UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Noviembre 2024

Estimación del porcentaje de afiliación a Seguridad Social en Salud en Bogotá, utilizando modelos de estimación en áreas pequeñas con información auxiliar medida con error

Estimation of the Health Social Security Affiliation Percentage in Bogotá Using Small Area Estimation Models with Measurement Error in Auxiliary Information

Paola Martínez[a]
paolamartinezs@usantotomas.edu.co

Andrés Felipe Ortiz[b]
afortizr@gmail.com

Resumen

La estimación en áreas pequeñas es una metodología utilizada para calcular estimaciones confiables de una característica de interés haciendo uso de información auxiliar que permite, a través de un modelo, mejorar las estimaciones directas realizadas. Dado que en muchos casos esta información auxiliar contiene un error de medición, se hace necesario incluir este componente en el modelo planteado. La aplicación se realiza con la Encuesta Multipropósito de Bogotá - EM (2021), de tal manera que se pueda obtener información relevante y confiable sobre el porcentaje de afiliación a la seguridad social en salud en las diferentes etapas de vida, que permita identificar grupos de población que puedan tener necesidades específicas de atención médica.

Estimar el porcentaje de afiliación a la seguridad social en salud es crucial para la toma de decisiones informadas y la formulación de políticas públicas en Colombia, especialmente en el contexto de los Objetivos de Desarrollo Sostenible (ODS), como lo son promover la salud y el bienestar, reducir las desigualdades y garantizar la cobertura universal de salud. La metodología propuesta permite hacer comparaciones entre las estimaciones directas y aquellas que arroje el modelo, para determinar cuáles de ellas son las más precisas y útiles. Lo anterior no solo mejora la calidad de los datos disponibles, sino que también fortalece la capacidad del Distrito para responder a las necesidades de su población de manera más efectiva y equitativa.

***Palabras clave*:** estimación en áreas pequeñas, modelos mixtos, error de medición, seguridad social en salud, ciclos de vida, EBLUP.

1. Introducción

El acceso universal a los sistemas de salud es un objetivo fundamental para asegurar el bienestar y la equidad social en cualquier país. En Colombia, como en muchos otros lugares del mundo, la estimación precisa del porcentaje de personas afiliadas a los sistemas de salud juega un papel importante para la formulación de políticas públicas efectivas. Esta necesidad es aún más urgente cuando se considera la población no afiliada, quienes enfrentan mayores barreras para acceder a servicios de salud esenciales.

En el marco de los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, específicamente el ODS 3 de salud y bienestar, se evidencia la importancia de lograr la cobertura sanitaria universal,

[a]Estudiante
[b]Director

garantizando que todas las personas puedan acceder a servicios de salud de calidad sin sufrir dificultades financieras. La medición de la afiliación a los sistemas de salud permite monitorear los avances hacia este objetivo global y diseñar intervenciones que cierren las brechas existentes.

El Departamento Administrativo Nacional de Estadísticas- DANE como ente rector del Sistema Estadístico Nacional - SEN, tiene como obligación generar estadísticas oficiales que permitan la formulación de políticas públicas a partir de la ejecución de operaciones estadísticas, entendidas como el conjunto de procesos y actividades que comprenden la identificación de necesiadades, diseño, construcción, recolección o aocpio, procesamiento, análisis, difusión y evaluación, el cual conduce a la producción de información estadística sobre un tema de interés nacional y/o territorial. DANE (2019)

Estas operaciones estadísticas se realizan utilizando diferentes metodologías de recolección, dentro de las cuales se encuentra el muestreo probabilístico; en muchos casos las estimaciones generadas a partir de estas operaciones estadísticas por muestreo probabilístico no son confiables, por lo que en ciertos niveles de desagregación, con tamaños de muestras pequeños, no es posible sacar conclusiones relevantes para el Distrito, que permitan, por ejemplo, plantear políticas públicas en salud que se ajusten a las necesidades de cada grupo poblacional, pues los coeficientes de variación asociados superan el 15 %.

La metodología de estimación en áreas pequeñas (SAE) tiene como objetivo solucionar la limitación mencionada anteriormente, y plantea una serie de modelos que incluyen información auxiliar correlacionada con la característica de interés, que permiten construir estimaciones que mejoran en comparación con las calculadas por el método directo y por lo tanto se consideran más confiables para la toma de desiciones.

El Departamento Administrativo Nacional de Estadísticas (DANE) en colaboración con la Secretaría Distrital de Planeación (SDP), ha desarrollado la Encuesta Multipropósito de Bogotá como una operación estadística relevante para comprender las necesidades y condiciones de la población en Bogotá, dentro de los cuales se encuentra el acceso a los sistemas de salud en la ciudad.

El objetivo, entonces, se centra en el cálculo de estimadores tomando como áreas pequeñas las localidades y los ciclos de vida, de tal manera que se elijan las mejores estimaciones a partir del error cuadrático medio estimado, mejorando las estimaciones que actualmente genera el DANE de manera directa, pero teniendo en cuenta que en muchos casos la información auxiliar que esta relacionada con la variable de interés proviene de otras operaciones estadísticas por muestreo probabilístico y se hace necesario incluir dentro del modelo el error con que es medida dicha información. Es por esto que se hace una revisión de literatura que inicia, sin lugar a dudas, desde el planteamiento de estos modelos con Fay & Herriot (1979) y continua con la incorporación del error de muestreo en el modelo. También es importante hacer un contexto sobre la información que se va a utilizar, en términos de sus características principales, así como la validación de la presencia de las áreas definidas para el alcance de este estudio que en este caso serán las localidades y los ciclos de vida.

2. Marco teórico

Actualmente el país cuenta con información relevante producida por el Sistema Estadítico Nacional - SEN, que permite a las entidades pertinentes implementar políticas públicas con los resultados. Las estadísticas oficiales son producidas por operaciones estadísticas de diversos tipos: a partir de resgistros administrativos, por censo, por muestreo probabilístico y no probabilístico.

La información estadística que proviene de muestreos probabilísticos en muchas ocasiones no alcanza a ser de calidad para niveles de desagreación bajos, en donde los tamaños de muestra son pequeños; por esto, los coeficientes de variación no son aceptables y por lo tanto, tampoco son confiables las estimaciónes en dichos dominios. Sin embargo, la implementación de métodos como la estimación en áreas pequeñas, por parte de las entidades públicas ha avanzado de manera gradual, aunque a un ritmo lento, limitándose únicamente a las estimaciones directas a partir de un diseño muestral establecido, tal como se menciona en Rico et al. (2024)

Surge entonces la motivación de entender ¿cómo se definen estos modelos?, ¿cuáles son sus supuestos?, ¿cuál es la mejor información auxiliar a utilizar?, ¿cuáles son los estimadores más eficientes?, entre otras incógnitas.

Fay & Herriot (1979) son quizás los primeros en abordar este problema y encontrar una manera de mejorar las estimaciones a partir de información auxiliar, que se conoce como estimación en áreas pequeñas (SAE, por sus siglas en inglés).

Para que el Departamento del Tesoro administrara los fondos a los diferentes estados en Estados Unidos, se requería de las estimaciones del ingreso per-cápita para cada uno. Dado que en muchos casos la muestra era muy pequeña para cada estado y las estimaciones tenían errores muestrales muy grandes y no eran confiables, se realizaba una sustitución del valor estimado por el promedio de la estimación de los condados.

La metodología propuesta por Fay & Herriot (1979) permitió hacer una estimación del índice per-cápita para las áreas pequeñas utilizando un modelo con información auxiliar correspondiente a los datos censales (modelo de FH), lo cual permitió concluir que dichas estimaciones tienen un error promedio menor que las estimaciones directas realizadas a partir de un diseño muestral.

Se estableció un estimador en el que se ajustaba una ecuación de regresión a las estimaciones de la muestra utilizando información auxiliar en las variables independientes (valores del condado, datos de la declaración de impuestos y datos sobre viviendas del censo de 1970). Además, realizaron una evaluación del estimador resultante y midieron la bondad de ajuste entre la ecuación de regresión y los datos de la muestra, mostrando resultados satisfactorios para esta nueva metodología.

Posterior a esta nueva posibilidad dentro del campo del muestreo, de construir mejores estimaciones a partir de información censal, diferentes investigadores realizan trabajos sobre el tema de SAE. Uno de ellos es Pfeffermann (2002) quien realiza una revisión de los principales avances en este tema y presenta los los desarrollos recientes de la época. El repaso se concentra en los métodos dependientes del modelo, con énfasis especial en puntos de predicción de las cantidades del área objetivo y una evaluación de los errores cuadráticos. Así mismo, se examinan las posibles ganancias de modelar la correlación entre los efectos aleatorios de áreas pequeñas que se utilizan para representar la variación no completada de las cantidades objetivo de áreas pequeñas, para una revisión y una evaluación de los métodos anteriores utilizados para SAE.

Uno de los aspectos importantes, relevantes para el objetivo de este trabajo es responder a la pregunta ¿Utilizar información auxiliar con error de medición ayuda a mejorar las estimaciones?. Es importante resaltar que en muchos casos no se cuenta con información censal o recolectada por registros administrativos y que, como se mencinó anteriormente, mucha de la información estadística producida en nuestro país proviene de operaciones estadísticas por muestreo probabilístico. Teniendo en cuenta lo anterior, es importante garantizar que las estimaciones realizadas por métodos SAE que utilizan información con error muestrales y no muestral, son confiables.

Ybarra & Lohr (2008) introducen un modelo de FH con error de medición. Este se asumió con error de medición funcional donde los valores verdaderos de las covariables son cantidades fijas pero desconocidas. Entonces, se introdujo un nuevo estimador de área pequeña que tiene en cuenta la variabilidad muestral en la información auxiliar mostrando que es aproximadamente insesgado. En el desarrollo del estimador se tuvo en cuenta dentro del modelo el error que proviene del diseño muestral, y presentan la estimación de los parámetros de regresión y el cálculo del error cuadrático medio. A su vez, el modelo propuesto por Burgard et al. (2020) amplía el marco de Fay-Herriot para abordar la incorporación del error de medición en las variables auxiliares.

Este estimador se aplicó para predecir características medidas en la Encuesta nacional de examen de salud y nutrición de EE. UU., con información auxiliar proveniente de la Encuesta Nacional de Entrevistas de Salud de EE.UU, que esta correlacionada positivamente con las estimaciones de la primera encuesta. Investigaciones anteriores concluyen que si la información esta correlacionada, se mejoran las estimaciones en áreas pequeñas, tal como se menciona en Lohr & Prasad (2003).

Uno de los trabajos más recientes sobre la temática del trabajo es presentado por Burgard et al. (2020) en el que desarrollan un modelo de FH en tres etapas. En la primera etapa se establece el modelo de muestreo, en la segunda el modelo de enlace y en la tercera el modelo de error de medición. Al hacer los desarrollos correspondientes, se presenta una ampliación del modelo de FH con el error aleatorio de las variables auxiliares distribuido normal multivariado. Además presentan la aplicación para estimar proporciones de pobreza para la Encuesta de Condicionales de vida de España utilizando como información auxiliar, la proveniente de la Encuesta de Población Activa.

2.1. Modelo de Fay & Herriot

Inicialmente el modelo propuesto por Fay & Herriot (1979) se podía escribir de la siguiente manera, como una ecuación asociada a un modelo mixto, que incluye efectos fijos y aleatorios:

$$\hat{y_i} = {x_i}^T \beta + u_i + e_i \tag{1}$$

Donde $\hat{y_i}$ es la estimación directa de la característica de interés, x_i es la matriz de covariables o información auxiliar, β es el vector de coeficientes del modelo, u_i es el efecto aleatorio o error de la regresión que no esta asociado a las variables auxiliares y e_i es el error de muestreo; lo anterior para cada área $i = 1...m$. Además $u_i \sim (0, \sigma_u^2)$ y $e_i \sim (0, \psi_i)$.

Entonces el mejor predictor lineal insesgado empírico (EBLUP) bajo el modelo propuesto por FH es:

$$\hat{y}_i^{FH} = \hat{\gamma_i}\hat{y_i} + (1 - \hat{\gamma_i}) x_i^T \hat{\beta} \tag{2}$$

Donde $\hat{\gamma_i} = (\hat{\sigma_u}^2/(\hat{\sigma_u}^2 + \hat{\psi}_i))$ y $\hat{\beta} = \left(\sum_{i=1}^m \hat{\gamma}_i x_i x_i^\top\right)^{-1} (\sum_{i=1}^m \hat{\gamma}_i x_i \hat{y}_i)$, el cual corresponde a la estimación de los coeficientes del modelo a través de métodos como el de máxima verosimilitud.

Adicionalmente, el Error Cuadrático Medio (ECM) del predictor de Fay-Herriot, según Rico et al. (2024), se define como:

$$\widehat{\mathrm{ECM}}(\hat{y}_i^{FH}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \tag{3}$$

donde:

$$g_{1i}(\hat{\sigma}_u^2) = \hat{\gamma}_i \hat{\psi}_i,$$

$$g_{2i}(\hat{\sigma}_u^2) = (1 - \hat{\gamma}_i)^2 x_i^\top \left(\sum_{i=1}^m (\hat{\sigma}_u^2 + \hat{\psi}_i)^{-1} x_i x_i^\top \right)^{-1} x_i,$$

$$g_{3i}(\hat{\sigma}_u^2) = (1 - \hat{\gamma}_i)^2 (\hat{\sigma}_u^2 + \hat{\psi}_i)^{-1} \overline{\mathrm{Var}}(\hat{\sigma}_u^2),$$

siendo $\overline{\mathrm{Var}}(\hat{\sigma}_u^2)$ la varianza asintótica del estimador $\hat{\sigma}_u^2$.

2.2. Variables auxiliares con error de medición

Dado que el modelo FH no incluía el error de medición para las variables auxiliares, Ybarra & Lohr (2008) proponen un estimador que tiene en cuenta los errores de medición (EM) en las covariables $\hat{x}_i$ estimadas:

$$\hat{y}_i^{EM} = \hat{\gamma_i}\hat{y_i} + (1 - \hat{\gamma_i}) \hat{x}_i^T \hat{\beta} \tag{4}$$

Donde $\hat{\gamma}_i = \frac{\hat{\sigma_v^2}+\hat{\beta^T} C_i \hat{\beta}}{\hat{\sigma_v^2}+\hat{\beta^T} C_i \hat{\beta} \psi_i}$, $\hat{\beta} = \left(\sum_{i=1}^{m} \frac{\hat{x}_i \hat{x}_i^T}{\hat{\sigma}_u^2+\hat{\psi}_i+\hat{\beta}^T C_i \hat{\beta}}\right)^{-1} \left(\sum_{i=1}^{m} \frac{\hat{x}_i \hat{y}_i}{\hat{\sigma}_u^2+\hat{\psi}_i+\hat{\beta}^T C_i \hat{\beta}}\right)$ y $C_i = \text{Var}(\nu_i)$, que corresponde a la matriz de covarianzas del error de medición de las covariables $\hat{x}_i$, siendo ν_i el vector de errores de medición.

El Error Cuadrático Medio (ECM) del predictor ajustado $\hat{y}_i^{EM}$ se expresa como:

$$\widehat{\text{ECM}}(\hat{y}_i^{EM}) = g_{1i} + g_{2i} + g_{3i}, \tag{5}$$

donde:

$$g_{1i} = \hat{\gamma}_i^2 \hat{\psi}_i,$$

$$g_{2i} = (1-\hat{\gamma}_i)^2 \hat{x}_i^\top \left(\sum_{j=1}^{m} (\hat{\sigma}_u^2 + \hat{\psi}_j + \hat{\beta}^T C_j \hat{\beta})^{-1} \hat{x}_j \hat{x}_j^T\right)^{-1} \hat{x}_i,$$

$$g_{3i} = (1-\hat{\gamma}_i)^2 (\hat{\sigma}_u^2 + \hat{\psi}_i + \hat{\beta}^T C_i \hat{\beta})^{-1} \text{Var}(\hat{\sigma}_u^2),$$

3. Objetivos

3.1. Objetivo general

Estimar bajo la metodología SAE el porcentaje de afiliación al Sistema de Salud de las personas residentes en Bogotá, para cada localidad y por ciclo de vida, contemplando información auxiliar con error de medición.

3.2. Objetivos específicos

- Evaluar la eficiencia de las estimaciones del porcentaje de afiliación a los sistemas de salud en Bogotá basadas en estimación en áreas pequeñas, en comparación con las estimaciones directas generadas por el DANE.
- Analizar cómo el error de medición impacta las estimaciones por metodología SAE.
- Proponer mejoras metodológicas para la medición de afiliación a los sistemas de salud en áreas pequeñas, para ser implementadas por el DANE.

4. Descripción de los datos

Este ejercicio explora el ajuste de un modelo de Fay Herriot con errores de medición en las covariables, propuesto por Ybarra & Lohr (2008) para el porcentaje de afiliados a seguridad social en salud por localidad y por ciclos de vida en Bogotá. Se utilizan los datos de la Encuesta Multipropósito - EM para la variable respuesta y como variable explicativa la *tasa de informalidad*[1]; como esta covariable proviene de una encuesta por muestreo, dado que se toma de la misma Encuesta Multipropósito, entonces se debe considerar dentro del modelo el error de medición asociado.

La elección de la variable auxiliar de la EM, se hace teniendo en cuenta la relación con el porcentaje de afiliación; dado que las dinámicas laborales influyen en el acceso a seguridad social, la población ocupada *informal*[2] puede explicar el comportamiento de la variable de interés, pues esta población no es afiliada

[1] La tasa de informalidad se define como el total de personas informales sobre el total de personas ocupadas

[2] Son todos los asalariados o empleados domésticos que no cuentan con cotizaciones de salud ni a pensión por concepto de su vínculo laboral con el empleador que los contrató

por el empleador ni está obligada a afiliarse a los servicios de salud; además y de acuerdo con J. (2010-09) la estructura del mercado laboral en Colombia tiene un efecto en la probabilidad de que una persona cotice al régimen de salud.

y . Sin embargo, se decide incluir en el modelo la tasa de informalidad, ya que esta tiene una relación más fuerte con la variable de interés, y está construida teniendo en cuenta la población ocupada.

A partir de la revisión de los inventarios de información del país, se identifican dos registros administrativos relacionados con la temática en salud y que se encuentran disponibles para los dominios definidos en este ejercicio (en total 95 dominios, los cuales corresponden a cada una de las 19 localidades por cada uno de los cinco ciclos de vida); uno de ellos contiene la cantidad de atenciones al servicio de salud y el otro, el registro de afiliaciones reportadas por las entidades administradoras de planes de beneficios.

4.1. Encuesta Multipropósito Bogotá - Cundinamarca - 2021

Es una operación estadística por muestreo probabilístico producida por el DANE y la Secretaría Distrital de Planeación - SDP, cuyos resultados son relevantes para la comprensión de las necesidades y condiciones de la población en Bogotá lo que permite la planificación y ejecución efectiva de políticas públicas orientadas al desarrollo social, económico y urbano de la ciudad.

Esta encuesta tiene como objetivo: Obtener información estadística que permita hacer seguimiento a las variables necesarias para el diseño, monitoreo y evaluación de políticas públicas, sobre aspectos sociales, económicos y de entorno urbano de los hogares y habitantes de Bogotá y los municipios seleccionados de Cundinamarca. DANE (2022)

Como características generales de la EM se puede mencionar que su diseño muestral definido es probabilístico, estratificado y de conglomerados, la población objetivo corresponde a la población civil no institucional, residen de la cabecera, centro poblado y rural disperso de Bogotá y presenta resultados de diferentes temáticas como son: demografía y población, educación, empleo y mercado laboral, salud, vivienda y servicios básicos, condiciones socioeconómicas, tecnología y cultura y recreación.

4.2. Información Auxiliar - Registros Administrativos

Dado que incluir información auxiliar al modelo de estimación en áreas pequeñas permite mejorar la precisión de las estimaciones, se toma como covariables la información relacionada con los siguientes registros administrativos y se acopiarán de las fuentes mencionadas:

- Registro Individual de Prestación de Servicios y de Salud - RIPS del Ministerio de Salud y Protección Social: este registro se implementó normativamente por medio de la Resolución 3374 de 2000 y contiene los datos mínimos y básicos que el Sistema de General de Seguridad Social en Salud requiere para los procesos de dirección, regulación y control para las entidades que reportan la información como lo son las instituciones prestadoras de servicios de salud - IPS, Empresas Administradoras de Servicios de Salud - EAPB y prestadores de atenciones particulares. Los RIPS contienen información demográfica del paciente, como edad, sexo y residencia y tienen un cobertura que incluye al Distrito.

- Base de Datos Única de Afiliados - BDUA del Ministerio de Salud y Protección Social: La BDUA es un sistema de información que consolida y administra los datos de los afiliados al Sistema General de Seguridad Social en Salud (SGSSS) en Colombia. Esta base de datos esta regulada por la Resolución 4622 de 2016, que establece los lieamientos para la administración, actualización y uso de la información de afiliados al SGSSS. La BDUA contiene información detallada sobre la afiliación de los individuos, incluyendo datos demográficos como nombre, edad, sexo, y lugar de residencia, así como información sobre la entidad administradora de su plan de beneficios y el

estado de su afiliación. Al ofrecer una cobertura a nivel nacional, incluida Bogotá, la BDUA es una fuente de información auxiliar que permite su uso para el ejercicio de estimación en áreas pequeñas.

4.3. Enfoque diferencial - Ciclo de vida

La inclusión del enfoque diferencial e interseccional en la producción de información, busca identificar y abordar las desigualdades y brechas existentes entre diferentes grupos de población, por lo que se hace necesario y pertinente divulgar información con mayores niveles de desagregación, que promuevan la equidad, la inclusión y el respeto de los derechos humanos de todos los ciudadanos. DANE (2020)

El enfoque diferencial de ciclo de vida parte de reconocer que en cada etapa del ciclo de vida, se tienen diferentes necesidades vitales, expectativas sociales, capacidades y responsabilidades.

A continuación se presentan las definiciones de los ciclos de vida que hacen parte del alcance de este artículo, tomados de la Guía para la Inclusión de Enfoque Diferencial e Interseccional, DANE (2020)

- Adolescencia: comprendida entre los 12 y los 17 años, en esta etapa se consolidan aspectos físicos, mentales, emocionales y relacionales que permiten el desarrollo de capacidades y los aprendizajes necesarios para lograr la autonomía.

- Juventud: comprendida entre los 18 y los 28 años. En esta etapa la persona se encuentra en proceso de consolidación de su autonomía intelectual, física, moral, económica, social y cultural que hace parte de una comunidad política y en ese sentido ejerce su ciudadanía

- Adultez: comprendida entre los 29 y 59 años. Es una etapa de la vida caracterizada, en general, por el ejercicio de la libertad, la capacidad de agencia, la participación en el mercado laboral y las responsabilidades de cuidado a otras personas y se puede dividir en adulto joven de los 29 a los 44 años y adulto maduro de los 45 a los 59 años.

- Adulto/a mayor: aquella persona que cuenta con sesenta (60) años o más.

5. Aplicación

Como se mencionó anteriormente, se estima el porcentaje de afiliación al Sistema de Salud, a partir de tres variables auxiliares: tasa de informalidad, porcentaje de personas registradas en BDUA y porcentaje de personas atendidas en el sistema de salud, presentando los resultados de la aplicación del modelo de Fay Herriot, para diferentes escenarios.

Inicialmente, se presenta la relación de las variables auxiliares con el porcentaje de afiliación:

- Porcentaje de personas registradas en BDUA

  En la figura 1 se puede ver la fuerte relación lineal que existe entre estas dos variables, lo que tiene sentido, ya que uno corresponde al registro administrativo que contiene todas las personas afiliadas y que por obligatoriedad deben ser reportadas y el otro corresponde a la estimación realizada por la EM. El coeficiente de correlación es del 0.82.
  Si bien se tienen porcentajes de personas registradas en BDUA superiores a 100, para varios dominios, esto se debe a que los totales en el BDUA superan los valores proyectados por el DANE para el año 2021.

Figura 1: Porcentaje de afiliación vs Porcentaje BDUA

- Porcentaje de personas registradas en RIPS

  El porcentaje de personas registradas en RIPS no presenta una relación lineal fuerte con el porcentaje de afiliación; esto se debe a que las atenciones en servicios de salud pueden ocurrir en reiteradas ocasiones para la misma persona; además existen factores que influyen en el comportamiento de esta variable como lo son la prevalencia de enfermedades, la demanda de servicios médicos, la cultura de la prevención, entre otros. Adicionalmente, y de acuerdo a las aclaraciones presentadas por la Secretaría Distrital de Salud, este registro administrativo tiene inconsistencias asociadas a la localidad de residencia, pues es la localidad de atención la que se acopia en el registro, y la otra es construida por el equipo de la entidad.

Figura 2: Porcentaje de afiliación vs Porcentaje BDUA

- Tasa de informalidad

  En la figura 3 se pueden identificar los ciclos de vida y la relación de la tasa de informalidad con el porcentaje de afiliación. Si bien el coeficiente de correlación es 0.19, cuando se calcula por ciclo de vida, se observan correlaciones negativas reflejando una relación lineal inversa, pues a mayor tasa de informalidad menor es el porcentaje de afiliación; sin embargo, esta relación no se comporta de la misma manera para el grupo de los adolescentes, ya que ellos tienen las tasas de informalidad más altas, pero pueden estar afiliados al Sistema de Salud como beneficiarios.

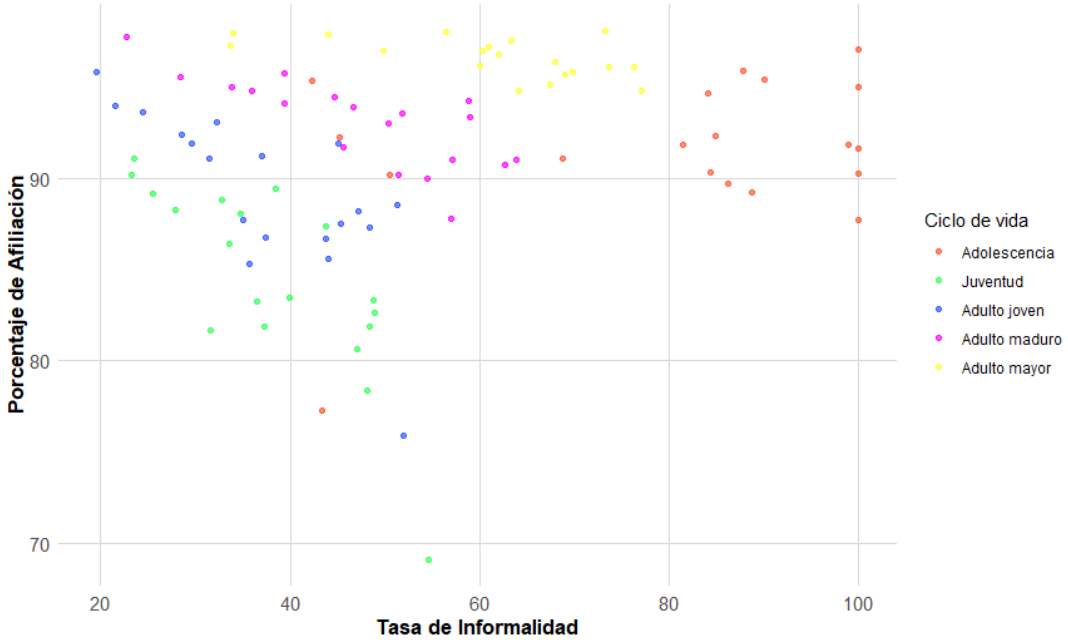

Figura 3: Relación entre la tasa de informalidad y el porcentaje de afiliación por grupo de edad

Para la implementación de los modelos, se utiliza el paquete *SAE* y *saeME*; con el primero, de acuerdo con Molina & Marhuenda (2015), es posible hacer el cálculo del estimador EBLUP y con el segundo se pueden realizar estimaciones en áreas pequeñas utilizando información auxiliar que es medida con error, a través de la implementación del modelo de Fay-Herriot propuesto por Ybarra y Lohr. Ubaidillah & Mubarak (2023) presentan las funciones `FHme` y `mse_FHme` para el cálculo del EBLUP, así como del Error Cuadrático Medio usando el método de Jacknife.

Los modelos que se plantean se presentan en la tabla 1:

Modelo	**Covariables Incluidas**
1	Porcentaje de personas registradas en BDUA Porcentaje de personas registradas en RIPS
2	Porcentaje de personas registradas en BDUA Porcentaje de personas registradas en RIPS Tasa de informalidad
3	Tasa de informalidad
4	Porcentaje de personas registradas en BDUA
5	Porcentaje de personas registradas en BDUA Tasa de informalidad
6	Porcentaje de personas registradas en RIPS
7	Porcentaje de personas registradas en RIPS Tasa de informalidad

Tabla 1: Combinaciones de covariables incluidas en los modelos

En la figura 4 se presentan los coeficientes de variación para la estimación directa, para el EBLUP generado sin tener en cuenta la variable con error de medición (modelo 1) y para el EBLUP que incluye el error de medición en la covariable tasa de informalidad (modelo 2). Se presentará el detalle de los resultados de estos dos modelos, pues las conclusiones para los demás siguen el mismo sentido.

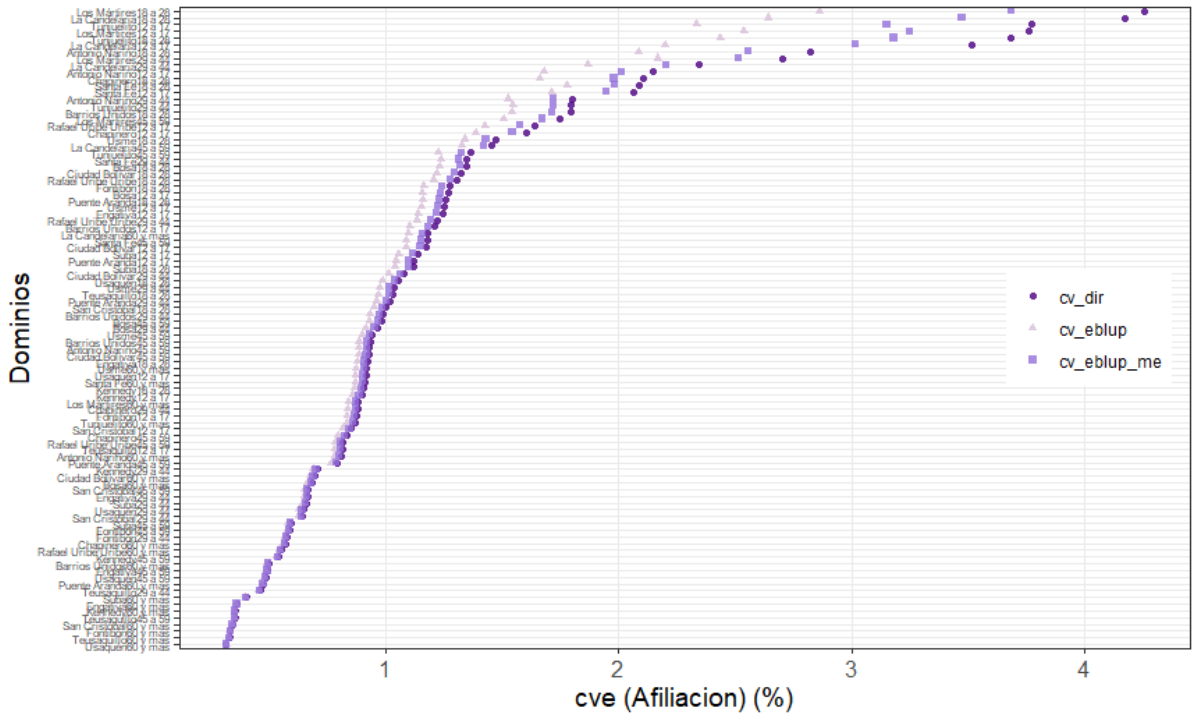

Figura 4: Coeficientes de variación para cada estimador del modelo 1 y modelo 2, para todos los dominios

Para ver el detalle para cada uno de los ciclos de vida, se presentan en las figuras 5, 6, 7, 8 y 9 los mismos coeficientes de variación incluidos en la figura 4.

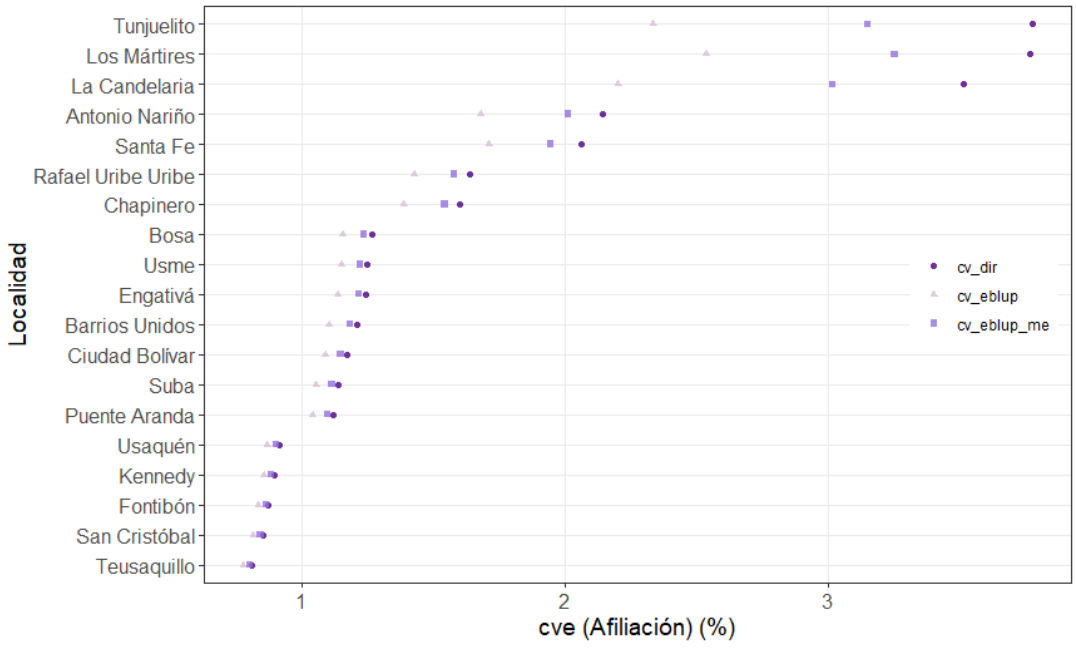

Figura 5: Coeficientes de variación para cada estimador para el ciclo de vida Adolescencia

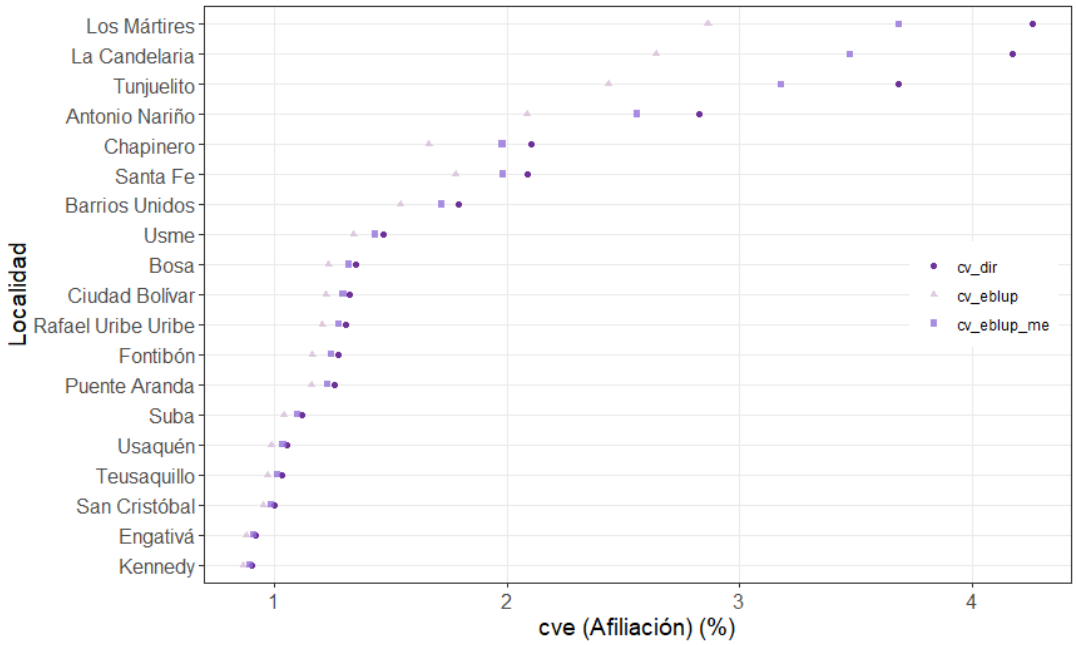

Figura 6: Coeficientes de variación para cada estimador para el ciclo de vida Juventud

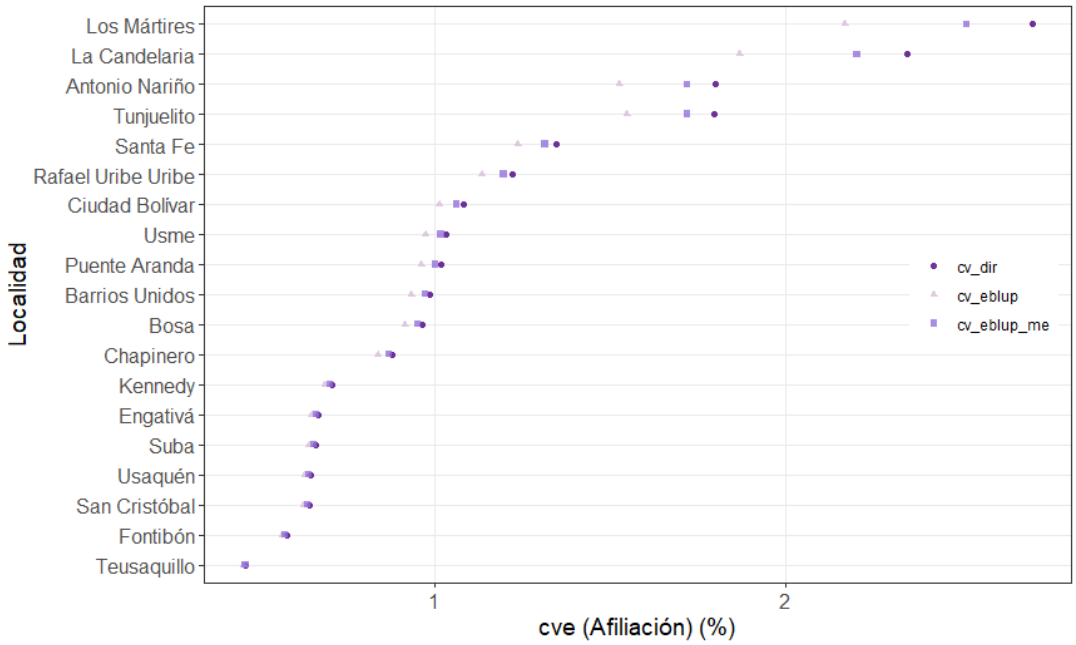

Figura 7: Coeficientes de variación para cada estimador para el ciclo de vida Adulto joven

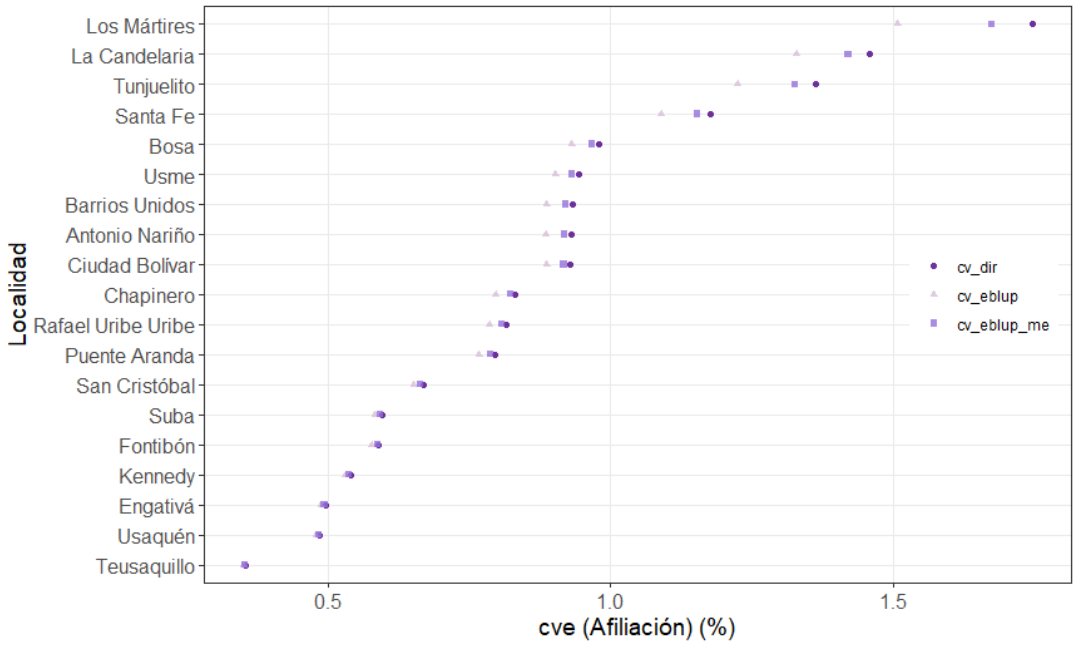

Figura 8: Coeficientes de variación para cada estimador para el ciclo de vida Adulto maduro

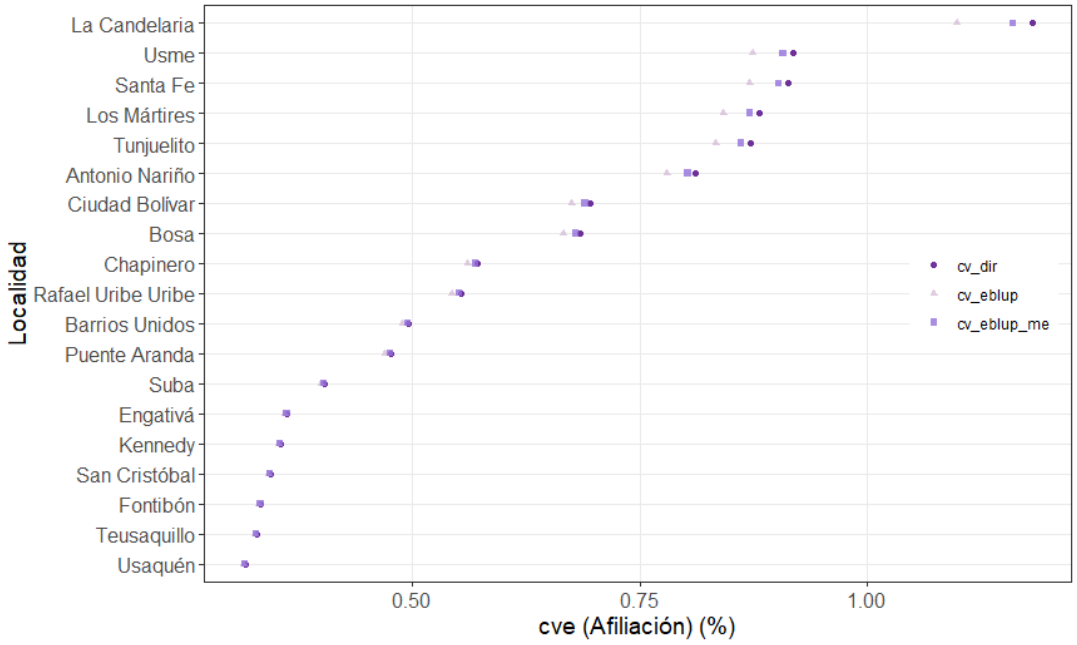

Figura 9: Coeficientes de variación para cada estimador para el ciclo de vida Adulto mayor

De manera general, e independientemente del método de estimación, los coeficientes de variación se hacen más pequeños a medida que se incrementa la edad; además la mejoría de esta medida también es más evidente para la población de adolescentes y jóvenes. Con respecto a las localidades, aquellas con mayores coeficientes de variación se mantienen en todos los ciclos de vida y son Tunjuelito, Los Mártires, La candelaria y Santa Fe.

Para todos los dominios definidos, es mejor el coeficiente de variación asociado al EBLUP resultante del modelo que no incluye la variable tasa de informalidad (modelo 1); sin embargo, también en todos los casos el coeficiente de variación de la estimación directa es el peor que aquellos asociados al modelo 2 que incluye la variable medida con error; lo anterior es de esperarse, ya que cuando se ajusta el modelo con error de medición, se debe agregar ese error al ECM a calcular por dominio, lo que implica un incremento inevitable en el error total.

Específicamente para los adolescentes y jóvenes se puede observar cómo la estimación del porcentaje de afiliación en salud cambia para las localidades más pequeñas y con cv's más altos, tal como lo muestra la figura 10

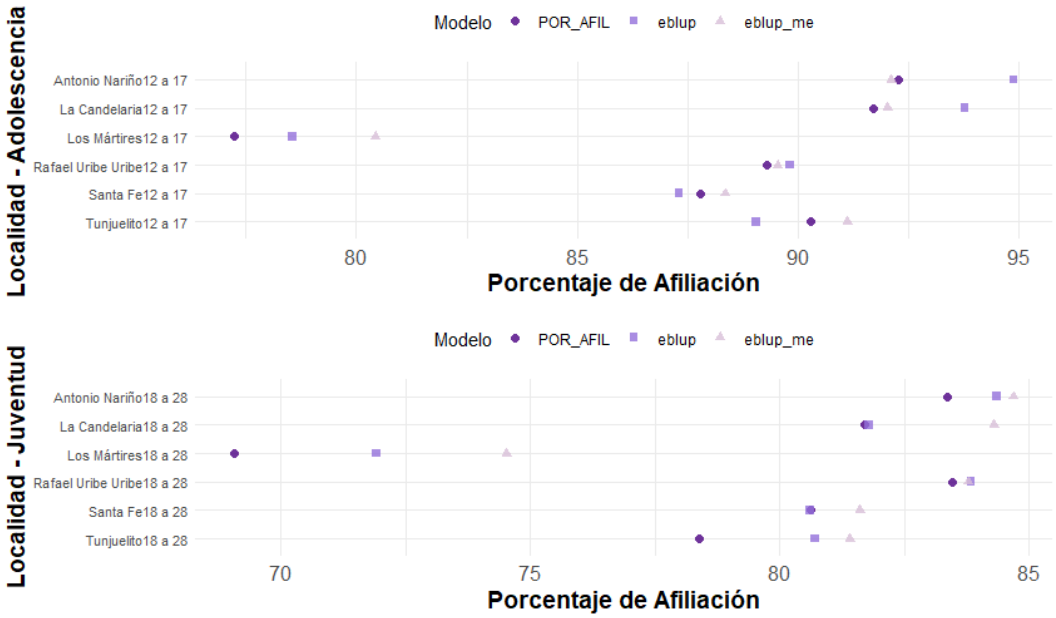

Figura 10: Estimaciones de porcentaje de afiliación para adolescentes y jóvenes

Con respecto a los demás modelos, se mantiene el mismo comportamiento, es decir, siguen mejorando los coeficientes de variación generados con cualquiera de los modelos planteados, con respecto al calculado por el método directo.

6. Conclusiones

Actualmente, las entidades del SEN tienen un amplio inventario de registros administrativos que son aprovechados para la generación de información. Sin embargo, en muchos casos, no se ha realizado un diagnóstico de calidad, lo que dificulta determinar la confiabilidad de esta información. Además, la falta de documentación asociada impide que el usuario entienda las características de los datos acopiados. Aunque el país debe continuar avanzando en el fortalecimiento de registros administrativos, mientras esto sucede, se dispone de métodos que permiten mejorar las estimaciones haciendo uso de información auxiliar, como la proveniente de encuestas. A pesar de que estas encuestas pueden incluir un error de medición, su incorporación ofrece una solución efectiva a las limitaciones antes mencionadas, siempre que se consideren las implicaciones metodológicas asociadas.

Este tipo de ejercicios permiten identificar localidades y ciclos de vida específicos que merecen especial atención en los procesos de muestreo, dado que suelen ser las que presentan los coeficientes de variación

más bajos. Por ejemplo, la estimación de los porcentajes de afiliación muestra variaciones notables en localidades como Santa Fe, Los Mártires y La Candelaria a lo largo de los diferentes ciclos de vida. Estas diferencias pueden representar un cambio importante en la definición de políticas públicas orientadas a estas poblaciones, fortaleciendo así la toma de decisiones para la implementación de planes y programas.

Los resultados obtenidos evidenciaron que, para todos los dominios analizados, el coeficiente de variación asociado al estimador EBLUP fue menor al de la estimación directa, lo que demuestra la mejora que ofrece el uso de información auxiliar. No obstante, cuando se incluyen variables medidas con error, como la tasa de informalidad, el coeficiente de variación se incrementa inevitablemente, ya que el error de medición debe ser agregado al cálculo del error cuadrático medio (ECM) por dominio. Estos resultados evidencian la importancia de evaluar cuidadosamente la inclusión de variables medidas con error, considerando tanto los beneficios en la capacidad explicativa como el impacto en la precisión del estimador.

La inclusión de variables auxiliares correlacionadas con la variable de interés es fundamental para mejorar las estimaciones. Identificar la oferta de información disponible en el país, especialmente por parte de las entidades públicas, permite elegir variables que expliquen el fenómeno de interés, apoyándose en un análisis estadístico y temático por parte de expertos para garantizar su relación. Sin embargo, cuando se incluyen variables auxiliares medidas con error, es necesario evaluar si su incorporación es realmente beneficiosa. Esto se debe a que el error de medición puede aumentar el ECM total del estimador, lo que podría contrarrestar las ventajas de incluir estas variables. Por otro lado, ignorar el error de medición puede llevar a una subestimación considerable del ECM final, generando un exceso de confianza en los resultados y afectando la validez de las inferencias.

Finalmente, para trabajos futuros, en casos en los que se utilicen múltiples variables auxiliares provenientes de encuestas, es necesario considerar la covarianza entre estas para reflejar adecuadamente la estructura de los datos y garantizar que el modelo capture todas las fuentes de variabilidad de manera precisa.

Referencias

Alvarado, F., Meléndez, M. & Pantoja, M. (2021), Mercados laborales fragmentados y el sistema de protección social en colombia, Undp lac working paper no. 14, Programa de las Naciones Unidas para el Desarrollo (PNUD).
*https://www.undp.org/es/latin-america/publicaciones/mercados-laborales-fragmentados-y-sistemas-de-proteccion-social-en-colombia

Burgard, J. P., Esteban, M. D., Morales, D. & Pérez, A. (2020), 'A fay–herriot model when auxiliary variables are measured with error', *TEST* **29**(1), 166–195.

Camacho, A., Conover, E. & Hoyos, A. (2014), 'Effects of colombia's social protection system on workers' choice between formal and informal employment', *The World Bank Economic Review* **28**(3), 446–466.

DANE (2019), 'Decreto 2404 de 2019', `https://www.sen.gov.co/sites/default/files/migracion-files/sen/Decreto_2404_diciembre_27_2019.pdf`. Consultado en junio de 2024.

DANE (2020), 'Guía para la inclusión del enfoque diferencial e interseccional'. Documento del DANE.
*https://www.dane.gov.co/files/investigaciones/genero/guia-inclusion-enfoque-difencias-intersecciones-produccion-estadistica-SEN.pdf

DANE (2022), 'Metodología general encuesta multiproósito bogotá - cundinamarca em 2021'. Documento interno del DANE.
*https://microdatos.dane.gov.co/index.php/catalog/743/download/21328

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

García-Ubaque, J. C., Riaño-Casallas, M. I. & Benavides-Piracón, J. A. (2012), 'Informalidad, desempleo y subempleo: Un problema de salud pública', *Revista de salud pública* **14**, 138–150.

George E. Battese, R. M. H. & 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.
*https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478561

Ghosh, M., Sinha, K. & Kim, D. (2006), 'Empirical and hierarchical bayesian estimation in finite population sampling under structural measurement error models', *Scandinavian Journal of Statistics* **33**(3), 591–608.

Guao, M. & Ghosh, M. (2012), 'Empirical bayes prediction in random effects models with covariates subject to measurement error', *Calcutta Statistical Association Bulletin* **64**(3-4), 265–276.

Harmening, S., Kreutzmann, A.-K., Pannier, S., Skarke, F., Rojas-Perilla, N., Salvati, N., Schmid, T., Templ, M., Tzavidis, N. & Würz, N. (2024), *emdi: Estimación en Áreas Pequeñas basada en Indicadores y Datos Mapa (SAE for Indicators and Maps)*. R package version 2.2.2.
*https://cran.r-project.org/web/packages/emdi/index.html

Ivanescu, E., Morales, D. & Mínguez, C. (2023), *saeRobust: Robust Small Area Estimation*. R package version 1.1.3.
*https://cran.r-project.org/web/packages/saeRobust/index.html

J., P. T. F. (2010-09), 'Evolución de la afiliación al sgsss en colombia, informalidad laboral y desempleo'.
*https://repositorio.unal.edu.co/handle/unal/69927

JL, C. A., Martínez, R., Cañón JR, L. et al. (2017), 'The informal economy: an occupational health issue', *Archivos de Prevencion de Riesgos Laborales* **20**(1), 30–32.

Lohr, S. L. & Prasad, N. G. N. (2003), 'Small area estimation with auxiliary survey data', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **31**(4), 383–396.

Medina-Gómez, O. & López-Arellano, O. (2019), 'The informality of labor and the right to health in mexico: a critical analysis', *Ciência & Saúde Coletiva* **24**, 2583–2592.

Molina, I. & Marhuenda, Y. (2015), 'sae: An R Package for Small Area Estimation', *The R Journal* **7**(1), 81–98. *https://doi.org/10.32614/RJ-2015-007

Molina, I. & Rao, J. N. K. (2010), 'Small area estimation of poverty indicators', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **38**(3), 369–385.

Pfeffermann, D. (2002), 'Small area estimation: New developments and directions', *International Statistical Review / Revue Internationale de Statistique* **70**(1), 125–143.

Porter, A. T., Wikle, C. K. & Holan, S. H. (2015), 'Small area estimation via multivariate fay–herriot models with latent spatial dependence', *Australian & New Zealand Journal of Statistics* **57**(1), 15–29.

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

Rao, J. N. (2014), 'Small-area estimation', *Wiley StatsRef: Statistics Reference Online* pp. 1–8.

Rico, F. O., Piñerez, C. F. T. & Vargas, N. R. (2024), 'A review of the use of small area estimation in colombia', *Revista Colombiana de Estadística - Applied Statistics* **47**(2), 407–422. *http://doi.org/10.15446/rce.v47n2.112779

Torabi, M., DATTA, G. S. & RAO, J. N. K. (2009), 'Empirical bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates', *Scandinavian Journal of Statistics* **36**(2), 355–368.

Ubaidillah, A. & Mubarak, M. R. (2023), *saeME: Small Area Estimation with Measurement Error*. R package version 0.1.0. *https://cran.r-project.org/web/packages/saeME/index.html

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

Ybarra, L. M. R. & Lohr, S. L. (2008), 'Small area estimation when auxiliary information is measured with error', *Biometrika* **95**(4), 919–931.

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

A. Anexo: Tablas de coeficiente de variación por ciclo de vida

Localidad	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Antonio Nariño	2.15	1.68	2.01	2.01	1.70	1.75	2.00	2.01
Barrios Unidos	1.21	1.10	1.18	1.18	1.11	1.13	1.18	1.18
Bosa	1.27	1.16	1.24	1.24	1.17	1.17	1.23	1.24
Chapinero	1.60	1.39	1.54	1.54	1.40	1.41	1.54	1.54
Ciudad Bolívar	1.17	1.09	1.15	1.15	1.10	1.10	1.15	1.15
Engativá	1.25	1.14	1.22	1.22	1.15	1.15	1.21	1.22
Fontibón	0.88	0.84	0.87	0.87	0.84	0.84	0.86	0.87
Kennedy	0.90	0.86	0.89	0.89	0.86	0.86	0.89	0.89
La Candelaria	3.51	2.20	3.02	2.99	2.20	2.28	2.96	3.02

Dominio	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Los Mártires	3.76	2.54	3.25	3.25	2.57	2.62	3.09	3.25
Puente Aranda	1.12	1.04	1.10	1.10	1.05	1.05	1.10	1.10
Rafael Uribe Uribe	1.64	1.43	1.58	1.58	1.44	1.45	1.57	1.58
San Cristóbal	0.85	0.82	0.84	0.84	0.82	0.82	0.84	0.84
Santa Fe	2.06	1.71	1.95	1.95	1.73	1.74	1.93	1.95
Suba	1.14	1.06	1.12	1.12	1.06	1.07	1.11	1.12
Teusaquillo	0.81	0.78	0.81	0.81	0.78	0.78	0.80	0.81
Tunjuelito	3.77	2.33	3.15	3.15	2.39	2.42	3.09	3.15
Usaquén	0.91	0.87	0.90	0.90	0.87	0.87	0.90	0.90
Usme	1.25	1.15	1.22	1.22	1.16	1.16	1.22	1.22

Tabla 2: Coeficientes de Variación para el ciclo de vida Adolescentes

Localidad	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Antonio Nariño	2.83	2.09	2.56	2.56	2.12	2.13	2.50	2.56
Barrios Unidos	1.79	1.54	1.72	1.72	1.55	1.56	1.70	1.72
Bosa	1.35	1.23	1.32	1.32	1.24	1.25	1.31	1.32
Chapinero	2.10	1.66	1.98	1.98	1.69	1.73	1.96	1.98
Ciudad Bolívar	1.32	1.22	1.30	1.30	1.23	1.23	1.29	1.30
Engativá	0.92	0.88	0.91	0.91	0.88	0.88	0.91	0.91
Fontibón	1.27	1.16	1.24	1.24	1.17	1.18	1.24	1.24
Kennedy	0.90	0.87	0.89	0.89	0.87	0.87	0.89	0.89
La Candelaria	4.17	2.64	3.47	3.43	2.56	2.59	3.34	3.47
Los Mártires	4.26	2.86	3.68	3.68	2.87	2.89	3.36	3.68
Puente Aranda	1.26	1.16	1.23	1.23	1.17	1.17	1.23	1.23
Rafael Uribe Uribe	1.31	1.21	1.28	1.28	1.21	1.21	1.27	1.28
San Cristóbal	1.00	0.95	0.99	0.99	0.95	0.95	0.98	0.99
Santa Fe	2.09	1.78	1.98	1.98	1.79	1.79	1.95	1.98
Suba	1.12	1.04	1.10	1.10	1.05	1.05	1.09	1.10
Teusaquillo	1.03	0.97	1.01	1.01	0.97	0.98	1.01	1.01
Tunjuelito	3.68	2.44	3.18	3.18	2.48	2.50	3.05	3.18
Usaquén	1.05	0.99	1.04	1.04	0.99	1.00	1.03	1.04
Usme	1.47	1.34	1.43	1.43	1.35	1.35	1.42	1.43

Tabla 3: Coeficientes de Variación para el ciclo de vida Juventud

Localidad	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Antonio Nariño	1.80	1.53	1.72	1.72	1.54	1.55	1.71	1.72
Barrios Unidos	0.99	0.93	0.97	0.97	0.94	0.94	0.97	0.97
Bosa	0.97	0.92	0.95	0.95	0.92	0.92	0.95	0.95
Chapinero	0.88	0.84	0.87	0.87	0.84	0.84	0.87	0.87
Ciudad Bolívar	1.08	1.01	1.06	1.06	1.02	1.02	1.06	1.06
Engativá	0.67	0.65	0.66	0.66	0.65	0.65	0.66	0.66
Fontibón	0.58	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Kennedy	0.71	0.69	0.70	0.70	0.69	0.69	0.70	0.70
La Candelaria	2.35	1.87	2.20	2.18	1.82	1.89	2.17	2.20
Los Mártires	2.70	2.17	2.52	2.51	2.17	2.18	2.42	2.52
Puente Aranda	1.02	0.96	1.00	1.00	0.96	0.97	1.00	1.00
Rafael Uribe Uribe	1.22	1.13	1.20	1.20	1.14	1.14	1.19	1.20

Dominio	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
San Cristóbal	0.64	0.63	0.64	0.64	0.63	0.63	0.64	0.64
Santa Fe	1.35	1.24	1.31	1.31	1.24	1.24	1.31	1.31
Suba	0.66	0.64	0.66	0.66	0.64	0.64	0.66	0.66
Teusaquillo	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
Tunjuelito	1.80	1.55	1.72	1.72	1.56	1.57	1.71	1.72
Usaquén	0.65	0.63	0.64	0.64	0.63	0.63	0.64	0.64
Usme	1.03	0.98	1.02	1.02	0.98	0.98	1.02	1.02

Tabla 4: Coeficientes de Variación para el ciclo de vida Adulto Joven

Localidad	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Antonio Nariño	0.93	0.88	0.92	0.92	0.89	0.89	0.92	0.92
Barrios Unidos	0.93	0.89	0.92	0.92	0.89	0.89	0.92	0.92
Bosa	0.98	0.93	0.97	0.97	0.93	0.94	0.96	0.97
Chapinero	0.83	0.80	0.82	0.82	0.80	0.80	0.82	0.82
Ciudad Bolívar	0.93	0.89	0.92	0.92	0.89	0.89	0.92	0.92
Engativá	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49
Fontibón	0.59	0.58	0.59	0.59	0.58	0.58	0.59	0.59
Kennedy	0.54	0.53	0.54	0.54	0.53	0.53	0.54	0.54
La Candelaria	1.46	1.33	1.42	1.41	1.29	1.31	1.42	1.42
Los Mártires	1.75	1.51	1.67	1.67	1.51	1.52	1.66	1.67
Puente Aranda	0.79	0.77	0.79	0.79	0.77	0.77	0.79	0.79
Rafael Uribe Uribe	0.82	0.79	0.81	0.81	0.79	0.79	0.81	0.81
San Cristóbal	0.67	0.65	0.66	0.66	0.65	0.65	0.66	0.66
Santa Fe	1.18	1.09	1.15	1.15	1.09	1.10	1.15	1.15
Suba	0.59	0.58	0.59	0.59	0.58	0.58	0.59	0.59
Teusaquillo	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Tunjuelito	1.36	1.23	1.33	1.33	1.23	1.24	1.32	1.33
Usaquén	0.49	0.48	0.48	0.48	0.48	0.48	0.48	0.48
Usme	0.94	0.90	0.93	0.93	0.90	0.91	0.93	0.93

Tabla 5: Coeficientes de Variación para el ciclo de vida Adulto Maduro

Localidad	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Antonio Nariño	0.81	0.78	0.80	0.80	0.78	0.78	0.80	0.80
Barrios Unidos	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49
Bosa	0.68	0.67	0.68	0.68	0.67	0.67	0.68	0.68
Chapinero	0.57	0.56	0.57	0.57	0.56	0.56	0.57	0.57
Ciudad Bolívar	0.69	0.68	0.69	0.69	0.68	0.68	0.69	0.69
Engativá	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
Fontibón	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Kennedy	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
La Candelaria	1.18	1.10	1.16	1.16	1.09	1.10	1.16	1.16
Los Mártires	0.88	0.84	0.87	0.87	0.84	0.85	0.87	0.87
Puente Aranda	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47
Rafael Uribe Uribe	0.55	0.54	0.55	0.55	0.54	0.54	0.55	0.55
San Cristóbal	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34

Dominio	Directo	Modelo_1	Modelo_2	Modelo_3	Modelo_4	Modelo_5	Modelo_6	Modelo_7
Santa Fe	0.91	0.87	0.90	0.90	0.87	0.87	0.90	0.90
Suba	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
Teusaquillo	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Tunjuelito	0.87	0.83	0.86	0.86	0.84	0.84	0.86	0.86
Usaquén	0.32	0.31	0.32	0.32	0.31	0.31	0.32	0.32
Usme	0.92	0.87	0.91	0.91	0.88	0.88	0.91	0.91

Tabla 6: Coeficientes de Variación para el ciclo de vida Adulto Mayor