
Estimación de la tasa de desempleo e ingreso medio del departamento de Cundinamarca utilizando SAEDaniel Mendoza Castrillón^a
danielmendoza@usantotomas.edu.coJosé Fernando Zea^b
josezea@usantotomas.edu.co

1. Resumen

En el presente artículo se propone hacer uso del análisis de componentes principales para obtener covariables que se incorporen como información auxiliar de los estimadores de Fay-Herriot (FH) y Fay-Herriot (FH espacial) espacial. Se muestran las ventajas de esta propuesta en términos de la precisión de las estimaciones obtenidas (por medio de medidas de precisión basadas en el error cuadrático medio) y el mejor aprovechamiento de las variables auxiliares disponibles en un contexto de un número reducido de áreas pequeñas. Se realiza una aplicación empírica usando el software R (dplyr, ggplot2, survey y otros paquetes) con la información disponible en la encuesta multipropósito 2017 la cual contiene información socioeconómica de Bogotá y municipios aledaños, se estima en particular el promedio del ingreso percapita del hogar y la tasa de desempleo, se llevan a cabo comparaciones del estimador de FH, FH espacial con covariables sin transformar (Molina, 2015) y con covariables basadas en las coordenadas obtenidas a partir del análisis de componentes principales. En el trabajo se ilustra como el Fay – Herriot espacial con variables provenientes de las componentes principales tienen el mejor desempeño en términos de precisión de las estimaciones.

Palabras Clave: Estimación en áreas pequeñas, Componentes principales, Ingreso medio, tasa de desempleo, Fay Herriot, Encuesta Multipropósito, Cundinamarca.

2. Abstract

In this article, we propose using principal components analysis in a context of high-dimensional spaces to obtain orthogonal variables to use as auxiliary information in Fay-Herriot (FH) and spatial Fay-Herriot (FH). The accuracy are computed for direct, Fay-Herriot and proposed estimator. An empirical application with R ecosystem (dplyr, ggplot2, survey, and other R packages) is carried out using 2017 Bogotá Multipurpose Survey (EMB2017), a household survey with socioeconomic and demographic information of Bogotá and surrounding municipalities. In particular, percapita average household income and unemployment rate are estimated, then a benchmark of proposal estimators with classical FH estimator is carried out (Molina, 2015). The work illustrates how the spatial Fay - Herriot estimator with selected components from PCA has the best performance in terms of accuracy.

Keywords: Small Area Estimation, Principal Components, mean income, unemployment rate, Fay Herriot, multipurpose survey, Cundinamarca.

^aEstudiante pregrado Estadística, U. Santo Tomás, sede Bogotá^bDocente Facultad de Estadística, U. Santo Tomás, sede Bogotá

3. Introducción

En el ámbito político, es necesario realizar recogidas de información de la población, esto con el fin de conocer con certeza la composición y características principales que se encuentran en la totalidad del espacio territorial, para ello se realizan encuestas para satisfacer esta necesidad y posteriormente tener la posibilidad de tomar decisiones correspondientes, con el fin de realizar políticas públicas que mejoren la calidad de vida o satisfagan necesidades que se encuentren en el proceso. Una problemática que se presenta en este procedimiento son las limitaciones presupuestales o restricciones de tiempo existentes para obtener tamaños muestrales adecuadas que permitan obtener estimaciones lo suficientemente precisas para diferentes dominios o grupos de interés.

Dos medidas de interés muy útiles para la elaboración de políticas públicas son la tasa de desempleo, y el ingreso per cápita de los hogares, estas mediciones son usualmente obtenidas por encuestas de hogares especializadas y diseñadas de tal forma que se puedan obtener estimaciones lo suficientemente precisas y con unos niveles de desagregación pertinentes.

Sin embargo, los estudios de muestreo con diseños muestrales probabilísticos son en ocasiones bastante costosos y focalizados en áreas muy grandes o en grupos poblacionales grandes, no siempre es viable disponer de encuestas que responda a las necesidades de información de las regiones o de grupos particulares, por ejemplo, la encuesta multipropósito de Bogotá, encuesta diseñada para obtener resultados precisos para las 20 localidades de Bogotá, no se encuentra diseñada para obtener estimaciones a nivel de UPZ, tampoco es posible con esta encuesta obtener estimaciones precisas al interior de los grupos conformados por cruces que involucra varias variables categóricas (estrato, sexo, grupos de edad). La estimación de áreas pequeñas (SAE) permite generar estimaciones para estos dominios de interés que no logran ser bien estudiados por medio de diseños muestrales probabilísticos. La estimación de áreas pequeñas hace uso de diferentes metodologías estadísticas basadas en la inferencia estadística, el uso de información auxiliar altamente correlacionada con los parámetros de interés y el desarrollo de modelos predictivos. La estimación de áreas pequeñas se realiza sobre dominios para los cuales no se cuenta con un número suficientemente de encuestas, estos dominios se conocen como áreas pequeñas.

El objetivo principal de este trabajo es obtener estimaciones la tasa de desempleo y del ingreso medio de los municipios de Cundinamarca y Bogotá (Colombia). Aunque la encuesta multipropósito tiene un diseño muestral que permite obtener estimaciones precisas para cerca de 31 municipios, para algunos de estos el número de encuestas no es suficiente, por lo cual se hará uso de la metodología de estimación de áreas pequeñas. Además para los municipios no abordados por la EMB se llevarán a cabo estimaciones del ingreso y el desempleo haciendo uso de información auxiliar a nivel municipal, estas estimaciones se comparan con las estimaciones directas. Para hacer uso de la metodología de estimación de áreas pequeñas se recurre a las estimaciones obtenidas a partir de la encuesta multipropósito (Departamento Administrativo Nacional de Estadísticas - DANE) y de información a nivel de municipio del portal Terridata (portal del Departamento Nacional de Planeación que cuenta con información auxiliar a nivel municipal), la cual brindará los datos necesarios para la aplicación de la metodología SAE y así poder comparar los resultados obtenidos y presentar las posibles mejoras que de esta resulten.

4. Antecedentes

En esta sección se hará una búsqueda exhaustiva de la literatura relacionada con el tema de la aplicación de la metodología de estimación en áreas pequeñas (SAE) en estudios socioeconómicos, entre los trabajos más relevantes se relacionan los siguientes.

Molina(2019) detalla diferentes alternativas de estimación en áreas pequeñas para poder satisfacer la necesidad que se presenta en el proceso de estimación de diferentes variables socio demográficas, además de ello, se explica de manera clara y concisa las situaciones en las cuales es necesario hacer uso de esta

metodología, se ilustran varios estudios de casos, en diferentes países del mundo.

En el artículo escrito por J.D. DREW, M.P. SINGH G.H. CHOUDERY (1982), expresan con claridad distintas técnicas de muestreo en áreas pequeñas entre ellas, el estimador de dominio Post-estratificado el cuál se hace mayor énfasis ya que es el estimador que ofrece la mayor precisión en el desarrollo del artículo. Este documento muestra la aplicación del estimador anteriormente nombrado y otros estimadores como los compuestos, sintéticos y de muestra dependiente. El objetivo del artículo es estimar el total de la población empleada y desempleada obteniendo información de la encuesta de fuerza laboral de Canadá y teniendo como información auxiliar los datos de las viviendas de los sectores evaluados, este artículo ofrece diferentes manera de obtener estimaciones precisas que puede ser de gran utilidad en el momento de enfrentarse a problemáticas que se resuelven mediante la aplicación de la metodología SAE.

Felipe Ortiz y Fernando Zea (2018), exponen de manera detallada la implementación del modelo Fay Herriot para la estimación del desempleo y del ingreso medio para el departamento de Cundinamarca usando como variables auxiliares el promedio de las pruebas saber 11, energía per capita por cada municipio, entre otras, en este se explica claramente el proceso llevado a cabo por medio del software R y se compara la precisión en base al MSE del modelo comparada con el Coeficiente de Variación de los estimadores directos. Este trabajo proporciona información de gran utilidad ya que se trata la información contenida en la encuesta multipropósito del DANE realizada en 2014. s

5. Marco Teórico

5.1. Encuesta multipropósito

La Encuesta multipropósito surge por la necesidad de dar continuidad a las encuestas de Calidad de Vida (1991, 1993, 2003 y 2007) y a la Encuesta de Capacidad de Pago (2004) que la Secretaría Distrital de Planeación (SDP) aplicó para el distrito. Dado que Bogotá requería las temáticas tanto de calidad de vida como de capacidad de pago, la SDP consideró conveniente integrar las dos encuestas en una Encuesta Multipropósito.¹

77.025 hogares en Bogotá y 32.086 hogares en municipios de Cundinamarca hicieron parte de la Encuesta multipropósito (EM) 2017, ubicadas en 19 localidades urbanas con sus Unidades de Planeamiento Zonal UPZ, en la ruralidad de Bogotá y en la zona urbana de 37 municipios de Cundinamarca.²

El diseño de la muestra propuesto es un diseño probabilístico, multietápico, estratificado y de conglomerados en lo que respecta a las UPZ de Bogotá y los municipios de Cundinamarca.³.

5.2. Muestreo

El muestreo es una metodología estadística que permite estimar variables de interés obteniendo una muestra aleatoria específica de la población objetivo, el alcance de esta metodología es principalmente obtener valores precisos de dichas variables sin incurrir en mayores costos. Los estimadores mas utilizados en los diferentes diseño muestrales para el total poblacional son:

- Estimador Horvitz-Thompson (1952):

¹recuperado de: <https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/encuesta-multiproposito>

²recuperado de: <https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/encuesta-multiproposito/encuesta-multiproposito-2017>

³Recuperado de http://www.sdp.gov.co/sites/default/files/disenom_etodologicoem2017.pdf

$$\hat{t}_y = \sum_S \frac{y_k}{\pi_k} \quad (1)$$

Donde π_k hace referencia a la probabilidad de inclusión para el k-ésimo elemento.

- Estimador Hansen-Hurwitz (1952):

$$\hat{t}_y = \frac{1}{m} \sum_S \frac{y_k}{p_k} \quad (2)$$

Se usa este estimador cuando el diseño de muestreo se realiza con reemplazo, donde p_k hace referencia a la probabilidad de inclusión para el k-ésimo elemento y m a la cantidad de escogencias de los elementos.

- Estimador Háyek:

$$\hat{y} = \frac{t_y}{\hat{N}} \quad (3)$$

donde:

$$\hat{N} = \sum_S \frac{1}{\pi_k} \quad (4)$$

Se usa este estimador de la media cuando el valor de N es desconocido, se procede entonces a realizar el cociente entre el total y la sumatoria de los factores de expansión asociados a cada individuo.

- Estimador Razón:

$$\hat{R} = \frac{\sum_S w_k y_k}{\sum_S w_k z_k y_k} \quad (5)$$

Donde z_k hace referencia al dominio de interés de la muestra.

5.3. Diseños muestrales

Un diseño muestral es aquella metodología que se usa para llevar a cabo la selección de la muestra y posteriormente la estimación de los parámetros de las variables de interés como lo es el total poblacional, entre los diseños muestrales mas utilizados se tienen, el muestreo aleatorio simple (MAS), muestreo estratificado y muestreo por conglomerados.

5.3.1. Muestreo Aleatorio Simple

La característica principal del muestro aleatorio simple radica en que las probabilidades de inclusión se calculan de la misma manera para cada muestra, debido a que se asume homogeneidad de la variable de interés en la población, por ende la probabilidad de inclusión y el estimador de Horvitz-Thompson, estarían dados de la siguiente manera:

$$\pi_k = \frac{n}{N} \quad (6)$$

$$t_y = \frac{N}{n} \sum_S y_k \quad (7)$$

5.3.2. Muestreo Estratificado

El muestreo estratificado se caracteriza principalmente por la obtención de subgrupos de la población, esto con el fin de aumentar la certeza de las estimaciones conociendo con ayuda de información auxiliar, aquellos grupos de la población que tienen alta varianza entre sí, pero poca varianza dentro de sí. Después de la división en subgrupos se procede a realizar la extracción de la muestra para cada subgrupo. El estimador de Horvitz-Thompson quedaría de la siguiente manera:

$$t_{y,\pi} = \sum_{h=1}^H t_{yh,\pi} \quad (8)$$

$$t_{y,\pi} = \sum_{k \in S_h} \frac{y_k}{\pi_k} \quad (9)$$

Donde h representa el estrato específico de la población.

5.4. Muestreo por conglomerados

En el momento de realizar la selección de la muestra existe la posibilidad de que no se pueda contar con un marco de muestreo determinado, este problema acarrea grandes dificultades debido a que no es posible obtener la muestra. Una opción para corroborar esta situación es optar por los conglomerados de la población, se lleva a cabo una selección de los posibles grupos que naturalmente se encuentren en todo el territorio y después de esto se realiza un censo a dicho grupos. Existe una gran problemática al realizar esta metodología y es principalmente la precisión de las estimaciones de los parámetros de la variable de interés, ya que al hacer censo en los grupos seleccionados se está considerando un fuerte homogeneidad de los individuos por ende la información proporcionada puede llegar a ser redundante o respectiva y por ende no genera utilidad a la hora de analizar la información obtenida.

5.5. Muestreo bietápico

Cuando no se dispone de un marco de muestreo determinado, es necesario recurrir a diferentes conglomerados que puedan encontrarse en un espacio específico, estos conglomerados ayudan a realizar las estimaciones respectivas de las variables de interés. En dichos conglomerados es posible que la varianza sea muy homogénea en este tipo de situaciones es mejor optar por obtener una muestra con el fin de reducir costos evitando así realizar un posible censo, este proceso se conoce como muestreo en varias etapas, el cual me permite obtener muestras de cada conglomerado, usualmente llamados unidades de muestreo. Para dicho diseño de muestreo el estimador de Horvitz-Thompson está dado por:

$$\hat{t}_{y,\pi} = \sum_{k \in S_I} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \quad (10)$$

Donde,

$$\hat{t}_{yi,\pi} = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}$$

S_I hace referencia al universo de unidades de muestreo, π_{Ii} es la probabilidad de inclusión de la unidad de muestreo, S_i es la unidad de i -ésima unidad de muestreo, $\pi_{k|i}$ es la probabilidad de inclusión del k -ésimo elementos de la i -ésima unidad de muestreo

5.6. Estimación en áreas pequeñas

Como se ha hablado anteriormente para el buen desarrollo de las encuestas es necesario obtener muestras de la población objetivo esto con el fin de obtener información valiosa para luego generar estimaciones de diferentes parámetros como el total poblacional de la variable de interés. En algunas ocasiones estas muestras no pueden ser lo suficientemente grandes para realizar las estimaciones directas (entiéndase por estimaciones directas las que se realizan solo con los datos obtenidos de la muestra) estos espacios donde no se obtiene una muestra significativa suelen denominarse "áreas pequeñas", en este caso en específico es necesario recurrir a metodologías diferentes que permitan obtener estas estimaciones, una de ellas es la estimación en áreas pequeñas, que otorga por medio de inferencia estadística estas estimaciones de dichos dominios con poca muestra por medio de modelos que relacionan variables auxiliares que proporcionan información de gran utilidad para realizar este proceso. Uno de los mecanismo mas utilizados para realizar estimaciones en áreas pequeñas es el modelo Fay-Herriot.

5.6.1. Modelo Fay-Herriot

Es necesario trabajar con el concepto de los modelos en el momento de realizar estimaciones indirectas de un área en específico, como dice Molina & Rao (2015) estos modelos ofrecen varias ventajas:

- Los estimadores "óptimos" se pueden derivar bajo el supuesto modelo.
- Se pueden asociar medidas de variabilidad específicas del área con cada estimador a diferencia de las medidas globales (promediadas en áreas pequeñas) a menudo utilizadas con estimadores indirectos tradicionales.
- Los modelos pueden validarse a partir de los datos de la muestra.
- Una variedad de los modelos pueden ser entretenidos dependiendo de la naturaleza de las variables de respuesta y la complejidad de las estructuras de datos (como la dependencia espacial y las estructuras de series de tiempo).

En situaciones en las cuales no se cuenta con la información necesaria para realizar las estimaciones directas por medio de las metodologías tradicionales y por ende las zonas no están valoradas en la encuesta, es necesario optar por la utilización del modelo de área más básico, el cual está dado por:

$$\bar{y}_i = \mathbf{z}^T \boldsymbol{\beta} \quad (11)$$

El cual solo cuenta con los valores de las variables auxiliares de las áreas en específico ($\mathbf{z}$) y con los coeficientes de regresión $\boldsymbol{\beta}$

La estructura del modelo Fay Herriot está dada por la siguiente ecuación:

$$\bar{y}_i = \mathbf{z}^T \boldsymbol{\beta} + v_i + e_i, e_i \sim (0, \Psi), v_i \sim (0, \sigma_v^2) \quad (12)$$

Para la implementación de este modelo es necesario contar con las estimaciones directas de la variable poblacional de interés, es decir, las áreas que están incluidas en la encuesta. Donde $\mathbf{z}$ hace referencia a la información auxiliar que se obtiene del área en específico, $\boldsymbol{\beta}$ contiene los coeficientes del modelo que generan un producto junto con la información auxiliar. v_i son los efectos aleatorios asociados al área específica, estos son independientes e idénticamente distribuidos, por último e_i hace referencia al error muestral.

5.6.2. Estimador BLUP y EBLUP

Como dicen Molina & RAO (2015), Con el fin de obtener un error cuadrático medio (MSE) lo más pequeño posible, es necesario hacer uso del estimador BLUP, ya que este ofrece esa posibilidad y además de ello no depende de la normalidad de los efectos aleatorios, está expresado de la siguiente manera:

$$\bar{y}_i^{BLUP} = \mathbf{z}^T \boldsymbol{\beta} + \gamma_i (\bar{y}_i - \mathbf{z}^T \boldsymbol{\beta}), \quad (13)$$

Donde,

$$\gamma_i = \frac{\sigma_v^2}{\Sigma + \sigma_v^2}$$

En esta ecuación σ_v^2 hace referencia a la varianza de los efectos del área específica v_i que es desconocida, y Σ hace referencia a la varianza de los errores muestrales, esta varianza si es conocida.

Debido a que es necesario realizar la estimación correspondiente de los coeficientes de la regresión ($\boldsymbol{\beta}$) y la varianza de los efectos de cada área (σ_v^2) por medio de Máxima Verosimilitud o Máxima Verosimilitud Restringida, nace entonces el estimador EBLUP, el cual está dado por:

$$\bar{y}_i^{EBLUP} = \mathbf{z}^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\hat{y}_i - \mathbf{z}^T \hat{\boldsymbol{\beta}}) \quad (14)$$

Donde $\hat{y}_i$ hace referencia a la estimación directa de la variable de interés para cada una de las áreas incluidas.

5.6.3. Cálculo del MSE

Para el cálculo del error cuadrático medio se usa la siguiente fórmula:

$$MSE(\hat{y}_i) = \begin{cases} g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) & \text{Si } i \in A \\ \mathbf{z}_i(\mathbf{ZV}^{-1}\mathbf{Z})^{-1}\mathbf{z}_i^t + \hat{\sigma}_v^2 & \text{Si } i \notin A \end{cases} \quad (15)$$

Donde $g_{1i}(\hat{\sigma}_v^2) = \frac{\hat{\sigma}_v^2 \Sigma_i}{\hat{\sigma}_v^2 + \Sigma_i}$ y $g_{2i}(\hat{\sigma}_v^2) = \frac{\Sigma^2}{(\hat{\sigma}_v^2 + \Sigma)^2} \mathbf{z}_i(\mathbf{ZV}^{-1}\mathbf{Z})^{-1}\mathbf{z}_i^t$

y

$$g_{3i}(\hat{\sigma}_v^2) = \frac{2\Sigma^2}{m(\hat{\sigma}_v^2 + \Sigma)^3} (\hat{\sigma}_v^4 + 2\hat{\sigma}_v^2 \sum_{i=1}^I \frac{\Sigma_i}{\Sigma} + \sum_{i=1}^I \frac{\Sigma_i^2}{\Sigma}),$$

$$\mathbf{V} = \text{diag}(\sigma_v^2 + \Sigma_i, \dots, \sigma_v^2 + \Sigma_I)$$

5.6.4. Fay Herriot Espacial

El modelo Fay Herriot Espacial se caracteriza por tener en cuenta la proximidad geográfica que puede encontrarse en un estudio de poblaciones. Esto eventualmente puede mejorar las estimaciones que se realicen. Se tiene en cuenta la misma estructura del modelo Fay Herriot en la ecuacion (9), el cambio se da en el vector de los efectos aleatorios asociados a cada área construyendo así un modelo autorregresivo simultáneo (SAR), que estaría dado por:

$$\mathbf{v} = \rho \mathbf{W} \mathbf{v} + \mathbf{e}_i, \quad \mathbf{e}_i \sim (0, \sigma^2 \mathbf{I}) \quad (16)$$

Donde ρ hace referencia al peso que se le da a la matriz de proximidad en la estructuración del vector de efectos aleatorios por dominio. La matriz cuadrada $\mathbf{W}$ hace referencia a la proximidad que existe entre

todos los dominios donde la diagonal está rellena de ceros y los demás valores serán 0 si no hay proximidad y la proporción de individuos próximos en otro caso. Un ejemplo de esta matriz es el siguiente:

	M1	M2	M3	M4	M5
M1	0.00	0.50	0.50	0.00	0.00
M2	0.33	0.00	0.33	0.33	0.00
M3	0.25	0.25	0.00	0.25	0.25
M4	0.00	0.33	0.33	0.00	0.33
M5	0.00	0.00	0.50	0.50	0.00

En este caso por ejemplo el municipio 1 (M1), tiene proximidad con el municipio 2 (M2) y el municipio 3 (M3), entonces el procedimiento para obtener los valores de la matriz es el siguiente $\frac{1}{\#Proximidades}$, y así se rellena sucesivamente la matriz de proximidad.

6. Objetivos

6.0.1. Objetivo General

Comparar las estimaciones generadas por el estimador directo y diferentes variaciones del estimador de Fay - Herriot para la tasa de desempleo y el ingreso medio realizadas para los municipios de Cundinamarca encuestados en la encuesta multipropósito de Bogotá.

6.0.2. Objetivos Específicos

- Escoger las variables necesarias para el ajuste del modelo propuesto mediante análisis de componentes principales.
- Construir las estimaciones directas, Fay - Herriot y FayHerriot espacial, de la tasa de desempleo e ingreso medio del hogar para los municipios observados en la Encuesta Multipropósito de Bogotá, comparando la precisión de los dos estimadores utilizados.
- Generar estimaciones para los municipios no observados en la Encuesta Multipropósito de Bogotá con sus respectivas medidas de error.

7. Marco Metodológico

Para este estudio se realizará un análisis exhaustivo de los datos recolectados los cuales harán referencia a la información obtenida de la encuesta multipropósito para el año 2017, además de la información auxiliar que se pueda encontrar para los municipios de Cundinamarca específicamente, esto con el fin de realizar las estimaciones pertinentes con las metodologías específicas y así comparar la precisión de los resultados.

Las variables que serán de utilidad para el análisis serán las encontradas en la base de datos referente a la encuesta multipropósito que hacen referencia a las preguntas principales realizadas en dicha encuesta esto con el fin de construir las estimaciones directas. Se usarán también variables auxiliares de los municipios tales como resultados de las pruebas ICFES, gasto per cápita de energía, etc, con el objetivo de obtener las estimaciones inferenciales en las cuales se hará uso de la metodología SAE específicamente el modelo Fay Herriot.

Las variables auxiliares se recogerán de los datos proporcionados por Terridata que hace referencia al sistema de estadísticas oficiales del Departamento Nacional de Planeación.

La base de datos de la encuesta multipropósito se encuentra disponible y con libre acceso en el repositorio del DANE, por ende se descargará y se realizará el respectivo análisis.

7.1. Varianza muestral

Debido a que en la información proporcionada por el DANE respecto a los resultados de la encuesta multipropósito, solo se cuenta con los factores de expansión como variable útil para calcular la varianza, se sigue el siguiente proceso de estimación:

- Se obtienen los factores de expansión asociados a cada municipio.
- Se obtiene una muestra con reemplazo del tamaño de la cantidad de municipios teniendo en cuenta los factores de expansión.
- Para cada individuo de la muestra con reemplazo se calcula su respectiva estimación ya sea para el ingreso medio o Tasa de desempleo
- Se repite este proceso 5000 veces y después de obtiene la varianza de las 5000 estimaciones.

7.2. Componentes principales

Existe una gran cantidad de variables recolectadas en Terridata, referentes a todos los municipios, muchas de ellas guardan una correlación importante frente al ingreso medio y la tasa de desempleo, sin embargo la dimensionalidad que se encuentra es muy grande, debido a esta problemática se recurre a realizar un análisis de componentes principales con el fin de reducir esta dimensionalidad sin perder información importante para el modelo. El proceso como tal que se llevará a cabo, se compone de una selección de algunas variables con la mayor correlación, en este caso se habla de una cantidad de 10 variables como máximo, posteriormente, las variables restantes se le realizará una reducción de dimensionalidad generando una combinación lineal por medio de la metodología de componentes principales. En el ajuste del modelo se tendrá en cuenta entonces, aquellas variables con la mayor correlación y las coordenadas de los componentes que también tienen una buena correlación con el ingreso, posteriormente se visualizará el ajuste del modelo sin las coordenadas y con ellas, con el fin de comparar el coeficiente de determinación y así escoger el mejor modelo para realizar las respectivas estimaciones.

Para obtener los resultados finales del estudio se realizarán las estimaciones pertinentes para cada uno de los municipios de Cundinamarca con las metodologías antes mencionadas, después se comparará la estimaciones de ambas metodologías con el fin de conocer la precisión, además de ello, para aquellos municipios en los cuales no se puedan realizar estimaciones directas debido a la cantidad de información se aplicará la estimación por medio del modelo Fay Herriot con el uso directo de la información auxiliar.

8. Resultados

8.1. Ingreso Medio

Se implementó entonces, el modelo Fay Herriot espacial y Fay Herriot con dos variaciones, la primera con un conjunto de variables auxiliares seleccionadas con procedimientos estadísticos y coherencia temática y la segunda variación mediante la selección mediante el uso de componentes principales resultantes de cerca de 40 variables auxiliares correlacionadas con la variable ingreso medio, esto con el fin de obtener las estimaciones pertinentes para cada municipio y así poder compararlas con las estimaciones directas. Además de hacer esta comparación se decidió extrapolar los modelos para aquellos municipios que están fuera de la muestra. Los resultados se mostrarán posteriormente.

8.1.1. Variables Escogidas

Las variables que se escogieron con base a la correlación y coherencia temática frente a la estimación directa son las siguientes:

- Avalúo catastral rural (2012)
- Cobertura bruta en educación media (2017)
- Puntaje promedio Pruebas Saber 11 - Lectura crítica (2017)
- Puntaje promedio Pruebas Saber 11 - Matemáticas (2017)
- Penetración de banda ancha (2017)
- Indicador de desempeño fiscal (2017)
- Gastos totales per cápita (2017)
- Numero de personas validadas en el Sisben (2016)

Es válido recalcar que al realizar un análisis de la correlación existente entre las variables, se puede encontrar multicolinealidad si esta es muy alta. Para las variables, "Puntaje promedio Pruebas Saber 11 - Lectura crítica" y "Puntaje promedio Pruebas Saber 11 - Matemáticas", se encontró una correlación de casi el 90 % , por ende se optó por realizar un promedio entre las dos variables para cada municipio y se tomó esta variable en lugar de las dos antes mencionadas.

8.1.2. Resultados modelos

Para el modelo con solo las variables correlacionadas tenemos la salida en la Tabla 1.

Tabla 1: Resultados	
	<i>Dependent variable:</i>
	IncomeMean
‘Avalúo catastral rural‘	332,206.900** (155,929.200)
‘Cobertura bruta en educación media‘	861.728 (1,612.807)
‘Penetración de banda ancha‘	6,037.500 (14,409.370)
‘Indicador de desempeño fiscal‘	−150.998 (6,333.676)
‘Gastos totales per cápita‘	8,767.324 (54,698.760)
Numero.de.personas.validadas	0.015 (0.057)
PromPuntajesPruebas	−4,273.914 (17,441.720)
Constant	824,855.600 (951,179.000)
Observations	38
R ²	0.661
Adjusted R ²	0.582
Residual Std. Error	153,689.200 (df = 30)
F Statistic	8.360*** (df = 7; 30)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Por otro para el modelo con las coordenadas de las primeras 10 componentes de las variables restantes los resultados se encuentran en la Tabla 2. Se destaca como el modelo con componentes brinda un ajuste mucho mayor en términos de coeficiente de variación, esta mejora en el ajuste permitirá obtener mejores estimaciones en el modelo.

Se observa una clara diferencia entre los dos coeficientes de variación, teniendo un coeficiente mas alto el modelo que usa las coordenadas de las componentes principales, sin embargo se realizará una comparación entre los coeficientes de variación de las estimaciones de cada modelo que se muestra en la Figura 1.

Tabla 2: Resultados

	<i>Dependent variable:</i>
	IncomeMean
‘Avalúo catastral rural‘	310,107.800** (116,945.900)
PromPuntajesPruebas	−11,992.970 (13,393.170)
‘Gastos totales per cápita‘	64,818.380 (42,299.460)
Dim.2	29,368.250 (23,046.780)
Dim.3	33,096.840 (21,675.710)
Dim.4	−49,179.010* (24,687.970)
Dim.5	101,932.400*** (25,717.820)
Dim.8	68,366.180* (34,026.420)
Numero.de.personas.validadas	−0.094 (0.064)
Constant	1,264,930.000* (711,651.100)
Observations	38
R ²	0.812
Adjusted R ²	0.752
Residual Std. Error	118,442.400 (df = 28)
F Statistic	13.449*** (df = 9; 28)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

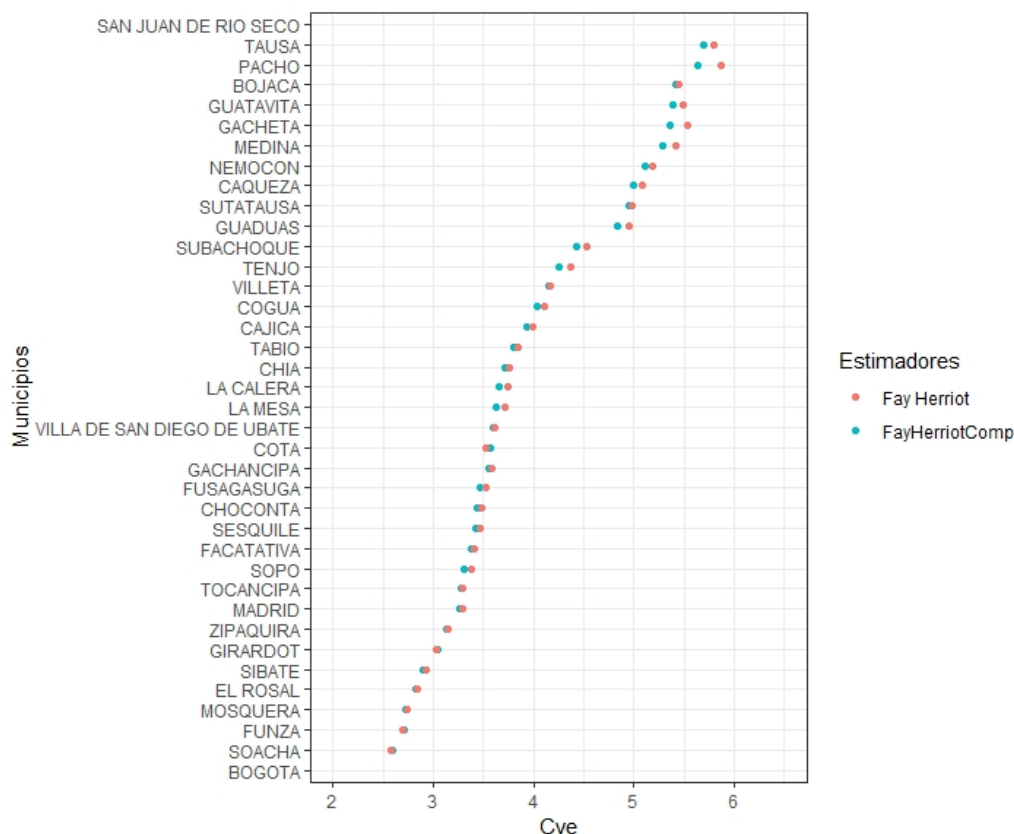


Figura 1: Comparación Coeficientes ambos modelos

8.1.3. Comparación CVE's Estimación Directa, modelo Fay Herriot y Fay Herriot espacial

En la Figura 2, se detalla la comparación entre el modelo Fay Herriot, Fay Herriot Espacial y las estimaciones directas.

Se puede observar que no hay una diferencia significativa y realizando una comparación exhaustiva, se puede identificar una leve mejora en los CVE's del modelo Fay Herriot Espacial, sin embargo se realizará una comparación de la media de los errores cuadráticos medios por cada municipio.

Fay Herriot	Fay Herriot Espacial
4.40	4.21

Se concluye en general que el uso de la metodología de estimación de áreas pequeñas por medio del modelo Fay Herriot y sus variaciones garantiza mejores estimaciones de la variable de interés, en este caso el ingreso medio, que usando estimaciones directas. Para este resultado en particular el modelo Fay Herriot espacial tuvo un mejor ajuste, por ende este será el modelo que se usará para la posterior extrapolación.

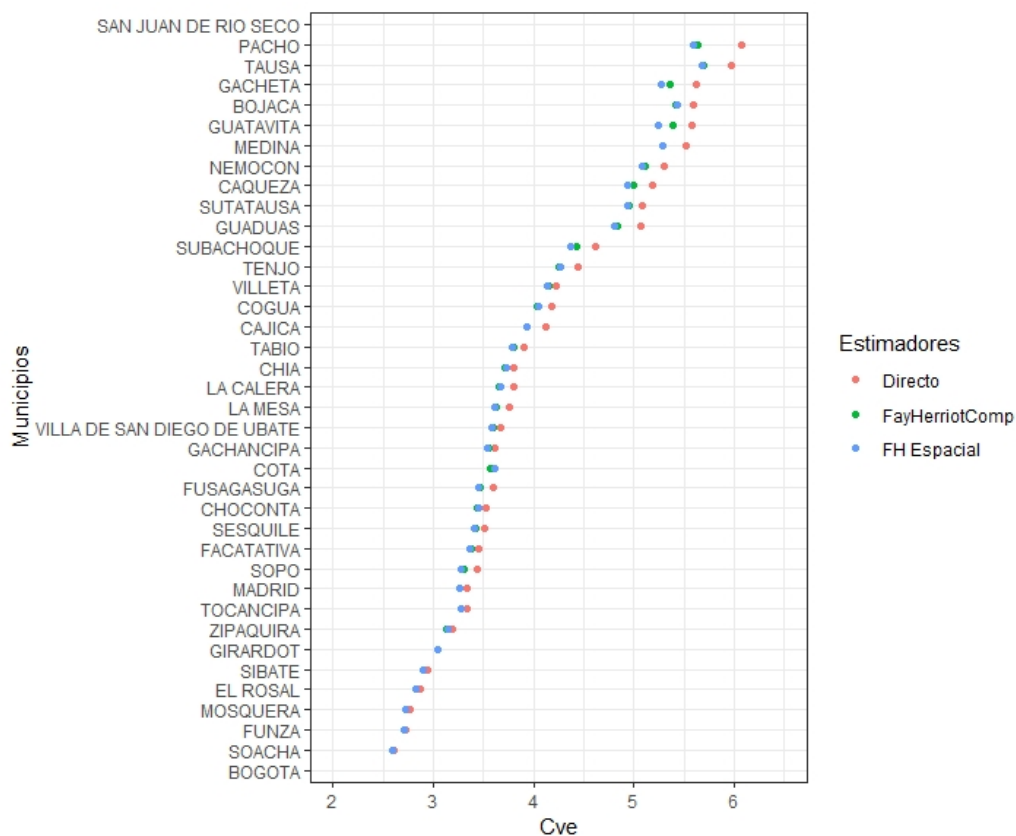


Figura 2: Comparación Coeficientes

8.2. Extrapolación

En el siguiente gráfico se observa el comportamiento del ingreso medio en todos los municipios de Cundinamarca, con las estimaciones proporcionadas por el modelo Fay Herriot Espacial.

Se observa claramente la diferencia de Bogotá frente a los otros municipios, sin embargo hay municipios en los que el ingreso medio es alto también, como en Ricaurte, Chía, Cota y La Calera, respectivamente.

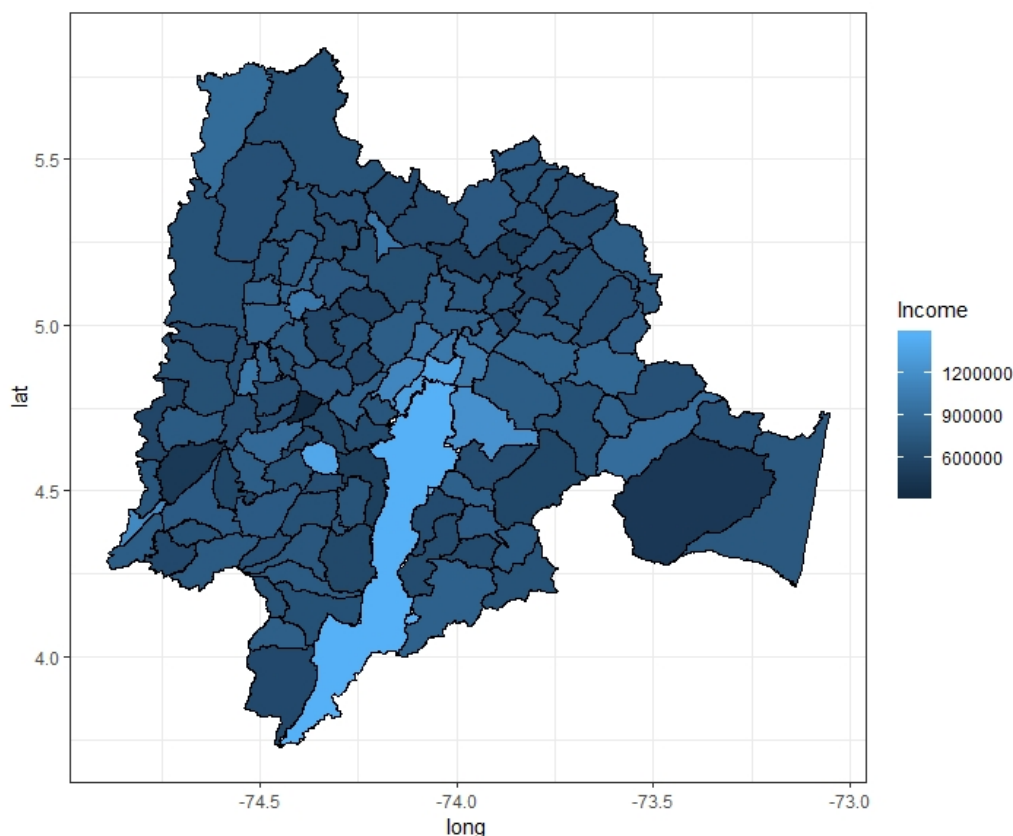


Figura 3: Extrapolation

9. Tasa de Desempleo

El caso de la tasa de desempleo presentó mayores dificultades en su estimación debido al comportamiento de las variables, por ello, se decidió aplicar un modelo basado solo en las componentes principales de todas las variables que tenían una correlación más alta que el 50 % positiva y negativa. Los resultados fueron los siguientes

9.0.1. Variables Escogidas

Las variables que se escogieron con base a la correlación frente a la estimación directa son las siguientes:

- Porcentaje de predios urbanos (2017)
- Porcentaje de inversión - Transporte (2017)
- Porcentaje de predios rurales (2017)
- Porcentaje población rural (2017)
- Numero de personas validadas en el Sisben (2016)

9.0.2. Resultados modelos

Para el modelo con solo las variables correlacionadas tenemos la información en la Tabla 3.

Tabla 3: Resultados

	<i>Dependent variable:</i>
	TDD
‘Porcentaje de predios urbanos’	0.0001 (0.0003)
‘Porcentaje de inversión - Transporte’	−0.001* (0.001)
‘Porcentaje de predios rurales’	
‘Porcentaje población rural’	−0.0001 (0.0003)
Numero.de.personas.validadas	−0.000 (0.000)
Constant	0.078*** (0.024)
Observations	38
R ²	0.336
Adjusted R ²	0.256
Residual Std. Error	0.018 (df = 33)
F Statistic	4.180*** (df = 4; 33)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Debido a que el ajuste de las variables para la tasa de desempleo se dificulta en gran medida, se optó por realizar un modelo solo con las dimensiones del análisis de componentes principales para todas las variables con mayor correlación, los resultados se encuentran en la Tabla 4:

Tabla 4: Resultados Coordenadas

	<i>Dependent variable:</i>
	TDD
Dim.2	0.007*** (0.001)
Dim.4	-0.004 (0.003)
Dim.5	-0.005* (0.003)
Dim.9	-0.008 (0.005)
Constant	0.067*** (0.002)
Observations	38
R ²	0.589
Adjusted R ²	0.539
Residual Std. Error	0.014 (df = 33)
F Statistic	11.802*** (df = 4; 33)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Se observa una clara diferencia entre los dos coeficientes de variación, en la figura 4 se puede detallar el comportamiento de los coeficientes de variación para cada municipio usando ambos modelos, consecuente a esto se escoge entonces el modelo con las coordenadas de las dimensiones para realizar los posteriores análisis.

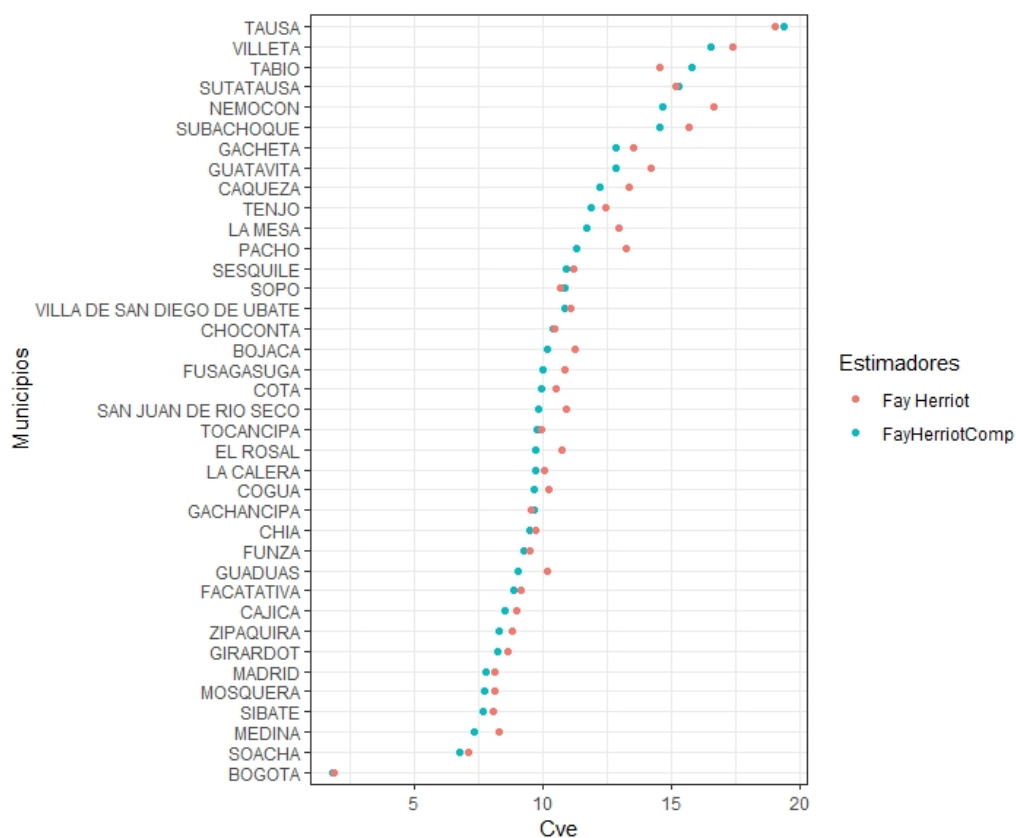


Figura 4: Elaboración propia

9.1. Comparación CVE's Estimación Directa, modelo Fay Herriot y Fay Herriot espacial

Al observar el gráfico de la Figura 5, debido a la notable diferencia, se asume que las estimaciones proporcionadas por el modelo Fay Herriot Espacial son mejores que las del modelo Fay Herriot, sin embargo se realiza una comparación de la media del coeficiente de variación con el fin de esclarecer la diferencia.

FayHerriotComp	Fay Herriot Espacial
11.11	10.22

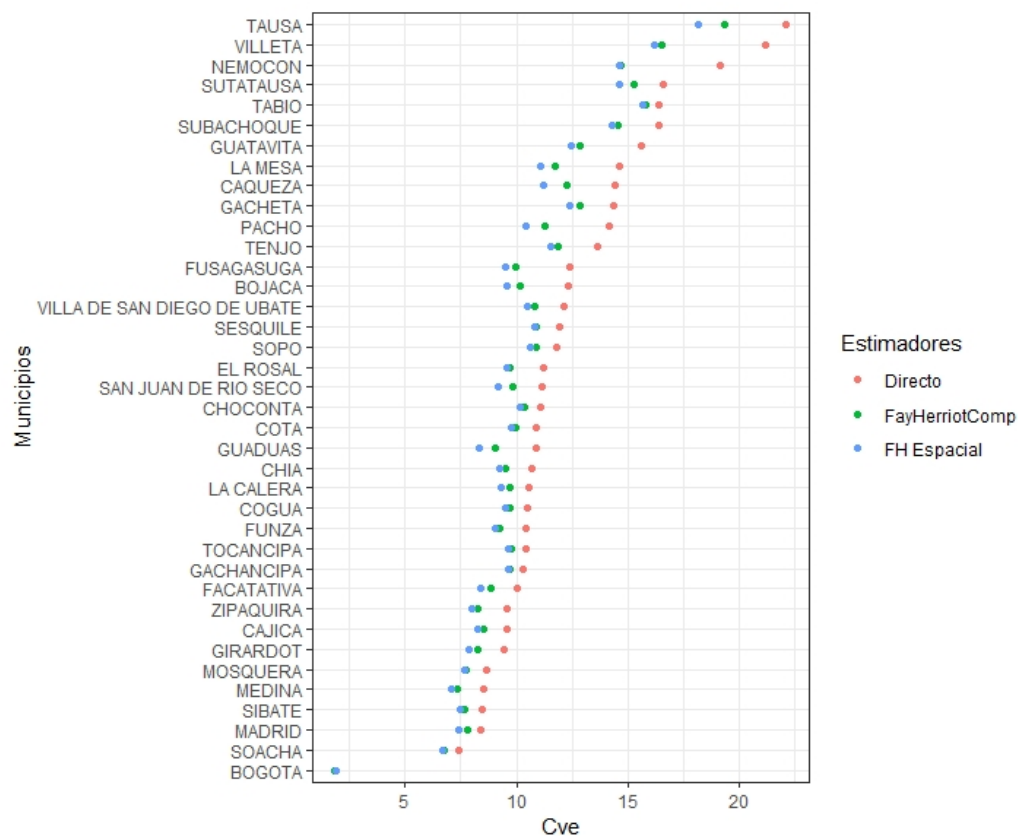


Figura 5: Elaboración propia

9.2. Extrapolación

En el siguiente gráfico se observa el comportamiento de la tasa de desempleo en todos los municipios de Cundinamarca, con las estimaciones proporcionadas por el modelo Fay Herriot Espacial.

Se puede observar que los departamentos en los que hay mayor tasa de desempleo son Tabio, Nemocón, Villeta y Subachoque, mientras en los municipios don la tasa es mas alta son Puerto Salgar, Paratebueno, Soacha y Madrid.

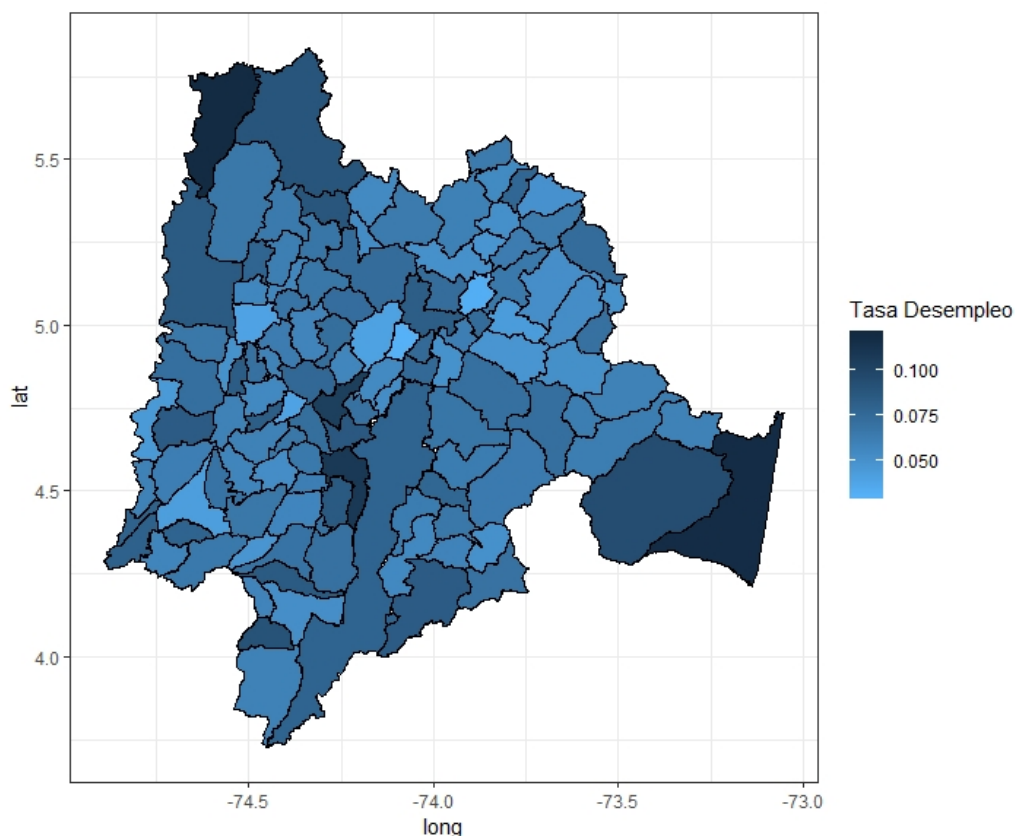


Figura 6: Extrapolation

10. Conclusiones

Se concluye entonces que para la selección de variables de un modelo, en particular un modelo Fay Herriot, con el uso de componentes principales se obtiene un mejor ajuste en el caso en que la dimensionalidad sea una problemática, que en efecto, en el estudio de encuestas multipropósitos o estudios censales la cantidad de variables es muy alta. Por ende con el uso de componentes principales se aprovecha la mayor cantidad de información posible con las coordenadas de cada individuo de un número determinado de componentes que dependen de la variable respuesta del modelo.

Además en la comparación del modelo Fay Herriot y su varación Espacial, se llega a la conclusión de que (contando con las mismas variables en ambos modelos) el efecto espacial en el modelo genera mejores predicciones por medio de medidas de precisión basadas en el error cuadrático medio, por ende es importante contar con la correlación espacial en un estudio de poblaciones.

Referencias

- [1] Gutierrez. H. A., *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*, 2009, Bogotá, Colombia, Ediciones de la U.
- [2] Molina. I., “Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas”, Series Estudios Estadísticos, No 97, (LC/TS.2018/82/Rev.1), Santiago, Comisión Económica para América Latina y el Caribe, (CEPAL), 2019
- [3] Molina. I. & RAO J.N.K., *Small Area Estimation Second Edition*, 2015, Hoboken, New Jersey, Wiley.
- [4] Ortiz, Felipe & F. Zea, José. (2018). Small Area Estimation Methodology (SAE) applied on Bogota Multipurpose Survey (EMB). *Romanian Statistical Review*
- [5] Morales, Domingo. (2015). ESTIMACION EN ÁREAS PEQUEÑAS: METODOS BASADOS EN MODELOS
- [6] Husson, Francois & Le Sebastien & Pages Jerome. *Exploratory Multivariate Analysis by examples using R.*, 2017, New York, Taylor & Francis Group
- [7] Pushpal K. Mukhopadhyay & Allen McDowell. *Small Area Estimation for Survey Data Analysis Using SAS Software*, 2011, SAS Institute Inc.
- [8] Särndal, C. E., B. Swensson, and Jan Wretman. *Model Assisted Survey Sampling*, 1992, 3rd ed.