB}(\mathbf{x}_{ij})) \quad (2.10)$$

donde $\hat{f}^{OOB}(\mathbf{x}_{ij})$ es la estimación de la variable objetivo en los datos de validación. Con $\hat{\mu}_{ij}^{MERF}$ se puede predecir las medias condicionales de una variable dependiente continua. Basados en esta ecuación, se propone el estimador de la media a nivel de dominio para cada área i , el cual viene dado por (Krennmair et al., 2022):

$$\hat{\mu}_i^{MERF} = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{\mu}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + \hat{\sigma}_e^2/n_i} * \frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{OOB}(\mathbf{x}_{ij})) \quad (2.11)$$

$$\bar{\hat{f}}_i(\mathbf{x}_{ij}) = \frac{1}{N_i} \sum_{j \in U_i} \hat{f}(\mathbf{x}_{ij})$$

Para áreas no muestreadas, el estimador propuesto para la media a nivel de área se reduce a la parte fija del modelo: $\hat{\mu}_i^{MERF} = \hat{f}_k(\mathbf{x}_{ij})$

Contar con información auxiliar a nivel de individuo no siempre es posible. A diferencia del modelo BHF, los totales o medias a nivel de área no pueden utilizarse directamente, ya que al modelar la variable dependiente con un MERF, se tiene que $f(\bar{\mathbf{x}}_i) \neq \bar{f}_i(\mathbf{x}_{ij})$. Por esta razón, los autores proponen calibrar las estimaciones basadas en modelos MERF mediante el cálculo de ponderaciones que dependen exclusivamente de covariables agregadas a nivel censal. De este modo, disponiendo de microdatos a nivel de encuesta y medias a nivel de área, se plantea el siguiente estimador para la media del área i (Krennmair et al., 2022):

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i} \hat{w}_{ij} \left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i \right]$$

Siguiendo la técnica descrita por (Owen, 1990) y (Qin y Lawless, 1994) los pesos w_{ij} se estiman bajo las siguientes 3 condiciones:

- $\sum_{j=1}^{n_i} w_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) = 0$, donde $\bar{\mathbf{x}}_i$ es la media poblacional del área i
- $w_{ij} \geq 0$
- $\sum_{j=1}^{n_i} w_{ij} = 1$

w_{ij} se estima por medio de la siguiente formula:

$$\hat{w}_{ij} = \frac{1}{n_i} \frac{1}{1 + \hat{\lambda}_i^\top (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)}$$

donde $\hat{\lambda}_i$ es la solución por multiplicadores de Lagrange de

$$\sum_{j=1}^{n_i} \frac{\mathbf{x}_{ij} - \bar{\mathbf{x}}_i}{1 + \hat{\lambda}_i^\top (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)} = 0$$

(Krennmair y Schmid, 2022) proponen una metodología de Bootstrap paramétrico para la estimación el error cuadrático medio de los estimadores $\hat{\mu}_i^{\text{MERF}}$ y $\hat{\mu}_i^{\text{MERFagg}}$, en la cual se deben seguir los siguientes pasos:

1. Con los datos de la encuesta ajustar el modelo MERF, obteniendo $\hat{f}, \hat{\sigma}_e, \hat{\sigma}_\mu$, en el caso de tener covariables a nivel agregado calcular \hat{w}_{ij}
2. Para cada valor de la encuesta calcular los residuos $\hat{r}_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij})$ y la media a nivel de área de estos $\bar{r}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{r}_{ij}$ para $i = 1, \dots, D$
3. Para replicar la estructura jerárquica se utilizan los residuos marginales, los cuales se centran obteniendo $\hat{r}_{ij}^c = \hat{r}_{ij} - \bar{r}_i$
4. para $b = 1, \dots, B$:

- a) Aplicar un muestreo aleatorio simple con remplazo para cada área i a partir de la distribución empírica de los residuos y se obtiene:

$$r_{ij}^{*(b)} = \text{srswr}(r_{ij}^c, n_i), \bar{e}_i^{*(b)} = \text{srswr}\left(r_{ij}^c \frac{\hat{\sigma}_{bc,\epsilon}}{\sqrt{N_i - n_i}}, 1\right), \text{ y } \mu_i^{*(b)} = \text{srswr}(\bar{r}^c, 1).$$

- b) Calcular los valores (pseudo)-verdaderos para la población basándose en los efectos fijos a partir de las estimaciones de $\hat{\mu}_i^{\text{MERF}}$ o $\hat{\mu}_i^{\text{MERFagg}}$

$$\bar{y}_i^{(b)} = \sum_{j=1}^{n_i} \hat{w}_{ij} \hat{f}(\mathbf{x}_{ij}) + \mu_i^{*(b)} + \bar{E}_i^{(b)}, \quad \text{donde} \quad \bar{E}_i^{(b)} = \frac{n_i}{N_i} \bar{r}_{ij}^{*(b)} + \frac{N_i - n_i}{N_i} \bar{e}_i^{*(b)}$$

- c) Con las covariables de la muestra \mathbf{x}_{ij} generar la muestra Bootstrap de la variable respuesta:

$$y_{ij}^{(b)} = \hat{f}^{\text{OOB}}(\mathbf{x}_{ij}) + \mu_i^{*(b)} + r_{ij}^{*(b)}$$

Se usan las predicciones OOB de \hat{f} para imitar las variaciones de las covariables x_{ij} a través de las predicciones de las observaciones no utilizadas dentro de cada árbol en el proceso de ajuste que varían a lo largo de las replicaciones en cada muestra Bootstrap.

- d) Estimar $\hat{\mu}_i^{\text{MERF}}$ o $\hat{\mu}_i^{\text{MERFagg}}$ con los valores de $y_{ij}^{(b)}$. Es importante resaltar que las ponderaciones \hat{w}_{ij} permanecen constantes en las B muestras, pues las covariables \mathbf{x}_{ij} de la encuesta original y las covariables a nivel de población \bar{x}_i no cambian.

5. Finalmente, calcular el ECM estimado para cada media de los dominios $i = 1, \dots, D$

$$ECM(\hat{\mu}_i^{\text{MERF}}) = \frac{1}{B} \sum_{b=1}^B \left[\left(\hat{\mu}_i^{\text{MERF}(b)} - \bar{y}_i^{(b)} \right)^2 \right]$$

$$ECM(\hat{\mu}_i^{\text{MERFagg}}) = \frac{1}{B} \sum_{b=1}^B \left[\left(\hat{\mu}_i^{\text{MERFagg}(b)} - \bar{y}_i^{(b)} \right)^2 \right]$$

Dado que estos métodos se basan en el modelo se debe verificar su correcta especificación y cumplimiento de los supuestos en los datos disponibles de la encuesta; sin embargo, no es posible verificar el modelo en los dominios no muestreados, por lo que no se puede asegurar que las estimaciones resultantes en las áreas no muestreadas realmente satisfacen los supuestos del modelo.

2.4. Árbol de decisión con efectos mixtos para la estimación de proporciones

El modelo MERF descrito en la sección anterior, fue propuesto únicamente para la estimación de medias, la gran mayoría de metodologías en SAE se basan únicamente en este indicador. Cuando la variable objetivo es dicotómica y se desea estimar una proporción a nivel de área el procedimiento a seguir es por medio de un modelo lineal generalizado mixto expuesto en la sección (2.2.4), sin embargo siguiendo la idea del uso del modelo MERF para la estimación de medias se puede sustituir la función $f()$ que corresponde al modelo GLMM, por un modelo de aprendizaje automático que tenga en cuenta los efectos aleatorios de las áreas y de esta forma obtener una alternativa semi-paramétrica para la estimación de una proporción en áreas pequeñas. Los métodos tradicionales de (Fay y Herriot, 1979) y de (Battese et al., 1988) no son eficientes cuando se usan para estimar proporciones puesto que no se puede asegurar que las predicciones estén en el intervalo $[0,1]$. Algunos autores que han estudiado el problema de estimar una proporción usando métodos de estimación de áreas pequeñas usando modelos lineales generalizados son (Chandra et al., 2018); (Molina y Strzalkowska-Kominiak, 2020) y (Parker et al., 2023) bajo modelos de unidad. (Salvati et al., 2012) propone un estimador para la proporción con modelos de área.

En el ámbito del aprendizaje automático, se distinguen dos enfoques principales: el aprendizaje supervisado y el no supervisado. En este trabajo nos enfocamos en el primero, en el cual los datos disponibles están etiquetados y el objetivo es encontrar una función que relacione las variables de entrada con sus respectivas etiquetas. En este contexto, la información auxiliar X representa las variables de entrada del modelo, mientras que la variable respuesta binaria Y corresponde a la etiqueta o *output*. El algoritmo se entrena utilizando una muestra de la base de datos original, con el propósito de aprender las relaciones entre las variables X para asignar correctamente la etiqueta a cada observación. Posteriormente, se evalúa el desempeño del modelo comparando las etiquetas asignadas con los valores reales de la variable respuesta.

El árbol de decisión es una metodología de aprendizaje supervisado cuyo objetivo es segmentar la muestra en distintos grupos en función de las variables explicativas X . En el caso de problemas de clasificación, la predicción de la variable respuesta Y se realiza asignando la categoría modal de cada grupo. Este método, desarrollado en la década de 1970, es uno de los más simples y fáciles de interpretar, especialmente en tareas de clasificación binaria. Los algoritmos de aprendizaje automático basados en árboles de decisión dividen recursivamente la muestra de entrenamiento en subgrupos homogéneos mediante reglas basadas en las covariables. Estas divisiones generan estructuras jerárquicas denominadas nodos. El proceso comienza en un nodo raíz que contiene toda la muestra de entrenamiento y, a través de reglas de partición, se generan nodos intermedios hasta alcanzar los nodos

terminales, donde finalmente se asigna la predicción correspondiente a cada observación.

(Dangeti, 2017) presenta las principales diferencias entre los modelos logísticos y los modelos basados en árboles para la clasificación de individuos:

Modelo Logístico	Modelo basado en árboles
Se expresa por medio de una ecuación que modela la variable dependiente a partir de las independientes.	Se establecen reglas en diferentes niveles (nodos) para establecer la clasificación.
Es un modelo paramétrico que se define como una combinación lineal entre las variables independientes multiplicadas por los parámetros.	Es un modelo no paramétrico en el que no existen parámetros. Implícitamente se realiza un análisis de importancia y selección de principales variables.
Se realiza una suposición distribucional sobre la variable respuesta	No se hacen suposiciones sobre la distribución de la variable respuesta
La forma del modelo está predeterminada (función logística)	La forma del modelo no está predeterminada. El modelo se ajusta para proveer la mejor clasificación según los datos.
Proporciona buenos resultados cuando las variables independientes son continuas y se cumple la linealidad.	Proporciona mejores resultados cuando la mayoría de las variables son de naturaleza categórica.
Es difícil encontrar interacciones complejas entre variables (relaciones no lineales) y además los valores atípicos y faltantes afectan el rendimiento del modelo.	No es necesario asumir linealidad, estos modelos captan interacciones complejas entre las covariables. Pueden adaptarse ante distribuciones altamente sesgadas, valores atípicos y valores faltantes.

Figura 2.2: Comparación de técnicas
Fuente: Dangeti (2017).

En el caso de los árboles de decisión basados en modelos lineales generalizados mixtos (GLMM), el objetivo es proporcionar una alternativa en la que, al igual que en el modelo MERF, se mantengan los supuestos distribucionales, pero con una mayor flexibilidad en términos de la linealidad de la relación y una mayor robustez frente a datos atípicos en la muestra. A partir de los datos de entrenamiento, se estiman los parámetros del modelo, los cuales se evalúan sobre la muestra y se actualizan de manera recursiva hasta cumplir un criterio de detención, como un tamaño mínimo de muestra o la inestabilidad de los parámetros debido a la presencia de valores atípicos.

Existen diversos algoritmos para la construcción de árboles basados en modelos GLMM, los cuales se abordan en detalle en la Sección 3.1.

Entrenar un modelo con múltiples muestras en lugar de un único conjunto de datos puede aumentar su robustez y reducir la varianza de las estimaciones. Sin embargo, en la práctica, generalmente se dispone de un único conjunto de datos de entrenamiento. Para abordar esta limitación, surgen los algoritmos de ensamblaje, los cuales construyen múltiples modelos simples, como los árboles de decisión, entrenándolos con distintas particiones o subconjuntos del conjunto de datos original. Este enfoque permite que la estimación final sea más estable y precisa. Los métodos de ensamblaje son ampliamente utilizados, ya que han demostrado ser efectivos en la reducción de la varianza y en la mejora del rendimiento predictivo del modelo.

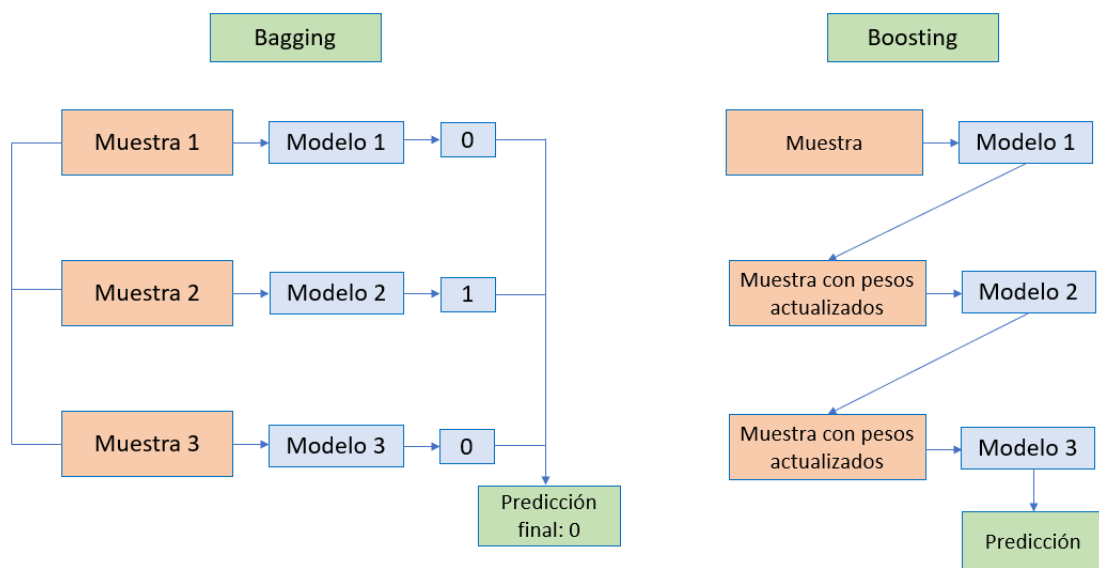


Figura 2.3: Bagging y Boosting
Fuente: Elaboración propia.

En la figura 2.3 se ilustran los métodos de ensamblaje que se usarán en este trabajo, en primer lugar el algoritmo bagging consiste en entrenar modelos independientes usando submuestras del conjunto de entrenamiento. Dado que el objetivo es estimar proporciones, el modelo a ejecutado, corresponde a un árbol de clasificación; cada modelo se utiliza para hacer la estimación de la variable Y y la predicción final está dada por la moda de la predicción de todos los árboles entrenados. En el algoritmo boosting no se corren modelos

independientes sino que se usa una metodología secuencial, en la cual se ajusta un primer modelo con la base de entrenamiento y se realiza la estimación de la variable Y , se crea una ponderación de la base de entrenamiento en la cual a los datos de etiquetado correcto por el modelo se les asigna un peso mayor que a los datos en los cuales el modelo se equivocó, luego se ajusta un nuevo modelo usando los pesos calculados. De esta forma, el modelo aprenderá más de los datos en los cuales acertó y mejorará las estimaciones.

3. Metodología propuesta

Como se expuso en la sección 2.3.1 el modelo Random Forest de efectos mixtos provee ventajas frente a los modelos lineales, no solo en términos de supuestos distribucionales, también en cuanto a la robustez a los datos atípicos, la no linealidad de las relaciones y la selección de variables que se realiza con el algoritmo. La inclusión de este modelo se ha explorado para la estimación de la media de una variable continua Y , sin embargo, la revisión de la literatura realizada muestra que aún no se ha explorado cuando el parámetro a estimar es una proporción, es decir cuando el modelo a nivel de individuo no es una regresión sino un modelo de clasificación. Así como en el caso de la regresión el modelo MERF contribuye a la reducción de las medidas de precisión de las estimaciones, en un modelo de clasificación se espera que los algoritmos de ensamblaje, como bagging y boosting, disminuyan la varianza de las estimaciones de la proporción a nivel de área. Además, estos métodos proporcionarían ventajas inherentes a los enfoques no paramétricos en comparación con el modelo lineal logístico.

Para la aplicación de los modelos presentados en la Sección 2.4 en la estimación de áreas pequeñas, es fundamental considerar la estructura jerárquica de los datos, derivada de la partición en dominios. La correlación entre las observaciones dentro de cada dominio se modela mediante efectos aleatorios, mientras que las relaciones entre covariables se representan a través de efectos fijos, resultando en un modelo aditivo que combina ambos términos.

La metodología de este estudio se centra en la comparación entre el modelo lineal logístico de efectos mixtos (Generalized Linear Mixed Model, GLMM) y modelos de clasificación binaria de efectos mixtos, implementados mediante los algoritmos bagging (Mixed-Effects Bagging Algorithm, MEBA) y boosting (Mixed-Effects Boosting Algorithm, MEBO). Para ello, se entrenan los modelos utilizando los microdatos de la encuesta y, posteriormente, se estima la proporción a nivel de área mediante el estimador Plug-In utilizando los microdatos de la población. Además, para evaluar la precisión de las estimaciones, se emplea el método de remuestreo Bootstrap no paramétrico con el objetivo de

calcular el error cuadrático medio (ECM) de cada estimador. Finalmente, se comparan las metodologías en función de las estimaciones obtenidas, el ECM, el sesgo y las medidas de ajuste correspondientes a cada modelo.

3.1. Estimador Plug-in

(Molina y Strzalkowska-Kominiak, 2020) exponen que el mejor predictor empírico bajo el modelo con errores anidados (EB) tiene error cuadrático medio mínimo y es insesgado bajo el modelo, sin embargo, ajustar el modelo GLMM y calcular la aproximación Monte Carlo de $\hat{\theta}_i^{EB}$, requiere un tiempo computacional considerable y teniendo en cuenta que se debe realizar en cada muestra Bootstrap, realizar este procedimiento para poblaciones grandes puede ser inviable. Las autoras proponen una alternativa, la cual consiste en que a partir de los estimadores de β y μ_i se calcula un estimador plug-in simplemente prediciendo los valores no muestreados a través del modelo ajustado:

$$\hat{\theta}_i^{PI} = \frac{1}{N_i} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} \hat{p}_{ij} \right),$$

donde $\hat{p}_{ij} = h(\mathbf{x}_{ij}\hat{\beta} + \hat{u}_i)$ es el valor predicho de la observación fuera de la muestra. Para $\alpha = (\beta, \sigma_\mu^2)'$ conocido, el estimador Plug-In, $\hat{\theta}_i^{PI}$, no puede presentar un error cuadrático medio (ECM) inferior al del mejor predictor empírico, $\hat{\theta}_i^{EB}$. Sin embargo, su principal ventaja radica en su simplicidad computacional. Cabe destacar que ambos estimadores coinciden cuando la función de enlace $g(\cdot)$ es lineal.

Usando la función logística $g(p) = \log(p/(1-p))$, la cual se aplica en este trabajo, se evidencia que es aproximadamente lineal para $p \in (0.2, 0.8)$ como se muestra en la figura 3.1, así que el estimador plug-in basado en el modelo GLMM debe ser muy similar al estimador EB, $\hat{\theta}_i^{EB}$, en términos de ECM, para valores $p \in (0.2, 0.8)$, esto también hace que ambos estimadores, sean cercanos al EBLUP, $\hat{\theta}_i^{EBLUP}$, basado en el modelo con errores anidados, por lo que para estimar proporciones de características ni muy poco ni muy frecuente, también tiene sentido utilizar el estimador EBLUP.

Molina y Strzalkowska-Kominiak mediante ejercicios de simulación concluyen que el estimador plug-in basado en un modelo GLMM a nivel de unidad con un enlace logístico funciona de manera muy similar a los estimadores EB bajo el mismo modelo, en términos de las estimaciones obtenidas y el error cuadrático medio. Por lo que dada su simplicidad en la implementación computacional es una excelente alternativa cuando se tienen poblaciones grandes.

El estimador plug-in combina la información proveniente de la muestra $\sum_{j \in s_i} y_{ij}$ con

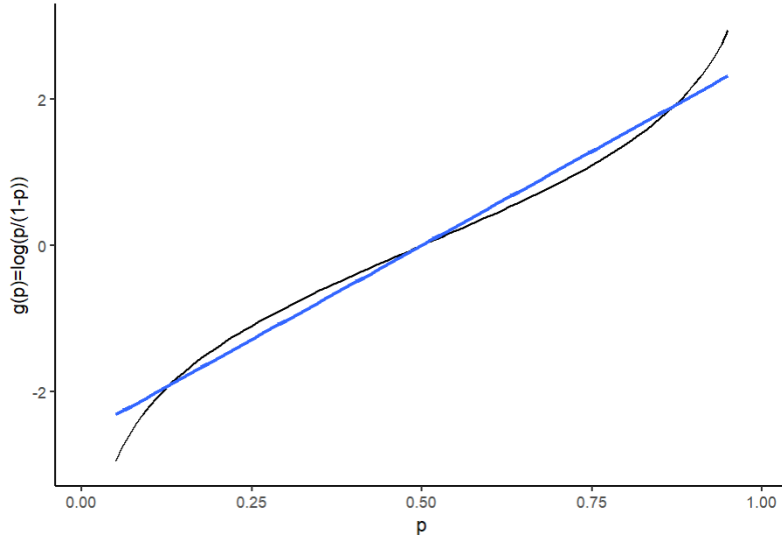


Figura 3.1: Función de enlace logístico
Fuente: Elaboración propia.

la estimación de la variable de interés en las unidades no muestreadas, obtenida a partir de la información auxiliar y el modelo ajustado $\sum_{j \in \bar{s}_i} \hat{p}_{ij}$. En este trabajo, además del estimador propuesto por (Molina y Strzalkowska-Kominiak, 2020), se plantean dos alternativas adicionales de modelamiento para la estimación de las probabilidades \hat{p}_{ij} , con el objetivo de evaluar su desempeño y comparar su eficacia en la estimación de áreas pequeñas:

- i. Metodología propuesta en (Molina y Strzalkowska-Kominiak, 2020) por medio de un modelo lineal generalizado mixto (GLMM), expuesto en la sección 2.2.4:

$$Y_{ij} \mid \mu_i \sim \text{Bern}(p_{ij})$$

$$g(p_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta} + \mu_i, \mu_i \stackrel{iid}{\sim} N(0, \sigma_\mu^2), j = 1, \dots, N_i, i = 1, \dots, D$$

donde μ_i es el efecto aleatorio del área i , $\boldsymbol{\beta}$ es el vector de coeficientes de la regresión y $g(p) = \log(p/(1-p))$.

Para la implementación computacional de este modelo se usa la librería *glmmTMB* (Mollie et al., 2017) del software *R*, donde se estiman los parámetros $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \sigma_\mu^2)'$ por log-verosimilitud, cuya función está dada por:

$$\begin{aligned}
l &= l(\boldsymbol{\beta}, \sigma_\mu^2) \\
&= -\frac{D}{2} \ln(2\pi\sigma_\mu^2) + \beta \sum_{i=1}^D \sum_{j=1}^{n_i} y_{ij} X_{ij} + \sum_{i=1}^D \ln \left[\int e^{\mu_i \sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \ln(1+e^{X_{ij}\beta+u_i}) - \frac{\mu_i^2}{2\sigma_\mu^2}} d\mu_i \right]
\end{aligned}$$

Para obtener la estimación de $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_\mu^2)'$ se usa aproximación de Laplace para integrar sobre los efectos aleatorios.

- ii. **Mixed Effects Tree-Bagging (MEBA):** se hace uso de la librería *glmertree* (Fokkema et al., 2018) del software *R*, el cual es un algoritmo basado en árboles generalizados mixtos, haciendo un particionamiento recursivo para la obtención de las estimaciones, los efectos aleatorios se asumen gaussianos y se estiman por medio de un GLMM. Con esta función se obtiene un árbol de decisión de efectos mixtos, sobre el cual se aplica el método de ensamblaje bagging para obtener resultados más robustos y disminuir la estimación de la varianza.

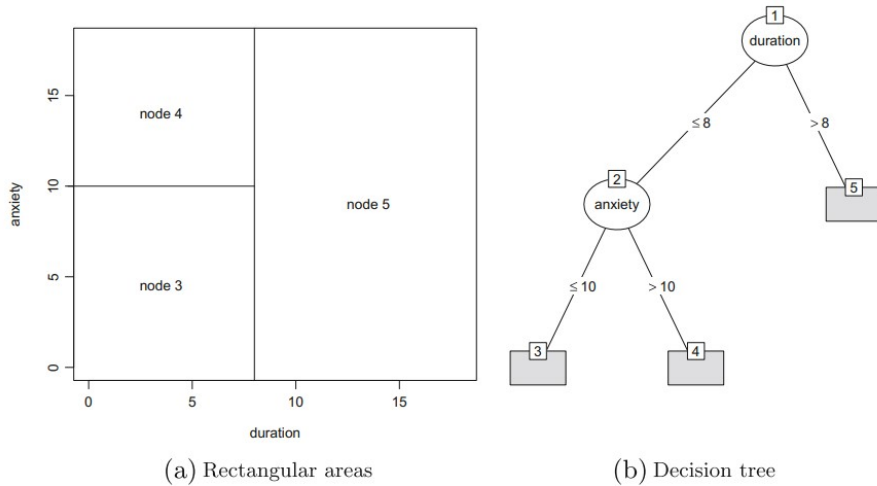


Figura 3.2: Ejemplo partición recursiva

Fuente: (Fokkema et al., 2018)

Las regiones rectangulares de una partición recursiva se pueden representar gráficamente como nodos en un árbol de decisión, como se muestra en la figura (3.2). La primera división del árbol separa las observaciones en dos subgrupos, por medio de la variable duración y un valor de división en cual es ocho, esto produce dos regiones rectangulares que también se representan el nodo 2 y el nodo 5. El nodo 2 es un nodo interno, ya que las observaciones en este nodo se dividen a su vez en los

nodos terminales 3 y 4. Las observaciones en el nodo 5 no se dividen más y, por lo tanto, este es un nodo terminal.

Si usamos el árbol de la figura (3.2) para predecir sobre una observación, esta se asigna a uno de los nodos terminales según sus valores en las variables de división. Entonces la predicción se basa en la distribución estimada de la variable Y dentro del nodo terminal. Por ejemplo, la predicción puede ser la media específica del nodo de una única variable continua o la moda en una variable cualitativa. (Fokkema et al., 2018) presenta una metodología en la cual los nodos terminales del árbol de decisión resultan ser un modelo lineal generalizado, en donde el valor predicho para una nueva observación se determina mediante las estimaciones de parámetros específicos del nodo, en donde también se ajustan los efectos aleatorios.

(Fokkema et al., 2018) expone que en el modelo *glmertree* los efectos fijos β_m son parámetros locales y sus valores dependen del nodo terminal m , mientras que los efectos aleatorios μ_i son globales, es decir estos son constantes para todos los nodos terminales. Los autores exponen el siguiente procedimiento para la estimación del árbol:

- Paso 0:** Iniciar estableciendo un valor r (numero de iteraciones) y todos los valores de los efectos aleatorios $\hat{u}_{(r)} = 0$.
- Paso 1:** Establecer $r = r + 1$ y estimar un árbol GLM usando la estimación de los efectos aleatorios $\hat{u}_{(r-1)}$.
- Paso 2:** Ajustar el modelo de efectos mixtos $g(p_{ijm}) = \mathbf{x}_{ij}\beta_m + \mu_i$, donde m corresponde al nodo terminal y u del paso anterior.
Con el modelo ajustado extraer $\hat{u}_{(r)}$
- Paso 3:** Repetir los pasos 1 y 2 hasta lograr convergencia.

iii. Mixed Effects Tree-Boosting (MEBO): el modelo a implementar es denominado *GPBOOST* el cual combina los modelos basados en arboles con procesos gaussianos y de efectos mixtos por medio del método de ensamblaje Boosting. Para la implementación se usa la librería *gpboost* (Sigrist et al, (2021), en donde la variable respuesta Y es la suma de una función media potencialmente no lineal $F(X)$ y de efectos aleatorios μ_i :

$$y_{ij} = F(\mathbf{x}_{ij}) + \mu_i + e_{ij} \quad (3.1)$$

donde $F(\mathbf{x}_{ij})$ es una suma o conjunto de árboles, construidos con la información auxiliar \mathbf{x}_{ij} y e_{ij} es un término de error independiente. Los efectos aleatorios μ_i actualmente pueden consistir en procesos gaussianos (incluidos los procesos con

coeficientes aleatorios), los cuales se usan en este trabajo; efectos aleatorios agrupados (incluidos los efectos anidados, cruzados y de coeficientes aleatorios) y combinaciones de estos.

La función marginal de y en la ecuación (3.1) es:

$$y \sim \mathcal{N}(F(X), \Psi), \quad \Psi = Z\Sigma Z' + \sigma_\mu^2 I_n$$

En un modelo de proceso gaussiano, los efectos aleatorios $u = (u_1, \dots, u_D)'$ corresponden a un proceso gaussiano de dimensión finita con una función de covarianza $cov(\mu_i, \mu_{i'}) = c(i, i')$

Normalmente se supone que la función de covarianza es de la forma:

$$c(i, i') = \sigma_1^2 r(\|i - i'\| / \rho)$$

donde r es una función de autocorrelación con $r(0) = 0$. $\sigma_1^2 = V(u)$ y ρ es un parámetro que determina qué tan rápido decae la autocorrelación con respecto a la distancia.

Bajo un proceso gaussiano, la matriz de covarianza se aproxima por (Sigrist et al., 2021):

$$(\Sigma)_{jk} = \sigma_1^2 r(d_{jk} / \rho)$$

con $d_{lk} = \|u_l - u_k\|$, $l, k = 1, \dots, D$

El modelo *gpboost* busca encontrar:

$$(\hat{F}(\cdot), \hat{\alpha}) = \underset{(F(\cdot), \alpha) \in (\mathcal{H}, \Theta)}{\operatorname{argmin}} R(F(\cdot), \alpha),$$

con

$$R(F(\cdot), \alpha) : (F(\cdot), \alpha) \mapsto L(y, F, \alpha)|_{F=F(X)},$$

y $L(y, F, \alpha)$ es una función de pérdida correspondiente a la log-verosimilitud de 3.1.

(Sigrist et al., 2021) proponen que para un α fijo, el método boosting permite obtener el mínimo $R(F(\cdot), \alpha)$ de un forma secuencial en la que se agrega una actualización $f_m(\cdot)$ en la estimación de $F_m(\cdot)$:

$$F_m(\cdot) = F_{m-1}(\cdot) + f_m(\cdot), \quad m = 1, \dots, M,$$

donde $f_m(\cdot)$ se escoge de tal forma que su adición resulte en la minimización de la función. Esta minimización normalmente no se puede hacer analíticamente y por lo tanto se utiliza una aproximación.

Estimar o entrenar el modelo *gpboost* implica que el algoritmo aprende de los parámetros de covarianza, de los efectos aleatorios y de la función predictora $F(X)$ en cada iteración, mejorando la predicción resultante. Dentro de las principales ventajas que provee este método se encuentra:

- Busca reducir el sesgo por lo que provee una gran ventaja en la precisión de la predicción.
- Modela las relaciones no lineales, discontinuas o con interacciones complejas en las variables predictoras.
- Es robusto ante valores atípicos y ante multicolinealidad entre las variables predictoras.

3.2. Estimación de la incertidumbre

Ya definido el estimador a usar el siguiente paso es determinar que tan confiable es la estimación resultante, esto se determina por medio de las medidas de precisión, en donde se tendrá en cuenta la variabilidad por medio del Error Cuadrático Medio (ECM) y la exactitud evaluado por el Sesgo absoluto (B).

Error cuadrático medio

Para la estimación del error cuadrático medio del estimador plug-in, independientemente del modelo que se use para la estimación en los datos no muestreados, seguimos el procedimiento planteado en (Molina, 2019) allí se exponen dos alternativas Bootstrap para la estimación del ECM:

Bootstrap paramétrico:

1. Ajustar el GLMM a los datos de la muestra s , obteniendo estimadores $\hat{\sigma}_\mu^2$ y $\hat{\beta}$ de los parámetros del modelo.
2. Generar efectos aleatorios independientes Bootstrap

$$\mu_i^{*(b)} \sim N(0, \hat{\sigma}_\mu^2), \quad i = 1, \dots, D$$

3. Generar un censo Bootstrap $\mathbf{y}_i^{*(b)} = (Y_{i1}, \dots, Y_{iN_i})'$, de la forma

$$Y_{ij}^{*(b)} \stackrel{ind}{\sim} \text{Bern}(p_{ij}^{*(b)}), p_{ij}^{*(b)} = h(\mathbf{x}_{ij}\hat{\beta} + \mu_i^{*(b)}), j = 1, \dots, N_i, i = 1, \dots, D,$$

y calcular los verdaderos valores de los indicadores $\bar{Y}_i^{*(b)}$, $i = 1, \dots, D$.

4. Para cada área $i = 1, \dots, D$, extraer del censo Bootstrap $\mathbf{y}_i^{*(b)}$ los elementos de la muestra de esa área, Y_{ji} , $j \in s_i$, construyendo el vector $\mathbf{y}_{is}^{*(b)}$.

Sea $\mathbf{y}_s^{*(b)} = \left(\left(\mathbf{y}_{1s}^{*(b)} \right)', \dots, \left(\mathbf{y}_{Ds}^{*(b)} \right)' \right)'$ el vector con los valores en la muestra de todas las áreas, siendo $s = s_1 \cup \dots \cup s_D$ la muestra original.

5. Ajustar el modelo GLMM a los datos Bootstrap $\mathbf{y}_s^{*(b)}$ y calcular los predictores Bootstrap $\hat{\theta}_i^{PI^{*(b)}}$, $i = 1, \dots, D$.
6. Repetir los pasos 2-5, para $b = 1, \dots, B$.

El estimador Bootstrap del ECM del predictor $\hat{\theta}_i^{PI}$ viene dado por

$$ECM \left(\hat{\theta}_i^{PI} \right) = B^{-1} \sum_{b=1}^B \left(\hat{\theta}_i^{PI^{*(b)}} - \bar{Y}_i^{*(b)} \right)^2$$

Bootstrap no paramétrico:

1. Con los puntos de la muestra (x_j, y_j) se genera una población replicando estos puntos con el factor de expansión $\{(x_j^*, y_j^*)\}$ para $j = 1, \dots, N_i^*$ donde $N_i^* = [\hat{N}_i]$
2. Calcular la media de cada dominio para la población Bootstrap

$$\bar{Y}_i^* = \frac{1}{N_i^*} * \sum_{j=1}^{N_i^*} y_{ij}^*$$

3. Se toman B muestras con remplazo de la población Bootstrap, para cada muestra b se calcula el estimador Plug-In $\hat{\theta}_i^{PI^{*(b)}}$
4. El estimador Bootstrap del ECM del predictor \hat{Y}_i^{PI} viene dado por

$$ECM \left(\hat{\theta}_i^{PI} \right) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\theta}_i^{*(b)} - \bar{Y}_i^* \right)^2$$

Uno de los principales objetivos de este trabajo es obtener una metodología flexible en comparación a la metodología actual de estimación de proporción en áreas pequeñas, por lo tanto en los ejercicios de simulación y en la aplicación se hará uso de la metodología Bootstrap no paramétrico para la estimación del ECM. Como se comparan diferentes proporciones, para evaluar el desempeño de los estimadores propuestos se utilizarán también las siguientes medidas:

Error cuadrático medio relativo

$$RECM \left(\hat{\theta}_i^{PI} \right) = \frac{ECM \left(\hat{\theta}_i^{PI} \right)}{\hat{\theta}_i^{PI}}$$

Sesgo absoluto

$$B(\hat{\theta}_i^{PI}) = | \hat{\theta}_i^{PI} - \theta_i |$$

Sesgo relativo

$$RB(\hat{\theta}_i^{PI}) = \frac{B(\hat{\theta}_i^{PI})}{\hat{\theta}_i^{PI}}$$

3.3. Escenario de simulación

Para evaluar y comparar los estimadores propuestos, se implementa un esquema de simulación que considera distintos tamaños muestrales, este ejercicio se basa en el trabajo de Krennmair et al (2022) donde se busca replicar el escenario de simulación propuesto con el fin de comparar los resultados obtenidos. Esto permite representar un escenario más realista, donde los dominios presentan variaciones en el tamaño de la muestra. Los valores asignados a los parámetros fueron seleccionados de manera que las proporciones estimadas se encuentren en el intervalo $(0,3, 0,7)$. Esta elección se justifica en que, al tratarse de valores centrales, el error de muestreo tiende a ser mayor, lo que permite evaluar el desempeño de las técnicas en condiciones más exigentes. Si el método demuestra ser eficaz bajo estas circunstancias, también lo será en situaciones con menor error. Cabe destacar que la función $p(1 - p)$, que determina la varianza de un estimador de proporción, alcanza su valor máximo en $p = 0,5$. Por esta razón, este valor es comúnmente utilizado en el cálculo del tamaño muestral, garantizando un enfoque conservador en la estimación de errores.

Este escenario se desarrolla utilizando el método MonteCarlo, en el cual se ejecuta el siguiente procedimiento 100 veces:

1. Se genera una población compuesta por treinta dominios ($D = 30$), con tamaños poblacionales $N_i = 1000$ para $i = 1, \dots, D$.
2. Se simulan los efectos aleatorios μ_i para cada dominio mediante la distribución normal, $\mu_i \sim N(0, 2)$.
3. Se generan las variables auxiliares independientes $x_{ij1} \sim N(\mu_i, 1)$ y $x_{ij2} \sim N(\mu_i, 1)$.
- 4 Se calcula la variable dependiente:

$$p_{ij} = \alpha_0 + \alpha_1 * x_{ij1} + \alpha_2 * x_{ij2} + e_{ij}$$

considerando los siguientes parámetros:

- $\alpha_0 = 0,7$
- $\alpha_1 = 0,6$
- $\alpha_2 = -0,3$
- $e_{ij} \sim N(0, 0,25)$

5. Se obtiene la probabilidad de éxito y luego la variable dicotómica mediante el vínculo logístico:

$$Y_{ij} \sim \text{Bernoulli} \left(\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) \right)$$

6. A partir de la población generada, se extrae una muestra mediante un diseño de muestreo estratificado, en el cual las unidades son seleccionadas mediante un muestreo aleatorio simple dentro de cada estrato (ESTMAS). En este ejercicio, los estratos corresponden a los dominios y , con el propósito de evaluar el desempeño de los estimadores bajo distintos escenarios muestrales, se asigna un tamaño de muestra diferente a cada dominio. Específicamente, los tamaños muestrales se establecen de la siguiente manera: $n_1 = 10$, $n_2 = 20$, $n_3 = 30$, y así sucesivamente, hasta el último dominio, cuyo tamaño muestral es $n_{30} = 300$.

7. Para cada iteración $a = 1, \dots, 100$ se calculan a nivel de dominio:

- i. Las proporciones poblacionales en cada muestra generada

$$\theta_i^a = \frac{1}{N_i} \sum_{U_i^a} y_{ij}$$

- ii. Las estimaciones directas por medio del estimador Horvitz-Thompson, haciendo uso de los pesos muestrales $w_{ij} = \frac{N_i}{n_i}$, con N_i y n_i tamaños poblacionales y muestrales del dominio i respectivamente

$$\hat{\theta}_i^{\text{Directo}^a} = \frac{1}{N} \sum_{j \in s_i^a} w_{ij} * y_{ij}$$

- iii. Se toma una muestra de entrenamiento correspondiente al 80 % de la muestra, con esta se ajustan los 3 modelos (GLMM, MEBA, MEBO) y se calcula el estimador Plug-In para cada uno

$$\hat{\theta}_i^{\text{PI}^a} = \frac{1}{N_i} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} \hat{p}_{ij} \right)$$

-
- iv. Para cada estimación $\hat{\theta}_i^{PI-GLMM^a}$, $\hat{\theta}_i^{PI-MEBA^a}$ y $\hat{\theta}_i^{PI-MEBO^a}$ se realiza la estimación del error cuadrático medio por medio del método Bootstrap descrito en la sección anterior, junto con el sesgo absoluto y las medidas relativas de estas estimaciones.
8. Al finalizar las 100 iteraciones, para cada uno de los 3 métodos (GLMM, MEBA, MEBO) se calcula el promedio de cada una de las estimaciones $\hat{\alpha}^*$ de la siguiente forma:

$$\hat{\theta}^* = \frac{1}{a} \sum_{l=1}^{100} \hat{\alpha}^a$$

donde $\hat{\alpha}^* \in \{\hat{\theta}_i^{PI}, ECM(\hat{\theta}_i^{PI}), RECM(\hat{\theta}_i^{PI}), B(\hat{\theta}_i^{PI}), RB(\hat{\theta}_i^{PI})\}$

En primer lugar se evalúa el sesgo absoluto y el sesgo relativo en cada uno de los tres métodos propuestos, para esto se presenta la tabla 3.1 correspondiente a la estimación de los valores de estas medidas para cada uno de los 30 dominios, según las simulaciones. La figura 3.3 correspondiente al boxplot para el sesgo absoluto $B(\hat{\theta}_i^{PI})$ y el sesgo relativo $RB(\hat{\theta}_i^{PI})$, evidenciando un menor sesgo en los estimadores basados en el modelo GLMM y MEBO. Un hallazgo a resaltar es como el estimador basado en el modelo MEBA tiene un sesgo alto, incluso en el caso del sesgo relativo llega a ser mayor que en los estimadores directos.

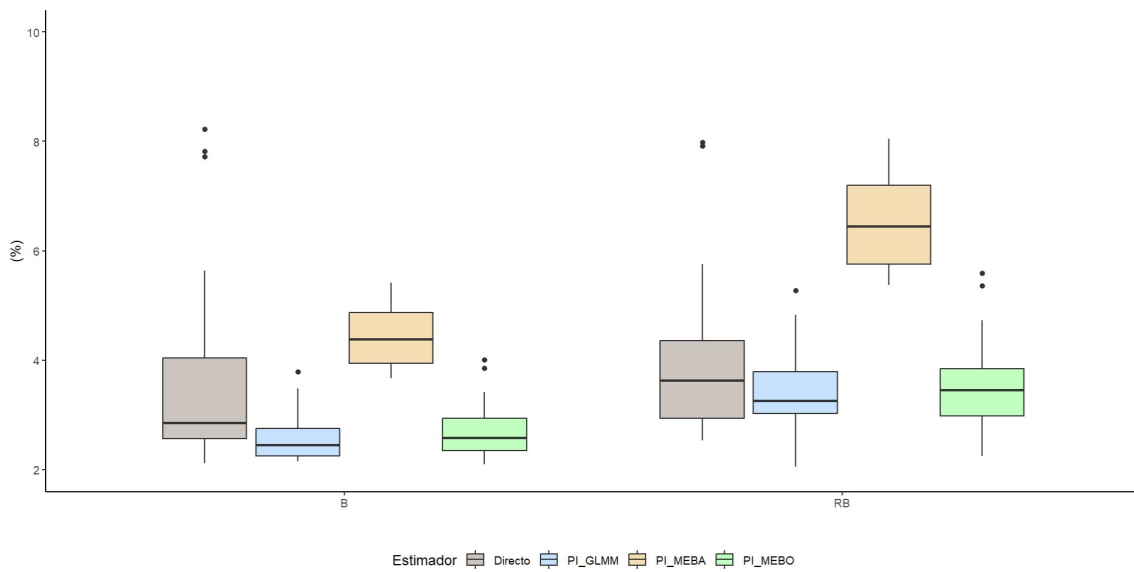


Figura 3.3: Comparación del sesgo en la simulación
Fuente: elaboración propia.

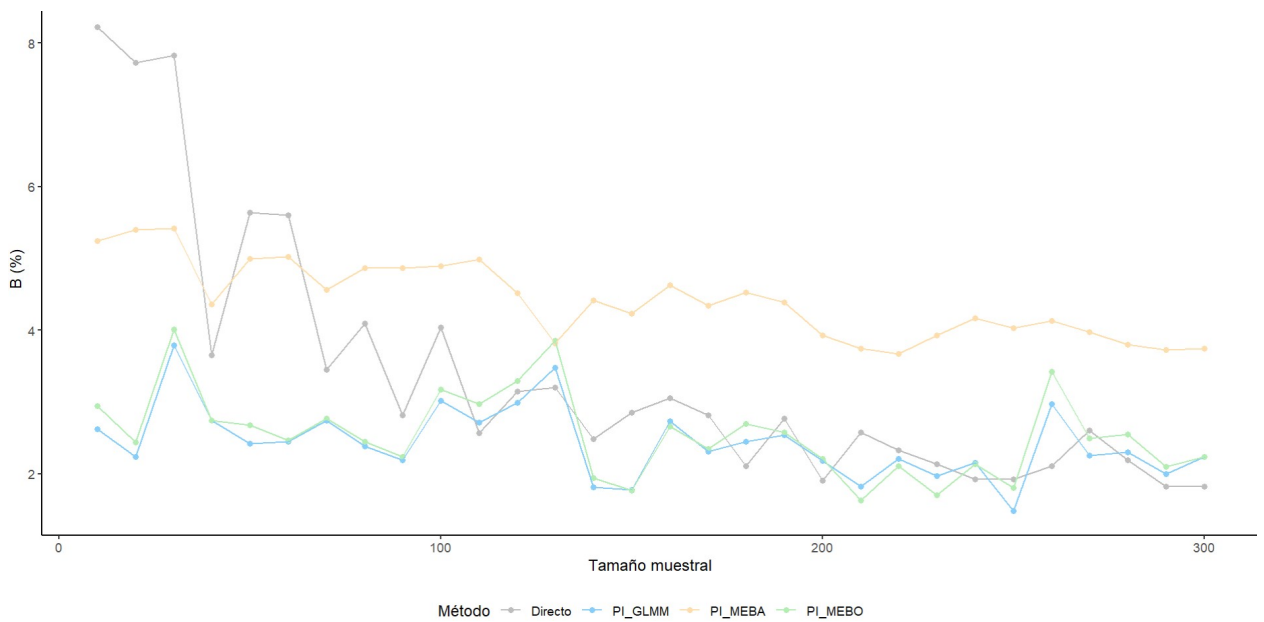


Figura 3.4: Comportamiento del sesgo según el tamaño muestral
Fuente: elaboración propia.

Dominio	Muestra	Sesgo absoluto (B %)				Sesgo relativo (RB %)			
		Directo	PI-GLMM	PI-MEBA	PI-MEBO	Directo	PI-GLMM	PI-MEBA	PI-MEBO
1	10	8,2	2,6	5,2	2,9	10,9	3,6	7,8	4,1
2	20	7,7	2,2	5,4	2,4	10,9	3,1	8,0	3,4
3	30	7,8	3,8	5,4	4,0	11,3	5,3	8,0	5,6
4	40	3,7	2,7	4,4	2,7	5,1	3,8	6,5	3,8
5	50	5,6	2,4	5,0	2,7	8,0	3,4	7,4	3,7
6	60	5,6	2,4	5,0	2,5	7,9	3,4	7,4	3,4
7	70	3,5	2,7	4,6	2,8	4,8	3,8	6,8	3,9
8	80	4,1	2,4	4,9	2,4	5,8	3,3	7,2	3,4
9	90	2,8	2,2	4,9	2,2	3,9	3,0	7,2	3,1
10	100	4,0	3,0	4,9	3,2	5,6	4,2	7,2	4,4
11	110	2,6	2,7	5,0	3,0	3,6	3,8	7,4	4,1
12	120	3,1	3,0	4,5	3,3	4,3	4,1	6,6	4,6
13	130	3,2	3,5	3,8	3,9	4,4	4,8	5,6	5,4
14	140	2,5	1,8	4,4	1,9	3,5	2,5	6,5	2,7
15	150	2,9	1,8	4,2	1,8	4,0	2,5	6,2	2,5
16	160	3,1	2,7	4,6	2,7	4,3	3,8	6,8	3,7
17	170	2,8	2,3	4,3	2,3	3,9	3,2	6,4	3,2
18	180	2,1	2,4	4,5	2,7	2,9	3,4	6,6	3,7
19	190	2,8	2,5	4,4	2,6	3,8	3,5	6,4	3,6
20	200	1,9	2,2	3,9	2,2	2,6	3,0	5,7	3,1
21	210	2,6	1,8	3,7	1,6	3,5	2,5	5,5	2,3
22	220	2,3	2,2	3,7	2,1	3,2	3,1	5,4	2,9
23	230	2,1	2,0	3,9	1,7	2,9	2,7	5,7	2,4
24	240	1,9	2,2	4,2	2,1	2,7	3,0	6,1	3,0
25	250	1,9	1,5	4,0	1,8	2,7	2,1	5,9	2,5
26	260	2,1	3,0	4,1	3,4	2,9	4,1	6,0	4,7
27	270	2,6	2,3	4,0	2,5	3,6	3,1	5,8	3,5
28	280	2,2	2,3	3,8	2,6	3,0	3,2	5,5	3,5
29	290	1,8	2,0	3,7	2,1	2,5	2,8	5,4	2,9
30	300	1,8	2,2	3,7	2,2	2,5	3,1	5,5	3,1

Tabla 3.1: Comparación del sesgo en la simulación

En la figura 3.4 se compara el comportamiento del sesgo según el tamaño muestral, recordando que se fijó un n_i diferente en cada dominio, iniciando en 10 hasta el último dominio un tamaño de muestra 300. Se evidencia como el sesgo del estimador directo va disminuyendo a medida que aumenta el tamaño muestral, llegando a ser menor que los estimadores basados en el modelo en tamaños muestrales cercanos a 300. El estimador basado en el modelo MEBA también tiende a disminuir a medida que aumenta el tamaño de muestra, sin embargo es el que presenta el peor comportamiento respecto al sesgo. Los estimadores basados en el modelo GLMM y MEBO presentan valores similares a lo largo de las estimaciones, siendo MEBO ligeramente más sesgado que el estimador PI-GLMM.

En segundo lugar se evalúa el error cuadrático medio y el error cuadrático medio rela-

tivo en cada uno de los tres métodos propuestos, la tabla 3.2 presenta los valores de estas medidas para cada uno de los 30 dominios, evidenciando valores altos para el estimador directo y el estimador basado en el modelo MEBA, por esto no se grafican en el boxplot presentado en la figura 3.5 correspondiente al boxplot para el $ECM(\hat{\theta}_i^{PI})$ y $RECM(\hat{\theta}_i^{PI})$, aquí se ve un comportamiento similar en ambos estimadores, teniendo una mediana ligeramente mayor en el estimador basado en el modelo MEBO y además se ve la presencia de un mayor número de estimaciones atípicas.

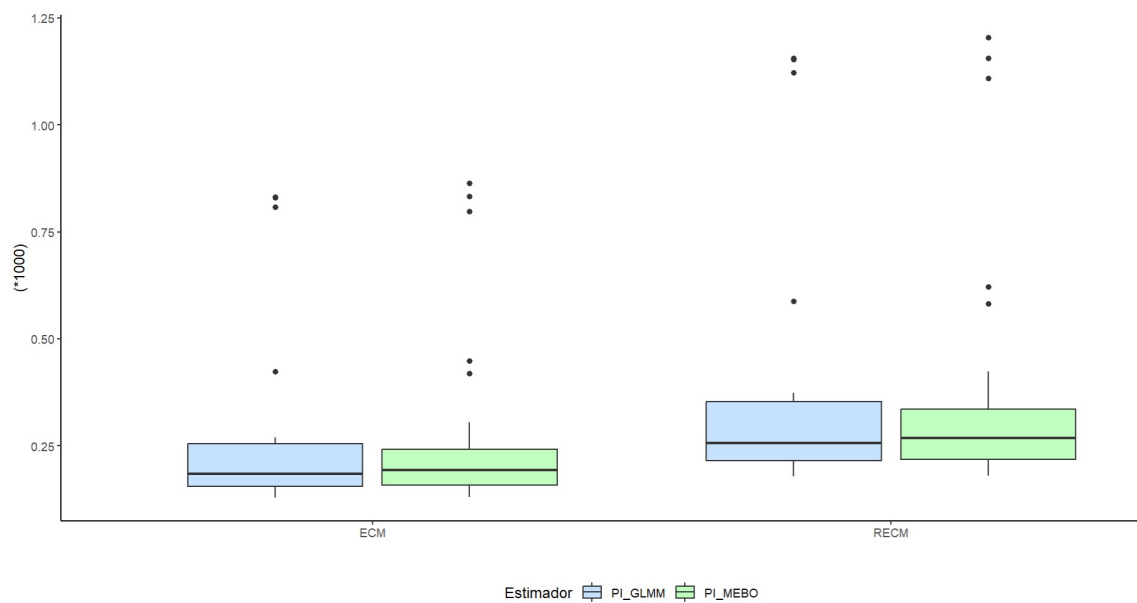


Figura 3.5: Comparación del error cuadrático medio en la simulación
Fuente: elaboración propia.

Dominio	Muestra	(10.000*) ECM				(10.000*) RECM			
		Directo	PI-GLMM	PI-MEBA	PI-MEBO	Directo	PI-GLMM	PI-MEBA	PI-MEBO
1	10	291,7	8,3	29,2	8,3	385,8	11,5	43,4	11,6
2	20	192,7	8,1	30,9	8,0	27,2	11,2	45,9	11,1
3	30	168,2	8,3	32,5	8,6	24,4	11,6	48,3	12,0
4	40	71,2	2,5	21,7	2,4	10,0	3,4	32,2	3,3
5	50	83,9	4,2	26,5	4,5	11,9	5,9	39,2	6,2
6	60	72,5	4,2	27,4	4,2	10,2	5,9	40,6	5,8
7	70	49,4	2,7	22,9	2,9	6,9	3,7	33,9	4,0
8	80	46,9	2,6	26,0	2,4	6,6	3,6	38,5	3,4
9	90	34,1	2,3	25,0	2,3	4,7	3,1	36,9	3,2
10	100	45,3	2,6	25,5	3,0	6,3	3,6	37,7	4,2
11	110	27,2	2,0	26,8	2,0	3,8	2,7	39,5	2,8
12	120	26,9	1,8	22,8	1,9	3,7	2,5	33,5	2,7
13	130	28,9	1,9	17,6	1,9	4,0	2,6	26,0	2,7
14	140	22,5	1,8	20,6	1,8	3,2	2,5	30,4	2,6
15	150	22,1	1,9	19,1	2,1	3,1	2,7	28,2	2,8
16	160	26,7	2,2	24,6	2,2	3,7	3,1	36,2	3,0
17	170	21,1	1,7	20,6	1,8	2,9	2,4	30,2	2,5
18	180	17,4	1,5	22,7	1,4	2,4	2,1	33,3	2,0
19	190	20,7	1,9	20,8	1,9	2,9	2,7	30,5	2,7
20	200	14,4	1,3	17,0	1,3	2,0	1,8	24,9	1,8
21	210	16,2	1,6	15,1	1,5	2,2	2,2	22,0	2,1
22	220	15,0	1,6	14,4	1,5	2,1	2,2	21,1	2,0
23	230	13,2	1,6	16,5	1,6	1,8	2,2	24,1	2,2
24	240	11,8	1,5	19,0	1,5	1,6	2,1	27,7	2,1
25	250	10,9	1,5	17,3	1,6	1,5	2,0	25,3	2,3
26	260	11,3	1,4	19,1	1,3	1,6	1,9	27,9	1,8
27	270	15,0	1,6	17,5	1,9	2,1	2,2	25,5	2,7
28	280	11,9	1,5	16,0	1,5	1,6	2,0	23,2	2,0
29	290	10,2	1,7	15,0	1,7	1,4	2,4	21,9	2,4
30	300	9,3	1,5	15,6	1,6	1,3	2,1	22,7	2,2

Tabla 3.2: Comparación del error cuadrático medio en la simulación

En el gráfico 3.6 se compara el comportamiento del error cuadrático medio según el tamaño muestral. Nuevamente se ve como el ECM del estimador directo va disminuyendo a medida que aumenta el tamaño muestral, sin embargo no llega a ser menor a los estimadores basados en el modelo, muy seguramente por que la población generada fue simulada siguiendo una distribución. Los estimadores basados en el modelo GLMM y MEBO presentan valores similares a lo largo de las estimaciones, presentando un comportamiento similar en todos los diferentes tamaños de muestra.

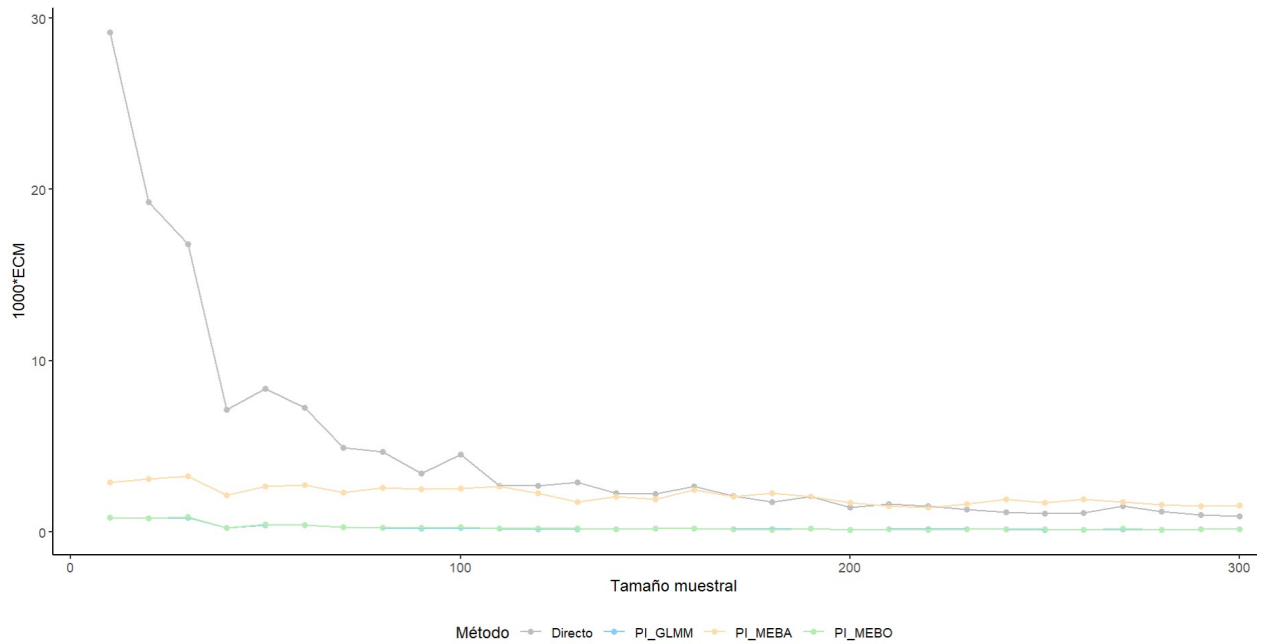


Figura 3.6: Comportamiento del ECM según el tamaño muestral
Fuente: elaboración propia.

En tercer lugar dado que:

$$ECM(\hat{\theta}_i^{PI}) = V(\hat{\theta}_i^{PI}) + B(\hat{\theta}_i^{PI})^2$$

Se evalúa la varianza estimada en cada uno de los tres métodos propuestos, en la figura 3.7 se evidencia un hallazgo importante respecto al estimador basado en el modelo MEBA, este es el que presenta la menor varianza dado el método de ensamblaje Bagging cuyo principal objetivo es la reducción de la varianza, a comparación del método Boosting que busca reducir también sesgo sacrificando un poco la varianza. Respecto a los estimadores basados en el modelo GLMM y MEBO se observa un comportamiento similar entre ellos, presentando pequeñas diferencias en la estimación de algunos dominios.

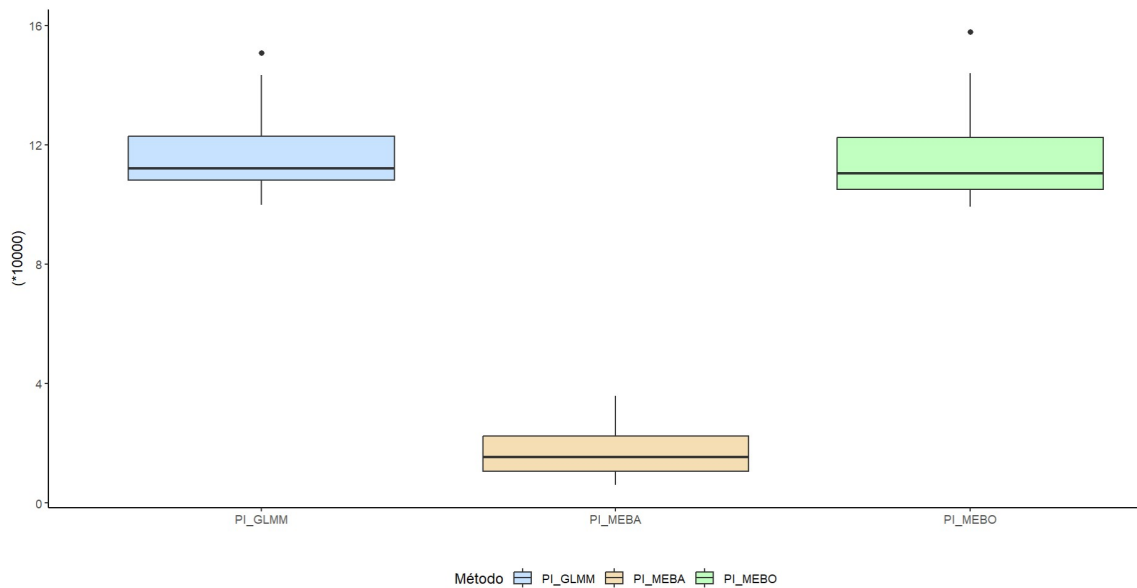


Figura 3.7: Comparación de la varianza en la simulación
Fuente: elaboración propia.

Finalmente se presentan las estimaciones por los 3 métodos propuestos y se comparan con las estimaciones directas, en donde se ve la volatilidad de las estimaciones directas en los dominios con tamaños de muestra menores a 100 cuyos errores muestrales superan un margen de error mayor a 15 %; a medida que aumenta el tamaño de muestra estas tienden a estabilizarse y comportarse de forma similar a las estimaciones basadas en el modelo GLMM y MEBO, las cuales permanecen sobre el mismo intervalo en los diferentes escenarios de muestra. En el caso de las estimaciones basadas en el modelo MEBA aunque tienden a ser estables sin importar el tamaño muestral, nuevamente se ve que introducen un gran sesgo.

Dominio	Muestra	Estimación (%)			
		Directo	PI-GLMM	PI-MEBA	PI-MEBO
1	10	75,6	72,2	67,4	72,1
2	20	70,8	72,1	67,4	72,0
3	30	69,0	71,8	67,3	71,7
4	40	71,5	72,1	67,4	72,1
5	50	70,6	72,0	67,5	72,2
6	60	70,7	72,1	67,5	72,0
7	70	71,7	72,1	67,6	71,9
8	80	71,2	72,0	67,6	72,0
9	90	72,5	72,1	67,8	72,0
10	100	71,5	72,0	67,7	72,0
11	110	71,8	72,0	67,8	71,9
12	120	73,3	72,2	68,0	72,1
13	130	72,4	72,1	67,9	72,0
14	140	71,5	72,1	67,9	72,1
15	150	71,7	72,2	68,0	72,1
16	160	71,3	71,9	67,9	71,9
17	170	72,8	72,4	68,3	72,3
18	180	72,2	72,2	68,2	72,2
19	190	72,4	72,3	68,3	72,3
20	200	72,5	72,3	68,3	72,2
21	210	73,3	72,5	68,6	72,3
22	220	71,8	72,2	68,3	72,1
23	230	72,7	72,3	68,5	72,3
24	240	72,5	72,3	68,6	72,2
25	250	72,0	72,2	68,5	72,1
26	260	72,4	72,3	68,6	72,3
27	270	71,8	71,8	68,5	71,9
28	280	72,7	72,4	68,8	72,3
29	290	71,9	71,9	68,6	71,8
30	300	71,8	72,1	3,4	72,2

Tabla 3.3: Comparación de las estimaciones en la simulación

3.4. Aplicación en datos reales

El programa PNIS (Programa Nacional Integral de Sustitución de Cultivos de Uso Ilícito) surgió por el acuerdo de paz en Colombia que se desarrolló en el año 2016. El principal objetivo de este es tener un país libre de cultivos ilícitos acorde a los derechos humanos, la protección del medio ambiente y el bienestar. Según la Unión Temporal IPSOS-Uniandes (2023) este programa fue implementado en 56 municipios y benefició a 99.097 hogares relacionados con los cultivos de uso ilícito en calidad de cultivadores, no cultivadores en zonas de influencia de estos cultivos y recolectores.

En el año 2022, el Departamento Nacional de Planeación (DNP) contrató a la Unión Temporal Ipsos-Uniandes para realizar la evaluación institucional y de resultados de la evaluación. En el archivo nacional de datos del DNP (anda.dnp.gov.co) se encuentran los documentos referentes a esta evaluación y además los microdatos de las encuestas recolectadas que son de acceso público. Según la Unión Temporal IPSOS-Uniandes (2023), esta evaluación se divide en dos componentes principales:

- El primer componente tuvo como objetivo validar y analizar el diseño institucional del programa para cumplir con objetivos alineados con el Acuerdo Final de Paz y la Política Integral de Drogas Ruta Futuro, en relación con la articulación entre actores y políticas estatales dispuestos para tal fin.
- El segundo componente tuvo como objetivo evaluar los resultados del PNIS en relación con los distintos tipos de beneficios contemplados y sus líneas de intervención. Para ello, se identificaron los resultados de cada acción implementada en el marco del programa, al tiempo que se analizaron los factores internos y externos que han influido en su implementación y en el cumplimiento de sus objetivos. Entre estos factores, se consideraron las características de los territorios y las problemáticas asociadas al conflicto armado..

Para dar cumplimiento a estos objetivos se plantearon diferentes investigaciones, haremos énfasis en el análisis cuantitativo. En este contexto, se realizó un estudio del programa centrado directamente en los beneficiarios, para lo cual se diseñó una estrategia muestral orientada a obtener información confiable que permitiera evaluar el cumplimiento de los objetivos del programa. La población objetivo del estudio está conformada por las familias beneficiarias del PNIS, ubicadas en los territorios donde se ha implementado el programa, que abarcan 14 departamentos, 56 municipios y aproximadamente 3.785 veredas.

Para la implementación del diseño muestral, se usó como marco muestral el registro administrativo de los hogares inscritos en el programa, el cual se compone de diferentes variables, las cuales fueron tomadas como información auxiliar en el análisis. El diseño muestral planteado como se muestra en la figura 3.8 fue probabilístico de dos etapas, en la primera etapa es un diseño estratificado en donde se seleccionan aleatoriamente por medio

de un muestreo aleatorio simple dentro de cada estrato las veredas, las cuales corresponden a las unidades primarias de muestreo (UPM), los estratos se construyeron en función del número de beneficiarios por veredas. En la segunda etapa por medio del mismo diseño (ESTMAS) se seleccionan aleatoriamente por tipo de actividad (Cultivador, No cultivador y Recolector) los hogares beneficiarios dentro de cada vereda seleccionada en la primera etapa.

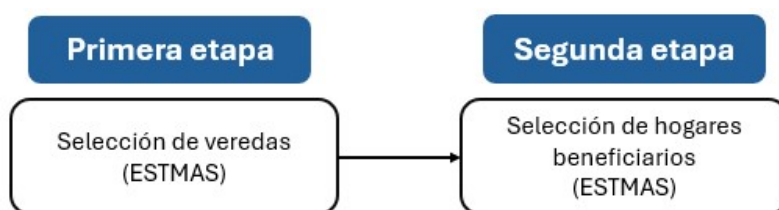


Figura 3.8: Esquema del diseño
Fuente: elaboración propia.

Se recolectó un total de 1.730 encuestas, las cuales fueron aplicadas en 121 veredas, 42 municipios y 12 departamentos. En la tabla 3.4 se presentan los valores descriptivos del número de encuestas a nivel municipal y departamental y la figura 3.9 se presenta un mapa de calor correspondiente al tamaño de muestra por departamento. Dado que el cuartil 3 correspondiente al 75 % en la muestra de municipios es 32, se evidencia que la gran mayoría cuentan con una muestra pequeña, por lo que no es posible obtener estimaciones haciendo uso del diseño muestral allí. A nivel de departamento el escenario es similar pues 2 de los 14 departamentos donde se ha implementado el programa no cuentan con muestra, la mediana es 93, indicando también que es muy posible que al realizar las estimaciones a nivel de departamento sean muy pocos en los que se pueda publicar la cifra.

Dominio	Municipio	Departamento
Minimo	0	0
Q1	8	28
Mediana	18	93
Media	41	144
Q3	32	186
Maximo	591	591

Tabla 3.4: Descripción de la muestra

El objetivo es determinar la efectividad que tuvo el programa en términos de la sustitución

ción de cultivos ilícitos en los territorios donde fue implementado, dentro del cuestionario se preguntó: Después de empezar la implementación del PNIS, ¿ha habido erradicaciones forzadas en la vereda en la que está inscrito su predio o finca? (Si/No). Esta variable dicotómica es la variable objetivo. Se estimará la proporción de hogares beneficiarios del programa en los que ha habido erradicaciones forzadas después de implementar el programa, este parámetro se toma como un indicador de éxito del programa.

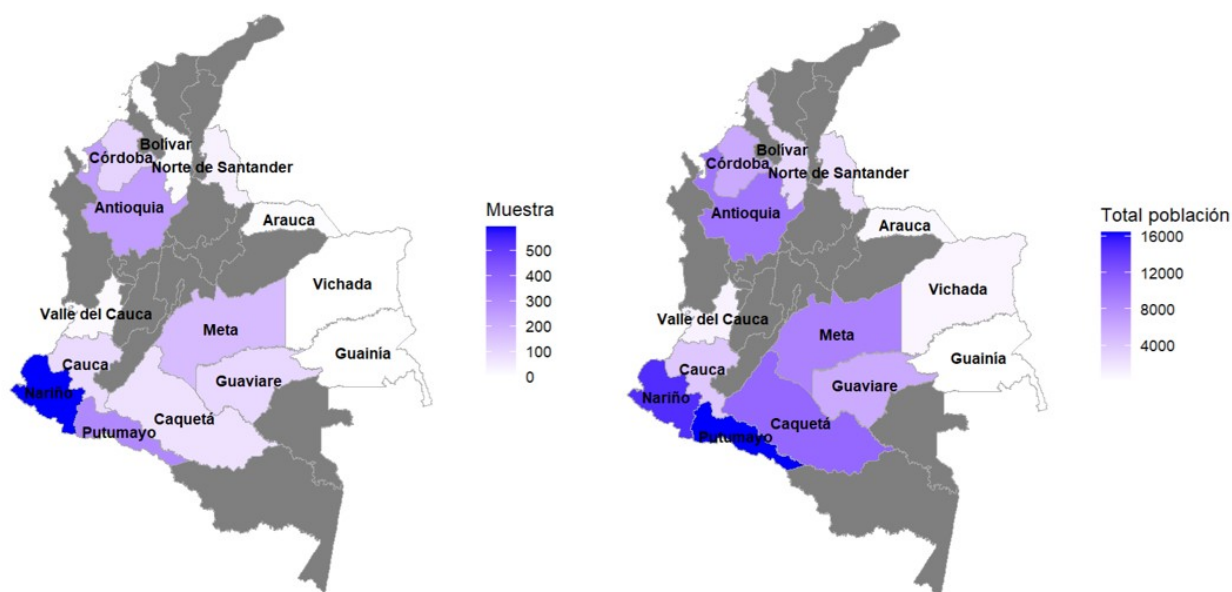


Figura 3.9: Tamaño muestral y poblacional por departamento
Fuente: elaboración propia.

En primer lugar se calculan las estimaciones directas a nivel de departamento y siguiendo los pasos establecidos por CEPAL (2023), tomando en lugar del coeficiente de variación el margen de error dado que se está estimando una proporción, se determina si las estimaciones son o no publicables. En la tabla 3.5 se presentan los resultados, en donde solo 4 de los 14 departamentos cumplen con los criterios de calidad para publicar la cifra estimada. En los departamentos Guainía y Vichada no se cuenta con estimaciones dado que no se recolectó muestra por las dificultades de acceso a estos territorios. Es de suma importancia obtener el indicador de éxito del programa a nivel de departamento, dado que las estimaciones directas no son suficientes se hará uso de la metodología propuesta para lograr el objetivo.

Departamento	Estimación	n	n_{eff}	Error relativo	Margen de error
Antioquia	9.7%	247	124	6,5%	12,7%
Arauca	0,0%	8		0,0%	0,0%
Bolívar	58.4%	8	4	8,1%	15,9%
Caquetá	19.2%	73	37	8,1%	15,9%
Cauca	9.4%	90	45	2,5%	4,9%
Córdoba	39,0%	107	54	7,3%	14,3%
Guainía					
Guaviare	24,4%	97	49	6,4%	12,5%
Meta	7,6%	166	83	2,3%	4,5%
Nariño	23,8%	591	296	4,2%	8,2%
Norte de Santander	31,9%	33	17	16,6%	32,5%
Putumayo	57.5%	298	149	6,7%	13,1%
Valle del Cauca	40,6%	12	6	5,9%	11,6%
Vichada					

Tabla 3.5: Estimaciones directas por departamento

Para aplicar cualquier técnica de áreas pequeñas es fundamental contar con información auxiliar, cuando se establece un modelo a nivel de unidad como es el caso de este trabajo, es necesario que esta información este a nivel de individuo. El marco muestral se construyó a partir de los registros administrativos del programa, donde se registra la información de los hogares inscritos en este, por lo tanto el marco muestral cuenta con diferentes variables que serán consideradas como la información auxiliar para aplicar la metodología, las variables consideradas son:

- Sexo y edad del jefe de hogar.
- Área de la propiedad de la vivienda sustituida en metros cuadrados.
- Tipo de actividad.
- Año de ingreso al programa.
- Variable indicadora si recibe algún beneficio o ingreso adicional del programa.

Se hace uso del estimador Plug-In basado en los modelos GLMM y MEBO, no se considera el modelo MEBA dados los resultados de la simulación en donde se evidencia el sesgo que introduce este modelo.

Se ajusta el modelo GLMM para la totalidad de la muestra y dado que esta técnica sigue ciertos supuestos, se valida el supuesto distribucional de los efectos aleatorios μ_i , donde la prueba de normalidad Shapiro-Wilk da un valor p de 0.98, concluyendo que los efectos aleatorios se ajustan a la distribución normal.

Grafico QQ-Plot Efectos aleatorios

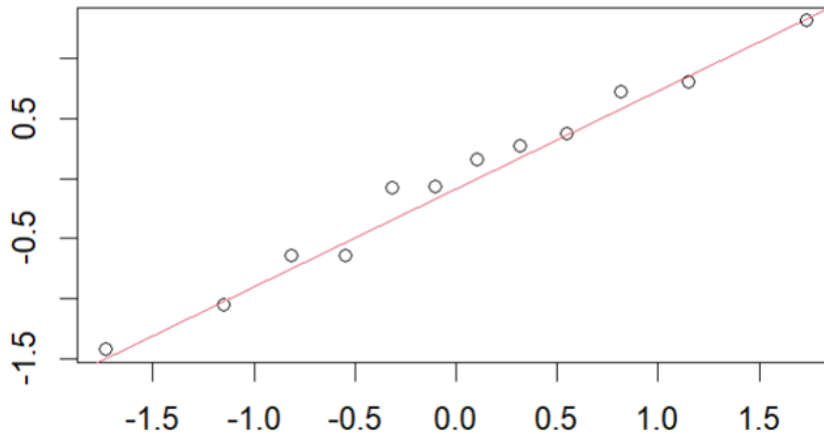


Figura 3.10: QQ-plot de los efectos aleatorio del modelo GLMM

Fuente: elaboración propia.

Para ajustar el modelo MEBO se hace una partición de la muestra creando una muestra de entrenamiento correspondiente al 80 % y una muestra de validación correspondiente al 20 % restante. Dentro de las medidas de ajuste para evaluar el desempeño de los modelos se considera:

- Accuracy: tasa correcta de clasificación.
- Curva ROC: es un gráfico que presenta la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte.
- AUC: es el área bajo la curva ROC y representa que tan bueno es el modelo para clasificar a los individuos según la variable respuesta (éxito o no) a lo largo de todo el rango de puntos de corte posibles.

Respecto a la tasa correcta de clasificación, accuracy, el modelo MEBO es ligeramente mejor que el modelo GLMM. La curva ROC y el valor AUC reflejan que el rendimiento de los modelos es practicamente igual, por los valores de AUC y el valor donde la curva alcanza su punto de corte (donde la sensibilidad y especificidad alcanzan el punto mas alto). También se presenta el histograma de las probabilidades predichas por los modelo, mostrando una similitud en la distribución de estas con pequeñas diferencias en el intervalo que toman.

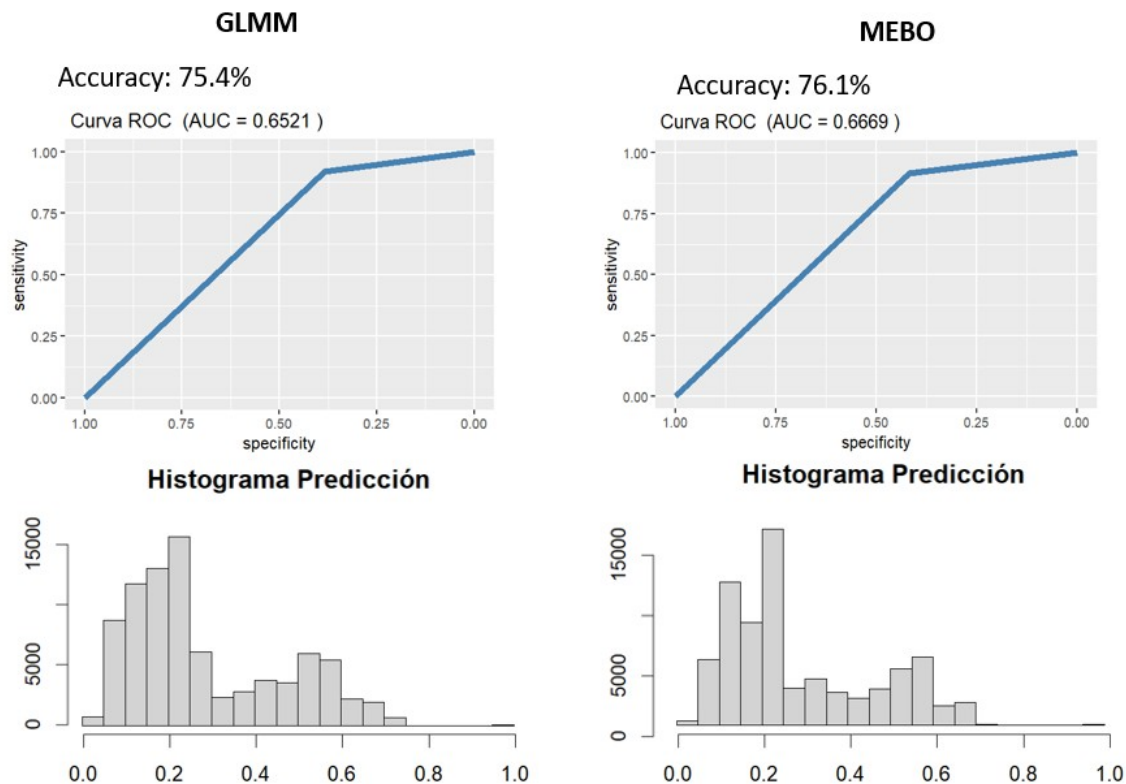


Figura 3.11: Comparación medidas de ajuste
Fuente: elaboración propia.

Después de validar los supuestos requeridos en cada modelo y el rendimiento predictivo de estos, se calcula el estimador Plug-In usando los dos modelos ajustados, estas estimaciones se comparan con las estimaciones directas. En la figura 3.12 se presenta el mapa de calor con la estimación del indicador de éxito del programa a nivel de departamento, en el caso de las estimaciones directas solo se grafican las que cumplen con los criterios de calidad, en el caso de las estimaciones basadas en los modelos, se calculan las que tiene un ECM estimado menor al 5%.

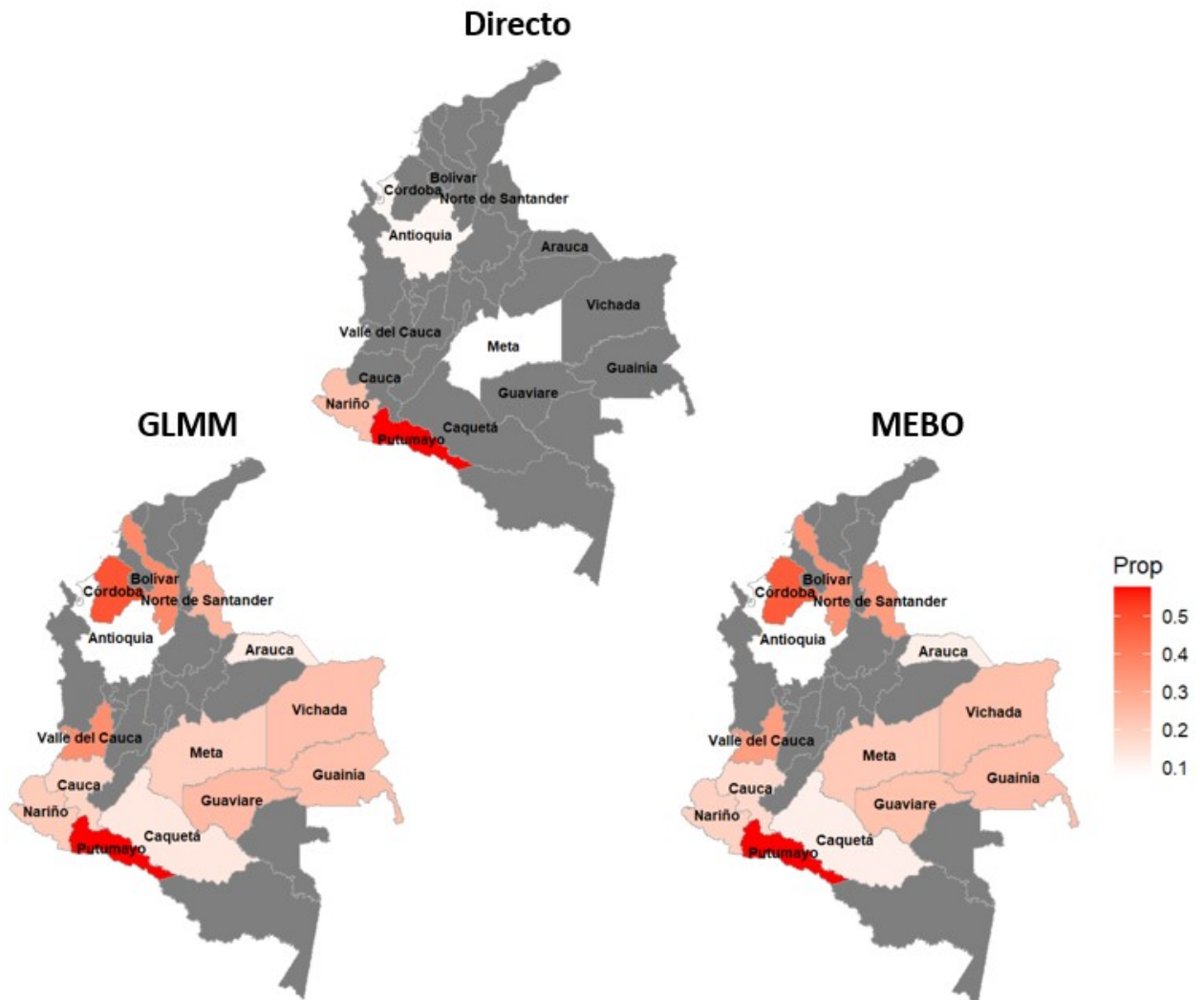


Figura 3.12: Comparación de las estimaciones
Fuente: elaboración propia.

En la tabla 3.6 se presentan los valores de las estimaciones del indicador de éxito, en rojo se resaltan los valores que no cumplen los criterios de calidad y estos solo se toman de referencia para comparar las estimaciones resultantes de cada uno de los modelos, donde se ve una similitud en estas y también comparándolas con las estimaciones directas.

Departamento	Estimación (%)			10000*ECM	
	Directo	GLMM	MEBO	GLMM	MEBO
Antioquia	9,66	8,61	9,92	1,22	0,08
Arauca	0,00	13,22	13,91	155,21	193,18
Bolívar	58,40	37,40	35,58	359,69	423,89
Caquetá	19,22	14,49	13,78	17,07	24,67
Cauca	9,41	19,25	19,19	103,91	93,82
Córdoba	39,00	48,27	47,39	57,01	43,38
Guainía		24,35	25,62	1,87	1,63
Guaviare	24,38	25,74	24,73	7,33	7,97
Meta	7,58	20,68	22,88	151,56	211,64
Nariño	23,82	21,57	20,90	4,36	7,80
Norte de Santander	31,88	27,27	33,72	3,77	6,74
Putumayo	57,47	56,91	56,90	4,30	7,07
Valle del Cauca	40,60	36,36	33,44	42,14	103,81
Vichada		24,18	25,14	4,62	3,63

Tabla 3.6: Comparación estimaciones

En la Figura 3.13 se presenta la comparación del error cuadrático medio (ECM) para el estimador Plug-In basado en los dos modelos. La mediana indica un mejor desempeño del modelo GLMM en términos de error, aunque se observan valores atípicos asociados al ECM de la estimación en Bolívar. Es importante destacar que, a diferencia del escenario de simulación, en la aplicación práctica se evidencian diferencias más marcadas en las estimaciones y en la precisión de los dos modelos, lo que permite analizar su comportamiento en un contexto donde la muestra no fue generada bajo un modelo.

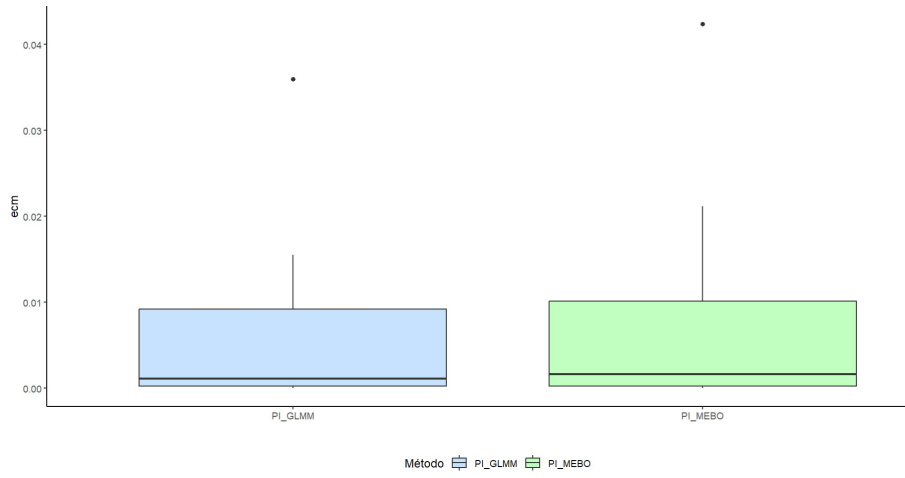


Figura 3.13: Comparación error cuadrático medio
Fuente: elaboración propia.

4. Conclusiones y futuros trabajos

Este trabajo explora el potencial de los métodos de aprendizaje automático basados en árboles para estimar proporciones en áreas pequeñas, las conclusiones se establecen según los resultados del ejercicio de simulación y la aplicación en datos reales.

- Si bien existe literatura sobre modelos de aprendizaje automático con efectos mixtos y enfoques no paramétricos que incorporan la estructura jerárquica de los datos, las aplicaciones de estos modelos en la estimación en áreas pequeñas son aún limitadas, lo que representa una oportunidad para futuras investigaciones.
- Los modelos basados en árboles para clasificación presentan varias ventajas frente al modelo lineal logístico. En particular, no requieren asumir linealidad ni una distribución específica para los errores, pueden detectar interacciones de alto orden entre las covariables, realizan una selección implícita de variables y son robustos ante datos atípicos y valores faltantes.
- Con el modelo MEBO, basado en el método de ensamblaje Boosting, se tiene un sesgo y un error cuadrático medio menor que los otros cuatro estimadores, sin embargo, siguen siendo muy similares a los del modelo logístico mixto GLMM cuando los datos provienen de un modelo.
- Aunque los modelos GLMM y MEBO son similares, este último puede considerarse como una alternativa no paramétrica en caso de fallas del modelo (por ejemplo, brindando seguro contra especificaciones erróneas del modelo, selección de variables válidas y manejo efectivo de valores atípicos).
- Es importante resaltar que en el ejercicio de aplicación se ven diferencias en cuanto a las estimaciones y el error cuadrático medio de los modelos GLMM y MEBO,

indicando que cuando los datos no provienen de una distribución basada en un modelo, el estimador Plug-In basado en el modelo MEBO podría llegar a tener mejores resultados.

- El método de ensamblaje Bagging aunque disminuye la varianza de las estimaciones no es una buena alternativa por el sesgo que introduce, a diferencia del método Boosting que no solo disminuye la varianza sino que dada su estructura secuencial disminuye el error en las estimaciones y por lo tanto tiene un sesgo menor.
- Este trabajo sirve como inspiración para considerar la exploración de diferentes tipos de modelos como máquinas de soporte vectorial, árboles de regresión aditiva bayesiana, otros modelos de bagging o boosting adicionales a los considerados en este trabajo y muchos más que permitan conservar la estructura jerárquica de los datos y por lo tanto ser aplicados en áreas pequeñas.

Finalmente se resalta la aplicación con datos reales, en la que se obtuvo un modelo de clasificación para la variable indicadora del éxito del Programa y con este se pasó a una inferencia basada en el modelo que permitió obtener información a nivel de departamento, estos resultados ayudan a comprender los factores que afectan el éxito del programa y brindan información que puede ser relevante para priorizar las áreas rurales y más vulnerables.

Referencias

Anderson, W., Guikema, S., Zaitchik, B., & Pan, W. (2014). Methods for estimating population density in data-limited areas: Evaluating regression and tree-based models in Peru. *PloS one*, 9(7), e100037.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Ann. Statist.* 47 (2) 1148 - 1178, April 2019. <https://doi.org/10.1214/18-AOS1709>.

Avila, J.L; Huerta, M; Leiva, V; Riquelme, M. & Trujillo L. (2020). The Fay-Herriot model in small area estimation: EM algorithm and application to official data. *REVSTAT – Statistical Journal*, 18(5), 613-635.

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.

Bilton, P., Jones, G., Ganesh, S., & Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, 115, 53-66.

Capitaine, L., Genuer, R., & Thiébaud, R. (2021). Random forests for high-dimensional longitudinal data. *Statistical methods in medical research*, 30(1), 166-184.

Casas-Cordero, C; Encina, J & Lahiri P. (2016). Poverty mapping for the Chilean comunas. In: Pratesi, M. (ed.) “Analysis of Poverty Data by Small Area Estimation”, volume 20, pages 379-404, Wiley, Chichester, UK.

Chandra, H; Kumar, S & Aditya, K. (2018). Small area estimation of proportions with different levels of auxiliary data. *Biometrical Journal*, 60(2), 395-415.

Comisión Económica para América Latina y el Caribe (CEPAL), Diseño y análisis estadístico de las encuestas de hogares de América Latina, Metodologías de la CEPAL, N°5 (LC/PUB.2023/14-P), Santiago, 2023.

Correa, L., Molina, I. y Rao, J.N.K., (2012). Comparison of methods for estimation of poverty indicators in small areas. Unpublished report.

Dagdoug, M., Goga, C., & Haziza, D. (2023). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542), 1234-1251.

De Moliner, A., & Goga, C. (2018). Sample-based estimation of mean electricity consumption curves for small domains. *Survey Methodology*, 44(2), 193-215.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Diallo, M. S., & Rao, J. N. K. (2018). Small Area Estimation of Complex Parameters Under Unit-Level Models with Skew-Normal Errors. *Scandinavian Journal of Statistics*.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. y Santamaría, L. (2008). Bootstrap Mean Squared Error of a Small-Area EBLUP, *Journal of Statistical Computation and Simulation*, 75, 443–462.

Gutiérrez, H. A. (2009). Estrategias de muestreo: Diseño de encuestas y estimación de parámetros. Facultad de Estadística, Universidad Santo Tomás.

Fay III, 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(366a), 269-277.

Fokkema M, Edbrooke-Childs J & Wolpert M (2021). “Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data.” *Psychotherapy Research*, 31(3), 329-341.

Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H (2018). “Detecting Treatment-Subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees.” *Behavior Research Methods*, *50*, 2016-2034. doi:10.3758/s13428-017-0971-x <<https://doi.org/10.3758/s13428-017-0971-x>>.

Hall, P., & Maiti, T. (2006). On Parametric Bootstrap Methods for Small Area Prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2), 221-238.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters*, 81(4), 451-459.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6), 1313-1328.

Hajjem, A., Larocque, D., & Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126, 114-118.

Jiang, J., & Rao, J. S. (2020). Robust Small Area Estimation: An Overview. *Annual Review of Statistics and its Application*, 7(1), 337-360.

Krennmair, P., & Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. (Working Paper).

Krennmair, P., Würz, N., Schmid, T. (2022). Tree-Based Machine Learning in Small Area Estimation. *The Survey Statistician*, 2022, Vol. 86, 22–31.

Krennmair, P., Würz, N., & Schmid, T. (2022). "Analysing opportunity cost of care work using mixed effects random forests under aggregated census data. (Working Paper).

Lahiri, P., & Pramanik, S. (2019). Evaluation of synthetic small-area estimators using design-based methods. *Austrian Journal of Statistics*, 48(4), 43-57.

Li, H. & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

Marchetti, S; Giusti, C; Pratesi M; Salvati, N; Giannotti, F; Pedreschi, D; Rinzivillo, R; Pappalardo, L & Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31, 263-281.

Molina, I. (2019), "Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas", *Series Estudios Estadísticos*, No 97, (LC/TS.2018/82/Rev.1), Santiago, Comisión Económica para América Latina y el Caribe, (CEPAL).

Molina, I. & Marhuenda, Y. (2015), *sae: An R Package for Small Area Estimation*, *The R Journal*, 7, 81–98.

Molina, I., & Strzalkowska-Kominiak, E. (2020). Estimation of proportions in small areas: application to the labour force using the Swiss Census Structural Survey. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(1), 281-310.

Molina, I., & Rao, J. N. (2010). Small area estimation of poverty indicators. *Canadian Journal of statistics*, 38(3), 369-385.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler and Benjamin M. Bolker (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2), 378-400. doi: 10.32614/RJ-2017-066.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18 (1), 90–120.

Parker, P; Janicki, R & Scott H. (2023). Comparison of unit-level small area estimation modeling approaches for survey data under informative sampling. *Journal of Survey Statistics and Methodology*, 11(4), 858-872.

Prasad, N.G.N. y Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators, *Journal of the American Statistical Association*, 85, 163–171.

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22 (1), 300 – 325.

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.

Seibold, H., Hothorn, T., & Zeileis, A. (2019). Generalised linear model trees with global additive effects. *Advances in Data Analysis and Classification*, 13(3), 703-725.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86, 169-207.

Sigrist, F., Gyger, T., Kuendig, P. (2021). “gpboost: Combining Tree-Boosting with Gaussian Process and Mixed Effects Models.” R package version 1.5.1, <<https://github.com/fabsig/GPBoost>>.

Sigrist F. (2022) Gaussian process boosting, *Journal of Machine Learning Research*, 23, 1-46

Särndal, C. (1992) *Model Assisted Survey Sampling*, Springer.

Tellez, C; Rico, I; Guerrero, S & Trujillo L. (2021). Estimation of educational esta-

blishments performance in Saber 5o tests in Colombia. An approach from small area estimation. BEIO – Boletín de Estadística e Investigación Operativa, volume 37(3), 169-182.

Tellez, C; Trujillo, L; Pedraza, A.F. (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, volumen 13(2), 62-78.

Tellez, C; Trujillo, L; Sosa, J.C; Gutiérrez, A. (2024). Small area estimation using multiple imputation in three-parameter logistic models. Chilean Journal of Statistics, volume 15(1), 1-26.

Unión Temporal IPSOS-Uniandes & Departamento Nacional de Planeación - DNP - (2023). Evaluación Institucional y de Resultados del Programa Nacional Integral de Sustitución de Cultivos Ilícitos (PNIS) en el marco de la política integral de drogas del estado colombiano. <<https://anda.dnp.gov.co/index.php/catalog/165/study-description>>.