Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Octubre 2023

Estimación de los indicadores de Ingreso promedio y la tasa de desempleo municipal en Cundinamarca en el 2017 utilizando áreas pequeñas y transformaciones funcionales

Estimation of average income indicators and municipal unemployment rate in Cundinamarca in 2017 using small areas and functional transformations.

Jonnathan Guerrero Beltrán[a]
jonnathanguerrero@usantotomas.edu.co

Andres Felipe Ortiz Rico[b]
andresortiz@usantotomas.edu.co

Resumen

La Estimación en Áreas Pequeñas (SAE) tiene como objetivo mejorar las estimaciones en situaciones donde los errores de muestreo son significativamente altos, lo que dificulta la obtención de conclusiones fiables. En este estudio, se emplea la metodología SAE para estimar el ingreso promedio y la tasa de desempleo a nivel municipal en Cundinamarca, utilizando datos de la encuesta multipropósito 2017. Se aplican transformaciones a los datos y se analizan las ventajas y desventajas de estas transformaciones. Basado en estas transformaciones, se seleccionan las variables auxiliares más pertinentes para cada tipo de transformación, y se evalúa la calidad de las estimaciones mediante el coeficiente de variación. Como resultado, se concluye que el uso del modelo Fay-Herriot con transformaciones en los datos reduce el valor del coeficiente de variación, lo que sugiere que la transformación de los datos es una opción viable para mejorar la calidad de las estimaciones de los indicadores de interés. Además, se utiliza el software estadístico R como entorno de trabajo para el procesamiento de datos.

***Palabras clave*:** Coeficiente de variación, encuesta multiproposito 2017, información auxiliar, ingreso promedio, modelo Fay-Herriot, tasa de desempleo, transformaciones en los datos.

[a]Estudiante Maestría en Estadística Aplicada
[b]Asesor y Profesor Facultad de estadística USTA

Índice

1. Introducción **5**

2. Marco teórico **6**

2.1. Antecedentes 6

2.2. Métodos Basados en el diseño. 7

2.3. Métodos basados en el modelo 9

2.3.1. Modelo Fay- Herriot 10

2.3.2. Modelo de Battese-Harter-Fuller 11

2.4. Transformaciones funcionales 13

3. Objetivos **14**

3.1. Objetivo General 14

3.2. Objetivos específicos 14

4. Metodología **14**

4.1. Encuesta multipropósito 2017 15

4.2. Librería emdi 15

4.3. librería Trafo 16

4.4. Información auxiliar 16

5. Análisis de resultados **18**

5.1. Tasa de desempleo en Cundinamarca 18

5.1.1. Estimación de la tasa de desempleo sin realizar transformaciones 19

5.1.2. Estimación de la tasa de desempleo con la transformación logaritmo 20

5.1.3. Estimación de la tasa de desempleo con la transformación Log -Shift 20

5.1.4. Estimación de la tasa de desempleo con la transformación Box - Cox 21

5.1.5. Estimación de la tasa de desempleo con la transformación Dual 21

5.1.6. Análisis de los residuales para la tasa de desempleo 21

5.1.7. Análisis de los CVs para las estimaciones de la tasa de desempleo 24

5.2. Ingreso Promedio 26

5.2.1. Estimación del ingreso promedio sin realizar transformaciones 27

5.2.2. Estimación del ingreso promedio con la transformación logaritmo 27

5.2.3. Estimación del ingreso promedio con la transformación Log-Shift 27

5.2.4. Estimación del ingreso promedio con la transformación Box-Cox 28

5.2.5. Estimación del ingreso promedio con la transformación Dual 28

5.2.6. Análisis de los CVs para las estimaciones del ingreso promedio 29

5.2.7. Análisis de los residuales para el ingreso promedio 31

6. Conclusiones **34**

7. Anexos **35**

7.1. Descomposición de Varianza 35

7.2. Error Cuadrático medio transformado 35

7.3. Características de las funciones de transformación 36

7.3.1. Transformación logaritmo 36

7.3.2. Transformación log - shif 36

7.3.3. Transformación Box - Cox 37

7.3.4. Transformación doble potencia 37

7.4. Estimación del ingreso promedio y tasa de desempleo (Estrategia 2) 38

7.4.1. Tasa de desempleo en Cundinamarca 2017 (Estrategia 2) 38

7.4.2. Ingreso promedio en Cundinamarca en 2017 (Estrategia 2) 41

Índice de tablas

1. Información auxiliar 17

2. λ que maximiza la función 19

3. Valores P, prueba de normalidad 19

4. Variables auxiliares para el modelo FH sin transformación 19

5. Variables auxiliares para el modelo FH, logaritmo 20

6. Variables auxiliares para el modelo FH, Log - Shift 20

7. Variables auxiliares para el modelo FH, Box - Cox 21

8. Variables auxiliares para el modelo FH, Dual 21

9. CVs Tasa de desempleo con las diferentes transformaciones 24

10. Tasa de desempleo en Cundinamarca 2017 26

11. λ que maximiza la función 26

12. Valores P, prueba de normalidad 27

13. Variables auxiliares para el modelo FH, sin transformaciones 27

14. Variables auxiliares para el modelo FH, logaritmo 27

15. Variables auxiliares para el modelo FH, log-Shift 28

16. Variables auxiliares para el modelo FH, Box-Cox 28

17. Variables auxiliares para el modelo FH, Dual 28

18. CVs Ingreso Promedio con las diferentes transformaciones 29

19. Ingresos promedio 31

20. Modeolo FH con las variables definitivas 38
21. λ que maximiza la función 39
22. Tasa de desempleo en Cundinamarca 2017 40
23. Modelo de FH definitivo ingresos promedio 41
24. λ que maximiza la transformación 41
25. Ingresos promedio 43

1. Introducción

En el contexto social y político, resulta fundamental adquirir un conocimiento profundo de las características esenciales de una población y de la evolución de un país. Esto se torna crucial para llevar a cabo el seguimiento de propuestas públicas y considerar nuevas estrategias que permitan abordar situaciones que requieran intervención externa en las condiciones del territorio estudiado. En otras palabras, se trata de la formulación de políticas públicas destinadas a mejorar las condiciones existentes y contribuir al mejoramiento de la calidad de vida de los ciudadanos. Como se menciona en el trabajo de Romero et al. (2012), dos indicadores cruciales para evaluar la calidad de vida son la tasa de desempleo y el ingreso per cápita de los hogares.

En Rojas-Perilla et al. (2017) indican que se han explorado estrategias para estimar ciertos indicadores en niveles de agregación específicos. Sin embargo, no siempre es factible realizar estimaciones para todas las desagregaciones que una población pueda implicar, ya sea debido a la falta de muestras representativas o la posibilidad de obtener estimaciones inexactas debido a su insuficiencia.

Una solución planteada por Rojas-Perilla et al. (2017) para abordar este problema consiste en aumentar el tamaño de la muestra utilizando los datos de la misma encuesta. Sin embargo, es importante tener en cuenta que esta medida conlleva costos adicionales para la entidad involucrada. Además, es esencial considerar que en algunas ocasiones, la realización de la encuesta puede verse dificultada por cuestiones como el acceso a áreas de interés, que pueden ser complicadas debido a factores como la logística de llegada al lugar específico, problemas de transporte, seguridad, entre otros.

En Bogotá por ejemplo, la encuesta multipropósito 2017 (EM 2017) actualiza la información estadística de las condiciones sociales, económicas y del entorno de los hogares y habitantes de Bogotá y 37 municipios de Cundinamarca, sin embargo en las versiones anteriores de la encuesta, años 2011 y 2014, la desagregación de Bogotá solo estaba establecida a nivel de localidad y 31 municipios de Cundinamarca, y actualmente la desagregación de Bogotá aumento a nivel de UPZ y 6 municipios adicionales, sin embargo con esta encuesta no es posible obtener buenas estimaciones con algunas características de cada área, como estrato, edad, nivel educativo, sexo; es por esto que SAE, permite obtener estimaciones confiables de los indicadores de interés por medio de información auxiliar para las áreas en las cuales los diseños probabilísticos no son convenientes.

Con frecuencia, los indicadores que se estiman mediante métodos basados en el modelo, como menciona Rojas-Perilla et al. (2017), incluyen mediciones de pobreza, desigualdad, tasas de recuento e incidencia de la pobreza, entre otros. Por tanto, resulta relevante emplear estos métodos para calcular los indicadores propuestos en este estudio, que se centran en el ingreso promedio y la tasa de desempleo a nivel municipal en Cundinamarca. Los autores sugieren que, para obtener estimaciones precisas, es necesario que los términos de error del modelo cumplan con ciertos supuestos. En caso de que no se cumplan, es necesario considerar estrategias para mejorar estas estimaciones.

En este contexto, Molina y Rao (2013) (Molina and Rao (2013)) proponen tres posibles soluciones para abordar la falta de cumplimiento de los supuestos. Por ejemplo, cuando los términos de error del modelo divergen significativamente de una distribución normal, el estimador del mejor predictor empírico (EBP) puede sesgarse. Las alternativas sugeridas por estos autores incluyen la formulación del EBP bajo supuestos paramétricos alternativos y más flexibles como una primera opción. Sin embargo, esta elección implica el desarrollo de nuevas herramientas para la estimación y capacitación del analista de datos.

La segunda opción es la utilización de metodologías que minimicen la dependencia de supuestos paramétricos. Aunque esta alternativa tiene la ventaja de no requerir supuestos rígidos, no garantiza necesariamente una mejora en la predicción empírica. Además, su implementación conlleva el desarrollo de nuevas estimaciones y herramientas de inferencia.

La tercera propuesta consiste en encontrar transformaciones basadas en los datos que sean apropiadas para garantizar la validez de los supuestos del modelo. El estudio llevado a cabo por Rojas-Perilla et al. (2017) señala que, en el caso particular analizado, las transformaciones basadas en los datos mejoraron de manera significativa las predicciones del modelo y permitieron una aproximación más adecuada al

cumplimiento de los supuestos del mismo.

Considerando lo mencionado anteriormente, en el marco de este estudio, se propone llevar a cabo estimaciones que se basen tanto en el modelo como en transformaciones de los datos. Específicamente, se emplearán transformaciones de tipo logaritmo, siguiendo la recomendación de Molina and Rao (2010), especialmente en el análisis de datos relacionados con los ingresos. Esta estrategia será aplicada de manera particular a los datos de para la estimación de la tasa de desempleo y al ingreso promedio en Cundinamarca, utilizando la metodología de Estimación en Áreas Pequeñas.

2. Marco teórico

El estudio y aplicación de las estimaciones en áreas pequeñas, también conocido como SAE por sus siglas en inglés, ha sido una disciplina en desarrollo desde el siglo XIX. Inicialmente, se centró en mejorar las estimaciones de parámetros de poblaciones, como totales y medias, específicamente en la estimación per cápita de ingresos en áreas pequeñas de los Estados Unidos en esa época, y estos métodos se abreviaron como "modelos FH.[en] referencia al estudio inicial Fay and Herriot (1979).

El desafío en la estimación en áreas pequeñas radica en la insuficiencia de muestras para obtener estimaciones precisas. Esto puede ocurrir cuando no se planifica adecuadamente una encuesta para estimar a un nivel tan detallado. En la literatura, las áreas afectadas por esta problemática se denominan "áreas pequeñas". Es importante destacar que el término "pequeñas"no se refiere necesariamente al tamaño de la población en esas áreas, ya que incluso áreas con poblaciones considerablemente grandes pueden tener estimaciones deficientes debido a muestras insuficientes. Esto se debe a que el tamaño de la muestra no siempre se planifica adecuadamente para cada área específica.

SAE comienza a ser útil cuando las estimaciones muestrales tienen coeficientes de variación altos, lo que indica una alta variabilidad en las estimaciones debido a muestras pequeñas o no representativas Ghosh and Rao (1994). SAE permite obtener estimaciones precisas y confiables para áreas pequeñas a pesar de contar con muestras pequeñas o incluso nulas, como se describe en Pfeffermann and Sverchkov (2007). El objetivo principal de SAE es producir estimadores de los parámetros de interés en estas áreas o dominios, que reduzcan el error de estimación y proporcionen una medida del error asociado.

En términos de estrategias para la estimación en áreas pequeñas, existen dos enfoques principales: los métodos basados en el diseño y los métodos basados en el modelo. Los métodos basados en el diseño incluyen estimadores compuestos, estimadores directos y estimadores sintéticos. Los métodos basados en el modelo, por otro lado, se dividen en modelos a nivel de área y modelos a nivel de unidad Lohr (2019), Lehtonen and Veijanen (2009), Bell et al. (2013). Estos enfoques se utilizan para abordar el desafío de obtener estimaciones precisas en áreas pequeñas con limitaciones en el tamaño de la muestra.

2.1. Antecedentes

Rojas-Perilla et al. (2017) propone un estudio hecho en México, particularmente en el estado de Guerrero, la motivación para desarrollar el proyecto en dicho país y en ese estado es porque México se destaca por tener una de las economías más grandes de América Latina e internamente se clasifican como una de las economías más desiguales del mundo. Además que el programa de las Naciones Unidas para el Desarrollo (PNUD) indica que el estado de Guerrero presenta una de las tasas más elevadas de pobreza y deficiencia en el desarrollo de su infraestructura. Los datos fueron tomados de la Encuesta de Ingresos y Gastos de los Hogares (ENIGH) 2010 y el Censo de población y vivienda de 2010, eligiendo la variable de respuesta como el ingreso familiar total per capita del trabajo, la cual tiene como unidad de medida pesos mexicanos y como variables explicativas se tienen las sociodemográficas disponibles en los hogares. Los resultados obtenidos al transformar los datos son satisfactorios, ya que permitieron reducir los errores de estimación, y tener buenas estimaciones en los estados en los que no fue posible obtener una muestra.

Kreutzmann et al. (2019a) muestra como la librería "emdi" del software estadístico R, brinda herramien-

tas que permiten simplificar la aplicación de métodos SAE, además destacan que esta librería genera los parámetros de transformación basada en los datos automáticamente. La incertidumbre de las estimaciones se obtiene por medio del uso del bootstrap paramétrico y bootstrap semiparamétrico y la estimación de de la incertidumbre adicional que se genera por la estimación del parámetro la recoge la estimación del error cuadrático medio, MSE por sus siglas en inglés. Lo nombrado anteriormente lo ilustran con algunos estados de Austria y se evidencian las mejoras en la calidad de los errores de medición además de la muestra de los resultados del paquete.

Fornieles (2013) realiza un estudio en el cual indica que las transformaciones en los datos han permitido dar solución a inconvenientes que surgían cuando se pretendían realizar algunas predicciones de la variable en estudio, las diferentes opciones de transformaciones que se usaron permitieron mejorar la curtosis y la asimetría de las variables del modelo, además de asociar linealmente las dos variables de interés lo que permitió convertir el modelo en uno de regresión lineal simple.

Zea y Ortiz (2018), muestran la aplicación del modelo Fay Herriot en la estimación a nivel de municipio de la tasa de desempleo y el ingreso promedio en Cundinamarca, usando el software estadístico R como herramienta fundamental para el calculo de los indicadores. En el estudio usan como principal fuente de información la encuesta multiproposito 2014, usando como variables auxiliares los resultados de las pruebas saber 11, tasa de beneficiarios del sistema de selección de beneficiarios de programas sociales (SISBEN), Avalúos catastrales y urbanos, tasa de abandono escolar, entre otras, muestran la comparación del error cuadrático medio de los modelos y logran identificar la precisión de las estimaciones, concluyendo que el uso de la metodología SAE se desarrollo con el fin de reducir el error de muestreo de las estimaciones directas, además de obtener unas predicciones plausibles para los municipios en los cuales no fueron cubiertos por la Encuesta Multiproposito 2014.

2.2. Métodos Basados en el diseño.

Estimadores directos

Los estimadores directos son aquellos que usan solamente los datos de la encuesta para el área específica. Independientemente del diseño muestral que se proponga la mayoría de veces los estimadores más usados para el total poblacional son el Estimador Horvitz-Thompson (1952) y el estimador Hansen-Hurwitz (1952), enseguida se describen

Estimador Horvitz-Thompson

El estimador del total poblacional propuesto por Horvitz-Thompson, $\tilde{t}_y$, es uno de los más usados cuando el diseño de muestreo se realiza sin reemplazo y esta definido como sigue

$$\tilde{t}_y = \sum_s \frac{y_k}{\pi_k} \tag{1}$$

Definiendo a π_k como la probabilidad de inclusión para el k-ésimo elemento, siendo $\frac{1}{\pi_k}$ el factor de expansión, y_k el valor de y para el k-esímo individuo de la muestra s de la población $\mathbb{U}$.

Estimador Hansen-Hurwitz

El estimador de Hansen-Hurwitz (1952) se usa frecuentemente para realizar las estimación del total poblacional cuando el diseño de muestreo se realiza con reemplazo,

$$\hat{t}_y = \frac{1}{m} \sum_s \frac{y_k}{p_k} \tag{2}$$

Donde m indica la cantidad de veces que se selecciona el elemento, p_k hace referencia a la probabilidad de inclusión para el k -ésimo elemento, siendo $\frac{1}{\pi_k}$ el factor de expansión, y_k el valor de y para el k-esímo

individuo de la muestra s de la población $\mathbb{U}$.

Los estimadores que se mostraron en las expresiones 1 y 2 son insesgados, además que no necesitan asumir ningún modelo o hipótesis sobre las variables, esto indicaría que son no paramétricos, pero el tamaño de la muestra podría producir una varianza grande no aceptable, particularmente si el área no tiene observaciones de la muestra.

Ahora bien, ajustando el estimador directo del total visto en la ecuación 1, se tiene que el estimador directo del total por dominio, definiendo como dominio a las subdivisiones o subgrupos específicos de la población que se eligen en función de características específicas de interés, como ubicación geográfica, edad, género, etnia, ingresos, nivel educativo u otras características demográficas, socioeconómicas o cualquier atributo relevante. $t_{d,dir}$, esta dado por la expresión 3.

$$\tilde{t}_{d,dir} = \sum_{k \in S_d} w_k y_k \delta_{kd}, \tag{3}$$

Donde $w_k = \pi_d^{-1}$ son los respectivos pesos muestrales para los individuos de la muestra en el área d, $\delta_{kd} = 1$ si el elemento k está en el domino d y $\delta_{kd} = 0$ en otro caso.

Estimadores sintéticos

Teniendo en cuenta las complicaciones que surgen en los estimadores directos, especialmente aquellos propuestos por Horvitz-Thompson o Hansen-Hurwitz, una alternativa viable es emplear el método sintético. Este enfoque establece relaciones entre medidas (cocientes) en distintos dominios. Cuando estas medidas se asemejan al cociente de los totales de la población, se puede mejorar la precisión de las estimaciones en comparación con el método directo. Sin embargo, es importante destacar que si estas medidas no son uniformes, el estimador sintético podría introducir sesgos adicionales. Este fenómeno sugiere que el método directo podría proporcionar estimaciones imparciales pero con varianzas inaceptables, mientras que el estimador sintético, al reducir la varianza, podría introducir sesgos en las estimaciones, como lo señalan (CostaJJ and VenturaJJJ, 2001). A continuación, se presenta la construcción del estimador sintético.

Sea N_d la cantidad que representa el total en cada dominio d, si los cocientes $\dfrac{\tilde{t}_{d,dir}}{N_d}$ son similares en dominios diferentes y si cada cociente es similar a $\dfrac{t_y}{t_u}$, siendo t_u un total de la población y la asociación con t_y corresponde al cociente de los totales de la población, entonces el estimador sintético $\tilde{t}_{d,sin}$ se puede definir como

$$\tilde{t}_{d,sin} = \left(\frac{\tilde{t}_y}{\tilde{t}_u}\right) N_d$$

Siendo $\frac{\tilde{t}_y}{\tilde{t}_u}$ el cociente entre los estimadores totales de la población. Formalmente el estimador sintético se define como:

$$\tilde{t}_{d,sin} = \sum_{j=1}^{J} N_{dj} \tilde{t}_{d,dir} \tag{4}$$

Estimadores compuestos

Con el fin de mejorar las estimaciones de los indicadores de interés escritos en las expresiones 3 y 4, disminuyendo la varianza del estimador directo y disminuir el sesgo del estimador a c sintético, se proponen los **estimadores compuestos**, estos proponen un peso α_d que depende del tamaño muestral de área, dando lugar al estimador dependiente del tamaño muestral, con el objetivo aumentar la eficiencia

del estimador directo y reducir el sesgo del estimador sintético, este estimador garantiza que no puede tener peor eficiencia que el estimador directo, ni mayor sesgo que el estimador sintético, la desventaja del estimador radica en que no es posible calcular estimaciones para áreas no muestreadas y no se conocen estimadores estables del error cuadrático medio (ECM)

Con el propósito de mejorar las estimaciones de los indicadores de interés expresados en las ecuaciones 3 y 4, con el objetivo de reducir la varianza del estimador directo y mitigar el sesgo del estimador sintético, se introduce la noción de **estimadores compuestos**. Estos estimadores proponen la incorporación de un peso α_d que depende del tamaño de la muestra en cada área, lo que da lugar a un estimador ponderado por el tamaño de la muestra. Este enfoque tiene como finalidad aumentar la eficiencia del estimador directo y disminuir el sesgo del estimador sintético. Es importante destacar que este estimador garantiza que no puede ofrecer una eficiencia inferior al estimador directo, ni un sesgo mayor al estimador sintético (Drew et al., 1982). No obstante, una desventaja de este enfoque es que no permite la generación de estimaciones para áreas no muestreadas y no se conocen estimadores robustos del error cuadrático medio (mse, por sus siglas en ingles). El estimador compuesto se escribe como:

$$\tilde{t}_d = \alpha_d \tilde{t}_{d,dir} + (1 - \alpha_d)\tilde{t}_{d,sin} \tag{5}$$

Definiendo α_d como sigue,

$$\alpha_d = \frac{mse(\tilde{t}_{d,sin}) - E[\tilde{t}_{d,sin} - t_d]}{mse(\tilde{t}_{d,sin}) + mse(\tilde{t}_{d,dir}) - 2E[\tilde{t}_{d,sin} - t_d][\tilde{t}_{d,dir} - t_d]}$$

Siendo α_d los pesos óptimos relativos. Es importante notar que si n_d es pequeño α_d tenderá a cero,por lo tanto se le dará mayor peso al estimador sintético. Para profundizar en los detalles, el lector puede consultar Molina et al. (2015) y Drew et al. (1982).

2.3. Métodos basados en el modelo

Los métodos basados en el modelo a nivel de área establecen una relación entre el estimador objetivo en las áreas específicas y la información auxiliar disponible para cada uno de estos dominios. El modelo a nivel de área más usado es el propuesto por Fay-Harriot (FH), (Fay and Herriot, 1979), este fue utilizado para estimar los ingresos per capita en áreas pequeñas de estados unidos utilizando datos de Censo de Población y Vivienda de 1970, datos del Servicio de Impuestos Internos (IRS) y otros, este modelo a revolucionado la estimación en áreas pequeñas, ya que a partir de esta propuesta se ha venido desarrollando estrategias para mejorar las estimaciones y aprovechar la información auxiliar que se pueda obtener de cada área. Cuando se implementa un diseño muestral muchas veces no se consideran resultados a nivel desagregados pequeños no planificados, sin embargo este modelo propone introducir estimadores compuestos basados en modelos para obtener resultados en estos niveles, cuando se dispone de variables auxiliares relacionadas con la variable objetivo a nivel de área pequeña. En el modelo FH las variables dependientes son estimaciones directas proveniente de la muestra y la variable auxiliar son valores censales que se pueden obtener de registros administrativos. En los casos que no se dispone de variables auxiliares censales, Ybarra and Lohr (2008) introducen un modelo FH con error de medición, el cuál permite incluir la covariables provenientes de encuestas y considerar su error de muestreo respectivo en el proceso de estimación. También introdujeron un nuevo estimador de área pequeña que tiene en cuenta la variabilidad muestral en la información auxiliar y derivan sus propiedades, en particular mostrando que es aproximadamente insesgado, aplican esta metodología para predecir las cantidades medidas en la Encuesta nacional de examen de salud y nutrición de EEUU.

Los modelos a nivel de unidad involucran los valores unitarios de la variable de respuesta con las variables auxiliares disponibles para cada individuo de la muestra, según Molina (2019) este tipo de modelos gana mayor eficiencia que los modelos a nivel de área, siempre y cuando existan variables auxiliares a nivel de

individuo que sean lo suficientemente informativas respecto a la variable de estudio(variable de respuesta), esto basado en el tamaño muestral n.

Los estimadores en áreas pequeñas basados en el modelo son mucho más sofisticados y más aplicados a las necesidades cotidianas que los modelos basados en el diseño, ya que involucran heterogeneidad entre áreas que no son explicadas por las variables auxiliares consideradas, aún teniendo en cuenta los errores de estimación de estas variables (Ybarra and Lohr, 2008).

Ahora bien las estrategias descritas anteriormente permiten realizar una exploración exhaustiva sobre el planteamiento del modelo, como análisis de residuales, comportamiento de las variables auxiliares respecto al indicador de interés, errores de medición, entre otras, lo que permite realizar diferentes tipos de transformaciones en sus métodos de estimación y comparar cual es la estrategia mas pertinente para lograr las mejores estimaciones, en particular la del ingreso promedio municipal en Cundinamarca en el año 2017.

2.3.1. Modelo Fay- Herriot

El modelo Fay- Herriot (FH), es un modelo a nivel de área propuesto por Fay and Herriot (1979), según Li and Lahiri (2010a) ha ganado amplio reconocimiento en el contexto de la Estimación de Áreas Pequeñas, debido a su simplicidad, capacidad para preservar la confidencialidad de los microdatos y su capacidad para generar estimaciones consistentes dentro del diseño. El modelo FH se define en dos etapas, la primer etapa se basa en el modelo muestral, donde describe el error de muestreo del estimador directo y en la segunda etapa es un modelo de enlace que asume las características de área, es decir, el modelo asume que existe un vector de tamaño p de variables auxiliares $x_d = (x_{d1}, x_{d2}, ..., x_{dp})$ que se encuentra disponible para cada dominio d y considerando que la variable dependiente como un escalar correspondiente al estimador directo $\hat{Y}_d$ para $d = 1, ..., D$, se tiene que

$$\hat{Y}_d = x_d^T \beta + u_d + e_d, \;\; e_d \sim N(0, \sigma_d^2), \;\; u_d \sim N(0, \sigma_u^2), \tag{6}$$

Asumiendo que β contiene los coeficientes de regresión del modelo que generan un producto junto con la información auxiliar. e_d independiente se refiere al error muestral asumiendo que las varianzas σ_d^2 son conocidas, ya que en la práctica dichas varianzas son estimadas con los datos muestrales. u_d son los efectos aleatorios asociados al área especifica, independientes e idénticamente distribuidos (iid) con varianza σ_u^2 desconocida e independientes de e_d.

Mejor predictor lineal insesgado (BLUP) para el modelo FH

Molina et al. (2015) sugieren hacer uso de los predictores BLUP, estos ofrecen la posibilidad de mejorar el MSE, pero no asegura una reducción necesaria del mismo. Además de no depender de la normalidad de los efectos aleatorios. Adicionalmente, el parámetro σ^2 es una medida de la homogeneidad de las áreas después de contar con las covariables x_d. Si σ^2 es conocida, β puede ser estimada por el estimador de mínimos cuadrados estándar $\tilde{\beta}(\sigma^2)$,así se obtiene el mejor predictor insesgado (BLUP) de $\hat{Y}_d$ que se expresa como sigue

$$\hat{Y_d}^{BLUP} = \mathbf{x_d}^T \beta + \gamma_d (\hat{Y}_d - \mathbf{x_d}^T \beta) \tag{7}$$

donde

$$\gamma_d = \frac{\sigma_u^2}{\sigma_d^2 + \sigma_u^2} \tag{8}$$

Siendo σ_u^2 la varianza de los efectos aleatorios del área en particular teniendo en cuenta que esta es desconocida, y σ_d^2 identifica la varianza de los errores muestrales, la cual es conocida, ya que en la

práctica dichas varianzas son estimadas con los datos muestrales y $\tilde{\beta}$ es el estimador de mínimos cuadrados ponderados de β bajo la expresión 6, se puede expresar como sigue:

$$\tilde{\beta} = \sum_{d=1}^{D} \gamma_d \mathbf{x_d} \mathbf{x'_d}^{-1} \sum_{d=1}^{D} \gamma_d \mathbf{x_d} \hat{Y}_d^{DIR}$$

Definiendo a $\hat{Y}_d^{DIR} = Y_d + e_d$ con $d = 1, ..., D$ y $Y_d = \mathbf{x_d}^t + u_d$.

Mejor predictor lineal insesgado empírico (EBLUP) para el modelo FH

Para construir el Mejor predictor lineal insesgado empírico (EBLUP, por sus siglas en inglés), es esencial considerar que en la definición del (BLUP) que se expreso en la ecuación (7), se requiere el conocimiento de σ_u^2, lo cual, en aplicaciones prácticas, es una incógnita. Por lo tanto, al sustituir σ_u^2 por un estimador, denotado como $\tilde{\sigma_u^2}$, se obtiene el EBLUP de $\tilde{Y}_d$. El EBLUP se expresa de la siguiente manera:

$$\hat{Y}_d^{EBLUP} = \mathbf{x_d}^T \tilde{\beta} + \tilde{\gamma}_d (\hat{Y}_d - \mathbf{x_d}^T \tilde{\beta}) \tag{9}$$

Siendo

$$\tilde{\beta} = \sum_{d=1}^{D} \hat{\gamma}_d \mathbf{x_d} \mathbf{x'_d}^{-1} \sum_{d=1}^{D} \hat{\gamma}_d \mathbf{x_d} \hat{Y}_d^{DIR} \tag{10}$$

Note que $\hat{Y}_d$ identifica a la estimación directa de la variable interés para cada área.

Cálculo del Error cuadrático medio MSE

Con el fin de calcular el coeficiente de variación, Prasad and Rao (1990) utilizan una aproximación del Error Cuadrático Medio del EBLUP para el modelo FH.

$$MSE(\hat{\tilde{Y}}_d) = \begin{cases} g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_d^2) & d \in A \\ \mathbf{x_d}^T (\mathbf{X^T V^{-1} X})^{-1} \mathbf{x_d} + \hat{\sigma}_u^2 & d \notin A \end{cases} \tag{11}$$

Donde A hace referencia al conjunto de dominios observados,

$g_{1d}(\hat{\sigma}_u^2) = \frac{\hat{\sigma}_u^2 \sigma_d^2}{\hat{\sigma}_u^2 + \sigma_d^2}$; y $g_{2d}(\hat{\sigma}_u^2) = \frac{\Sigma_d^4}{(\hat{\sigma}_u^2 + \sigma_d^2)^2} \mathbf{x}_d^T (\mathbf{X^T V^{-1} X}) \mathbf{x_d}^T$ y

$g_{3d}(\hat{\sigma}_u^2) = \frac{2\sigma_g^4}{D(\hat{\sigma_u}_v^2 + \sigma_d^2)^3} (\hat{\sigma}_u^4) + 2(\hat{\sigma}_u^2) \sum_{d=1}^{D} \frac{\sigma_d^2}{D} + \sum_{d=1}^{D} \frac{\sigma_i^4}{D}$

$V = diag(\sigma_u^2 + \sigma_1, ..., \sigma_u^2 + \sigma_D^2)$

2.3.2. Modelo de Battese-Harter-Fuller

El modelo de regresión con errores anidados o también conocido como modelo de Battese - Hater - Fuller fue propuesto por Battese et al. (1988), relaciona los valores unitarios de la variable respuesta teniendo en cuenta los efectos aleatorios específicos de cada unidad con el fin de mejorar las estimaciones logrando una mayor eficacia particularmente en el modelo a nivel de unidad,. Este modelo asocia de forma lineal los valores de una variable de interés Y_{di} para el individuo i dentro del área d con los valores de p variables auxiliares para ese mismo individuo, así:

$$Y_{di} = \mathbf{X}_{di}^t \beta + \mathbf{u}_d + \mathbf{e}_{di} \quad i = 1, ..., N_d, \quad d = 1, 2..., D, \tag{12}$$

Donde $\mathbf{X}_{di}$ es un vector $p \times 1$ de variables auxiliares, β es el vector de coeficientes auxiliares común para todas las áreas, u_d es el efecto aleatorio del área y e_{di} es el error a nivel de individuo.

$\mathbf{u}_d$ y $\mathbf{e}_{di}$ son independientemente distribuidas, con $u_d \sim$ iid $N(0, \sigma_d^2)$ y $e_i \sim$ ind $N(0, \sigma_e^2)$.

Mejor predictor lineal insesgado (BLUP) para el modelo de Battese - Harter - Fuller

Observe que la media del área d se puede descomponer en la suma de los valores observados en la muestra y los no muestreados, así

$$\bar{Y}_d = N_d^- 1 \left(\sum_{s_d} Y_{di} + \sum_{r_d} Y_{di} \right) \tag{13}$$

Por tanto el BLUP de $\bar{Y}_d$ bajo un modelo con errores anidados se obtiene ajustando el modelo a los datos de la muestra y prediciendo los valores de las variables Y_{di} fuera de la muestra del área d, como se ilustra en la expresión siguiente:

$$\hat{\bar{Y}}_d^{BLUP} = N_d^- 1 \left(\sum_{s_d} Y_{di} + \sum_{r_d} \hat{y}_{di}^{BLUP} \right) \tag{14}$$

ahora bien, seleccionando el estimador de mínimos cuadrados ponderados para $\tilde{\beta}$ bajo el modelo, los valores predichos se describen como sigue.

$$\tilde{Y}_{di}^{BLUP} = x_{di}^t \tilde{\beta} + \tilde{u}_d \tag{15}$$

$$\tilde{u_d} = \gamma_d (\bar{y}_{da} - \bar{x}_{da}^t \tilde{\beta}), \qquad \gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / a_d} \tag{16}$$

siendo $\bar{y}_{da} = a_{d.}^{-1} \sum_{s_d} a_{di} Y_{di}$ y $\bar{x}_{da} = a_{d.}^{-1} \sum_{s_d} a_{di} x_{di}$ las medias muestrales de la variable respuesta y las variables auxiliares respectivamente con pesos $a_{di} = k_{di}^{-2}$ y $a_{d.} = \sum_{s_d}$

Mejor predictor lineal insesgado empírico (EBLUP) para el modelo de Battese - Harter - Fuller

El BLUP presentado en la expresión 12 depende de los verdaderos valores de los componentes de varianza $\theta = (\sigma_u^2, \sigma_e^2)^t$ sustituyendo el verdadero valor de θ por un estimador consistente $\hat{\theta} = (\hat{\sigma_u^2}, \hat{\sigma_e^2})^t$ se obtiene el EBLUP dado por:

$$\hat{\bar{Y}}_d^{EBLUP} = N_d^- 1 \left(\sum_{s_d} Y_{di} + \sum_{r_d} \hat{y}_{di}^{EBLUP} \right) \tag{17}$$

si se llama $\hat{\beta}$ al resultado de sustituir θ por $\hat{\theta}$ en $\hat{\beta}$ los valores predicho ahora son:

$$\hat{Y}_{di}^{EBLUP} = x_{di}^t \hat{\beta} + \hat{u}_d \tag{18}$$

$$\hat{u}_d = \hat{\gamma}_d = (\bar{y}_{da} - \bar{x}_{da}^t \hat{\beta}), \quad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_d} \tag{19}$$

2.4. Transformaciones funcionales

El uso de transformaciones en los datos se recomienda teniendo en cuenta el comportamiento de la variable de interés. De hecho, Rojas-Perilla et al. (2017) señalan que, en especial para variables como los indicadores de ingreso y desempleo, es común utilizar la transformación logaritmo natural. La expresión correspondiente a esta transformación es la siguiente:

$$T(y_{di}) = log(y_d) \tag{20}$$

Esta transformación tiene la capacidad de convertir distribuciones que estén muy sesgadas hacia la derecha en distribuciones más simétricas. Como menciona Rojas-Perilla et al. (2017), esta transformación es ampliamente utilizada en una variedad de campos de investigación, lo que la convierte en una de las técnicas más comunes para abordar problemas de asimetría y desafíos relacionados con la falta de normalidad en los datos.

No obstante, existen otras transformaciones que potencialmente pueden mejorar la calidad de la estimación en el estudio, como se discute en Royston et al. (2011). Una de estas alternativas es la extensión de la transformación logaritmo, conocida como "Log-shift", la cual incorpora el parámetro de transformación λ, en la siguiente expresión se muestra la transformación Log-Shift:

$$T_\lambda(y_{di}) = log(y_d + \lambda) \tag{21}$$

Donde el parámetro λ permite ajustar la forma de la curva de la transformación logarítmica. Cuando $\lambda = 1$, la transformación "Log-shift.$^{\text{es}}$ equivalente al logaritmo natural estándar. Si $\lambda < 1$, la transformación tiende a comprimir valores más pequeños y estirar valores más grandes, lo que puede ser útil para reducir la asimetría en datos positivos sesgados a la derecha. Por otro lado, si $\lambda > 1$, la transformación tiene el efecto opuesto, comprimiendo valores más grandes y estirando valores más pequeños.

Ahora bien, existe otra familia de transformaciones de datos que incluye la transformación logaritmo como un caso especial, y se conoce como la transformación de **Box-Cox**. según Box and Cox (1964) , esta transformación se basa en los datos y tiene como objetivo transformar la variable de respuesta para lograr un modelo con estructura simple, errores que siguen una distribución normal y una varianza de error constante. La definición de esta transformación se escribe a continuación:

$$T_\lambda(y_d) = \begin{cases} \frac{(y_d+s)^\lambda - 1}{\lambda} & \lambda \neq 0; \\ \\ log(y_d + s) & \lambda = 0. \end{cases} \tag{22}$$

Donde y_d es la variable original, $T_\lambda(y_d)$ es la variable transformada y λ es el parámetro de transformación; Teniendo en cuenta que el logaritmo solo esta definido para valores positivos Box and Cox (1964) proponen el parámetro fijo s tal que $y_d + s > 0$, de lo contrario no es posible usar dicha transformación. Es de notar que si $\lambda = 0$ la transformación logarítmica se convierte en un caso particular de esta familia y si $\lambda = 1$ los datos únicamente se desplazan.

A pesar de que la transformación de Box- Cox posee varias propiedades presenta un inconveniente que es uno de los principales motivos para estudiar la siguiente transformación, **Transformación de doble potencia** o **Dual** por su definición en ingles,y es el truncamiento del parámetro λ Yang (2006), ya que este es acotado inferiormente por $\frac{1}{\lambda}$ si $\lambda > 0$ y acotado superiormente $\frac{-1}{\lambda}$ si $\lambda < 0$. La transformación de doble potencia se define como sigue.

$$T_\lambda(y_d) = \begin{cases} \frac{(y_d+s)^\lambda - (y_d+s)^{-\lambda}}{2\lambda} & \lambda > 0; \\ log(y_d + s) & \lambda = 0. \end{cases} \tag{23}$$

Con $y_d + s > 0$, y λ el parámetro de transformación. Además de los beneficios que tiene la **Transformación de doble potencia** esta garantiza que siempre transformará los datos sin importar el valor que tome λ, ya que a diferencia de la transformación de Box-Cox cuando $\lambda = 1$ deja los datos sin transformar, únicamente los traslada.

3. Objetivos

3.1. Objetivo General

Analizar y comparar los valores del Error Cuadrático Medio (MSE) vinculados a cada transformación aplicada en las estimaciones del ingreso promedio y la tasa de desempleo a nivel municipal en Cundinamarca en el año 2017. Este análisis se llevará a cabo utilizando técnicas de estimación en áreas pequeñas y diversas transformaciones de los datos.

3.2. Objetivos específicos

1. Realizar las estimaciones del ingreso promedio y tasa de desempleo municipal en Cundinamarca en el año 2017, usando las transformaciones en los datos, logaritmo, Log-Shift, Box-Cox y Doble Potencia.
2. Comparar los MSE de las transformaciones propuestas asociados a cada transformación, con el fin de evaluar cómo estas transformaciones mejoran la precisión de las estimaciones de los indicadores propuestos.
3. Caracterizar los municipios del departamento de Cundinamarca en base de sus ingresos y sus tasas de desempleo, con el fin de identificar los municipios que poseen mayor vulnerabilidad.

4. Metodología

El presente trabajo de grado realiza varias estimaciones del ingreso promedio y la tasa de desempleo a nivel municipal en Cundinamarca durante el año 2017. Se llevaron a cabo estimaciones directas y estimaciones basadas en modelos a nivel de área, específicamente, el modelo FH, con y sin transformaciones en los datos.

Las estimaciones propuestas en este estudio involucran varias transformaciones de datos, incluyendo el logaritmo, Log-Shift, Box-Cox y la transformación de Doble Potencia.

La secuencia de pasos se inicia con las estimaciones directas, seguidas de las estimaciones mediante el modelo FH sin transformaciones, se calcula el coeficiente de variación (CV) respectivo y se evalúa la mejora en la calidad de las estimaciones.

Posteriormente, se consideran todas las variables auxiliares del modelo FH y se lleva a cabo un análisis de su pertinencia en el modelo mediante un algoritmo paso a paso proporcionado por la librería .$^{\text{Em}}$di". En este proceso de selección de variables, se elige un modelo según diferentes criterios, en este caso, se utilizó el criterio de información de Akaike (AIC).

Luego, se aplican transformaciones a los datos, determinando el valor óptimo de λ cuando corresponde, para cada una de las transformaciones. Se evalúa el comportamiento de los residuales en cada caso y se

procede a calcular el modelo FH con las transformaciones en los datos y todas las variables auxiliares disponibles.

Después de esta etapa, se seleccionan las variables auxiliares más pertinentes para la estimación del indicador con cada transformación, siguiendo el algoritmo paso a paso. Estas variables seleccionadas se incorporan en un nuevo modelo FH con los datos transformados, se calcula el error cuadrático medio (MSE) y se evalúa la calidad del modelo a través del coeficiente de variación.

Finalmente, se realizan comparaciones de las medidas de calidad de cada uno de los modelos, indicando el modelo seleccionado, las variables utilizadas y, si se aplicó alguna transformación, se describen las características más relevantes de la misma. Esta información detallada la puede encontrar en la sección 5.

Es importante destacar que los datos de la encuesta multipropósito de 2017 son de acceso público y están disponibles en el repositorio del Departamento Administrativo Nacional de Estadística (DANE).

4.1. Encuesta multipropósito 2017

Las encuestas multipropósito son realizadas con el fin de dar continuidad a las encuestas de calidad de vida realizadas por la secretaria distrital de planeación (SDP), aplicadas en el distrito en los años (1991, 1993, 2003, 2007) y la aplicación de la encuesta de capacidad de pago en el 2004. La (SDP) consideró pertinente integrar las dos encuestas debido a que la ciudad de Bogotá requería de las temáticas tanto de calidad de vida como de capacidad de pago.

En el 2014 se realizó la encuesta multipropósito con el fin de obtener información estadística sobre aspectos sociales y económicos del entorno Urbano de los habitantes de Bogotá, contemplando las 19 localidades Urbanas con sus respectivos estratos económicos (6 estratos), las zonas urbanas de los 20 municipios de la sabana y de 11 cabeceras de provincia del departamento de Cundinamarca, allí el diseño muestral fue probabilístico, estratificado, de conglomerados y polietápico, obteniendo información de 46070 hogares, distribuidas en 20518 hogares de Bogotá y 25552 hogares en municipios de Cundinamarca.

En el 2017 la encuesta multipropósito actualiza la información estadística aumentando su cobertura, ya que esta presenta una desagregación a nivel de distrito por UPZ y a nivel municipal de forma rural, algo que no se tenia en cuenta en la encuesta inmediatamente pasada, además de añadir algunas preguntas para las mediciones de seguridad alimentaria. En el plan regional aumento de 31 municipios muestreados a 37, usando un diseño probabilístico, multietápico, estratificado y de conglomerados, en consecuencia se logró obtener la información de 319.952 personas, que corresponden a 109.111 hogares, particularmente en Bogotá se encuestaron 77025 hogares que representan 221.809 personas, mientras que en los 37 municipios de Cundinamarca se obtuvo información de 32086 hogares que corresponden a 98.143 personas aproximadamente. La Encuesta Multipropósito ha obtenido y desarrollado información estadística actualizada, que brinda a las diferentes instancias territoriales, tanto departamentales como distritales, herramientas que les facilite conocer y analizar la evolución económica y social de sus habitantes que permitan generar nuevos proyectos políticos, sociales y económicos.

Los datos y demás características de la encuesta se puede localizar abiertamente en la pagina de la Secretaria Distrital de Planeación, allí se encotrará información que será usada para realizar las estimaciones que tienen como fin este trabajo, específicamente el ingreso promedio y la tasa de desempleo en Cundinamarca en el 2017.

4.2. Librería emdi

Para el desarrollo de este trabajo se usa el software estadístico R, se enfatiza en la librería **emdi** Kreutzmann et al. (2019b), que propone herramientas fundamentales para cumplir con los objetivos de este estudio, ya que incluye mayores beneficios que otras librerías que trabajan con la estimación en áreas pequeñas.Una característica fundamental de esta librería es que permite elegir la transformación que se

desea y a su vez calcular los respectivos parámetros de transformación de forma automática y pertinente para modelos de Fay Harrriot a nivel de unidad, además de evaluar la medida de calidad de cada uno de los indicadores de interés Rojas-Perilla (2018), en particular se tomará como referencia el Error cuadrático medio, MSE, sin embargo se debe tener en cuenta que el MSE generalmente se calcula con un bootstrap paramétrico, pero esta librería permite incluir dos métodos, el bootstrap parámetrico bajo el modelo que proponen Molina and Rao (2010) y uno semiparamétrico (Flachaire, 2005), los dos métodos incorporan la incertidumbre que se genera al estimar el parámetro de transformación (Rojas-Perilla et al., 2017). Además se considera usar un algoritmo similar al de la función **ebp** de esta librería, por sus siglas en ingles que abrevian Empirical Best Prediction, para los modelos a nivel de área que se proponen en este documento, ya que esta función sigue la propuesta de Molina and Rao (2010) donde sugieren el cálculo de indicadores usando el enfoque ebp, realizando las predicciones puntuales por medio de las aproximaciones de Montecarlo, así mismo ajusta el modelo a nivel de unidad mediante el método de máxima verosimilitud restringida (REML) permitiendo transformar la variable según se requiera.

4.3. librería Trafo

Medina et al. (2019) indica que para mejorar los supuestos de un modelo de regresión lineal una estrategia válida consiste en establecer métodos de regresión más complejos, sin embargo también propone estrategias más simples como realizar transformaciones en los datos, de tal forma que le permita al usuario utilizar el mismo método de regresión, esta librería particularmente analiza cual de las diferentes transformaciones es la adecuada según los objetivos del usuario, análisis de los residuales del modelo de regresión, pruebas estadísticas y gráficas, asociación lineal de las variables transformada respecto a las variables auxiliares, entre otras que fueron de ayuda para definir cual era la mejor transformación y los mejores lambda para las respectivas transformaciones.

4.4. Información auxiliar

Las variables que se usaron como auxiliares fueron tomadas de los registros administrativos y censos realizados por el Departamento Nacional de Planeación DNP para el departamento de Cundinamarca en el 2016, de estos se destacan los registros de los hacinamientos de vivienda, tasa de cobertura bruta, base catastral, deserción educativa, referente al 2017 se tienen los registros del desempeño integral, Población total afiliada a EPS contributiva.

En la tabla 1 se describen algunas de las variables que fueron seleccionadas para la estimación de los indicadores propuestos.

Variable	Descripción
	Acceso a Bienes por Hogar los cuales son importantes para obtener un acercamiento al nivel de calidad de vida en los hogares
1.	(Automovil) Porcentaje de hogares del municipio que posee automóvil
2.	(Lavadora) Porcentaje de hogares del municipio que posee Lavadora
3.	Porcentaje de hogares del municipio que posee Computador
4.	(motocicleta) Porcentaje de hogares del municipio que posee Motocicleta
5.	Porcentaje de hogares del municipio que posee Televisión por cable
6.	(Calentador ducha) Porcentaje de hogares del municipio que posee Calentador de ducha eléctrica
7.	Porcentaje de hogares por tenencia de servicios de teléfono fijo e internet
	Superávit o déficit fiscal: Refleja la diferencia entre los ingresos y los gastos.
8.	(desempeno fiscal 2017) Medición del desempeño de la gestión financiera de las entidades territoriales de Cundinamarca
9.	(salario total) Salarios mensuales por hogar
	Deserción estudiantil durante los diferentes ciclos académicos. ()
10.	(Transición o m transicion) Hasta transición.
11.	(m primaria o Primaria) Primaria (Hasta Quinto Grado)
12.	(m secundaria o Secundaria) Secundaria (Hasta Séptimo grado)
13.	(m basica o Básica)Básica (Grado noveno)
14.	(m media o media) Media (Grado once)
15.	(c total) Deserción total
16.	(num est educ) Número de establecimientos educativos
17.	(hombres afliados a iss o nueva eps) Afiliación al régimen de salud contributivo o subsidiado
18.	(bienes raices) (No poseen bienes raices) Hogares que no poseen bienes raíces
19.	(Mujeres_trabajando) Porcentaje de mujeres trabajando
20.	Población en cabecera municipal
21.	Viviendas cercanas a lugares o establecimientos que pueden causar afectación
22.	Hogares por acceso a servicios públicos, privados o comunales
23.	Hogares por percepción del jefe(a) o cónyuge sobre el nivel de vida actual del hogar, con respecto al de hace 5 años
24.	Hogares por percepción del jefe(a) o cónyuge respecto a las condiciones actuales de vida de su hogar
25.	Niños y niñas menores de 5 años por sitio o persona con quien permanecen la mayor parte del tiempo entre semana, según localidad, área rural
26.	Personas de 10 años y más, por veces a la semana que en los últimos 30 días practicaron deporte o realizaron actividad física durante 30 minutos continuos o más, según localidad área rural
27.	Años promedio de educación para personas de 15 años y más, por grupos de edad, según localidad área rural
28.	(Avalúo rural)Valor comercial, el cual se encuentra ubicado en el área de encuesta.
29.	Personas de 5 años y más que usan internet, por tipo de uso, según localidad área rural
30.	(Empleado) Porcentaje de personas que trabajan como empleado.
31.	(área_cabecera_ha) Área geográfica que está definida por un perímetro urbano.

Tabla 1: Información auxiliar

5. Análisis de resultados

En este trabajo, se emplean las transformaciones estadísticas, Logaritmo, Log-Shift, Box-Cox y Dual, en modelos a nivel de área, específicamente el modelo FH. El objetivo principal es estimar los indicadores de interés, es decir, los ingresos promedio y la tasa de desempleo en Cundinamarca durante el año 2017. Además, se busca determinar, a través de la medida de calidad de cada modelo (coeficiente de variación), cuál es la transformación y las variables auxiliares más adecuadas para la estimación precisa de cada uno de estos indicadores,asumiendo el menor valor del CV.

Para la selección de las variables auxiliares pertinentes en cada modelo, se consideraron aquellas con un mayor poder predictivo. Se utilizó el criterio de información Akaike (AIC) para cada transformación y cada indicador, el cual evalúa la calidad relativa de los modelos, de manera similar a como se hace en los modelos lineales generalizados. Es importante destacar que este análisis se llevó a cabo de manera distinta en los modelos a nivel de área, específicamente en los modelos FH, debido a la necesidad de tener en cuenta los efectos aleatorios del modelo.

Para esta tarea, se empleó la librería .$^{\text{em}}$di", la cual analiza minuciosamente cada variable, realizando un algoritmo paso a paso y sugiriendo las variables finales más adecuadas para el modelo FH, junto con sus respectivos valores de AIC (Criterio de Información de Akaike) en este contexto particular. Además, es importante destacar que esta librería ofrece la flexibilidad de realizar un algoritmo paso a paso teniendo en cuenta otros criterios, lo que la convierte en una herramienta valiosa para futuros estudios derivados de este trabajo.

En seguida se escribe el paso a paso del calculo de de las estimaciones de cada indicador para cada transformación.

1. Se selecciona la transformación y se obtiene $T_\lambda = y_j^*(\lambda)$

Para la transformación de los datos y el cálculo del modelo es es necesario transformar la varianza, tal y como lo sugiere (Zhang and Rojas, 2010), es por ello que por medio de la descomposición de Taylor (ver ecuación 27) se realizó la respectiva transformación.

2. Se realiza la estimación del indicador con la variable transformada.

$y_j^*(\lambda) = x_d^T\beta + u_d + e_d, \;\; e_d \sim N(0, \hat{\sigma_d^2}), \;\; u_d \sim N(0, \sigma_u^2)$

2.1 Se aplica el algoritmo paso a paso para $y_j^*(\lambda)$ el cual determina las variables auxiliares más pertinentes para la transformación.

3. Se realiza la transfomación inversa $y_j^*(\lambda)$ a la escala original $y_j = T_\lambda^{-1}(y_j^*(\lambda))$

4. Se cálcula ECM haciendo uso de la generalización de la descomposición de la varianza de Taylor demostrando que

$$ECM(\hat{\theta}) = f^{-1}(\hat{y}_{j,eblup}) \cdot ECM(y_{j,\hat{eblup}}) \tag{24}$$

Donde f^{-1} es la función inversa de la transformación, j son los dominios, $\hat{y}_{j,eblup}$ es la estimación del parámetro con la transformación realizada y $ECM(y_{j,\hat{eblup}})$ es el ECM producido por esa estimación con la transformación usada.

Para mayores detalles de las ecuaciones y demostraciones consulte la sección 7

5.1. Tasa de desempleo en Cundinamarca

En la tabla 2 se presentan los valores correspondientes de λ que optimizan la función en relación a las transformaciones propuestas para la estimación de la tasa de desempleo en Cundinamarca.

Transformación	Lambda
Log - Shift	9.998
Box - Cox	2.998
Dual	6.610696e-05

Tabla 2: λ que maximiza la función

Se puede evidenciar que los valores de λ que se presentan en la tabla 2 los identifica como los valores pertinentes para cada una de las transformaciones identificando que para la transformación Log-Shift y Box-cox $\lambda > 1$, por lo tanto logra comprimir los valores más grandes y estira los valores más pequeños en los datos. Para el valor de la transformación Dual se tiene que el λ produce una transformación no lineal con un efecto que depende de ese valor específico.

Ahora bien, en la tabla 3 se muestra los valores p, que validan la normalidad en cada transformación.

Transformación	shapiro y Wilk P
Logaritmo	0.6551
Log- Shift	0.531
Box - Cox	0.403
Dual	0.6551

Tabla 3: Valores P, prueba de normalidad

En la tabla 3, se observa que los valores P asociados a las transformaciones estudiadas en esta investigación son superiores al nivel de significancia de 0.05, lo que indica que no hay evidencia suficiente para rechazar la hipótesis de normalidad. Los datos transformados parecen seguir una distribución normal.

5.1.1. Estimación de la tasa de desempleo sin realizar transformaciones

Realizando el algoritmo paso a paso, las variables seleccionadas para la estimación de la tasa de desempleo en Cundinamarca en el 2017 sin usar transformaciones se muestra en la tabla 4

Coefficients:	coefficients	std.error	t.value	p.value	
Intercepto	3.43E-02	5.32E-03	6.4528	1.10E-10	***
Empleado	9.13E-07	5.83E-07	1.5653	0.117516	
mujeres_trabajando	-2.18E-06	9.69E-07	-2.2503	0.024433	*
Transicion	8.82E-05	5.08E-05	1.7386	0.082105	.
Secundaria	8.26E-05	3.05E-05	2.7074	0.00678	**
Basica	-2.76E-05	1.27E-05	-2.1768	0.029493	*
Media	-7.29E-05	2.38E-05	-3.0561	0.002243	**
Primaria	1.74E-03	6.67E-04	2.607	0.009135	**
c_total	-2.68E-03	1.01E-03	-2.6524	0.007991	**
Calentador ducha	-1.01E-05	4.28E-06	-2.3563	0.018456	*
AIC	-247.16				

Tabla 4: Variables auxiliares para el modelo FH sin transformación

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 4, considerando el criterio de menor AIC.

5.1.2. Estimación de la tasa de desempleo con la transformación logaritmo

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación de la tasa de desempleo en Cundinamarca en el 2017 usando la transformación logaritmo se muestra en la tabla 5

coefficients	std.error	t.value	p.value		
(Intercept)	-5.50E+00	5.60E-01	-9.825848	8.71E-23	***
num_est_educ	-5.58E-04	1.57E-04	-3.552651	3.81E-04	***
hombres_afliados a iss_o_nueva_eps	-2.20E-04	1.02E-04	-2.16157	3.07E-02	*
No_poseen_bienes_raices	9.37E-05	2.66E-05	3.518856	4.33E-04	***
mujeres_trabajando	-1.32E-04	2.09E-05	-6.345596	2.22E-10	***
area_cabecera_ha	-3.64E-04	2.06E-04	-1.764352	7.77E-02	.
m_transicion	4.60E-03	1.24E-03	3.700196	2.15E-04	***
m_secundaria	4.03E-03	7.80E-04	5.17018	2.34E-07	***
m_basica	-1.70E-03	3.46E-04	-4.909029	9.15E-07	***
motocicleta	-3.65E+00	1.04E+00	-3.503905	4.58E-04	***
bienes_raices	1.55E+00	4.84E-01	3.211051	1.32E-03	**
m_media	-1.80E-03	5.82E-04	-3.094558	1.97E-03	**
m_primaria	8.04E-02	1.62E-02	4.946907	7.54E-07	***
c_total	-1.09E-01	2.51E-02	-4.36174	1.29E-05	***
Calentador de ducha	-5.26E-04	8.96E-05	-5.869988	4.36E-09	***
desempeno_fiscal_2017	2.11E-02	5.85E-03	3.602167	3.16E-04	***
Empleado	4.14E-05	1.48E-05	2.79175	5.24E-03	**
AIC	-20.36937				

Tabla 5: Variables auxiliares para el modelo FH, logaritmo

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 5, considerando el criterio de menor AIC.

5.1.3. Estimación de la tasa de desempleo con la transformación Log -Shift

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación de la tasa de desempleo en Cundinamarca en el 2017 usando la transformación Log - Shift se muestra en la tabla 6

coefficients	std.erro	r t.valu	e p.value		
(Intercept)	2.30E+00	2.86E-03	803.2701	¡2.2e-16	***
hogares_que_no_poseen_bienes_raices	2.99E-07	6.99E-08	4.2833	1.84E-05	***
mujeres_trabajando	-3.05E-07	9.39E-08	-3.2453	0.001173	**
m_secundaria	4.81E-06	1.90E-06	2.531	0.011373	*
m_basica	-1.51E-06	6.73E-07	-2.2371	0.025281	*
motocicleta	-1.64E-02	5.17E-03	-3.1621	0.001567	**
bienes_raices	7.46E-03	2.31E-03	3.2246	0.001261	**
no_total_de_predios	-5.62E-08	2.95E-08	-1.9053	0.056735	.
m_media	-2.59E-06	1.25E-06	-2.0692	0.038528	*
c_primaria	8.89E-05	3.69E-05	2.4113	0.015894	*
c_total	-9.66E-05	3.99E-05	-2.4203	0.015509	*
Calentador ducha	-1.29E-06	4.27E-07	-3.0225	0.002507	**
desempeno_fiscal_2017	4.67E-05	3.20E-05	1.4584	0.144721	*
AIC	-407.087				

Tabla 6: Variables auxiliares para el modelo FH, Log - Shift

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 6, considerando el criterio de menor AIC.

5.1.4. Estimación de la tasa de desempleo con la transformación Box - Cox

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación de la tasa de desempleo en Cundinamarca en el 2017 usando la transformación Box-Cox se muestra en la tabla 7

coefficients	std.error	t.value	p.value		
(Intercept)	-3.33E-01	1.10E-05	-30279.8671	¡2.2e-16	***
Ingresos	-2.57E-11	1.67E-11	-1.5385	0.123921	.
num_est_educ	2.07E-08	9.18E-09	2.2533	0.024239	*
mujeres_afliados_a_iss_o_nueva_eps	9.53E-09	4.81E-09	1.9829	0.047375	*
m_secundaria	2.54E-08	9.16E-09	2.7673	0.005652	**
motocicleta	-9.18E-05	6.17E-05	-1.4878	0.136816	.
bienes_raices	4.66E-05	2.53E-05	1.8411	0.065602	.
m_media	-6.48E-08	2.28E-08	-2.8375	0.004547	**
Calentador ducha	-6.39E-09	3.75E-09	-1.7014	0.088861	.
AIC	-417.27				

Tabla 7: Variables auxiliares para el modelo FH, Box - Cox

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 7, considerando el criterio de menor AIC.

5.1.5. Estimación de la tasa de desempleo con la transformación Dual

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación de la tasa de desempleo en Cundinamarca en el 2017 usando la transformación Dual se muestra en la tabla 8

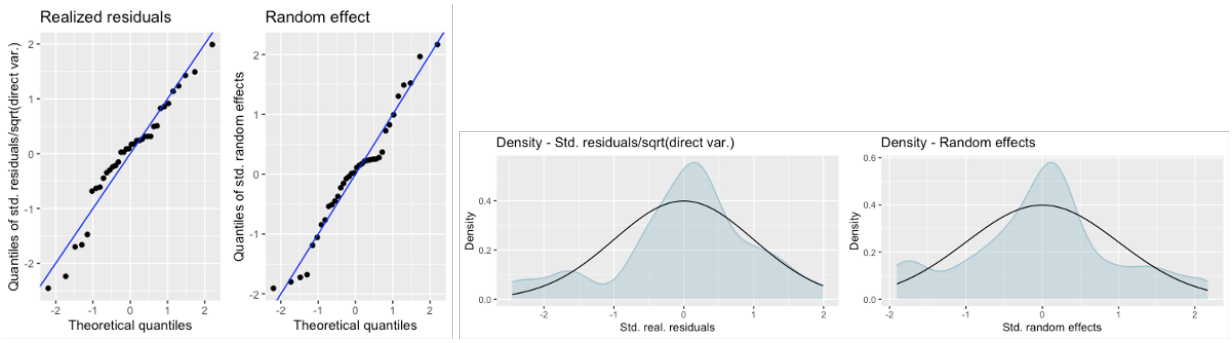
	coefficients	std.error	t.value	p.value	
(Intercept)	-3.38624	0.92975	-3.6421	0.0002704	***
AIC	196.04				

Tabla 8: Variables auxiliares para el modelo FH, Dual

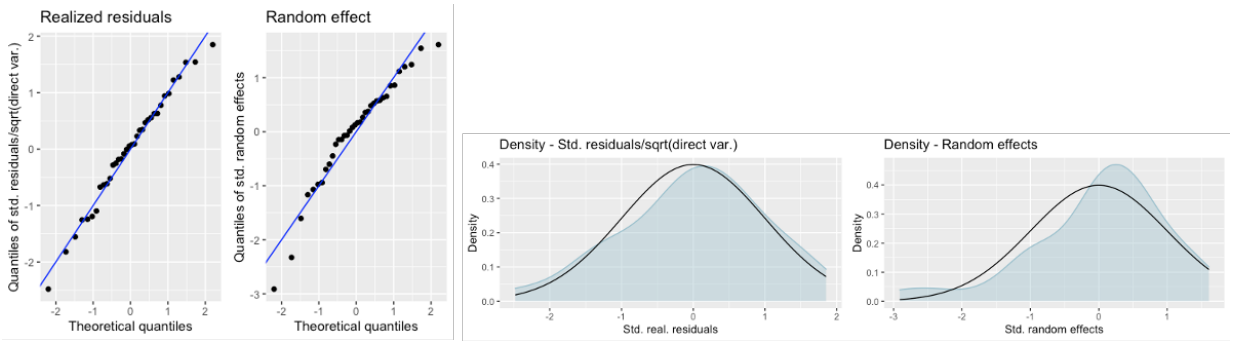
El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 8, considerando el criterio de menor AIC.

5.1.6. Análisis de los residuales para la tasa de desempleo

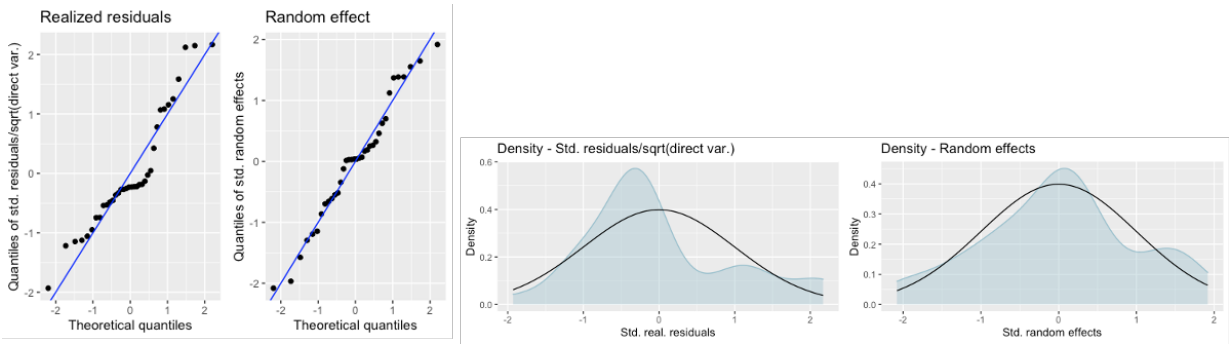
En la figura 1 se puede evidenciar que los residuales del modelo que se determina por medio de la transformación Box - Cox tienen mejor comportamiento que las demás transformaciones.



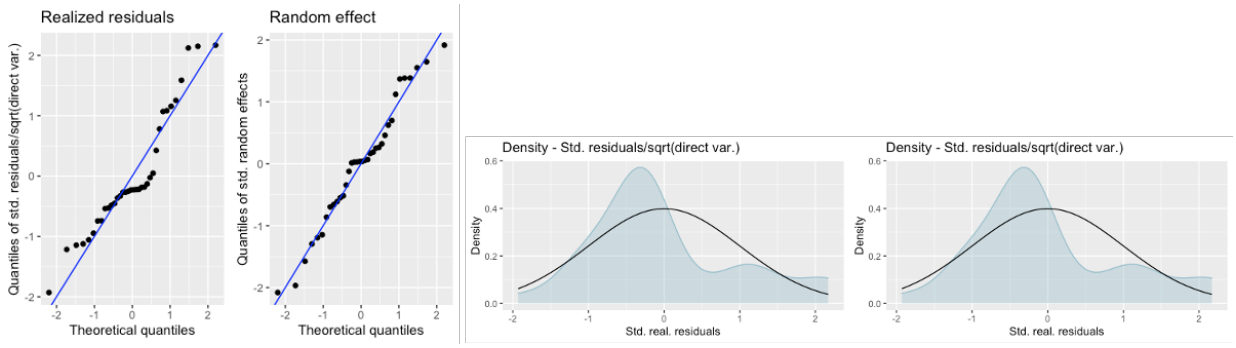
(a) Residuales logaritmo



(b) Residuales Log - Shif



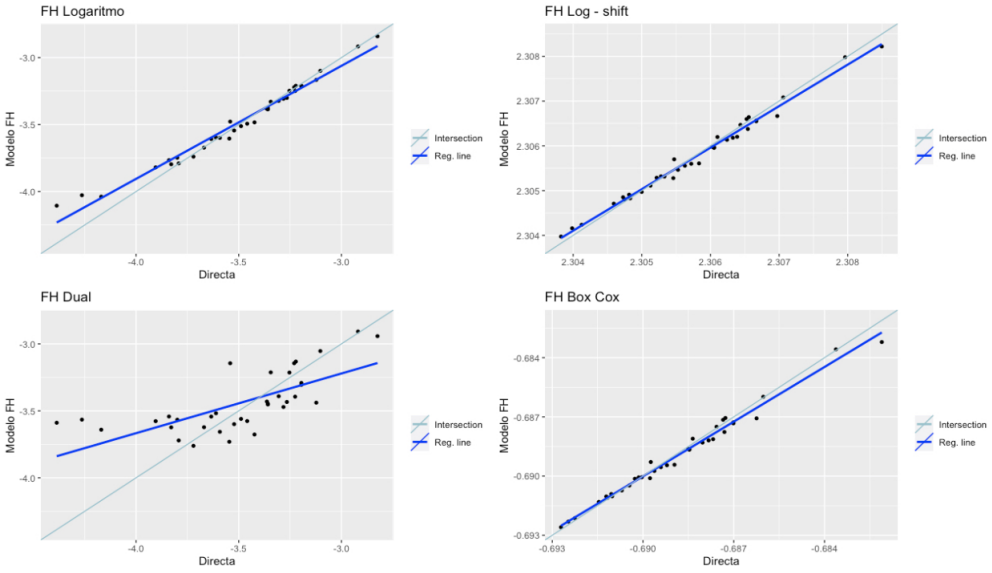
(c) Residuales Box - Cox



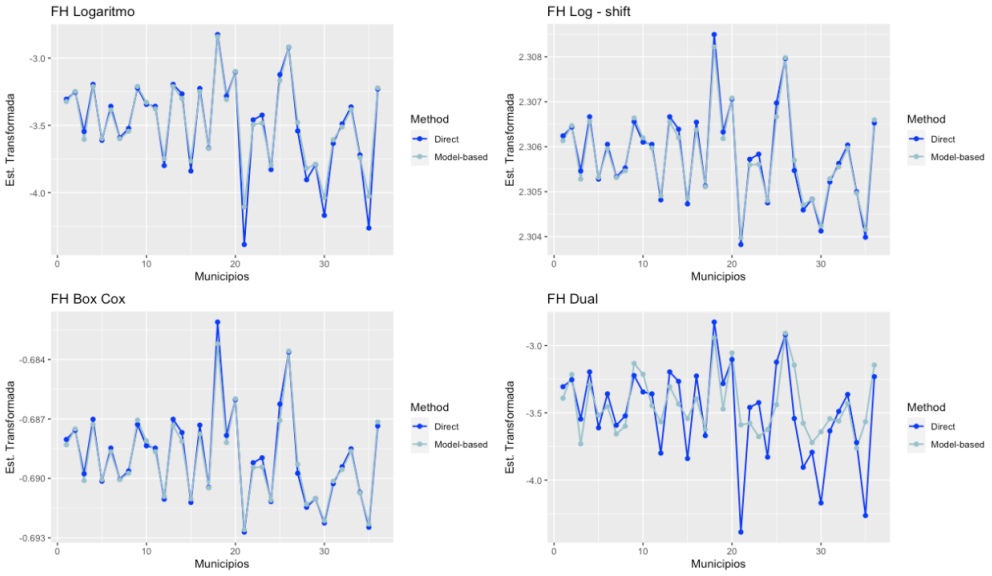
(d) Residuales Dual

Figura 1: Análisis de los residuales para la estimación de la tasa de desempleo según la transformación

Ahora bien en la figura 2, se puede evidenciar que el modelo ajustado a la transformación Box Cox, gráfica 2(b), es el que mejor se ajusta a los datos transformados, además en el diagrama 2(a) la linea de ajuste del modelo se asemeja mucho más a la linea de identidad de la transformación de Box Cox, por lo tanto es pertinente afirmar que el modelo FH con la transformación Box - Cox es el que tiene estimaciones más confiables.



(a) Diagrama de dispersión estimaciones directas y modelo FH



(b) Diagrama de lineas de la estimaciones directas y del modelo

Figura 2: Ajuste del modelo para la estimación de la tasa de desempleo según la transformación

5.1.7. Análisis de los CVs para las estimaciones de la tasa de desempleo

En la figura 3 se muestran las estimaciones de los CVs de cada modelo realizando las respectivas transformaciones en los datos.

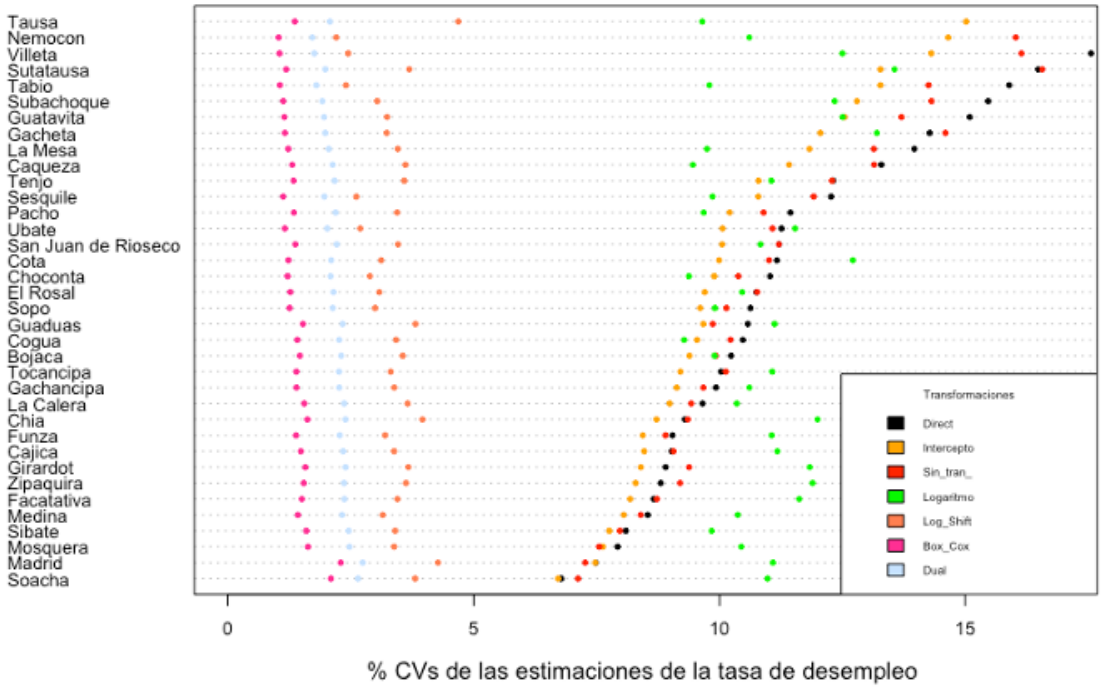


Figura 3: Estimaciones de los CV

Como se evidencia en la figura 3, los coeficientes de variación alcanzan su valor más bajo cuando se utiliza la transformación Box-Cox. Basándonos en los análisis previos, podemos concluir de manera apropiada que esta transformación es la que mejor se ajusta a los datos, lo que se traduce en una notable disminución de los valores de los coeficientes de variación.

En la tabla 9 se muestra las estimaciones de los CVs que presentan menor valor en su estimación.

Municipio	Dir_CV	int CV	Sin_Trans_CV	log_CV	log_shift_CV	Box_Cox_CV	Dual_CV
Soacha	6.78	6.72	7.12	10.98	3.81	2.10	2.65
Madrid	7.48	7.47	7.27	11.08	4.27	2.30	2.74
Mosquera	7.93	7.63	7.55	10.44	3.38	1.63	2.47
Sibate	8.09	7.76	7.97	9.84	3.40	1.60	2.46
Medina	8.54	8.05	8.40	10.37	3.15	1.42	2.32
Facatativa	8.66	8.18	8.73	11.62	3.45	1.51	2.37
Zipaquira	8.80	8.30	9.20	11.89	3.62	1.54	2.36
Girardot	8.90	8.40	9.38	11.83	3.67	1.58	2.39
Cajica	9.03	8.47	9.07	11.18	3.38	1.49	2.35
Funza	9.04	8.44	8.90	11.06	3.20	1.39	2.28

Tabla 9: CVs Tasa de desempleo con las diferentes transformaciones

En la siguiente lista se describen las abreviaturas de la tabla 9.

Dir CV: Coeficiente de la estimación directa para la tasa de desempleo.
Int CV: Coeficiente de variación del modelo FH solo con el intercepto.
Sin trans CV: Coeficiente de variación del modelo FH sin transformación.
Log CV: Coeficiente de variación del modelo FH con la transformación logaritmo.
Log-Shift CV:Coeficiente de variación del modelo FH con la transformación log - Shift.
Box-Cox CV: Coeficiente de variación del modelo FH con la transformación Box-Cox.
Dual CV: Coeficiente de variación del modelo FH con la transformación Dual.

Los datos sugieren que la transformación de Box-Cox es la que mejor se adapta a los datos, ya que se observa el coeficiente de variación más bajo en comparación con las otras transformaciones.

Teniendo en cuenta la conclusión anterior en la gráfica 4 se muestran las estimaciones de la tasa de desempleo en cundinamarca en el año 2017, realizando estimaciones por medio de los modelos de FH con la mejor transformación, Box-Cox.

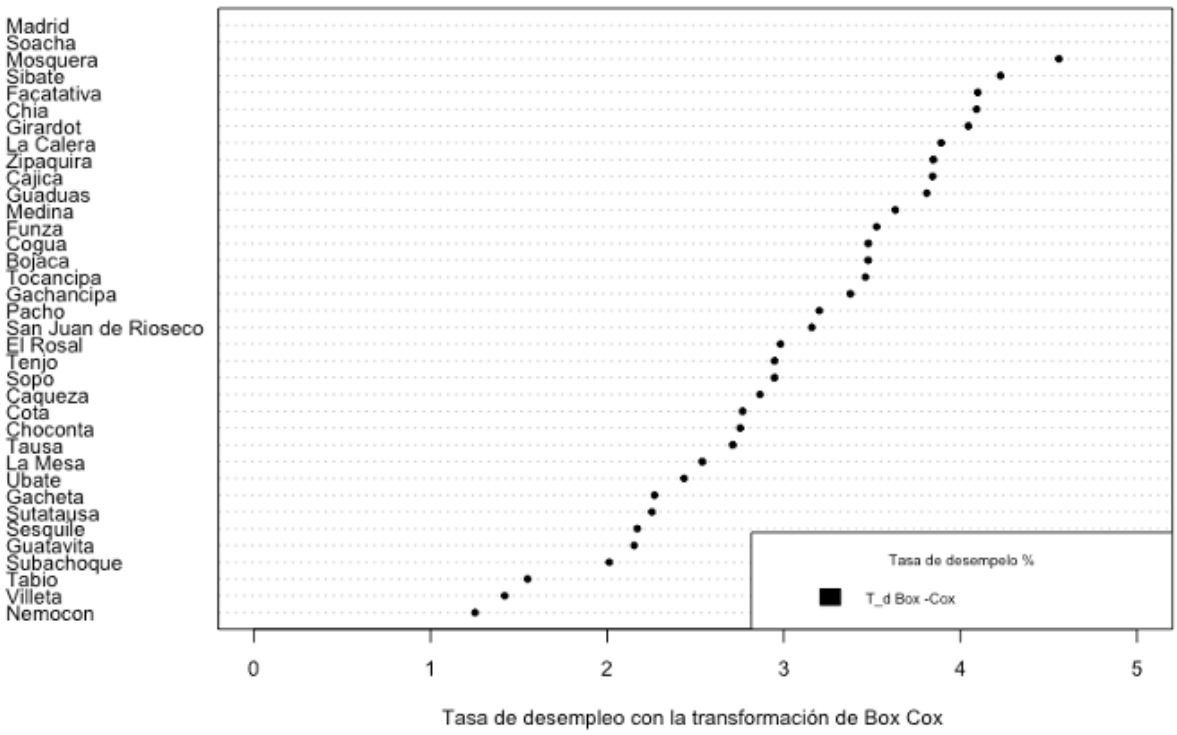


Figura 4: Tasa de desempleo en el 2017

La figura 4 muestra los municipios con los mayores porcentajes de tasa de desempleo en Cundinamarca durante el año 2017, destacándose Madrid, Soacha, Mosquera y Sibate. En contraste, los municipios con las tasas de desempleo más bajas en la misma región durante ese año son Nemocon, Villeta y Tabio.

Para obtener una visión más detallada de las estimaciones de la tasa de desempleo en estos municipios, se presenta la tabla 10, que incluye las estimaciones realizadas mediante la transformación Box-Cox en los datos correspondientes a los municipios con las tasas más altas de desempleo en Cundinamarca en

2017.

Municipio	Tasa Desempleo Box - Cox
Madrid	5.54654
Soacha	5.414495
Mosquera	4.509763
Facatativa	4.078354
Sibate	4.070749
Zipaquira	4.04026
Chia	3.98704
Girardot	3.966426
Cajica	3.894523
La calera	3.782311
Funza	3.63718
Guaduas	3.626184
Medina	3.599832
Bojaca	3.55244
Tocancipa	3.410413
Gachancipa	3.399196
Cogua	3.395286
Sopo	3.110858
San Juan de Rioseco	3.046995
Pacho	3.035118

Tabla 10: Tasa de desempleo en Cundinamarca 2017

La tabla 10 revela que los municipios listados muestran tasas de desempleo que se sitúan principalmente en el rango del 3.03 % al 5.55 % de su población municipal.

5.2. Ingreso Promedio

En la tabla 11 se presentan los valores correspondientes de λ que optimizan la función en relación a las transformaciones propuestas para la estimación del ingreso promedio en Cundinamarca.

Transformación	Lambda
Log - Shift	-566501.7
Box - Cox	0.408
Dual	4.837936e-05

Tabla 11: λ que maximiza la función

De la tabla 11 se puede evidenciar que la transformación Log-Shift tiene un valor de $\lambda < 0$ lo que indica que el parámetro λ esta comprimiendo los valores más pequeños y esta estirando los valores más grandes. Ahora bien, el valor de $\lambda > 0$ para la tranformación Box-Cox cumple con una función similar al paramétro de la transformación Log - Shift, es decir, comprime los valores más pequeños y estira los valores más grandes. Para la transformación Dual se tiene que el valor λ es diferente de 0, 1 y 2, por lo tanto λ produce una transformación no lineal con un efecto que depende de ese valor específico.

En la tabla 12 se muestra los valores P, que validan la normalidad en cada transformación.

En la tabla 12 , se observa que los valores P asociados a las transformaciones estudiadas en esta investigación son superiores al nivel de significancia de 0.05, lo que indica que no hay evidencia suficiente para rechazar la hipótesis de normalidad. Los datos transformados parecen seguir una distribución normal.

Transformación	shapiro y Wilk P
Logaritmo	0.9228
Log- Shift	0.7925
Box - Cox	0.7172
Dual	0.9228

Tabla 12: Valores P, prueba de normalidad

5.2.1. Estimación del ingreso promedio sin realizar transformaciones

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación del ingreso promedio en Cundinamarca en el 2017 sin usar transformaciones se muestra en la tabla 13

	coefficients	std.error	t.value	p.value	
(Intercept)	-4.18E+06	1.83E+06	-2.2928	0.021861	**
Salario total	1.55E+00	4.71E-01	3.2852	0.001019	**
lavadora	-2.95E+06	1.06E+06	-2.7745	0.005529	***
automovil	1.36E+07	3.02E+06	4.5072	6.57E-06	**
Avaluo rural	2.08E+05	7.61E+04	2.7398	0.006148	*
C_transicion	3.88E+03	2.30E+03	1.6841	0.092156	*
AIC	1028				

Tabla 13: Variables auxiliares para el modelo FH, sin transformaciones

El mínimo número de parámetros para estimar el ingreso promedio, se muestra en la tabla 13, considerando el criterio de menor AIC.

5.2.2. Estimación del ingreso promedio con la transformación logaritmo

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación del ingreso promedio en Cundinamarca en el 2017 usando la transformación logaritmo se muestra en la tabla 14

	coefficients	std.error	t.value	p.value	
(Intercept)	1.19E+01	6.81E-01	17.474	¡2.2e-16	***
calentador_ducha	4.13E-01	2.47E-01	1.67	0.0949219	.
salario total	6.75E-07	1.67E-07	4.0469	5.19E-05	***
lavadora	-1.68E+00	4.61E-01	-3.6388	0.0002739	***
automovil	4.57E+00	1.21E+00	3.7626	0.0001681	***
Avaluo rural	6.08E-02	3.07E-02	1.9796	0.0477531	*
Desempeño_fiscal	1.15E-02	5.98E-03	1.9283	0.0538218	.
AIC	1030.24				

Tabla 14: Variables auxiliares para el modelo FH, logaritmo

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 14, considerando el criterio de menor AIC.

5.2.3. Estimación del ingreso promedio con la transformación Log-Shift

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación del ingreso promedio en Cundinamarca en el 2017 usando la transformación Log-Shift se muestra en la tabla 15

	coefficients	std.error	t.value	p.value	
(Intercept)	1.07E+01	9.09E-01	11.8112	¡2.2e-16	***
calentador_ducha	5.89E-01	3.30E-01	1.7859	0.0741172	.
Salario total	8.68E-07	2.23E-07	3.9003	9.61E-05	***
lavadora	-2.21E+00	6.16E-01	-3.5823	0.0003406	***
automovil	5.80E+00	1.62E+00	3.5787	0.0003453	***
Avaluo rural	7.31E-02	4.11E-02	1.7808	0.0749381	.
Desempeño_fiscal	1.75E-02	8.02E-03	2.1849	0.0288971	*
AIC	-15.86				

Tabla 15: Variables auxiliares para el modelo FH, log-Shift

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 15, considerando el criterio de menor AIC.

5.2.4. Estimación del ingreso promedio con la transformación Box-Cox

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación del ingreso promedio en Cundinamarca en el 2017 usando la transformación Box - Cox se muestra en la tabla 16

	coefficients	std.error	t.value	p.value	
(Intercept)	2.44E+00	1.32E-03	1848.7666	¡2.2e-16	***
calentador_ducha	1.26E-03	8.56E-04	1.467	0.1423659	.
Salario total	1.94E-09	5.14E-10	3.778	0.0001581	***
lavadora	-4.21E-03	1.61E-03	-2.6227	0.0087245	**
automovil	1.15E-02	4.18E-03	2.7568	0.0058374	**
Desempeño_fiscal	4.43E-05	1.89E-05	2.3403	0.0192705	*
AIC	-462.0257				

Tabla 16: Variables auxiliares para el modelo FH, Box-Cox

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 16, considerando el criterio de menor AIC.

5.2.5. Estimación del ingreso promedio con la transformación Dual

Realizando el algoritmo paso a paso las variables seleccionadas para la estimación del ingreso promedio en Cundinamarca en el 2017 usando la transformación Dual se muestra en la tabla 17

	coefficients	std.error	t.value	p.value	
(Intercept)	1.19E+01	6.80E-01	17.5033	¡2.2e-16	***
calentador_ducha	4.10E-01	2.47E-01	1.6618	0.096549	.
Salario total	6.73E-07	1.66E-07	4.0451	5.23E-05	***
lavadora	-1.68E+00	4.60E-01	-3.6408	0.0002718	***
automovil	4.55E+00	1.21E+00	3.7623	0.0001684	***
Avaluo rural	6.07E-02	3.06E-02	1.9827	0.0474008	*
Desempeño_fiscal	1.16E-02	5.92E-03	1.9613	0.0498459	*
AIC	-29.76806				

Tabla 17: Variables auxiliares para el modelo FH, Dual

El mínimo número de parámetros para estimar la tasa de desempleo en Cundinamarca, se muestra en la tabla 17, considerando el criterio de menor AIC.

5.2.6. Análisis de los CVs para las estimaciones del ingreso promedio

Con base en los modelos previamente propuestos, la figura 18 presenta un resumen de los valores de los coeficientes de variación (CV) correspondientes a cada una de las transformaciones aplicadas a los datos.

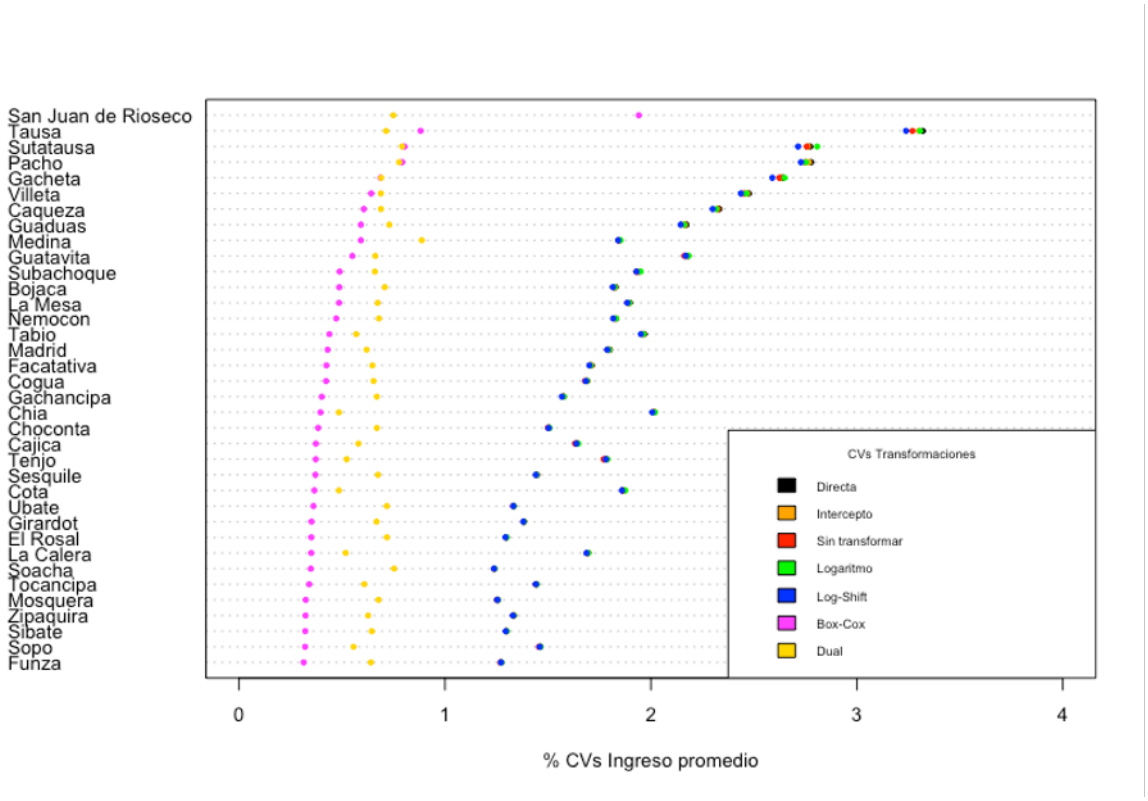


Figura 5: Estimaciones de los CVs ingreso promedio

De acuerdo a la figura 5, es evidente que la transformación que mejor se adecua a los datos, evaluada en función del coeficiente de variación, es la transformación de Box-Cox.

La tabla 18 presenta los municipios que muestran los mayores porcentajes en el coeficiente de variación, considerando cada una de las transformaciones aplicadas.

Municipio	CV_directa	CV_fh_intercepto	CV_sin_tran	CV_fh_log	CV_log_shift	CV_Box_Cox	CV_dual
Funza	1.274	1.273	1.269	1.274	1.272	0.314	0.641
Sopo	1.464	1.463	1.456	1.464	1.46	0.321	0.556
Sibate	1.299	1.298	1.297	1.3	1.295	0.322	0.645
Zipaquira	1.335	1.334	1.33	1.333	1.33	0.323	0.627
Mosquera	1.257	1.256	1.251	1.256	1.254	0.324	0.677
Tocancipa	1.445	1.444	1.44	1.445	1.441	0.341	0.608
Soacha	1.241	1.24	1.238	1.24	1.238	0.349	0.753
La Calera	1.696	1.695	1.69	1.695	1.687	0.351	0.518
El Rosal	1.299	1.298	1.297	1.299	1.294	0.351	0.719
Girardot	1.384	1.382	1.383	1.384	1.38	0.352	0.667

Tabla 18: CVs Ingreso Promedio con las diferentes transformaciones

En la siguiente lista se describen los titulos de las abreviaturas que se muestran en la tabla 18:

CV_directa : CV de la estimación directa.
CV_fh_intercepto:CV del modelo FH solo usando el intercepto en el modelo.
CV_sin_tran: CV del modelo FH sin usar transformaciones.
CV_fh_log: CV del modelo FH usando la transformación Logaritmo.
CV_log_shift: CV del modelo FH usando la transformación Log - Shift.
CV_Box_Cox: CV del modelo FH usando la transformación Box - Cox.
CV_dual: CV del modelo FH usando la transformación Dual.

Teniendo en cuenta los resultados anteriores, en la figura 6 se muestran las estimaciones realizadas por medio de la transformación Box - Cox, correspondiente a el ingreso promedio Cundinamarca identificando los municipios de menor ingreso.

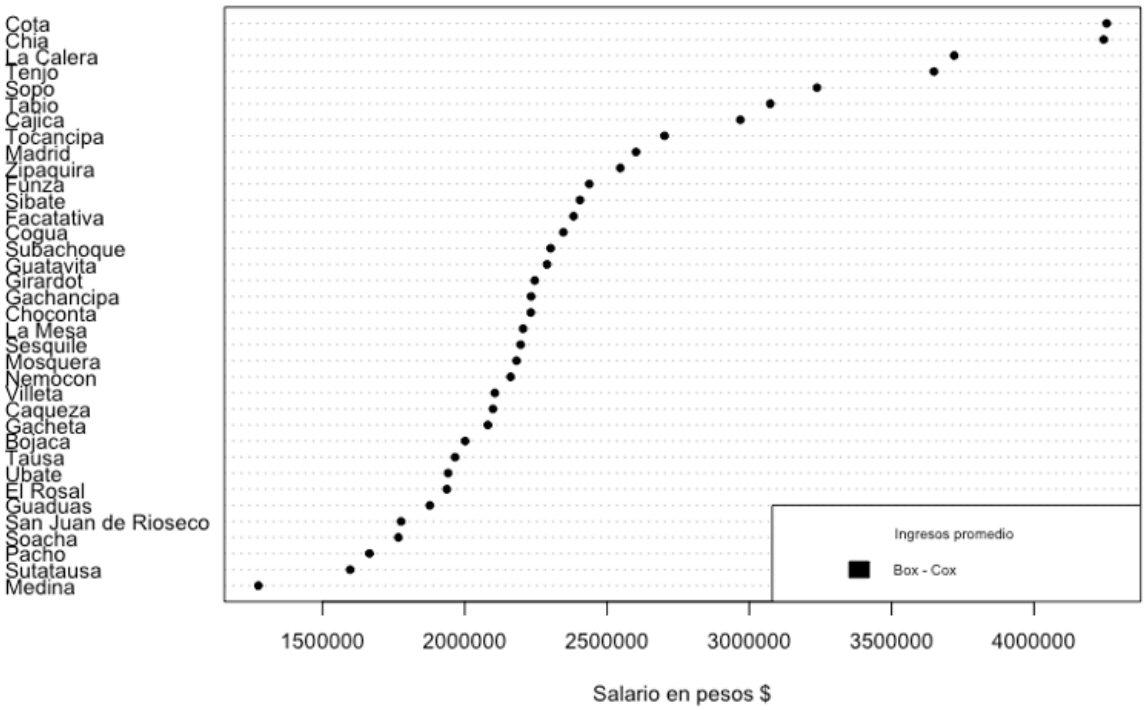


Figura 6: Ingresos promedio en Cundinamarca 2017

La figura 6 revela que, entre los municipios con los ingresos promedio más bajos, solo uno, Medina, registra un ingreso por debajo del salario mínimo. Por otro lado, es evidente que los municipios situados en las cercanías de la capital de Colombia son los que presentan los ingresos promedio más elevados, superando la cifra de $2,500,000.

Para una visualización detallada de los municipios con los ingresos promedio más bajos, se presenta la tabla. 25.

Municipio	fh_box_cox
Medina	1273859.30
Sutatausa	1598080.02
Pacho	1663156.87
Soacha	1766335.98
San Juan de Rioseco	1783751.05
Guaduas	1877029.15
El Rosal	1936578.79
Ubate	1941353.13
Tausa	1966763.95
Bojaca	2000787.68
Gacheta	2081936.66
Caqueza	2099434.17
Villeta	2105845.43
Nemocon	2160717.74
Mosquera	2181944.87
Sesquile	2196232.46
La Mesa	2203821.31
Choconta	2232221.64
Gachancipa	2232824.55

Tabla 19: Ingresos promedio

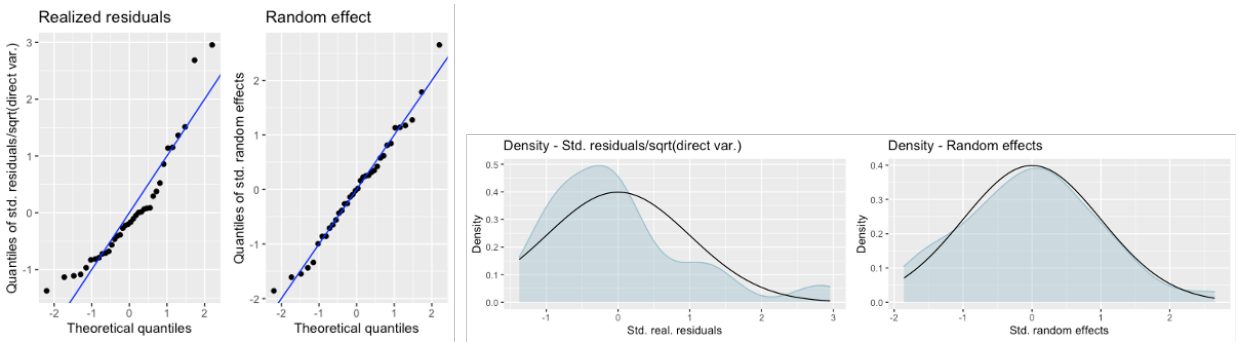
La tabla 6 muestra una diversidad considerable en los ingresos promedio de los municipios de Cundinamarca en 2017, con Medina teniendo el ingreso más bajo respecto a los demás municipios. El resto de los municipios por lo menos se estima que sus ingresos promedio superan el salario mínimo vigente.

5.2.7. Análisis de los residuales para el ingreso promedio

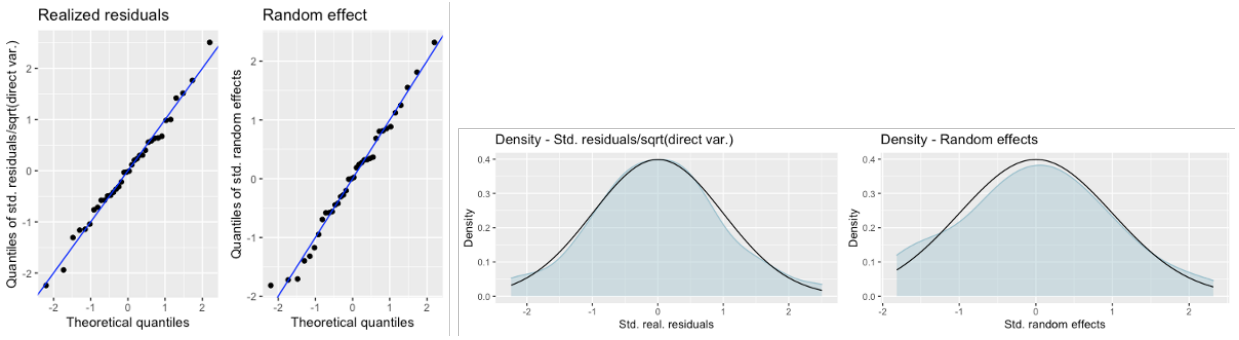
La figura 7 muestra el comportamiento de los residuales de cada modelo con el propósito de evaluar la estimación del indicador de interés en función de la transformación de los datos.

En la figura 7(c), se observa que los residuales del modelo basado en la transformación de Box-Cox presentan un comportamiento superior en comparación con las otras transformaciones.

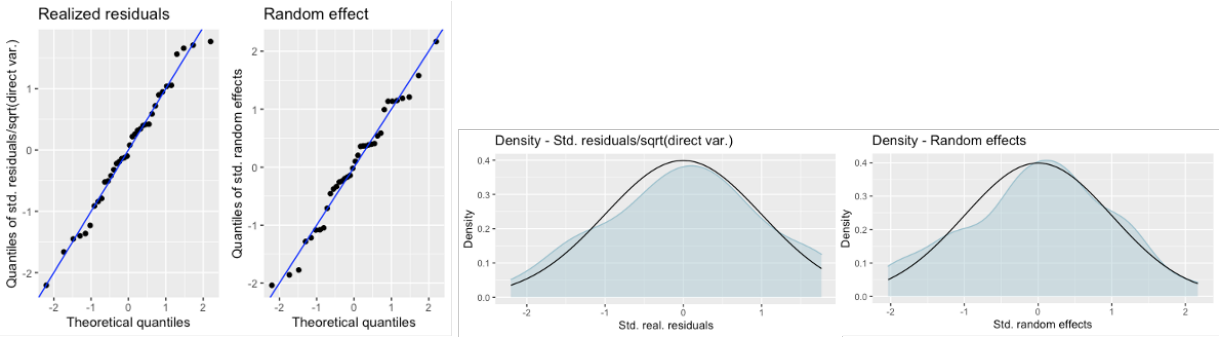
En la gráfica 8, se aprecia que el modelo ajustado a través de la transformación de Box-Cox, como se muestra en la figura 8(b), se ajusta mejor a los datos transformados. En consecuencia, es razonable suponer que este modelo tendrá un coeficiente de variación más bajo.



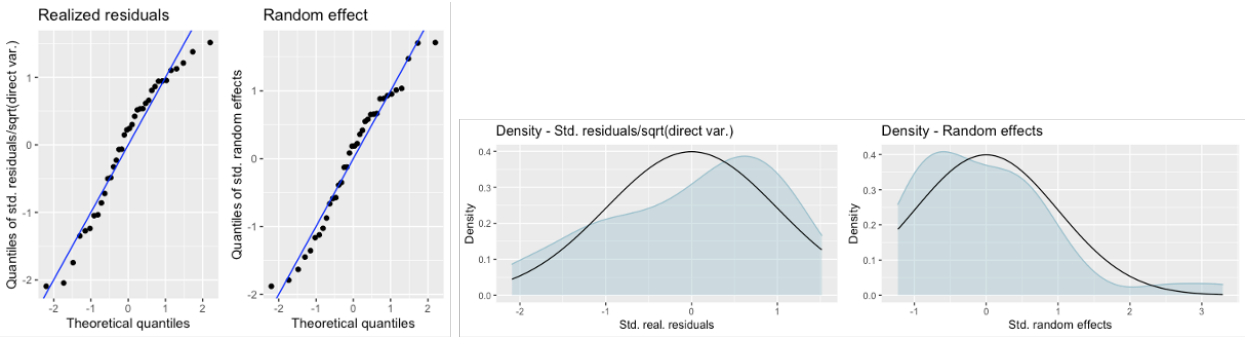
(a) Residuales logaritmo



(b) Residuales Log - Shif

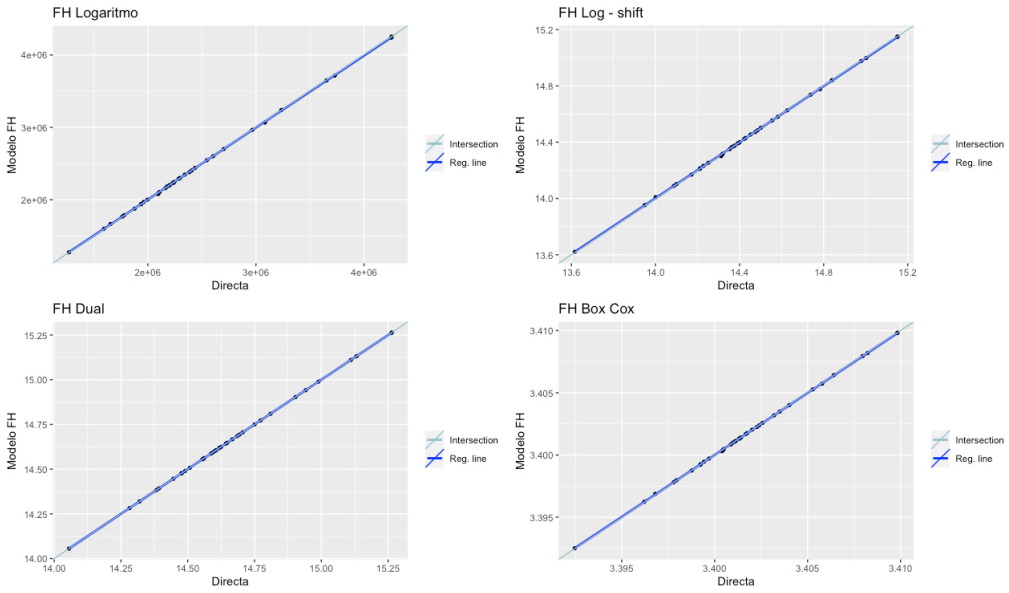


(c) Residuales Box - Cox

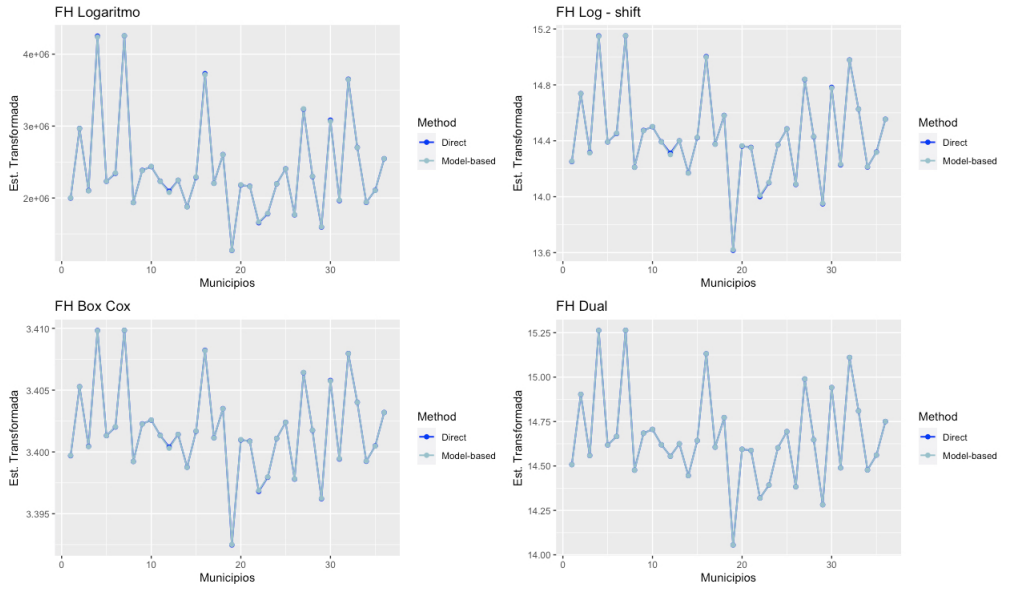


(d) Residuales Dual

Figura 7: Análisis de los residuales para la estimación del ingreso promedio según la transformación



(a) Diagrama de dispersión estimaciones directas y modelo FH



(b) Diagrama de lineas de la estimaciones directas y del modelo

Figura 8: Ajuste del modelo para la estimación del ingreso promedio según la transformación.

6. Conclusiones

Es evidente que las transformaciones aplicadas a los datos han mejorado varios aspectos de los modelos, incluida la calidad de las estimaciones y el comportamiento de los residuales. Esto se refleja claramente en la marcada disminución de los coeficientes de variación.

Las diversas estrategias utilizadas para evaluar el coeficiente de variación en función del error cuadrático medio de cada uno de los modelos transformados y la descomposición de la varianza, han fortalecido la confianza en las estimaciones realizadas en relación a los indicadores propuestos en este trabajo. Se ha determinado que la transformación Box-Cox es la transformación en los datos más adecuada tanto para estimar el ingreso promedio como la tasa de desempleo en los municipios de Cundinamarca en el 2017. Esta elección se basa en la observación de que los modelos transformados exhiben coeficientes de variación más bajos en comparación con las estimaciones directas y el modelo FH sin transformaciones, lo que resalta la calidad de los resultados obtenidos.

Para las transformaciones que necesitan el parámetro λ sería una nueva opción estimarlo con el fin de disminuir el error cuadrático medio u otros aspectos, ya que en este trabajo se estimo con el fin de mejorar los supuestos del modelo, tal vez para trabajos futuros pueda repercutir en la mejora de los coeficientes de variación de los modelos.

En cuanto a la selección de variables auxiliares para los modelos de Fay-Herriot, tanto transformados como no transformados, se destaca la utilidad del algoritmo paso a paso que ofrece la librería .$^{\text{em}}$di". Esta herramienta permite capturar los errores aleatorios del modelo, evaluar el AIC del modelo y evitar la saturación del mismo con variables que no contribuyen de manera significativa, lo que conduce a estimaciones más confiables al tener menor CV. Adicionalmente el algoritmo paso a paso de la librería permite tener en cuenta otros criterios de selección de la variable, lo que resultaría útil para próximos trabajos de investigación.

En relación a los indicadores de interés, se observa una relación inversa entre la tasa de desempleo y los ingresos promedio mensuales, sugiriendo una correlación negativa entre estas variables. Esto plantea la posibilidad de desarrollar un nuevo modelo basado en estas estimaciones que podría arrojar resultados mejorados en los indicadores de interés. Además, se propone una transformación adicional para la tasa de desempleo, utilizando la función arco seno, dado que esta variable se encuentra en el rango de 0 a 1, lo que la convierte en una opción interesante.

7. Anexos

7.1. Descomposición de Varianza

Zhang and Rojas (2010) indican que si T un estimador de θ, y g una función, entonces se tiene que

$$E(g(T)) \approx g(\theta) + g'(\theta)E(T-\theta) \tag{25}$$

$$Var(g(T)) \approx g'(\theta)^2 Var(T) \tag{26}$$

Por lo tanto se tiene que T es insesgado para θ, $E(g(T)) \approx g(\theta)$

Para la realización de las estimaciones se realizo un estudio exhaustivo del cálculo del MSE.

Para cada valor t que toma el estimador T, se hará una expansión de Taylor de primer orden para $g(t)$ alrededor del punto $t = \theta$, por ende se tiene que $g(t) \approx g(\theta) + g'\theta(T-\theta)$. Tomando la esperanza se tiene que $E(g(T)) \approx g(\theta) + g'(\theta)E(T-\theta)$. Es claro que cuando T es insesgado para $\theta, E(g(T)) \approx g(\theta)$.

Para la varianza se tiene que en $g(T) \approx g(\theta) + g'(\theta)(T-\theta)$

$$\begin{aligned} Var(g(T)) &\approx g'(\theta)^2 Var(T-\theta) \\ &= g'(\theta)^2 Var(T) \end{aligned} \tag{27}$$

7.2. Error Cuadrático medio transformado

Sea T un estimador de θ se quiere demostrar que

$$\begin{aligned} E(M(T)) &= Var(T) + [B(\theta)]^2 \\ &= g'(\theta)^2 Var(T) \end{aligned}$$

Ahora generalizando las transformaciones, $g(\cdot)$ una función

$$ECM(g(T)) = Var(g(T)) + [B(g(T))]^2$$

Por linealización de Taylor se tiene que

$$\begin{aligned} E[g(T)] &\approx g(\theta) + g'(\theta)E(T-\theta) \\ &= g'(\theta)^2 Var(T) \end{aligned}$$

$$\begin{aligned} Var[g(T)] &\approx [g'(\theta)]^2 Var(T) \\ B(g(T)) &= E[g(T)] - g(T) \\ &= g'(\theta)E(T-\theta) \\ &= g'(\theta)B(T) \end{aligned}$$

$$\begin{aligned} ECM(g(T)) &\approx [g'(\theta)]^2 Var(T) + [g'(\theta)]^2 B(T) \\ &\approx [g'(\theta)]^2 ECM(T) \end{aligned}$$

Ahora bien, al utilizar la transformación logaritmo se tiene un modelo de la forma $log(y) = x\beta + \epsilon$. Para obtener $ECM(\hat{y})$ se utiliza la transformación intermedia $g(\cdot) = exp(\cdot)$

$$ECM(exp\{log(\hat{y_i})\}) \approx [exp(\hat{y_i})]^2$$

$ECM(log\hat{y_i})$ viene del modelo ajustado $log(y_i) = x\beta + \epsilon$, análogamente pasa cuando se suma el error u_i para el modelo de Fay Harriot

7.3. Características de las funciones de transformación

7.3.1. Transformación logaritmo

Función inversa

$$\begin{aligned} T(y_i) &= log(y_i) \\ &= e_i^y \end{aligned}$$

Derivada de la función

$$\begin{aligned} T(y_i) &= log(y_i) \\ T'(y_i) &= \frac{1}{y_i} \end{aligned}$$

Varianza de la función

$$var(log(y_i)) = \frac{var(y_i)}{(y_i)^2}$$

Siendo y_i la estimación directa, $\hat{y}_i$ estimación del indicador bajo la transformación.

7.3.2. Transformación log - shif

Función inversa

$$\begin{aligned} T(y_i) &= log(y_i + \lambda) \\ &= e_i^y - \lambda \end{aligned}$$

Derivada de la función

$$\begin{aligned} T_\lambda(y_i) &= log(y_i + \lambda) \\ T'_\lambda(y_i) &= \frac{1}{y_i + \lambda} \end{aligned}$$

Varianza de la transformación

$$var(log(y_i + \lambda)) = \frac{var(y_i)}{(y_i + \lambda)^2}$$

7.3.3. Transformación Box - Cox

Función inversa

$$T_\lambda(y_i) = \begin{cases} \frac{(y_i+s)^\lambda - 1}{\lambda} & \lambda \neq 0; \\ log(y_i + s) & \lambda = 0. \end{cases}$$

$$\begin{aligned} T_\lambda(y_i) &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ y_i &= (\lambda y_i + 1)^{\frac{1}{\lambda}} - s \end{aligned}$$

Derivada de la función

$$\begin{aligned} T_\lambda(y_i) &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ T'_\lambda(y_i) &= (y_i + s)^{\lambda - 1} \end{aligned}$$

Varianza de la función

$$var\left(\frac{(y_i + s)^\lambda - 1}{\lambda}\right) = var(y_i)(y_i)^{2\lambda - 1}$$

7.3.4. Transformación doble potencia

Función inversa

$$T_\lambda(y_i) = \begin{cases} \frac{(y_i+s)^\lambda - (y_i+s)^{-\lambda}}{2\lambda} & \lambda > 0; \\ log(y_i + s) & \lambda = 0. \end{cases}$$

$$\begin{aligned} T_\lambda(y_i) &= \frac{(y_i + s)^\lambda - (y_i + s)^{-\lambda}}{2\lambda} \\ &= [\sqrt{1 + \lambda^2 y_i^2} + \lambda y_i]^{\frac{1}{\lambda}} \end{aligned}$$

Derivada de la función

$$\begin{aligned} T_\lambda(y_i) &= \frac{(y_i + s)^\lambda - (y_i + s)^{-\lambda}}{2\lambda} \\ T'_\lambda(y_i) &= \frac{1}{2}(y_i + s)^{\lambda - 1}[(y_i + s)^{2\lambda} + 1] \end{aligned}$$

Varianza de la función

$$var\left(\frac{(y_i + s)^\lambda - (y_i + s)^{-\lambda}}{2\lambda}\right) = \frac{1}{2} var(y_i)(y_i^{-1-\lambda})(y_i^{2\lambda} + 1)$$

7.4. Estimación del ingreso promedio y tasa de desempleo (Estrategia 2)

Para esta sección se realiza la estimación del ingreso promedio y tasa de desempleo en Cundinamarca teniendo en cuenta el siguiente algoritmo.

1. Se selecciona la transformación y se obtiene $T_\lambda = y_j^*(\lambda)$

Para la transformación de los datos y el cálculo del modelo es es necesario transformar la varianza, tal y como lo sugiere (Zhang and Rojas, 2010), es por ello que por medio de la descomposición de Taylor (ver ecuación 27) se realizó la respectiva transformación.

2. Se realiza la estimación del indicador con la variable transformada.

$y_j^*(\lambda) = x_d^T\beta + u_d + e_d, \;\; e_d \sim N(0, \hat{\sigma_d^2}), \;\; u_d \sim N(0, \sigma_u^2)$

2.1 Se aplica el algoritmo Paso a Paso para $y_j^*(\lambda)$ el cual determina las variables auxiliares más pertinentes para todas las transformaciones.

3. Se realiza la transfomación inversa $y_j^*(\lambda)$ a la escala original $y_j = T_\lambda^{-1}(y_j^*(\lambda))$

4. Se cálcula ECM haciendo uso de la generalización de la descomposición de la varianza de Taylor demostrando que

$$ECM(\hat{\theta}) = f^{-1}(\hat{y}_{j,eblup}) \cdot ECM(y_{j,\hat{eblup}}) \tag{28}$$

Donde f^{-1} es la función inversa de la transformación, j son los dominios, $\hat{y}_{j,eblup}$ es la estimación del parámetro con la transformación realizada y $ECM(y_{j,\hat{eblup}})$ es el ECM producido por esa estimación con la transformación usada.

Para mayores detalles de las ecuaciones y demostraciones consulte la sección 7

7.4.1. Tasa de desempleo en Cundinamarca 2017 (Estrategia 2)

Luego de aplicar el algoritmo paso a paso para el modelo FH sin transformaciones, las variables seleccionadas se muestran en la tabla 20.

Variables	coefficients	std.error	t.value	p.value	
(Intercept)	2.20E-02	2.25E-03	9.7691	2.20E-16	***
Hombres_afliados_a_iss_o_nueva_eps	6.09E-05	3.43E-05	1.7738	0.076099	*
Mujeres_afliados_a_iss_o_nueva_eps	-4.92E-05	3.12E-05	-1.5749	0.115268	*
M_transicion	4.87E-05	2.51E-05	1.938	0.052622	*
M_secundaria	-1.16E-05	5.15E-06	-2.2492	0.024502	*
Cinco_anos	8.67E-04	3.62E-04	2.3911	0.016798	*
Seis_diez_anos	-3.78E-04	1.40E-04	-2.6949	0.00704	**
Once_a_catorce	3.23E-04	1.13E-04	2.8479	0.004401	**
Quince_a_16	-1.28E-04	5.58E-05	-2.2853	0.022298	*

Tabla 20: Modeolo FH con las variables definitivas

La Tabla 20 presenta los valores de los coeficientes beta correspondientes a cada variable auxiliar en el modelo sin transformar, junto con la significancia de cada una de estas variables en el modelo.

En la tabla 21 se muestran los respectivos lambda que maximizan la función para las diferentes transformaciones.

Transformación	Lambda
log - shift	9.998
box - cox	1.441
Potencia	6.616e-07

Tabla 21: λ que maximiza la función

Teniendo en cuenta los resultados de la tabla 21 se plantean los respectivos modelos de FH.

Ahora bien, en la figura 9 se muestran las estimaciones de los CVs de cada modelo realizando las respectivas transformaciones en los datos.

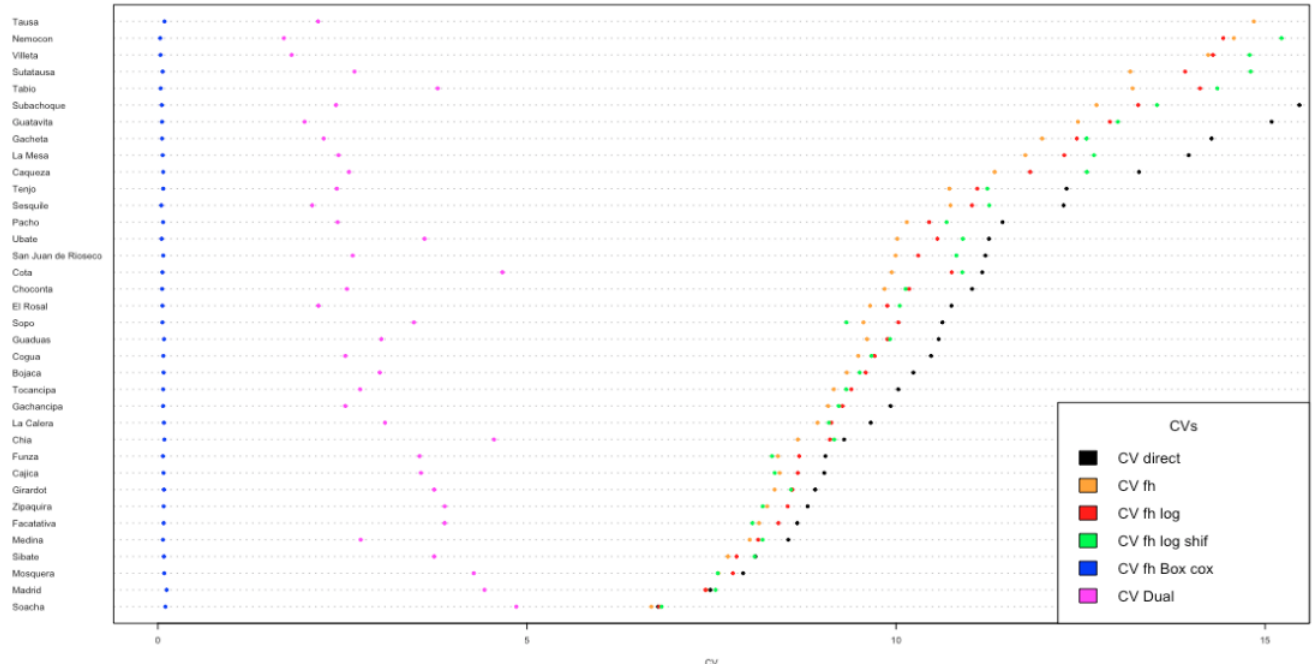


Figura 9: Estimaciones de los CV

En la figura 9 presentada, es evidente que los coeficientes de variación muestran una clara disminución al aplicar la transformación Box-Cox. Esto sugiere que esta transformación es la más adecuada para ajustar los datos. Además, los análisis previos respaldan esta conclusión, ya que demuestran que la transformación Box-Cox reduce significativamente los valores de los coeficientes de variación (CVs). Por lo tanto, se puede afirmar que esta transformación es la más apropiada para mejorar la consistencia de los datos y reducir la variabilidad en los CVs.

Teniendo en cuenta la conclusión anterior en la gráfica 10 se muestran las estimaciones de la tasa de desempleo en Cundinamarca en el año 2017, realizando estimaciones por medio de los modelos de FH con la mejor transformación, Box-Cox.

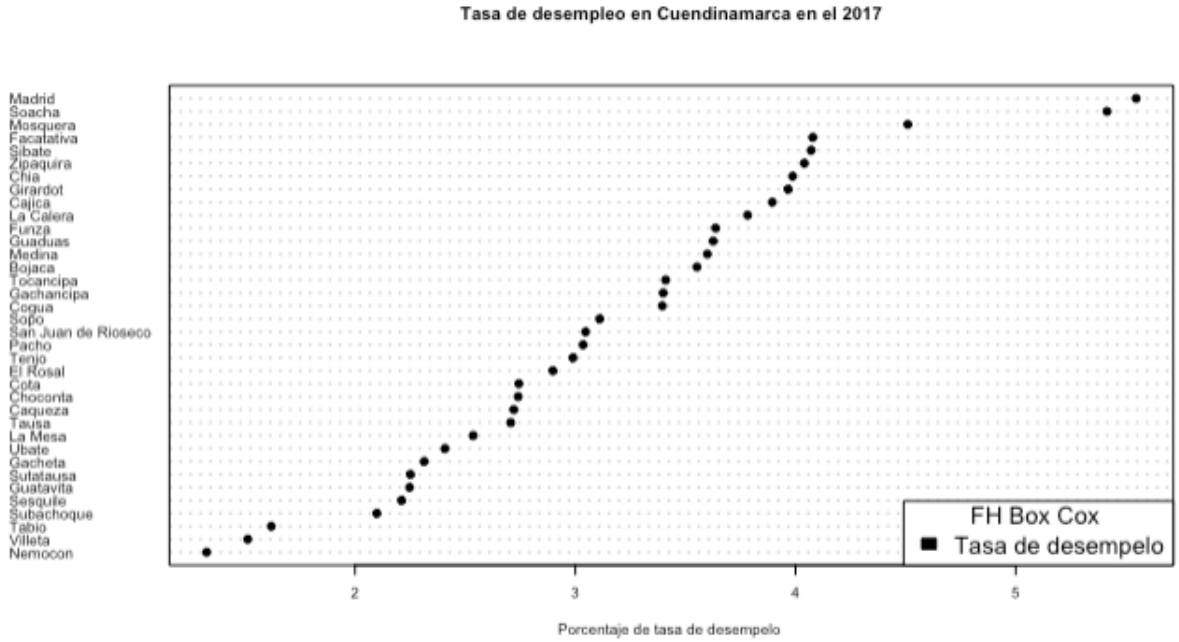


Figura 10: Tasa de desempleo en el 2017

En la Tabla 22 se detallan las estimaciones correspondientes a los municipios con los porcentajes más elevados de desempleo en el departamento de Cundinamarca para el año 2017. Estas estimaciones se llevaron a cabo utilizando el modelo FH, aplicando previamente una transformación de los datos utilizando el método de Box-Cox.

Municipio	Tasa Desempleo Box - Cox
Madrid	5.54654
Soacha	5.414495
Mosquera	4.509763
Facatativa	4.078354
Sibate	4.070749
Zipaquira	4.04026
Chia	3.98704
Girardot	3.966426
Cajica	3.894523
La calera	3.782311
Funza	3.63718
Guaduas	3.626184
Medina	3.599832
Bojaca	3.55244
Tocancipa	3.410413
Gachancipa	3.399196
Cogua	3.395286
Sopo	3.110858
San Juan de Rioseco	3.046995
Pacho	3.035118

Tabla 22: Tasa de desempleo en Cundinamarca 2017

7.4.2. Ingreso promedio en Cundinamarca en 2017 (Estrategia 2)

En la tabla 23 muestran la selección de las variables auxiliares seleccionadas luego de aplicar el algoritmo paso a paso.

Variables	coefficients	std.error	t.value	p.value	Sig
(Intercept)	-1.34E+06	1.19E+06	-1.128312	2.59E-01	.
calentador_ducha	1.29E+06	8.20E+05	1.572692	1.16E-01	***
monthly_wages	2.01E+00	5.08E-01	3.952562	7.73E-05	***
computador	3.92E+06	3.64E+06	1.075717	2.82E-01	*
lavadora	-6.33E+06	2.52E+06	-2.506381	1.22E-02	**
automovil	1.12E+07	4.20E+06	2.665996	7.68E-03	***
C_TRANSICIÓN	3.41E+03	2.67E+03	1.275622	2.02E-01	
Desempeño_fiscal	3.42E+04	1.76E+04	1.944938	5.18E-02	*

Tabla 23: Modelo de FH definitivo ingresos promedio

Teniendo como referencia las variables auxiliares definidas anteriormente, se procede a estimar los respectivos λ que maximizan la transformación con el objetivo de mejorar el comportamiento de los residuales, heterocedasticidad, linealidad y normalidad, ver tabla 24.

Transformación	Lambda
log - shift	-451959.4
box - cox	-0.289749
Potencia	4.837e-08

Tabla 24: λ que maximiza la transformación

Teniendo en cuenta los resultados de la tabla 24 se plantean los respectivos modelos con las respectivas transformaciones.

En la Figura 11 se presentan los coeficientes de variación (CVs) estimados para cada modelo, aplicando la transformación específica correspondiente. Esta representación gráfica permite visualizar y comparar la variabilidad de los datos después de la transformación en cada modelo, lo que es fundamental para evaluar la idoneidad de las transformaciones y determinar cuál de ellas se ajusta mejor a los datos.

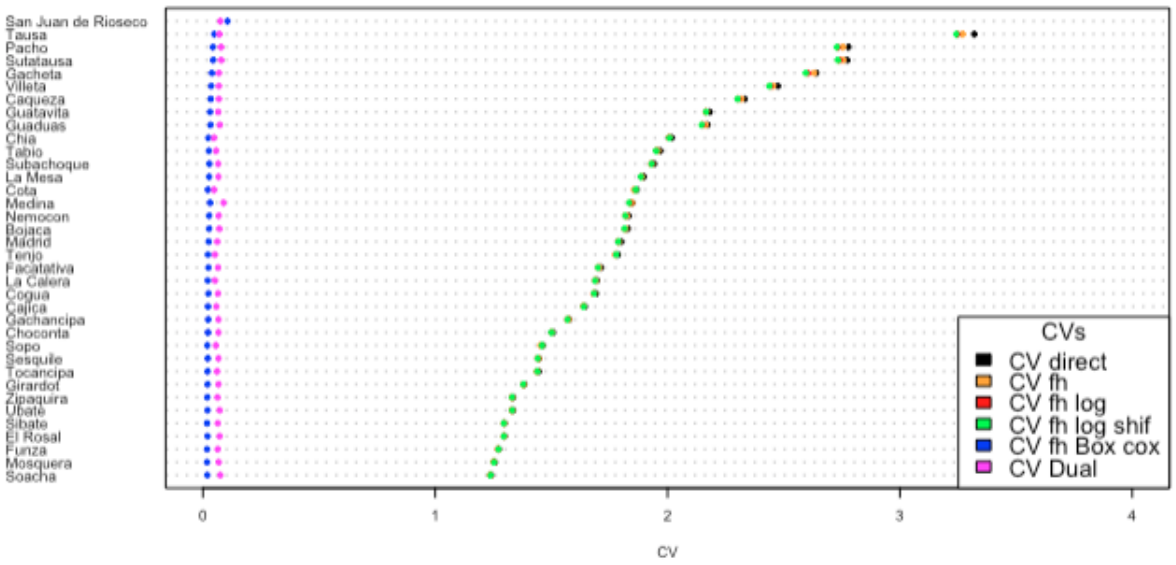


Figura 11: Estimaciones de los CV ingreso promedio

De acuerdo a lo representado en la Figura 11, se observa claramente que la transformación que mejor se adapta a los datos, evaluando el coeficiente de variación, es la transformación de Box-Cox. Esto se fundamenta en el hecho de que la transformación de Box-Cox exhibe el coeficiente de variación más bajo en comparación con las otras transformaciones, indicando una reducción significativa en la variabilidad de los datos.

Con base en los resultados previos, en la Figura 12 se presentan las estimaciones realizadas mediante la transformación de Box-Cox aplicada al ingreso promedio en el departamento de Cundinamarca. Esta representación permite identificar claramente los municipios que presentan los menores ingresos promedio, lo que es fundamental para comprender y analizar las disparidades económicas en la región.

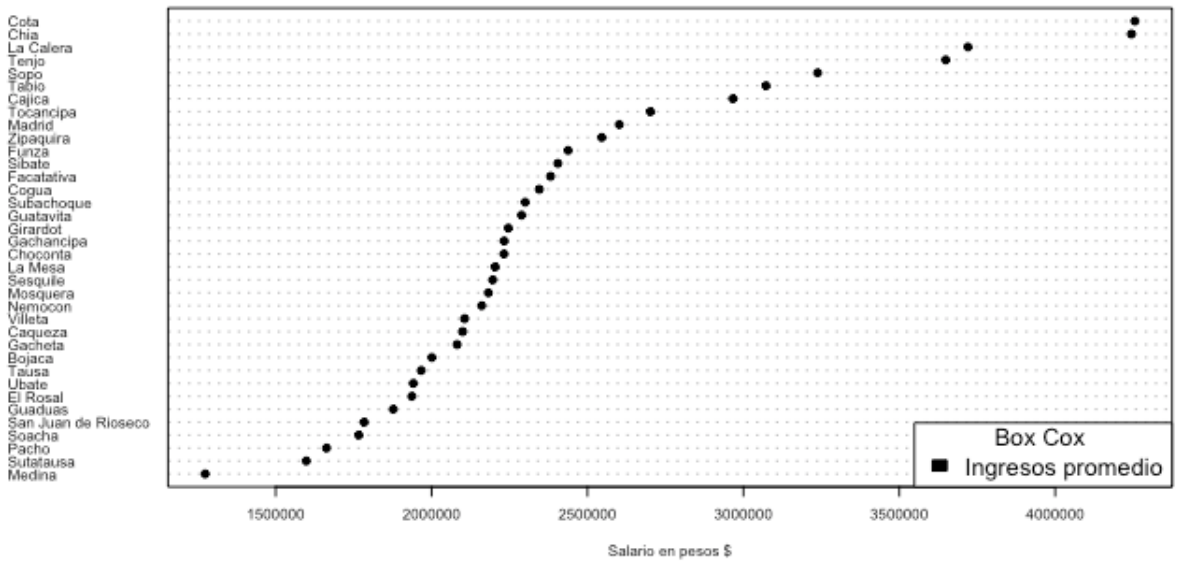


Figura 12: Ingresos promedio en Cundinamarca 2017

En la tabla 25 se muestran los municipios que obtuvieron el menor promedio en sus ingresos.

Municipio	fh_box_cox
Medina	1273859.30
Sutatausa	1598080.02
Pacho	1663156.87
Soacha	1766335.98
San Juan de Rioseco	1783751.05
Guaduas	1877029.15
El Rosal	1936578.79
Ubate	1941353.13
Tausa	1966763.95
Bojaca	2000787.68
Gacheta	2081936.66
Caqueza	2099434.17
Villeta	2105845.43
Nemocon	2160717.74
Mosquera	2181944.87
Sesquile	2196232.46
La Mesa	2203821.31
Choconta	2232221.64
Gachancipa	2232824.55

Tabla 25: Ingresos promedio

Referencias

Battese, G. E., Harter, R. M., and 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.

Bell, W. R., Datta, G. S., and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100(1):189–202.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

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

CostaJJ, A. and VenturaJJJ, A. S. E. (2001). Estimadores compuestos en estadóstica regional: Una aplicaciùn a la estimaciùn de la tasa de variaciùn de la ocupaciùn en la industriaj.

Drew, D., Singh, M., and Choudhry, G. (1982). Evaluation of small area estimation techniques for the canadian labour force survey. *Survey Methodology*, 8(1):17–47.

Fay, R. E. and 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.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49(2):361–376.

Fornieles, A. (2013). Transformaciones de datos en la elaboración de estudios salariales. *Revista de Psicología del Trabajo y de las Organizaciones*, 29(2):75–82.

Fuller, W. A. (1999). Environmental surveys over time. journal of agricultural. pages 331–345.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9(1):55–76.

Kordos, J. (2016). Development of small area estimation in official statistics. *Statistics in Transition. New Series*, 17:105–132 and 157.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019a). The r package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019b). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7):1–33.

Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In *Handbook of Statistics. Sample Surveys. Inference and Analysis*, volume 29, pages 219–249. Elsevier Scientific Publ. Co.

Li, H. and Lahiri, P. (2010a). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, 101(4):882–892.

Li, H. and Lahiri, P. (2010b). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101(4):882 – 892.

Lohr, S. L. (2019). *Sampling: Design and Analysis: Design And Analysis*. CRC Press.

Medina, L., Kreutzmann, A.-K., Rojas-Perilla, N., and Castro, P. (2019). The r package trafo for transforming linear regression models. *R Journal*, 9(2).

Molina, I. (2019). Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas.

Molina, I., Datta, G. S., and Rao, J. (2015). *Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects*. Statistics Canada.

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

Molina, I. and Rao, J. (2013). A review of poverty mapping procedures. *Maret*, 7:2015.

Neira, J. P. F. (2011). Estimación en dominios. *Universidad de la república-Uruguay, Facultad deficiencias económicas y de Administración-Licenciatura en estadística, Tutor: Guillermo Zoppolo*.

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

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480):1427–1439.

Prasad, N. N. and Rao, J. N. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171.

Rojas-Perilla, N. (2018). *The Use of Data-driven Transformations and Their Applicability in Small Area Estimation*. Freie Universitaet Berlin (Germany).

Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2017). Data-driven transformations in small area estimation. Technical report, Diskussionsbeitr‰ ge.

Romero, N. A. R., Estrada, J. G. S., and Fernández-Berrocal, P. (2012). Indicadores sociales, condiciones de vida y calidad de vida en jóvenes mexicanos. *Revista iberoamericana de psicología*, 5(1):71–80.

Royston, P., Lambert, P. C., et al. (2011). *Flexible parametric survival analysis using Stata: beyond the Cox model*, volume 347. Stata press College Station, TX.

Santiago Moreno, A. et al. (2012). *Aportaciones a la estimación en áreas pequeñas. Estimación de proporciones*. Granada: Universidad de Granada.

Yang, Z. (2006). A modified family of power transformations. *Economics Letters*, 92(1):14–19.

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

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

Zhang, H. and Rojas, H. A. G. (2010). *Teoría estadística: aplicaciones y métodos*. Hugo Andrés Gutiérrez Rojas.