Universidad de la República

Facultad de Ciencias Económicas y de Administración

# Mapas de Pobreza en Montevideo: una Aplicación de Estimación en Áreas Pequeñas

Trabajo final de grado presentado como requisito para la obtención del título

Licenciado en Estadística

*Estudiante:*
**Ignacio Javier Acosta Rodríguez**

*Tutora:*
**María Eugenia Riaño**

Montevideo, Marzo de 2023
Uruguay

El tribunal docente integrado por los abajo firmantes aprueba el trabajo final de grado:

# Mapas de Pobreza en Montevideo: una Aplicación de Estimación en Áreas Pequeñas

**Ignacio Javier Acosta Rodríguez**

Tutora: Lic. María Eugenia Riaño

Licenciatura en Estadística

**Puntaje** .....

**Tribunal**

Profesor Mag. Juan Pablo Ferreira.....

Profesor Mag. Fernando Massa.....

Profesora Lic. María Eugenia Riaño .....

**Fecha .....**

# Resumen

En el marco de la agenda 2030 para el desarrollo sostenible, la Organización de Naciones Unidas (ONU), en conjunto con sus países miembros se comprometen a trabajar para hacer frente a una serie de problemáticas que afectan al planeta tierra, entre ellas la pobreza. Poder identificar en el territorio los lugares en donde se concentra la pobreza es fundamental para la propuesta de políticas públicas y la intervención en el marco de los Objetivos de Desarrollo Sostenible (ODS).

Organismos internacionales como el Banco Mundial, CEPAL (Comisión Económica para América Latina y el Caribe) y FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura) recomiendan la elaboración de Mapas de Pobreza, de forma de identificar las regiones más vulnerables de un país o ciudad. La metodología recomendada para la construcción de los Mapas de Pobreza es la de **Estimación en Áreas Pequeñas**. En el presente trabajo se aplica esta metodología con el objetivo de estudiar la distribución de la pobreza a nivel de Barrios en Montevideo para el segundo semestre del año 2021 y primer semestre del año 2022.

Se presentan los aspectos conceptuales de **SAE** (Small Area Estimation) y los estimadores comúnmente utilizados cuando se trabaja con "áreas pequeñas". Se hace especial énfasis en la aplicación de modelos lineales mixtos (los cuales enmarcan los estimadores **SAE**). Seguido de ello, se introducen los modelos de área (en partícular el estimador **Fay-Herriot**). A continuación, se profundiza en modelos de área que toman en consideración la posible correlación espacial existente entre distintos dominios geográficos (más específicamente el estimador **Fay-Herriot espacial**). Finalmente, se elaboran los mapas con las alternativas mencionadas, utilizando como variables explicativas la cantidad de personas con nivel universitario o superior, cantidad de desocupados o inactivos y el número de hogares con 3 o más necesidades básicas insatisfechas.

Se tiene como resultado que todas las técnicas de estimación dentro del marco **SAE** logran mejorar la eficiencia de los estimadores calculados con un enfoque tradicional (es decir, las estimaciones basadas en el diseño). Se encontró que el estimador **Fay-Herriot espacial** con una matriz de vecindad definida por contigüidad es el que mejor se desempeña, con más del 65 % de los barrios con estimaciones de calidad.

**Palabras clave:** Muestreo, Muestreo en Áreas Pequeñas, Muestreo basado en modelos, Mapas de Pobreza, Modelos Mixtos, Modelos de área.

# Índice

# Índice de figuras

# 1. Introducción

En el marco de la agenda 2030 para el desarrollo sostenible, la Organización de Naciones Unidas (ONU), en conjunto con sus países miembros se comprometen a trabajar para hacer frente a una serie de problemáticas que afectan al planeta tierra. Es allí que surgen los Objetivos de Desarrollo Sostenible (ONU, 2016) . A partír de los mismos, el desarrollo de estadísticas confiables a niveles desagregados cobra vital importancia.

El primero de los 17 objetivos establece:

1) *"De aquí a 2030, erradicar para todas las personas y en todo el mundo la pobreza extrema (actualmente se considera que sufren pobreza extrema las personas que viven con menos de 1,25 dólares de los Estados Unidos al día)".*

2) *"De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales".*

Teniendo este objetivo en mente, el Banco Mundial (Bedi, Coudouel, y Simler, 2007) propone la utilización de mapas de pobreza para estudiar la distribución de la misma y diseñar así políticas públicas enfocadas en las regiones más vulnerables.

La implementación de encuestas para medir el fenómeno de la pobreza ha jugado un rol fundamental. A lo largo de los años, las encuestas por muestreo han sido utilizadas para obtener estimaciones de parámetros poblacionales de interés a partir de un subconjunto de individuos de la población objetivo.

Por lo general, los diseños muestrales logran una precisión aceptable cuando se trabaja a nivel poblacional, sin embargo, cuando se desea obtener estimaciones a un nivel más desagregado los errores de muestreo aumentan y en muchos casos esto lleva a obtener estimaciones poco confiables.

Cuanto más fina sea la partición de la población, mayor tienden a ser estos errores. En este paradigma en donde los tamaños muestrales son muy pequeños o incluso nulos, la inferencia tradicional basada en el diseño comienza a tener problemas.

Como estrategia para hacer frente a esta problemática surge la **inferencia basada en modelos**. A diferencia del caso anterior en donde la aleatoriedad proviene totalmente del mecanismo de selección de la muestra, en este último se supone la existencia de un

modelo superpoblacional "$\zeta$" que obra como mecanismo generador de los datos y es de este mismo de donde proviene la aleatoriedad.

En este trabajo se presentará el marco de **Small Area Estimation (SAE)** como herramienta metodológica para realizar estimaciones en áreas pequeñas.

Un "area pequeña" es un dominio, es decir, un subconjunto específico de la población. Esta subdivisión puede darse a partir de características más allá de lo geográfico, como pueden ser criterios sociodemográficos o combinaciones de ambos. Estos dominios cumplen que el tamaños de muestra efectivo no es lo suficientemente grande como para obtener estimaciones directas de calidad (Molina, 2019).

Los estimadores directos son aquellos que al momento de realizar estimaciones sobre un área, utilizan únicamente la información asociada a la misma (y no más) proveniente de la muestra. Por lo que el concepto de "área pequeña" no necesariamente tiene que ver con que tan extenso sea un territorio o que tantas personas integren un grupo, sino que refiere específicamente al tamaño de muestra esperado en el dominio.

Dentro de los modelos enmarcados en la metodología **SAE** existe una subdivisión entre aquellos a nivel de área o de individuo.

Los primeros construyen las estimaciones a partir de datos agregados a nivel de dominio. Ésta es una gran ventaja, a diferencia de los modelos de unidad, no es necesario que exista una relación biunívoca entre los individuos de la muestra y la información auxiliar. En países en donde la periodicidad de censos es baja y existe poco desarrollo de registros administrativos con fines estadísticos, esta alternativa permite integrar fuentes de información más versátiles como pueden ser información satelital provenientes de teledetección o información agrupada la cual por distintas razones no puede ser desglosada a nivel de persona. En este documento se desarrollará con mayor enfásis este tipo de análisis. Bajo el segundo enfoque (modelos a nivel de unidad), las estimaciones surgen de operar con las predicciones a nivel de los individuos.

A día de hoy no existen publicaciones haciendo uso de la metodología SAE para los barrios de Montevideo. Utilizando los microdatos publicados por el INE (Insituto Nacional de Estadística), tan solo se pueden obtener estimaciones directas cuya precisión es inaceptable en la mayoría de los barrios.

## 1.1. Objetivos

### 1.1.1. Objetivo General

Estimar la cantidad de personas por debajo del umbral de pobreza en cada barrio de Montevideo para el segundo semestre 2021 y primer semestre 2022.

### 1.1.2. Objetivos Específicos

- Comparar la eficiencia de los estimadores directos, Fay-Herriot (FH) y Fay-Herriot espacial (FH espacial).
- En el caso de FH espacial, construir los modelos en base a distintas $W$ (matriz de vecindad) con el fin de analizar el impacto de la elección de $W$ en las estimaciones finales.

## 1.2. Hipótesis

La hipótesis de trabajo es que los estimadores que consideran la correlación espacial en contexto SAE son más eficientes que los directos y que los estimadores FH que no incorporan la correlación espacial. A su vez, se espera que el estimador FH sea más eficiente que el estimador directo.

## 1.3. Justificación

La implementación de políticas públicas en grupos vulnerables de la sociedad es de gran importancia para los hacedores de políticas públicas. Consigo, conocer la distribución espacial de este conjunto de individuos es vital.

A su vez, desagregar esta información a niveles geográficos más pequeños permite visualizar de manera rápida y fácil lugares en donde se concentra la población más vulnerable. Un mapa de pobreza refleja la realidad territorial que no es vista cuando se estima la pobreza a nivel global.

# 2. Inferencia basada en el diseño vs basada en modelos

Se busca en este capítulo repasar brevemente los aspectos básicos del paradigma predominante, la inferencia basada en el diseño (por más detalle se recomienda leer (Särndal, Swensson, y Wretman, 1992)). También se introduce a la inferencia basada en modelos como alternativa.

## 2.1. Inferencia basada en el diseño

La inferencia basada en el diseño es aquella que supone que la aleatoriedad al momento de estimar un parámetro proviene pura y exclusivamente del mecanismo de recolección de los datos.

Dada una población $U = \{u_1, u_2, \ldots, u_N\}$ y una variable de interés $y$, la inferencia basada en el diseño establece que los valores poblacionales $y_1, y_2, \ldots, y_N$ (valores de la variable $y$ para los individuos de la población) no son más que valores fijos.

Bajo el paradigma de inferencia basada en el diseño, este es una medida de probabilidad definida sobre el conjunto $\mathcal{S} = \mathcal{P}(U)$ (conjunto de los subconjuntos incluidos en $U$, es decir, el conjunto de todas las posibles muestras). Por lo tanto, la probabilidad de que se seleccione la muestra $s$ es $\mathbb{P}(s \in \mathcal{S}) = p(s)$ a lo que se denomina como diseño muestral (Särndal y cols., 1992).

A partir de ello se derivan 2 propiedades relevantes:

1) $p(s) \geq 0, \forall s \in \mathcal{S}$

2) $\sum_{s \in \mathcal{S}} p(s) = 1$

Haciendo uso de $p(s)$ se definen las **probabilidades de inclusión**.

La inclusión de un individuo en la muestra puede ser resumido a la realización de una variable aleatoria Bernoulli con la siguiente función de cuantía:

$$\chi_k = \begin{cases} 1 & \text{si } k \in \mathcal{S} \\ 0 & \text{en otro caso} \end{cases} \tag{2.1}$$

Con esto, la probabilidad de inclusión del individuo k-ésimo en la muestra es:

$$\pi_k = \mathbb{P}(k \in \mathcal{S}) = \mathbb{P}(\chi_k = 1) = \sum_{s \ni k} p(s) \tag{2.2}$$

Otra definición relevante, especialmente al momento de estudiar la variabilidad de las estimaciones, es la probabilidad de inclusión de 2 individuos de manera simultánea. La misma se define como $\pi_{kl}$:

$$\pi_{kl} = \mathbb{P}(\{k \in \mathcal{S}\} \cap \{l \in \mathcal{S}\}) = \mathbb{P}(\chi_k \times \chi_l = 1) = \sum_{s \ni k \& l} p(s) \tag{2.3}$$

### 2.1.1. Estimador Horvitz-Thompson

En 1952, (Horvitz y Thompson, 1952) presenta el actualmente denominado **estimador Horvitz-Thompson**. Este estimador es el representante por excelencia de la inferencia basada en el diseño y el mismo se basa en el principio de $\pi$-expansión.

Este estimador define como ponderador al inverso de la probabilidad de inclusión del individuo, por lo que el estimador de Horvitz-Thompson de un total poblacional toma la siguiente expresión:

$$\hat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} \chi_k \ ^1 \tag{2.4}$$

Como puede verse, el valor del estimador depende de la composición de la muestra. Es por ello que la aleatoriedad del mismo proviene pura y exclusivamente del diseño $p(.)$. Este estimador es insesgado:

$$\mathbb{E}\left(\hat{t}_y\right) = \mathbb{E}\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right) = \mathbb{E}\left(\sum_{k \in U} \frac{y_k}{\pi_k} \chi_k\right) = \sum_{k \in U} \frac{y_k}{\pi_k} \mathbb{E}(\chi_k) = \tag{2.5}$$

$$\sum_{k \in U} \frac{y_k}{\pi_k} \left(\pi_k \times 1 + (1 - \pi_k) \times 0\right) = \sum_{k \in s} \frac{y_k}{\pi_k} \pi_k = \sum_{k \in U} y_k = t_y \tag{2.6}$$

[1] Se trabaja con la hipótesis de que el diseño es representativo, es decir, $\pi_k > 0, \forall k \in U$.

Se puede demostrar también que:

$$\mathbb{VAR}(\hat{t}_y) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l} \tag{2.7}$$

$$\widehat{\mathbb{VAR}}(\hat{t}_y) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_k y_l}{\pi_k \pi_l} \tag{2.8}$$

Bajo diseños "sencillos" (como el Simple o Estratificado Simple) el estimador de la varianza del estimador tiene una forma cerrada (esta es otra característica que hace que este enfoque sea fácilmente aplicable en algunos contextos).

Sin embargo, este enfoque no siempre es viable. Situaciones tales como la no respuesta, dominios de interés muy chicos, poblaciones raras (o inaccesibles) son algunos ejemplos en donde los supuestos de inferencia basada en el diseño comienzan a fallar.

Debido a lo mencionado anteriormente es que la inferencia basada en modelos brinda un marco alternativo para afrontar los problemas antes mencionados.

## 2.2. Inferencia basada en modelos

En el enfoque basado en modelos el tratamiento de las observaciones $y_i$ cambia por completo, las mismas pasan a ser consideradas como realizaciones de un modelo estocástico superpoblacional $\zeta$. Tanto las $y_i \in s$ como las $y_i \in U - s$ se consideran variables aleatorias.

Debido a ello, el total poblacional y el estimado ($t_y$ y $\hat{t}_y$) son también variables aleatorias. Lo único que es tomado como constante son las unidades incluidas en la muestra.

Este modelo puede ser definido de manera exhaustiva, es decir, se específica explícitamente al proceso estocástico generador de los valores poblacionales. Por otro lado, y más comunmente utilizado, el modelo se especifica "débilmente". Bajo esa coyuntura se especifican algunas de las características de la distribución de la variable de interés como pueden ser los momentos de primer y segundo orden. En cualquiera de los casos van a existir parámetros superpoblacionales que caracterizan al modelo y no serán conocidos (y que bajo el paradigma de inferencia basada en modelos, se buscan estimar).

Un ejemplo de modelo superpoblacional es el **modelo de regresión lineal simple**. Dadas 2 variables $x_i$ y $y_i$ con $i = 1, 2, \ldots, N$, se supone que la esperanza de $y_i$ condi-

cionada a $x_i$ es lineal en $x_i$, es decir que $\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$. A su vez se asume que $\mathbb{VAR}(y_i|x_i) = \sigma^2, \forall i$.

En este caso los parámetros superpoblacionales son $\beta_0$ y $\beta_1$. Se resalta nuevamente que estos parámetros son hipotéticos y desconocidos (incluso tras la realización de un censo). Tan solo se utilizan para caracterizar como los valores $Y$ tienden a cambiar con los valores de $X$.

Dada una muestra, existen varias formas de estimar los parámetros superpoblacionales de un modelo. En el caso del modelo de regresión podríamos utilizar por ejemplo MCO. Sin embargo, la mayoría de los métodos de estimación suponen que la muestra utilizada para estimar los parámetros del modelo proviene de una muestra aleatoria del modelo superpoblacional. Por lo tanto, no existiría relación entre los valores generados por el modelo y el mecanismo de selección de la muestra.

Por lo general, se tiende a utilizar diseños que favorecen la presencia de ciertas unidades en la muestra. Pasar por alto esto puede llegar a ser peligroso, ignorar el diseño al momento de estimar un modelo podría llevar a estimaciones sesgadas (Chambers, 2012).

### 2.2.1. Diseño no informativo

Este concepto es fundamental en muestreo de poblaciones finitas, tanto en la inferencia basada en modelos como en la inferencia basada en el diseño. Conceptualmente refiere a que un diseño es no informativo para realizar inferencia sobre el conjunto de parámetros superpoblacionales del modelo de una variable dada, si el modelo se mantiene válido sobre el conjunto de datos muestrados.

Formalmente, dado $\theta$ el parámetro de interés que es función de $Y$. Se denomina:

- $X_U$ a la matriz de variables auxiliares de la población ($X_{ij}$ conocida $\forall i = 1, 2, \ldots, N$ y $\forall j = 1, 2, \ldots, J$).
- $Y_U$ los valores de la variable de interés $Y$ para la población.
- $s$ el conjunto de individuos pertenecientes a la muestra.
- $Y_s$ a los valores que toma la variable $Y$ en los individuos de la muestra.

Se dice que un mecanismo de muestreo es no informativo para realizar inferencia sobre $\theta$ dado $X_U$ si la distribución de $Y_s$ dado $X_U$ es la misma que la de $Y_U$ dado $X_U$ restricto

a los individuos de s.

$$f_{Y_s|X_U}(y) = f_{Y_U|X_U}|_{\{y \in s\}}(y) \tag{2.9}$$

Esto se traduce en que el mecanismo de selección de los datos contiene información nula con respecto a los parámetros superpoblacionales. (Chambers, 2012)

Es esta condición la que permite realizar inferencia sobre el conjunto de individuos que se encuentran por fuera de la muestra haciendo uso del modelo ajustado con las observaciones de aquellos que sí pertenecen a la misma.

### 2.2.2. Estimación de los modelos

Basándose en (Chambers, 2012), volviendo al estudio de las propiedades del modelo, surge el interés de estudiar la media y varianza de $\left(t_y - \hat{t}_y\right)$. Valores cercanos a 0 en ambas cantidades son deseables.

Surge entonces la duda de:

- ¿Cuál será el mejor estimador $\hat{t}_y$ de $t_y$ dado el modelo especificado?
- ¿Cuál será la mejor manera de seleccionar la muestra?

Para el primer punto se consideran el sesgo de predicción, definido como $\mathbb{E}\left[(\hat{t}_y - t_y)\right]$ y la varianza de predicción definida como $\mathbb{VAR}\left[\left(t_y - \hat{t}_y\right)\right]$. A partir de ello surge el error cuadrático medio de $\hat{t}_y$ que toma la expresión:

$$\mathbb{ECM}(\hat{t}_y) = \mathbb{E}\left[(\hat{t}_y - t_y)^2\right] = \left[\mathbb{E}\left(\hat{t}_y - t_y\right)\right]^2 + \mathbb{VAR}\left(t_y - \hat{t}_y\right)$$

Se dice que el predictor $\hat{t}_y$ es insesgado bajo el modelo si el sesgo de predicción es nulo.

Luego, para llegar a una forma concreta del estimador se recuerda que dado un conjunto de variables conocido $X$, el valor que minimiza al error cuadrático medio será $\mathbb{E}(\hat{t}_y|X)$. Claramente esta esperanza dependerá (bajo el modelo escogido) de parámetros desconocidos cuya estimación derivará en aproximaciones "plug-in" de $t_y$.

Se verá en próximos capítulos la metodología adecuada para adaptar este concepto a SAE al estudiar el concepto de **Best Linear Unbiased Predictor (BLUP)** y **Empirical**

**Best Linear Unbiased Predictor (EBLUP)**.

## 2.3. Ventajas del enfoque basado en modelos

A nivel general, el enfoque basado en modelos se caracteriza por:

- Una **reducción de la varianza de las estimaciones** (en pos de una porción de sesgo).
- El uso de información auxiliar permite alcanzar estimaciones confiables con un **tamaño de muestra menor**.
- Permite tener estimaciones de grupos de la población difíciles de alcanzar.

## 2.4. Desventajas del enfoque basado en modelos

- La suposición de un modelo implica que el mismo sea verdadero. En caso de que esto no suceda, las estimaciones serán sesgadas.
- Se requiere de información auxiliar de calidad para construir modelos que logren describir el comportamiento de la variable de interés.

A lo largo del trabajo se seguirán explorando distintas propiedades de los estimadores enmarcados en el muestreo basado en modelos, más partícularmente aquellos utilizados en **SAE**.

# 3. Estimación en áreas pequeñas (SAE)

Por definición, un área pequeña es aquella cuyo tamaño de muestra efectivo es insuficiente para realizar estimaciones directas confiables. La metodología de SAE se basa fuertemente en los **estimadores indirectos**.

Los estimadores indirectos son aquellos que no solo hacen uso de la información relevada asociada al área de la cual se desea obtener una estimación, sino que se apoya de la información asociada al resto de los dominios para así ganar eficiencia.

La ganancia de eficiencia se logra suponiendo cierta estructura de homogeneidad entre las áreas.

(Chambers, 2012) establece que existen 2 modalidades para la aplicación de este enfoque:

- Existe un modelo global, válido en todas las áreas.
- Existe un modelo distinto por área que se construye combinando un componente que representa la homogeneidad entre áreas y otro para la heterogeneidad entre ellas.

La primer modalidad engloba a los **estimadores sintéticos** y la segunda a los modelos lineales mixtos utilizados en **Small Area Estimation**.

## 3.1. Estimadores sintéticos

Los estimadores sintéticos son aquellos que suponen una estructura de homogeneidad entre las áreas que proviene de un conjunto de parámetros en común. Es decir, los valores que toma la variable de interés fluctuan tan solo con base en un conjunto de variables auxiliares (Molina, 2019).

Este supuesto es muy restrictivo y lleva por lo general a grandes porciones de sesgo. Es por ello que estos estimadores (poco realistas) no son recomendados en la práctica.

A modo de ejemplo y por su estrecha relación con **SAE** se introducirán al **estimador sintético de regresión a nivel de área** y a los **estimadores compuestos**.

### 3.1.1. Estimador de regresión a nivel de área sintético (REG-SYN)

Los estimadores sintéticos de regresión pueden plantearse a nivel de área y a nivel de individuo. Si bien en esta tesis se presenta el **REG-SYN** a nivel de área para ser coherente con el enfoque escogido, la adaptación es directa. Esta subsección busca adaptar de manera resumida lo presentado por (Molina, 2019).

Se supone que existe información auxiliar a nivel de área, sea $x_d$ el vector de variables auxiliares asociadas al área $d$, con $x_d \in \mathbb{R}^p$.

Se supone también que la variable de interés asociada al dominio $d$, $\delta_d$, sigue un comportamiento lineal con respecto a las variables auxiliares. Es decir: $\delta_d = {x_d}'\beta,\ \forall d = 1, \ldots, D$.

Claramente, el problema inicial es el desconocimiento de esas cantidades, es por ello que se pasa a considerar las estimaciones directas $\hat{\delta}_d$ donde:

$$\hat{\delta}_d = {x_d}'\beta + \varepsilon_d$$

Bajo los supuestos:

- $\mathbb{E}(\varepsilon_d) = 0$
- $\mathbb{COV}(\varepsilon_{d1}, \varepsilon_{d2}) = 0, \forall d_1 \neq d_2$
- $\mathbb{VAR}(\varepsilon_d) = \psi_d$, con $\psi_d$ conocido.

Se estima $\beta$ a partir de **MCG** (mínimos cuadrados generalizados, ver (Rencher, 2008)), obteniendo:

$$\hat{\beta} = \left( \sum_{d=1}^{D} \psi_d^{-1} x_d {x_d}' \right)^{-1} \sum_{d=1}^{D} \psi_d^{-1} x_d \hat{\delta}_d \tag{3.1}$$

y finalmente:

$$\hat{\delta}^{\text{REG-SYN}} = {x_d}'\hat{\beta} \tag{3.2}$$

Este estimador es capaz de disminuir en gran medida la varianza en comparación con la estimación directa, sin embargo, debido a su potencial nivel de sesgo utilizar a la varianza como medida de variabilidad es en la mayoría de los casos errado. En el caso de los intervalos de confianza, la amplitud de los mismos en un contexto con mucho sesgo no refleja el nivel de cobertura real. Una alternativa es el uso del $\mathbb{ECM}$, sin embargo, no existen aproximaciones estables del mimsmo.

**Observación I:**

$$\mathbb{B}(\hat{\delta}_d|\beta) = \delta_d - \mathbb{E}(\hat{\delta}_d|\beta) = \delta_d - x_d{}'\beta \tag{3.3}$$

Esto quiere decir que dado el valor real de $\beta$, el sesgo no depende del tamaño muestral, por lo que el estimador no es asintóticamente insesgado.

**Observación II:**

Es posible obtener estimaciones en dominios con tamaño de muestra nulo.

Se detalla a continuación otra clase de estimadores que "ancla" a los estimadores sintéticos (en partícular el **REG-SYN**) con el estimador de **Fay-Herriot**.

## 3.2. Estimadores compuestos

Los estimadores compuestos buscan "absorber" las buenas propiedades de los estimadores directos (insesgadez aproximada, asintoticidad) y de los estimadores sintéticos (varianza pequeña) a partir de una combinación convexa de ambos.

Dado un estimador directo $\hat{\delta}_d^{\mathrm{DIR}}$ y un estimador sintético $\hat{\delta}_d^{\mathrm{SYN}}$ se define a un estimador compuesto asociado $\hat{\delta}_d^{\mathrm{C}}$ para el parámetro $\hat{\delta_d}$ como:

$$\hat{\delta}_d^{\mathrm{C}} = \phi_d \hat{\delta}_d^{\mathrm{DIR}} + (1-\phi_d)\hat{\delta}_d^{\mathrm{SYN}},\ \forall d = 1, 2, \ldots, D,\ 0 \leq \phi_d \leq 1 \tag{3.4}$$

La elección del valor de $\phi_d$ puede obtenerse o bien minimizando una aproximación del $\mathbb{ECM}$ o siguiendo el método propuesto por (Drew, Singh, y Choudhry, 1982) que define el valor del parámetro $\phi_d$ es función del tamaño de muestra efectivo del dominio.

**Método minimizar** $\mathbb{ECM}$:

Este método asigna a $\phi_d$ el valor que minimiza el error cuadrático medio del estimador compuesto en el área.

Una fórmula cerrada correspondiente a la estimación del parámetro se obtiene muy fácilmente derivando e igualando a 0.

$$\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{C}}\right) = \mathbb{ECM}\left(\phi_d\hat{\delta}_d^{\mathrm{DIR}} + (1-\phi_d)\hat{\delta}_d^{\mathrm{SYN}}\right) \tag{3.5}$$

$$= \phi_d^2\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{DIR}}\right) + (1-\phi_d)^2\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{SYN}}\right) + 2\phi_d(1-\phi_d)\mathbb{E}\left(\left[\hat{\delta}_d^{\mathrm{DIR}} - \delta_d\right]\left[\hat{\delta}_d^{\mathrm{SYN}} - \delta_d\right]\right) \tag{3.6}$$

Trabajando con la expresión se llega a que el valor de $\phi_d$ que minimiza el $\mathbb{ECM}$ es:

$$\phi_d^* = \frac{\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{SYN}}\right) - \mathbb{E}\left(\left[\hat{\delta}_d^{\mathrm{DIR}} - \delta_d\right]\left[\hat{\delta}_d^{\mathrm{SYN}} - \delta_d\right]\right)}{\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{DIR}}\right) + \mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{SYN}}\right) - 2\mathbb{E}\left(\left[\hat{\delta}_d^{\mathrm{DIR}} - \delta_d\right]\left[\hat{\delta}_d^{\mathrm{SYN}} - \delta_d\right]\right)} \tag{3.7}$$

Asumiendo que $\mathbb{E}\left(\left[\hat{\delta}_d^{\mathrm{DIR}} - \delta_d\right]\left[\hat{\delta}_d^{\mathrm{SYN}} - \delta_d\right]\right)$ es pequeño en relación al error cuadrático medio del estimador sintético se aproxima a $\phi_d$ como:

$$\phi_d^* \simeq \frac{\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{SYN}}\right)}{\mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{SYN}}\right) + \mathbb{ECM}\left(\hat{\delta}_d^{\mathrm{DIR}}\right)} \tag{3.8}$$

A partir de esta fórmula se explicita fácilmente el mecanismo de ponderación entre los 2 estimadores.

En caso de ser el $\mathbb{ECM}$ del estimador sintético significativamente más grande que el del estimador directo, $\phi_d^*$ tiende a 1. Esto se traduce a que el estimador compuesto se verá mayormente definido por la estimación directa. Esto sucede cuando el estimador sintético posee un nivel de sesgo elevado. Viceversa cuando la varianza del estimador directo es muy grande en comparación con el error del estimador sintético.

En la práctica se utiliza la siguiente aproximación:

$$\hat{\phi}_d^* = \frac{\widehat{\mathbb{ECM}}\left(\hat{\delta}_d^{\mathrm{SYN}}\right)}{\left(\hat{\delta}_d^{DIR} - \hat{\delta}_d^{SYN}\right)^2} \tag{3.9}$$

La misma surge de tomar en consideración que el $\mathbb{ECM}$ del estimador sintético puede ser estimado a partir del siguiente estimador aproximadamente insesgado:

$$\widehat{\mathbb{ECM}}\left(\hat{\delta}_d^{\text{SYN}}\right) = \left(\hat{\delta}_d^{SYN} - \hat{\delta}_d^{DIR}\right)^2 - \hat{\mathbb{V}}\left(\hat{\delta}_d^{SYN} - \hat{\delta}_d^{DIR}\right) + \hat{\mathbb{V}}\left(\hat{\delta}_d^{SYN}\right) \tag{3.10}$$

$$= \left(\hat{\delta}_d^{SYN} - \hat{\delta}_d^{DIR}\right)^2 - \hat{\mathbb{V}}\left(\hat{\delta}_d^{DIR}\right) + 2\widehat{\mathbb{COV}}\left(\hat{\delta}_d^{DIR}, \hat{\delta}_d^{SYN}\right) \tag{3.11}$$

Donde suponiendo $\mathbb{COV}\left(\hat{\delta}_d^{DIR}, \hat{\delta}_d^{SYN}\right) \approx 0$ (Lahiri y Pramanik, 2018) y reemplazando en la ecuación (3.8) se llega a la expresión planteada.

Por lo general, la estimación de $\phi_d$ a nivel de área suele ser inestable. Una alternativa es el uso de un $\Phi$ general que sea el resultado de promediar todas las estimaciones (Rao y Molina, 2014).

**Método basado en el tamaño de muestra efectivo**:

El método propuesto por (Drew y cols., 1982) plantea que el valor de $\phi_d$ dependa del tamaño muestral efectivo del área y define:

$$\phi_d = \begin{cases} 1, \text{si } \hat{N}_d \geq \eta.N_d \\ \frac{\hat{N}_d}{\eta.N_d}, \text{ si } \hat{N}_d < \eta.N_d \end{cases} \tag{3.12}$$

Siendo $\eta$ un número entre 0 y 1.

Esta estimación de $\phi$ suele asignar valores cercanos a 1, por lo que se le da la mayor parte del peso al estimador directo ganando muy poca eficiencia.

La relevancia de los estimadores compuestos radica en la explicitación del compromiso sesgo-varianza (estimador sintético - estimador directo). Las estrategias para definir $\phi_d$ pueden llegar a ser defectuosas.

Veremos que en el caso de las estimaciones basadas en modelos, la explicitación de un parámetro asociado a la heterogeneidad entre áreas incluido en un modelo de regresión, llevará a optimizar (siguiendo la distribución de probabilidad asumida por el modelo planteado) la elección de $\phi_d$.

## 3.3. Estimación en áreas pequeñas

Los estimadores de SAE surgen como alternativa mejorada a los estimadores sintéticos. En el sentido de que si bien suponen un cierto nivel de homogeneidad entre los dominios (que generalmente viene dada por la estructura paramétrica de los modelos) también agregan componentes que representan heterogeneidad en las áreas que no es explicada por el conjunto de variables auxiliares. Esto se llevará a cabo a partir de modelos mixtos (Molina, 2019).

### 3.3.1. Modelos lineales mixtos

En este texto se trabajará en el contexto de los modelos **lineales** mixtos. Los modelos lineales mixtos implican una generalización de los modelos lineales clásicos. Estos se caracterizan por la presencia de:

- **Efectos fijos:** un conjunto de covariables cuya relación con la variable de respuesta es homógenea a nivel poblacional, es decir, se comporta de la misma manera.
- **Efectos aleatorios:** factores categóricos cuyos niveles (en la muestra) pueden pensarse como una muestra de un espacio muestral. Una particularidad es que **no todos** los niveles son de interés. Los mismos son representados (a diferencia de los efectos fijos) como variables aleatorias (West, Welch, y Galecki, 2007).

A nivel general, un modelo lineal mixto puede escribirse como:

$$y = X\beta + Zu + \varepsilon$$

Donde $X, Z$ son matrices constantes y $u, \varepsilon$ son variables aleatorias.

### 3.3.2. BLUP para modelos lineales mixtos

Se buscará en esta sección estimar $\beta$ y $u$ siguiendo el esquema para modelos lineales mixtos generales propuestos por (Rao y Molina, 2014). Dado un modelo mixto general:

$$y = X\beta + Zu + \varepsilon$$

Con $X, Z \in \mathcal{M}_{D\times p}(\mathbb{R})$ constantes. $u, \varepsilon$ distribuidos independientemente cumpliendo:

- $u \sim (0_D, G)$
- $\varepsilon \sim (0_D, R)$

Se asume que existe un de conjunto parámetros asociados a la varianza $\omega = (\omega_1, \omega_2, ..., \omega_q)'$ tales que $\mathbb{VAR}(y) = V(\omega) = Z'GZ + R$.

Al momento de encontrar los *"Best linear unbiased predictors"* (es decir, los estimadores insesgados de mínimo error cuadrático medio) asumiremos estos parámetros conocidos.

Supongamos que nos interesa estimar una combinación lineal de $\beta$ y $u$ definida como: $\mu = I'\beta + m'u$ con $I$ y $m$ vectores constantes. Una manera de estimar esa cantidad a partir de un estimador lineal queda definida como $\hat{\mu} = a'y + b$. Con $a$ y $b$ vectores constantes.

En primer lugar, para que el estimador sea insesgado para el modelo debe cumplirse que:

$$\mathbb{E}(\mu) = \mathbb{E}(\hat{\mu}) \tag{3.13}$$

$$\mathbb{E}(\mu) = \mathbb{E}(I'\beta + m'u) = I'\beta + \mathbb{E}(u) = I'\beta \tag{3.14}$$

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(a'y + b) = a'Ey + b = a'X\beta + b \tag{3.15}$$

A partir de ello se deduce que $\hat{\mu}$ es insesgado si y solo si:

- $b = 0_D$
- $a'X = I'$

Por otro lado:

$$\mathbb{ECM}(\hat{\mu}) \overset{Insesgadez}{=} \mathbb{VAR}(\hat{\mu} - \mu) = \mathbb{VAR}(a'y - I'\beta - m'u) \tag{3.16}$$

$$= a'Va - 2a'ZGm + m'Gm \tag{3.17}$$

Aplicando extremos condicionados de Lagrange tenemos que:

$$\frac{\partial \mathbb{ECM}}{\partial a} = 2Va - 2ZGM \tag{3.18}$$

Luego, tomando como multiplicador $2\lambda$ (para simplificar cálculos) se llega a que:

$$\frac{\partial \mathbb{ECM}}{\partial a} + 2\lambda \frac{\partial (a'X - I')}{\partial a} = 0 \tag{3.19}$$

$$\Leftrightarrow 2Va - 2ZGM + 2\lambda X = 0 \tag{3.20}$$

$$\Leftrightarrow a = V^{-1}ZGM - V^{-1}X\lambda \tag{3.21}$$

Por otro lado, trabajando con la restricción $a'X = I'$ y sustituyendo por el valor de $a$ encontrado en el paso anterior se llega a que:

$$\lambda = -(X'V^{-1}X)^{-1}I + (X'V^{-1}X)^{-1}X'V^{-1}ZGM \tag{3.22}$$

Finalmente, remplazando el valor de $\lambda$ hayado en $a$ se obtiene que:

$$\hat{\mu}^* = a'y = I'\left(X'V^{-1}X\right)^{-1}X'V^{-1}y + m'GZ'V^{-1}\left(y - X(X'V^{-1}X)^{-1}X'V^{-1}y\right) \tag{3.23}$$

A partir de esto se deduce que los BLUP toman la siguiente expresión

$$\begin{aligned} \hat{\beta} &= (X'V^{-1}X)^{-1}X'V^{-1}y \\ \hat{u} &= GZ'V^{-1}(y - X\hat{\beta}) \end{aligned} \tag{3.24}$$

**Observación:** La estimación de los parámetros es independiente de la distribución de los mismos.

### 3.3.3. Error cuadrático medio del BLUP

Para derivar el $\mathbb{ECM}$ del **BLUP**, a quien en esta sección se denota como $t(\omega, y)$, se comienza notando que:

$$t(\omega, y) = b'\left(y - X\hat{\beta}\right) + I'\hat{\beta} \tag{3.25}$$

Donde $b' = m'GZ'V^{-1}$. Sumando y restando $I'\beta$ y $b'X\beta$, siendo $\beta$ el valor superpoblacional del parámetro (el "verdadero"), se llega a:

$$t(\omega, y) = b' \left(y - X\hat{\beta}\right) + I'\hat{\beta} + I'\beta - I'\beta + b'X\beta - b'X\beta \tag{3.26}$$
$$= b'y - b'X\hat{\beta} + I'\hat{\beta} + I'\beta - I'\beta + b'X\beta - b'X\beta \tag{3.27}$$
$$= b' (y - X\beta) + I'\beta - b'X\hat{\beta} + I'\hat{\beta} - I'\beta + b'X\beta \tag{3.28}$$
$$= t^* (\omega, \beta, y) + (I' - b'X) \hat{\beta} - (I' - b'X) \beta \tag{3.29}$$
$$= t^* (\omega, \beta, y) + (I' - b'X) \left(\hat{\beta} - \beta\right) \tag{3.30}$$
$$= t^* (\omega, \beta, y) + d' \left(\hat{\beta} - \beta\right) \tag{3.31}$$

Por lo que el estimador lineal dado el parámetro $\beta$ desconocido puede ser reescrito como un primer término asociado al valor del estimador cuando $\beta$ es conocido $[t^* (\omega, \beta, y)]$ y un segundo término que se basa en la diferencia entre el $\beta$ estimado y el poblacional.

A su vez, notando que:

$$t^* (\omega, \beta, y) - \mu = I'\beta + b' (y - X\beta) - I'\beta + m'u \tag{3.32}$$
$$= b' (Zu + \varepsilon) - m'u \tag{3.33}$$

$$d' \left(\hat{\beta} - \beta\right) = (u'Z' + \varepsilon') V^{-1} \tag{3.34}$$

Y especialmente que $t^* (\omega, \beta, y) - \mu$ y $d' \left(\hat{\beta} - \beta\right)$ son independientes a partir de que $\mathbb{E} ([b' (Zu + \varepsilon) - m'u] [(u'Z' + \varepsilon') V^{-1}]) = 0$.

Se llega a que el $\mathbb{ECM}(\hat{\mu})$ puede ser visto como:

$$\begin{aligned} \mathbb{ECM} (\hat{\mu}) &= \mathbb{ECM} \left(t^* (\omega, \beta, y) + d' \left(\hat{\beta} - \beta\right)\right) \\ &= \mathbb{ECM} (t^* (\omega, \beta, y)) + \mathbb{VAR} \left(d' \left(\hat{\beta} - \beta\right)\right) \\ &= g_1 (\omega) + g_2 (\omega) \end{aligned}$$

Desarrollando términos y reordenando se llega a que:

$$g_1(\omega) = \mathbb{VAR} (t^* (\omega, \beta, y) - \mu) = m' \left(G - GZ'V^{-1}ZG\right) m \tag{3.35}$$

$$g_2(\omega) = d' \left(X'V^{-1}X\right)^{-1} d \tag{3.36}$$

**Observación:** El segundo término refleja la variabilidad del estimador de $\beta$.

Tener una fórmula cerrada para la estimación de los parámetros y para el $\mathbb{ECM}$ del estimador de $\mu$ se corresponde con un primer acercamiento al estudio de los estimadores, sin embargo, no se debe olvidar que estas fórmulas fueron construidas con base en un supuesto que en la práctica es inalcanzable: **el conocimiento de los parámetros de varianza** $\omega$ (lo cual implica tener una especificación completa para las matrices $G$ y $R$).

Por lo que para "materializar" un estimador de $\mu$ se trabaja con un estimador en 2 fases. Primero se estima $\omega$, luego se remplazan estos parámetros en las fórmulas previamente calculadas. Es aquí que surge el estimador de tipo "plug-in" $\hat{t}(\hat{\omega}, y)$ denominado **Empirical Best Linear Umbiased Predictor (EBLUP)**, es decir, el mejor estimador lineal empírico insesgado.

Más adelante se estudiarán las técnicas propuestas en (Rao y Molina, 2014) para estimar los parámetros de varianza y el $\mathbb{ECM}$ del **EBLUP** entre otras cosas.

### 3.3.4. Estimación del EBLUP a partir de ML

El primer método propuesto para estimar los parámetros de varianza $\omega = (\omega_1, \omega_2, \ldots, \omega_q)$ es el de máxima verosimilitud. Para ello se asume la **normalidad de** $u$ **y** $\varepsilon$. Es con este supuesto con el que generalmente se trabaja al momento de utilizar estimadores de SAE basados en modelos mixtos. En lo que sigue del trabajo se dará por sentado el mismo.

Bajo este supuesto se tiene que:

$$y \sim \mathcal{N}(X\beta, Z'GZ + R) \tag{3.37}$$

La expresión de la función de verosimilitud es:

$$l(\beta, \sigma) = c - \frac{1}{2}\left(\ln(|V|) + (y - X\beta)'V^{-1}(y - X\beta)\right) \tag{3.38}$$

Aplicando las propiedades:

- $\frac{\partial \ln(|A|)}{\partial \omega_j} = tr\left(A^{-1}\frac{\partial A}{\partial \omega_j}\right)$
- $\frac{\partial A^{-1}}{\partial \omega_j} = -A^{-1}\frac{\partial A}{\partial \omega_j}A^{-1}$

se llega a que $\omega_j(\beta, \omega)$, definida como la derivada de la log-verosimilitud respecto a $\omega_j$, es:

$$\omega_j(\beta, \omega) = -\frac{1}{2}\left(\text{tr}(V^{-1}V_{(j)}) + (y - X\beta)'V^{(j)}(y - X\beta)\right) \tag{3.39}$$

- $V_{(j)} = \frac{\partial V}{\partial \omega_j}$
- $V^{(j)} = \frac{\partial V^{-1}}{\partial \omega_j} = -V^{-1}V_{(j)}V^{-1}$

La matriz de información de Fisher queda definida por:

$$\mathcal{I}_{jk} = \frac{1}{2}\text{tr}\left(V^{-1}V_{(j)}V^{-1}V_{(k)}\right) \tag{3.40}$$

Aplicando el método score de Fisher se itera hasta alcanzar convergencia:

$$\omega^{(n+1)} = \omega^{(a)} + \left[\mathcal{I}_{jk}(\omega^{(a)})^{-1}\right] S\left(\hat{\beta}(\omega^{(a)}), \omega^{(a)}\right) \tag{3.41}$$

Al alcanzar convergencia, se estiman todos los parámetros del modelo ($\beta, V, u$, etc) haciendo uso de $\hat{\omega}_{\text{ML}}$ y las fórmulas derivadas de los **BLUP**.

Es decir:

- $\hat{\beta}_{\text{ML}} = \hat{\beta}(\hat{\omega}_{\text{ML}}) = \left(X'V^{-1}(\hat{\omega}_{\text{ML}})X\right)^{-1} X'V^{-1}(\hat{\omega}_{\text{ML}})$
- $\hat{u}_{\text{ML}} = \hat{u}(\hat{\omega}_{\text{ML}}) = (GZ'V^{-1}(\hat{\omega}_{\text{ML}})\left(y - X\hat{\beta}(\hat{\omega}_{\text{ML}})\right)$

Sin embargo, la estimación a partir de máxima verosímilitud no toma en cuenta la pérdida de grados de libertad debido a la estimación de $\beta$ por lo que lleva a estimaciones sesgadas de $\omega$.

Como alternativa para obtener estimaciones insesgadas del conjunto de parámetros $\omega$ se utiliza el método de **máxima verosimilitud restringida**.

### 3.3.5. Estimación del EBLUP a partir de REML

Para tomar en consideración el sesgo potencial proveniente de no considerar la pérdida de grados de libertad debido a los parámetros *nuisance*, el **REML** aplica una transformación

a la variable de entrada $y$ tal que $y^* = A'y$ con $A \in \mathcal{M}_{n\times(n\times p)}(\mathbb{R})$, de rango completo y ortogonal a la matriz de diseño $X$.

Resulta que $y^* \sim \mathcal{N}(0, A'VA)$.

Por lo que el logaritmo de la verosimilitud de $y^*$, a quien denominaremos verosimilitud restringida, toma la siguiente expresión:

$$l_R(\omega) = c - \frac{1}{2}\left[\ln(|V|) + \ln(X'V^{-1}X) + y'Py)\right] \tag{3.42}$$

$$P = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1} \tag{3.43}$$

Derivando con respecto a $\omega_j$ se obtiene que las entradas del gradiente $S_R(\omega)$ se definen como:

$$S_{R_j}(\omega) = \frac{\partial l_R(\omega)}{\partial \omega_j} = -\frac{1}{2}\mathrm{tr}(PV_{(j)}) + \frac{1}{2}y'PV_{(j)}Py \tag{3.44}$$

Por otro lado, la matriz de información de Fisher tiene como elementos:

$$\mathcal{I}_{R_{jk}} = \frac{1}{2}\mathrm{tr}(PV_{(j)}PV_{(k)}) \tag{3.45}$$

Un aspecto destacado tanto de la función de verosimilitud restringida y de la matriz de información de Fisher, es que no se ven afectadas por la elección de $A$. Esto significa que dada una $A$ cualquiera, que cumpla los supuestos iniciales, los parámetros de varianza estimados serán idénticos.

Alcanzada la convergencia de los parámetros (trás aplicar el método score planteado en la subsección anterior), la obtención de $\hat{\beta}_{\mathrm{REML}}, \hat{V}_{\mathrm{REML}}, \hat{u}_{\mathrm{REML}}$ se realiza de la misma forma que en máxima verosimilitud clásica: reemplazando $\hat{\omega}_{\mathrm{REML}}$ en el lugar de $\omega$.

### 3.3.6. Error cuadrático medio del EBLUP

Estimados los parámetros, el siguiente paso al momento de estudiar un estimador de muestreo es analizar su variabilidad y sesgo. En este caso se estudiará el comportamiento conjunto a partir del $\mathbb{ECM}$.

Dada $t\left(\hat{\omega},\hat{\beta},y\right)-\mu$ comenzaremos sumando y restando $t(\omega,y)$ (quien se recuerda es el **BLUP**).

$$t\left(\hat{\omega},\hat{\beta},y\right)-\mu = t\left(\hat{\omega},\hat{\beta},y\right)-\mu+t(\omega,y)-t(\omega,y) \tag{3.46}$$

$$= [t(\omega,y)-\mu]+\left[t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)\right] \tag{3.47}$$

Aplicando $\mathbb{E}$ al cuadrado de la expresión obtiene:

$$\mathbb{E}\left[(t\left(\hat{\omega},\hat{\beta},y\right)-\mu)^2\right] = \mathbb{E}\left(\left([t(\omega,y)-\mu]+\left[t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)\right]\right)^2\right) \tag{3.48}$$

$$= \mathbb{ECM}(t(\omega,y))+\mathbb{E}\left[\left(t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)\right)^2\right]+2\mathbb{E}\left([(t(\omega,y)-\mu]+\left[t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)\right]\right) \tag{3.49}$$

$$\Rightarrow \mathbb{ECM}\left(t\left(\hat{\omega},\hat{\beta},y\right)\right) = \mathbb{ECM}(t(\omega,y))+\mathbb{E}\left[\left(t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)\right)^2\right] \tag{3.50}$$

Suponiendo que $\hat{\omega}$ es invariante ante traslaciones, se puede probar que la esperanza del producto entre $t(\omega,y)-\mu$ y $t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)$ es 0.

A partir de esto se deriva que:

$$\mathbb{ECM}(\hat{\mu}_{\text{EBLUP}}) = \mathbb{ECM}\left(t\left(\hat{\omega},\hat{\beta},y\right)\right) = \mathbb{ECM}(\hat{\mu}_{\text{BLUP}})+\mathbb{E}\left[\left(t\left(\hat{\omega},\hat{\beta},y\right)-t(\omega,y)\right)^2\right] \tag{3.51}$$

$$\Rightarrow \mathbb{ECM}(\hat{\mu}_{\text{EBLUP}}) \geq \mathbb{ECM}(\hat{\mu}_{\text{BLUP}}) \tag{3.52}$$

Con esto se confirma que el error cuadrático medio asociado al estimador empírico es mayor que el que asume conocimiento de los parámetros de varianza bajo distribución normal de $u$ y $\varepsilon$. Esto es esperable, a mayor nivel de información menor espacio a incurrir en errores.

Se deduce también que la práctica de estimar el $\mathbb{ECM}$ del **EBLUP** a partir del $\mathbb{ECM}$ del **BLUP** puede llevar a errores considerables ya que se menospreciaría al error. Especialmente en casos donde la estimación empírica varía en gran proporción respecto a cambios en $\omega$ (y aún peor cuando el mismo $\hat{\omega}$ tiene gran variabilidad).

Por otra parte, volviendo a la expresión del error cuadrático medio del **EBLUP**, el mismo se descompone como:

$$\Rightarrow \mathbb{ECM}(\hat{\mu}_{\text{EBLUP}}) = g_1(\omega) + g_2(\omega) + \mathbb{E}\left[\left(t\left(\hat{\omega}, \hat{\beta}, y\right) - t(\omega, y)\right)^2\right]$$

El último término debe ser aproximado, ya que a nivel general no tiene fórmula cerrada. (Kackar y Harville, 1984) propone una estimación a partir de aproximar $t\left(\hat{\omega}, \hat{\beta}, y\right) - t(\omega, y)$ con un polinomio de Taylor de órden 1 alrededor de $\hat{\omega}$.

$$\Rightarrow P_1(\omega) = \underbrace{t\left(\hat{\omega}, \hat{\beta}, y\right) - t(\hat{\omega}, y)}_{=0} + \frac{\partial t(\omega, y)}{\partial \omega}(\hat{\omega} - \omega) + r_2(\omega) \tag{3.53}$$

$$= \frac{\partial t^*(\omega, \beta, y)}{\partial \omega}(\hat{\omega} - \omega) + \frac{\partial d'(\hat{\beta} - \beta)}{\partial \omega}(\hat{\omega} - \omega) + r_2(\omega) \tag{3.54}$$

$$= \frac{\partial t^*(\omega, \beta, y)}{\partial \omega}(\hat{\omega} - \omega) + r_2^*(\omega) \tag{3.55}$$

$$= \frac{\partial (I'\beta + b'(y - X\beta))}{\partial \omega}(\hat{\omega} - \omega) + r_2^*(\omega) \tag{3.56}$$

$$= \frac{\partial b'}{\partial \omega}(y - X\beta)(\hat{\omega} - \omega) + r_2^*(\omega) \tag{3.57}$$

$$= d^*(\omega)(\hat{\omega} - \omega) + r_2^*(\omega) \tag{3.58}$$

Es de destacar que en uno de los pasos $\frac{\partial d'(\hat{\beta} - \beta)}{\partial \omega}$ fue agregado al error de aproximación ya que bajo normalidad ese término es despreciable (de menor órden que el principal).

Continuando con el problema de estimar el error cuadrático medio de $\hat{\mu}_{\text{EBLUP}}$:

$$\mathbb{E}\left[\left(t\left(\hat{\omega}, \hat{\beta}, y\right) - t(\omega, y)\right)^2\right] \approx \mathbb{E}\left(\left[d^*(\omega)(\hat{\omega} - \omega)\right]^2\right) \tag{3.59}$$

(Kackar y Harville, 1984) presenta en su paper una aproximación muy interesante:

$$\mathbb{E}\left(\left[d^*(\omega)(\hat{\omega} - \omega)\right]^2\right) \approx \text{tr}\left(\mathbb{E}(d^*(\omega)d^*(\omega))\bar{V}(\hat{\omega})\right) \tag{3.60}$$

$$= \text{tr}\left(\frac{\partial b'}{\partial \omega} V \frac{\partial b'}{\partial \omega}' \bar{V}(\hat{\omega})\right) =: g_3(\omega) \tag{3.61}$$

Finalmente una aproximación del error cuadrático medio (de segundo órden) del **EBLUP** queda definida como:

$$\mathbb{ECM}(t(\hat{\omega}, \hat{\beta}, y)) = g_1(\omega) + g_2(\omega) + g_3(\omega) \tag{3.62}$$

Al igual que en el caso del error cuadrático medio del **BLUP**, esta fórmula depende de los parámetros de varianza que en la práctica son desconocidos. Es por ello que se busca en la próxima subsección obtener estimaciones "tratables" en la práctica.

### 3.3.7. Estimación del error cuadrático medio del EBLUP

La obtención de medidas de calidad de los estimadores es sumamente importante. Existen distintas estrategias para llegar a una aproximación del $\mathbb{ECM}$ del **BLUP**:

- Hacer uso de la fórmula de $\mathbb{ECM}$ del **BLUP** y estimar remplazando en la fórmula por $\hat{\omega}$. Esta opción no se recomienda ya que como se mencionó anteriormente, se estaría depreciando un término del $\mathbb{ECM}$ que puede llegar a ser significativo.
- Sustituir en la aproximación del $\mathbb{ECM}$ del **EBLUP** a $\omega$ por su debida estimación.
- Sustituir en la aproximación del $\mathbb{ECM}$ **corregida** del **EBLUP** a $\omega$ por su debida estimación.
- Utilización de métodos de remuestreo

Las fórmulas asociadas al método (1) y (2) respectivamente no son más que:

$$\hat{\mathbb{ECM}}(t(\hat{\omega}, \hat{\beta}, y)) = g_1(\hat{\omega}) + g_2(\hat{\omega}) \tag{3.63}$$

$$\hat{\mathbb{ECM}}(t(\hat{\omega}, \hat{\beta}, y)) = g_1(\hat{\omega}) + g_2(\hat{\omega}) + g_3(\hat{\omega}) \tag{3.64}$$

Un problema de este segundo método surge de $g_1(\hat{\omega})$ quien a diferencia de $g_2(\hat{\omega})$ y $g_3(\hat{\omega})$ que son aproximadamente insesgados por construcción (de segundo orden), este no lo es.

Para obtener un estimador insesgado de $g_1(\hat{\omega})$ y por lo tanto del $\mathbb{ECM}$ del **EBLUP**, (Rao y Molina, 2014) propone en un principio evaluar el desarrollo de Taylor de segundo orden valuado en $\omega$.

$$g_1(\hat{\omega}) \approx g_1(\omega) + (\hat{\omega} - \omega)' \nabla g_1(\omega) + \frac{1}{2} \nabla^2 g_1(\omega)(\hat{\omega} - \omega) \tag{3.65}$$

$$:= g_1(\omega) + \Delta_1 + \Delta_2 \tag{3.66}$$

Luego, dado que $\hat{\omega}$ es insesgada para $\omega$ (en este caso se debe trabajar con **REML**), la esperanza de $g_1(\hat{\omega})$ es:

$$\mathbb{E}(g_1(\hat{\omega})) = g_1(\omega) + \mathbb{E}(\Delta_2) \tag{3.67}$$

Aplicando la aproximación propuesta por (Kackar y Harville, 1984) y bajo el supuesto de que $V$ tiene estructura lineal con respecto a $\omega_i, \forall i$, se demuestra que $\mathbb{E}(\Delta_2) = -g_3(\hat{\omega})$.

A partir de ello, un estimador aproximadamente insesgado del error cuadrático medio del **EBLUP** queda definido como:

$$\mathbb{E}\hat{\mathbb{C}}\mathbb{M}(t(\hat{\omega}, \hat{\beta}, y)) \approx g_1(\hat{\omega}) + g_2(\hat{\omega}) + 2g_3(\hat{\omega}) \tag{3.68}$$

# 4. Estimador de Fay-Herriot

El primer modelo que se estudiará es el de Fay-Herriot (Fay III y Herriot, 1979). Este modelo de área relaciona a los indicadores de interés de todos los dominios $\delta_d,\ d \in \{1, ..., D\}$ asumiendo que los mismos varían con respecto de un conjunto de $p$ variables auxiliares $x_d = (x_{d1}, x_{d2}, \ldots, x_{dp})'$ de manera constante. Los indicadores por área tienen una forma paramétrica definida por el siguiente modelo de regresión lineal:

$$\delta_d = x_d{}'\beta + u_d,\ d \in \{1, ..., D\}$$

donde $\beta$ es un vector de coeficientes fijo común a todas las áreas. Por otro lado, $u_d$ (quien es el término de error de la regresión) es distinto para cada área, el mismo es un efecto aleatorio que representa la heterogeneidad de los indicadores $\delta_d$ que no es explicada a partir del conjunto de variables auxiliares $x_d$.

El modelo más simple asume que $u_d \overset{iid}{\sim} (0, \sigma_u^2)$, con varianza $\sigma_u^2$ desconocida.

Dado que los verdaderos valores de los indicadores no son observables, es imposible ajustar el modelo de manera convencional. Es por ello que utilizando al estimador directo $\hat{\delta}^{DIR}$ y teniendo en cuenta que el mismo es insesgado bajo el diseño se llega a la siguiente expresión:

$$\hat{\delta}_d^{DIR} = \delta_d + e_d,\ d \in \{1, ..., D\}$$

en donde $e_d$ es el error de muestreo asociado al área $d$. A su vez, se asume que:

- Los errores son independientes entre sí y de los efectos aleatorios de su área $(u_d)$
- $\mathbb{E}(e_d) = 0$.
- $\mathbb{VAR}(e_d) = \psi_d = \mathbb{VAR}\left(\hat{\delta}^{DIR} | \delta_d\right)$.

Por lo que $e_d \overset{ind}{\sim} (0, \psi_d)$.

Las varianzas $\psi_d$ son las varianzas del estimador directo y por lo general son estimables sin dificultad. En la práctica se utilizan los microdatos de las encuestas.

Al combinar ambas expresiones se llega a que la estimación del estimador directo puede

ser formulada como:

$$\hat{\delta}_d^{DIR} = x_d{}'\beta + u_d + e_d,\ d \in \{1, ..., D\} \tag{4.1}$$

El mismo es un modelo lineal mixto (Molina, 2019). En primera instancia, puede verse como este estimador proporciona una mejora conceptual frente a los estimadores sintéticos. Si bien plantea una entructura de homogeneidad entre los dominios que se asocia a un conjunto de variables auxiliares, este modelo "renuncia" a cierto grado de ingenuidad al agregar un componente de heterogeneidad que capta diferencias entre los dominios que no están asociadas al conjunto de variables auxiliares.

Ajustando el modelo lineal mixto asociado a $\hat{\delta}^{DIR}$ se llega al modelo de Fay-Herriot.

$$\tilde{\delta}_d^{FH} = x_d{}'\tilde{\beta} + \tilde{u_d} \tag{4.2}$$

## 4.1. BLUP del modelo

Para el ajuste del modelo se hará uso de las fórmulas demostradas en la sección anterior. Para ello tan solo hace falta identificar al modelo y sus componentes.

En el caso del estimador de Fay-Herriot, es claro que la variable a estimar (en un área específica) también es de la forma $\mu = I'\beta + m'u$, donde :

- $m = \mathbb{I}_d$ (vector con un 1 en la posición d y 0 en el resto).
- $I_i = x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

Siguiendo las ecuaciones demostradas en la sección anterior se llega a que:

$$\tilde{\beta} = \left(\sum_{d=1}^{D} \gamma_d x_d x_d{}'\right)^{-1} \sum_{d=1}^{D} \gamma_d x_d \hat{\delta}_d^{DIR} \tag{4.3}$$

$$\tilde{u}_d = \gamma_d \left(\hat{\delta}_d^{DIR} - x_d{}'\tilde{\beta}\right) \ \text{ con } \ \gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d} \tag{4.4}$$

### 4.1.1. Expresión alternativa del BLUP

Remplazando el valor de $\tilde{u_d}$ en ecuación (4.2):

$$\tilde{\delta_d}^{FH} = x_d{}'\tilde{\beta} + \tilde{u_d} \tag{4.5}$$

$$= x_d{}'\tilde{\beta} + \gamma_d \left(\hat{\delta_d}^{DIR} - x_d{}'\tilde{\beta}\right) \tag{4.6}$$

$$= x_d{}'\tilde{\beta} + \gamma_d \hat{\delta_d}^{DIR} - \gamma_d x_d{}'\tilde{\beta} \tag{4.7}$$

$$= \gamma_d \hat{\delta_d}^{DIR} + (1 - \gamma_d)\, x_d{}'\tilde{\beta} \tag{4.8}$$

Puede verse como el estimador de Fay-Herriot es un estimador compuesto. El mismo se construye como una combinación lineal convexa del estimador directo y uno sintético.

El estimador de **FH** es capaz de regular el peso que le da al estimador directo y al estimador basado en el modelo de manera eficiente.

A nivel general, cuanto mayor sea la varianza del estimador directo, más cercano a 0 será el parámetro $\gamma_d$. Esto se traduce a un estimador de **FH** que asigna mayor peso al estimador sintético ($x_d{}'\tilde{\beta}$). De igual manera, a menor tamaño muestral efectivo en el dominio, mayor variabilidad tendrá el estimador directo y por ende, el estimador de **FH** se verá mayormente definido por el estimador sintético.

Otra manera de estudiar el comportamiento del estimador es a partir de $\sigma_u^2$. Cuando se tiene dominios homogéneos ($\sigma_u^2$ tendiendo a 0) el parámetro $\gamma_d$ tiende a 0, por lo que el estimador de **FH** da más peso al estimador sintético. Esto último es intuitivo, los dominios son parecidos por lo que un modelo global reforzará el tamaño muestral efectivo en el dominio.

### 4.1.2. Sesgo del BLUP

Estudiar el sesgo del estimador del estimador de **FH** nos permite demostrar que el mismo es consistente con el diseño. Esto quiere decir que el sesgo tiende a 0 cuando el tamaño de muestra crece. Recordemos que esto no sucedía con los estimadores sintéticos cuyo sesgo era invariante con respecto al tamaño muestral.

$$\mathbb{B}(\tilde{\delta}^{FH}) = \delta_d - \mathbb{E}(\tilde{\delta}^{FH}) \tag{4.9}$$
$$= \delta_d - \mathbb{E}\left(x_d{}'\tilde{\beta} + \gamma_d\left(\hat{\delta}_d^{\,DIR} - x_d{}'\tilde{\beta}\right)\right) \tag{4.10}$$
$$= \delta_d - x_d{}'\tilde{\beta} - \gamma_d\mathbb{E}\left(\hat{\delta}_d^{\,DIR}\right) + \gamma_d x_d{}'\tilde{\beta} \tag{4.11}$$
$$= \delta_d(1-\gamma_d) - x_d{}'\tilde{\beta}(1-\gamma_d) \tag{4.12}$$
$$= (1-\gamma_d)(\delta_d - x_d\tilde{\beta}) \tag{4.13}$$

Se puede apreciar que a medida que el tamaño de muestra aumenta, $\psi_d \to 0$ por lo que $\gamma_d \to 1$ y consigo el sesgo tiende a 0.

### 4.1.3. Error cuadrático medio del BLUP

Nuevamente, siguiendo las fórmulas planteadas en el capítulo anterior, se tiene que el error cuadrático medio del estimador de Fay-Herriot suponiendo los parámetros de varianza conocidos es:

$$\mathbb{ECM}(\tilde{\delta}_d^{\mathrm{FH}}) = g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2) \tag{4.14}$$

Con:

$$g_{1d}(\sigma_u^2) = \gamma_d\psi_d \tag{4.15}$$
$$g_{2d}(\sigma_u^2) = (1-\gamma_d)^2 x_d{}'\left(\frac{\sum_{d=1}^{D} x_d x_d{}'}{\psi_d + \sigma_u^2}\right)^{-1} x_d \tag{4.16}$$

## 4.2. EBLUP del modelo

Al igual que en el capítulo anterior, en la práctica los parámetros de varianza que definen el modelo son desconocidos. Es por ello que se deben desarrollar técnicas para primero obtener una estimación de los mismos y luego poder estimar un modelo.

En el caso del modelo de **Fay-Herriot** concentramos nuestro interés en la estimación de las varianzas de los efectos aleatorios.

### 4.2.1. Estimación de $\sigma_u^2$

En el caso del modelo de área (**FH**) se suma además del método de máxima verosimilitud clásica y restringida el estimador de (Prasad y Rao, 1990) que define a un estimador insesgado como el máximo entre $\hat{\sigma}^2_{\text{us}}$ y 0.

Donde:

$$\hat{\sigma}^2_{\text{us}} = \frac{1}{m-p}\left(\sum_{d=1}^{D}\left[\tilde{\delta}_d^{\text{FH}} - x_d{}'\hat{\beta}_{\text{WLS}}\right]^2 - \sum_{d=1}^{D}\psi_d(1-h_{dd})\right) \tag{4.17}$$

con

$$\hat{\beta}_{\text{WLS}} = \left(\sum_{d=1}^{D} x_d x_d{}'\right)^{-1}\left(\sum_{d=1}^{D} x_d \hat{\delta}_d^{\text{DIR}}\right) \tag{4.18}$$

$$h_{dd} = x_d{}'\left(\sum_{d=1}^{D} x_d x_d{}'\right)^{-1} x_d \tag{4.19}$$

Un aspecto a destacar es que esta fórmula y las resultantes de trabajar con **ML** o **REML** dependen de las varianzas del estimador directo. Estas se estiman previamente con sus respectivas técnicas, las cuales se asumen como conocidas para no exceder los límites de este trabajo.

#### Estimación de $\sigma_u^2$ por ML y REML

Nuevamente, tras operar con el conjunto de matrices planteado en el capítulo anterior, la estimación por medio de **ML** se lleva a cabo con las siguientes ecuaciones (aplicando el algoritmo score):

$$\mathcal{I}(\sigma_u^2) = \frac{1}{2}\sum_{d=1}^{D}\frac{1}{(\sigma_u^2+\psi_d)^2} \tag{4.20}$$

$$s(\tilde{\beta},\sigma_u^2) = -\frac{1}{2}\sum_{d=1}^{D}\frac{1}{\sigma_u^2+\psi_d} + \frac{1}{2}\sum_{d=1}^{D}\frac{(\hat{\delta}_d^{\text{FH}} - x_d{}'\tilde{\beta})^2}{(\sigma_u^2+\psi_d)^2} \tag{4.21}$$

De manera análoga a como se planteó anteriormente, alcanzada la convergencia basta con sustituir $\sigma_u^2$ por $\hat{\sigma}_u^2$ para obtener estimaciones de $\beta, u, V$.

Por otro lado, el método de **REML** se basa en:

$$\mathcal{I}(\sigma_u^2) = \frac{1}{2}\text{tr}\left(P^2\right) \tag{4.22}$$

$$s_R(\sigma_u^2) = -\frac{1}{2}\text{tr}(P) + \frac{1}{2}\left(\hat{\delta}_d^{DIR}\right)' P^2 \hat{\delta}_d^{DIR} \tag{4.23}$$

Con $P$ definida en (3.41).

### 4.2.2. Estimación del error cuadrático medio del EBLUP

El error cuadrático medio teórico del **EBLUP**,es decir aquel que supone conocimiento íntegro de los parámetros de la varianza; es análoga al planteado en el capítulo anterior.

Para estimarlo, tan solo hace falta evaluar alguna de las expresiones teoricas:

$$\mathbb{ECM}(t(\hat{\omega}, \hat{\beta}, y)) = g_{1d}(\omega) + g_{2d}(\omega) + g_{3d}(\omega) \tag{4.24}$$

$$\mathbb{ECM}(t(\hat{\omega}, \hat{\beta}, y)) = g_{1d}(\omega) + g_{2d}(\omega) + 2g_{3d}(\omega) \tag{4.25}$$

sobre los parámetros de varianza estimados.

$$g_{3d} = \psi_d^2(\psi_d + \sigma_u^2)\bar{V}(\hat{\sigma_u^2}) \tag{4.26}$$

### 4.2.3. Estimación del $\mathbb{ECM}$ a partir de Bootstrap paramétrico

Una alternativa a las fórmulas demostradas con anterioridad es el uso de métodos de resmuestreo tales como el **Bootstrap** o **Jackknife**. En esta sección se presenta a modo de ejemplo el método de bootstrap paramétrico propuesto por (González-Manteiga, Lombarda, Molina, Morales, y Santamaría, 2010) el cual es presentado en el libro de (Morales, Esteban, Pérez, y Hobza, 2021).

Suponiendo que el modelo empleado es correcto, el método sigue los siguientes pasos:

1) Obtener las estimaciones $\hat{\sigma}_u^2, \hat{\beta}$ con alguno de los métodos planteados.

2) Repetir B veces (se denomina $b$ a las repeticiones, $b = 1, 2, \ldots, B$) :

   - para $d = 1, 2, \ldots, D$, generar $u_d^{*(b)} \overset{iid}{\sim} \mathcal{N}(0, \hat{\sigma}_u^2)$
   - Construído el vector $u^{*(b)} = (u_1^{*(b)}, u_2^{*(b)}, \ldots, u_D^{*(b)})'$, se calcula los *valores reales Bootstrap* $\delta^{*(b)} = X\hat{\beta} + u^{*(b)}$ cuyos elementos denominamos $\delta_d^{*(b)}$.

- Luego, para cada una de las áreas se generan $e_d^{*(b)} \overset{iid}{\sim} \mathcal{N}\left(0, \hat{\psi}_u^2\right)$. Se construye también $e^{*(b)} = (e_1^{*(b)}, e_2^{*(b)}, \ldots, e_D^{*(b)})'$.
- Se calculan los datos Bootstrap $\delta_{Bootstrap}^{*(b)} = X\hat{\beta} + u^{*(b)} + e^{*(b)}$
- Se ajusta un modelo para los datos bootstrap y se obtienen las estimaciones $\hat{\sigma_u}^{2*(b)}, \hat{\beta}^{2*(b)}$.
- Para $d = 1, 2, \ldots, D$ se calcula el **EBLUP** de los datos bootstrap.
- $\hat{\delta_d}^{*(b)} = \frac{\hat{\sigma_u}^{2*(b)}}{\hat{\sigma_u}^{2*(b)} + \psi_d} \delta_{d-Bootstrap}^{*(b)} + \frac{\psi_d}{\hat{\sigma_u}^{2*(b)} + \psi_d} x_d{}' \hat{\beta}^{*(b)}$

3) Se estima el error cuadrático medio como:
$\widehat{\mathbb{ECM}} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\delta_d}^{*(b)} - \delta^{*(b)}\right)^2, d = 1, 2, \ldots, D.$

A partir de este método es también posible obtener estimaciones de la varianza del estimador.

## 4.3. Resumen

Se puede decir que el estimador Fay-Herriot ofrece una alternativa altamente provechosa para obtener estimaciones de mejor calidad en dominios con bajo tamaño de muestra. Esto se debe a:

- Agrega componentes de heterogeneidad para explicar diferencias entre áreas que no son explicadas por las covariables disponibles.
- No se ve afectado por datos individuales atípicos debido a que trabaja con información agregada a nivel de área.
- Se puede obtener predicciones para áreas no muestradas.
- (Molina, 2019) menciona también como se evitan problemas de confidencialidad de los datos. Al hacer uso de información auxiliar agregada a nivel de dominio, no existe necesidad de anonimizar los datos.
- El mismo es capaz de regular el peso asignado al estimador directo y al sintético basándose en un modelo superpoblacional y más específicamente es sus parámetros de varianza.

  Por ejemplo, en caso de que la varianza del efecto aleatorio tienda a 0, la mayor parte (si no es toda) del estimador será construida a partir de un estimador sintético. Esto se debe a que el estimador interpreta que la heterogeneidad entre áreas es prácticamente

nula, por lo que un modelo sintético podrá a partir de una muestra "reforzada" ganar fuerza al momento de estimar áreas con un tamaño de muestra pequeño (o en los casos más extremos, tamaño de muestra nulo).

Por otro lado, si la varianza del efecto aleatorio es muy alta con respecto a la del estimador directo, la situación será la opuesta. El estimador interpreta que las diferencias entre áreas no es captada completamente a partir del conjunto de variables auxiliares por lo que darle mucho peso a un estimador sintético podría incurrir en un gran nivel de sesgo.

Sin embargo, (Molina, 2019) menciona también varias limitantes de este modelo:

- Como todo modelo, es necesario que se cumplan los supuestos para que la capacidad predictiva sea buena. De no ser así se obtienen estimaciones sesgadas.
- Los estimadores del modelo se estiman con pocas observaciones (las áreas), lo cual suele ser menor a la cantidad de individuos a nivel poblacional. Los parámetros se estiman con menor eficiencia (una mejora con respecto a esto son los estimadores a nivel de individuo).
- Una vez ajustado el modelo, las estimaciones no se pueden desagregar en subdominios.

A su vez, existen distintas adaptaciones de este estimador que buscan tomar en cuenta un mayor nivel de información asociado a las áreas y a la relación existente entre ellas.

Dentro de estas otras versiones se encuentra el modelo de Fay-Herriot espacial, el temporal e incluso el espacio-temporal (entre otros). Todos siguen una estructura base análoga al presentado acá (modelo lineal mixto) con algunas modificaciones fácilmente analizables con lo estudiando en el capítulo 3.

En este trabajo se optó por introducir al estimador de Fay-Herriot espacial, el cual se implementará al final del TFG. Con el mismo se buscará captar una potencial autocorrelación espacial con respecto a la distribución de la pobreza en los barrios de Montevideo, Uruguay.

# 5. Estimador de Fay-Herriot espacial

En muchos casos prácticos el supuesto de independencia entre los efectos aleatorios correspondientes a las distintas áreas no se cumple. Existen casos en donde los mismos se encuentran correlacionados. En esta última sección se indagará el caso en donde los distintos $u_d$ presentan una correlación de tipo espacial.

(Diniz, Bini y Hawkins, 2003) describen a la autocorrelación espacial como:

*"La autocorrelación espacial es la correlación entre valores de una misma variable, estrictamente atribuible a la ubicación en una superficie bidimensional, introduciendo una desviación del supuesto de independencia entre las observaciones, fundamento de la estadística clásica".*

Se comenzará presentado al modelo de **FH espacial**, luego se presentarán aspectos esenciales del estudio de datos espaciales y finalmente se profundizará en las propiedades inherentes a este estimador.

El estimador supone que:

$$\hat{\delta}^{\text{DIR}} = X\beta + u + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \psi) \tag{5.1}$$

donde:

$$u = \rho W u + v,\ v \sim \mathcal{N}(0, \sigma_v^2 I_d) \tag{5.2}$$

donde $W$ es la matriz de vecindad, la misma es la encargada de definir la estructura de vecindad entre las distintas áreas. Como puede verse en la fórmula del modelo, el efecto aleatorio ahora depende del valor del resto de las áreas y de un término de error con media 0 y varianza constante entre los distintos dominios (a su vez, se supone que este error es no correlacionado). Es más, si se trabaja con matrices de vecindad estandarizadas (esto quiere decir que la suma de sus filas es igual a 1) el término de rezago $Wu$ representa el promedio (ponderado) de los valores del efecto aleatorio en las regiones vecinas para cada área.

Incorporando un modelo **SAR** (Spatial Autoregressive model) al estimador, se busca dejar libre de autocorrelación espacial al error del modelo $\varepsilon$.

## 5.1. Elección de W

La elección de W está sujeta a criterio del investigador y según como él/ella defina que dos áreas son vecinas.

Una primera opción es definir la vecindad a partir de una *matriz de conectividad binaria*. Bajo esta coyuntura dos áreas son vecinas si son limítrofes.

La matriz queda definida como:

$$w_{ij} = \begin{cases} 1, & \text{si las regiones i y j son limítrofes} \\ 0, & \text{en otro caso} \end{cases} \tag{5.3}$$

Otra estrategia para definir la vecindad entre áreas es la de $q$ vecinos más cercanos.

$$w_{ij} = \begin{cases} 1, & \text{si j está dentro de las } q \text{ regiones más cercanas a i} \\ 0, & \text{en otro caso} \end{cases} \tag{5.4}$$

Por lo general, las distancias se toman de un centroide a otro. A diferencia del primer caso, $W$ no necesariamente es simétrica.

Otro conjunto de estrategias para definir la estructura de vecindad de los distintos barrios (o a nivel más general, dominios geográficos) se basa en la distancia que mantienen los centroides.

Un ejemplo:

$$w_{ij} = \begin{cases} 1, & \text{si } d(i,j) < \alpha \\ 0, & \text{en otro caso} \end{cases} \tag{5.5}$$

Con $\alpha \in \mathbb{R}^+$, umbral a elección del investigador.

Otra visión más general, implica que todos los barrios son vecinos pero la fuerza en la que lo son depende pura y exclusivamente de la distancia a la que se encuentran y de un parámetro $\eta$.

$$w_{ij} = \begin{cases} d_{ij}^{-\eta}, & \forall\, \eta > 0 \\ 0, & \eta = 0 \end{cases} \tag{5.6}$$

En el capítulo de aplicación se estudiará como la elección de distintas matrices de vecindad afectan el peso asignado al estimador directo y al sintético.

## 5.2. Medición de la autocorrelación espacial

Antes de adentrarse a trabajar con modelos de tipo espacial, los cuales implican un nivel de dificultad mayor al momento de estimar, estudiar e interpretar, es deseable testear si la variable de interés efectivamente mantiene una estructura espacial.

Para medir el nivel de autocorrelación espacial entre polígonos existen 2 tipos principales de índices:

- Los **globales**: que resumen el nivel de autocorrelación espacial de la variable en todas las áreas.
- Los **locales**: que se centran en el nivel de autocorrelación espacial de la variable de interés entre un área dada y sus vecinas.

Se presenta a continuación el índice global de Moran.

### 5.2.1. Índice de Moran

Dadas D áreas (Bivand, Pebesma, y Gomez-Rubio, 2013), el índice se define como:

$$I = \frac{n}{\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}} \times \frac{\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{D}(y_i - \bar{y})^2} \tag{5.7}$$

Usualmente $|I| < 1$. Valores cercanos a uno indican autocorrelación espacial positiva. Esto quiere decir que polígonos cercanos toman valores parecidos. Niveles cercanos a $-1$ significan lo opuesto. Un valor del índice cercano a 0 refleja ausencia de correlación espacial.

Bajo las hipótesis de normalidad, independencia e idéntica distribución de $y$ en conjunto con la hipótesis de randomización (Cliff y Ord, 1973) demostraron que:

$$\mathbb{E}(I) = \frac{-1}{N-1} \tag{5.8}$$

$$\mathbb{VAR}(I) = \frac{DS_4 - S_3 S_5}{(D-1)(D-2)(D-3)\left(\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}\right)^2} - \mathbb{E}(I)^2 \tag{5.9}$$

$$\begin{aligned}
S_1 &= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij} + w_{ji})^2 \\
S_2 &= \sum_{i=1}^{D}\left(\sum_{j=1}^{D} w_{ij} + \sum_{j=1}^{D} w_{ji}\right) \\
S_3 &= \frac{D^{-1}\sum_{i=1}^{D}(y_i - \bar{y})^4}{\left(D^{-1}\sum_{i=1}^{D}(y_i - \bar{y})^2\right)^2} \\
S_4 &= (D^2 - 3D + 3)S_1 - DS_2 + 3\left(\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}\right)^2 \\
S_5 &= (D^2 - D)S_1 - 2DS_2 + 6\left(\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}\right)^2
\end{aligned}$$

De esta forma es posible definir pruebas de hipótesis y probar así la presencia de autocorrelación.

- $H_0$) no existe autocorrelación espacial
- $H_1$) los datos están espacialmente autocorrelacionados

En este trabajo se restringirá a trabajar con el índice global de Moran para analizar la efectividad de plantear un modelo mixto espacializado para el estudio de la pobreza en los barrios de Montevideo.

## 5.3. Estimación del modelo

Habiendo repasado los aspectos básicos que definen a un modelo espacial, se prosigue con el estudio de la estimación del estimador de **FH** espacial.

Para ello consideremos reescribir al efecto aleatorio de la siguiente manera:

$$u = \rho W u + v \Rightarrow u - \rho W u = v \tag{5.10}$$

$$\Rightarrow u(I_d - \rho W) = v \Rightarrow u = v(I_d - \rho W)^{-1} \tag{5.11}$$

Por lo que el modelo puede ser redefinido como:

$$\hat{\delta}^{\text{DIR}} = X\beta + v(I_d - \rho W)^{-1} + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \psi_d),\ v \sim \mathcal{N}(0, \sigma_v^2 I_d) \tag{5.12}$$

El modelo claramente se enmarca dentro del contexto de modelos lineales mixtos por lo que los aspectos estudiados con anterioridad y más particularmente en el capítulo 3 siguen siendo válidos.

Haciendo uso de esta expresión simplificada se tiene que:

$$G = \mathbb{VAR}(u) \tag{5.13}$$

$$= \mathbb{VAR}(v(I_d - \rho W)^{-1}) \tag{5.14}$$

$$= (I_d - \rho W)^{-1}\mathbb{VAR}(v)\big((I_d - \rho W)^{-1}\big)' \tag{5.15}$$

$$= (I_d - \rho W)^{-1}\sigma_v^2 I_d\big((I_d - \rho W)^{-1}\big)' \tag{5.16}$$

$$= \sigma_v^2 \left[(I_d - \rho W)'(I_d - \rho W)\right]^{-1} \tag{5.17}$$

Por otro lado, R sigue siendo la misma matriz que en el modelo clásico y:

$$V = \text{diag}(\psi_d) + \sigma_v^2 I_d{}' \left[(I_d - \rho W)'(I_d - \rho W)\right]^{-1} I_d \tag{5.18}$$

$$= \text{diag}(\psi_d) + \sigma_v^2 \left[(I_d - \rho W)'(I_d - \rho W)\right]^{-1} \tag{5.19}$$

Con ello el modelo queda completamente definido por lo que para obtener tanto el **SBLUP** (Spatial Best Linear Unbiased Predictor) y el **SEBLUP** (Spatial Empirical Best Linear Unbiased Predictor), basta con aplicar las fórmulas o técnicas (**ML** o **REML**) ya estudiadas.

La estimación de $\rho$ puede obtenerse a partir de **ML** o **REML** (el mismo es considerado un parámetro de varianza).

Finalmente, el estimador de Fay-Herriot espacial empírico toma la forma de:

$$\hat{\delta}_d^{\,SEBLUP} = x_d{}'\hat{\beta}_{(\hat{\omega})} + \hat{\sigma}^2_{v(\hat{\omega})}\left[(I_d - \hat{\rho}W)'(I_d - \hat{\rho}W)\right]^{-1}\hat{V}^{-1}_{(\hat{\omega})}(\hat{\delta}^{\mathrm{DIR}} - X\hat{\beta}_{(\hat{\omega})}) \tag{5.20}$$

con

$$\hat{\beta}_{(\hat{\omega})} = (X'V^{-1}_{(\hat{\omega})}X)^{-1}X'V^{-1}_{(\hat{\omega})} \tag{5.21}$$

## 5.4. Estimación del $\mathbb{ECM}$ del SEBLUP

Debido a la presencia de autocorrelación entre los efectos aleatorios, los estimadores del $\mathbb{ECM}$ en el Capítulo 3 dejan de ser insesgados (partícularmente el término $g_3(\omega)$). (Rao y Molina, 2014)

Como alternativa (Molina, Salvati, y Pratesi, 2008) presenta un método de Bootstrap paramétrico (el cual se encuentra ya implementado en R) que puede ser usado para obtener una estimación del $\mathbb{ECM}$. Para ello se presentan 2 opciones:

1) Estimar tan solo $g_3(\omega)$ a partir del método de Bootstrap paramétrico y combinarlo con las estimaciones aproximadamente insesgadas de $g_1(\omega)$ y $g_2(\omega)$ demostradas en el capítulo 3.
2) Estimar completamente el $\mathbb{ECM}$ con Bootstrap.

**Algoritmo:**

1) Ajustar el modelo de **FH-Espacial** obteniendo $\hat{\omega}, \hat{\rho}$ y por consiguiente $\hat{\beta}(\hat{\omega}), \hat{V}(\hat{\omega})$.
2) General un vector $t_1^*$ con $D$ copias independientes de $\mathcal{N}(0,1)$. A partir del mismo construír los **vectores bootstrap**:
   - $v^* = \hat{\sigma_v^2} t_1^*$.
   - $u^* = (I_d - \hat{\rho}W)^{-1}v^*$.
   - $\delta^* = X\hat{\beta}(\hat{\omega}) + u^*$ (cantidad de interés bootstrap).
3) Generar el vector $t_2^*$ con D copias independientes de $\mathcal{N}(0,1)$ que a su vez son independientes del vector $t_1^*$. Tras esto, construir el vector de errores aleatorios

$$\varepsilon^* = \mathrm{diag}(\hat{\psi}_1, \ldots, \hat{\psi}_D)^{\frac{1}{2}} t_2^* \tag{5.22}$$

4) Construir los datos del modelo bootstrap definidos como $\delta^{\text{DIR-B}} = \delta^* + \varepsilon^*$.

5) Considerando los valores estimados en un principio $\hat{\beta}(\hat{\omega}), \hat{V}(\hat{\omega}), \hat{\omega}$ como los reales, estimar nuevamente el modelo de **FH-Espacial** pero esta vez utilizando los datos del modelo bootstrap ($\delta^{\text{DIR-B}}$). Al estimar este modelo se obtendrá $\hat{\omega^*}, \tilde{\beta}(\hat{\omega^*})$. A estos se los denomina como **estimadores bootstrap**.

6) Se calcula el **Spatial BLUP**, es decir, haciendo uso de los datos bootstrap se calcula el estimador de **FH** haciendo uso de los parámetros de varianza "reales" (que son los estimados en el paso 1).

   Tras esto:

$$\tilde{\delta_d}^*(\hat{\omega}) = x_d'\tilde{\beta}(\hat{\omega}) + \mathbb{I}_d' G(\hat{\omega}) V(\hat{\omega})^{-1}[\delta^{\text{DIR-B}} - X\tilde{\beta}(\hat{\omega})] \tag{5.23}$$

7) Calcular el **Spatial EBLUP**, en este caso se ajusta el modelo de los datos bootstrap haciendo uso de las estimaciones bootstrap $\hat{\delta^*}, \tilde{\beta}(\hat{\delta^*})$.

$$\tilde{\delta_d}^*(\hat{\omega^*}) = x_d'\tilde{\beta}(\hat{\omega^*}) + \mathbb{I}_d' G(\hat{\omega^*}) V(\hat{\omega^*})^{-1}[\delta^{\text{DIR-B}} - X\tilde{\beta}(\hat{\omega^*})] \tag{5.24}$$

8) repetir B veces el paso 2 al 7.

Quedan definidos así:

- $\delta_d^{*(b)}$ es la cantidad de interés bootstrap, del área $d$ en el paso $b$.
- $\hat{\omega}^{*(b)}$ el estimador bootstrap de $\omega$.
- $\tilde{\delta_d}^{*(b)}(\hat{\omega})$ el **BLUP bootstrap**.
- $\tilde{\delta_d}^{*(b)}(\tilde{\omega}^{*(b)})$ el **EBLUP bootstrap**.

Con ello se estima:

$$\hat{g}_{3d}^{Bootstrap} = \sum_{b=1}^{B} B^{-1} \left[ \tilde{\delta_d}^{*(b)}(\tilde{\omega}^{*(b)}) - \tilde{\delta_d}^{*(b)}(\hat{\omega}) \right] \tag{5.25}$$

$$\hat{\mathbb{ECM}} \left( \hat{\delta}_d^{FH-SEBLUP} \right)^{\text{Bootstrap}} = B^{-1} \sum_{b=1}^{B} \left[ \delta_d^{*(b)} - \tilde{\delta_d}^{*(b)}(\tilde{\omega}^{*(b)}) \right]^2 \tag{5.26}$$

# 6. Metodología

## 6.1. Introducción

Con motivo de entender en detalle como fueron construidos los mapas de pobreza y de donde surgen los resultados a ser presentados en el capítulo próximo, se detalla a continuación la metodología empleada.

## 6.2. Etapas de la aplicación

1) Obtención de datos de la ECH y procesamiento.
   - Se seleccionan las variables de interés.
   - Se fusionan las bases (2do semestre 2021, 1er semestre 2022)[1] .
   - Se corrigen los ponderadores semestrales para ahora pasar a tener anuales.
2) Se construye la base de variables auxiliares.
3) Procesamiento de la cartografía digital de barrios de Montevideo.
   - Asignación de sistema de coordenadas geográfico **EPSG 32721**.
4) Obtención de las estimaciones directas.
   - Estimación.
   - $\mathbb{CV}$ y $\mathbb{VAR}$ por el método de Bootstrap Rao-Wu (Rao y Wu, 1988).
   - Contrucción de mapa de pobreza.
5) Cálculo del estimador de **FH**.
   - Estimación.
   - $\mathbb{CV}$ y $\mathbb{ECM}$
   - Contrucción de mapas de pobreza.
6) Estudio de la correlación espacial.
   - Contrucción de matrices de vecindad:
     - Matriz de conectividad binaria.
     - Matriz de vecinos más cercanos.

[1] Ver capítulo 7 para más detalles.

    - Matriz de vecindad que considera que los barrios son vecinos si se encuentran a menos de **5km**.
  - Cálculo del índice de Moran para cada matriz.

7) Cálculo del estimador de **FH-Espacial** para cada matriz de vecindad.
   - Estimación.
   - $\mathbb{CV}$ y $\mathbb{ECM}$
   - Construcción de mapas de pobreza.

8) Comparación de resultados y análisis.

9) Construcción de tablas anexas.
   - Resultados de las estimaciones.
   - Medidas de calidad

# 7. Datos

Como insumo principal, se hizo uso de datos provenientes de la **Encuesta Continua de Hogares (ECH)** del INE Uruguay.

La misma se caracteriza por un diseño tipo *cross-section* cuya muestra de un mes se compone por 6 grupos de rotación (hogares). Esto implica que una vez incluido en la muestra, el hogar permanecerá en la misma por un período de 6 meses.

En el primer mes, el hogar completa el **formulario de implantación** en donde se relevan características generales del hogar e individuos. Los 5 meses siguientes tan solo se releva información relacionada al mercado del trabajo para todos los integrantes del hogar que integran la población en edad de trabajar (Ferreira, 2022).

Es de la **base de implantación** de donde se obtienen los datos para estimar la cantidad de personas por debajo del umbral de pobreza en los barrio en Montevideo. Se hace uso de la variable de clasificación *"pobre"*. Las líneas de pobreza e indigencia son definidas por el INE a partir de la canasta básica alimentaria, un análisis exhaustivo de esto puede encontrarse en (Fuster, Glejberman, y Vernengo, 2006).

Se utilizaron las bases de implantación correspondientes al año móvil disponible más cercano (segundo semestre 2021 y primer semestre 2022).

Además, se utilizó la **base de personas del Censo de Población y Vivienda del año 2011** para confeccionar el conjunto de variables auxiliares.

Como variables auxiliares se utilizaron:

- Número de personas con nivel universitario o similar (**NUOS**).
    - Cuenta la cantidad de personas que tienen Niveledu_r con valor 6 a 11 en el barrio.
- Cantidad de desocupados e inactivos (**DOI**).
    - Cantidad de personas con pobpcoac con valor 3 a 6.
- Cantidad de hogares con tres o más necesidades básicas insatisfechas (**VNBI**).
    - Hogares cuyo valor en NBI_CANTIDAD es 3.

Los mapas fueron creados con base en los archivos vectoriales dispuestos por el INE.

Se presenta a continuación el mapa de Montevideo subdividio a nivel barrial (a partir del mismo es que se contruirán los mapas de pobreza con las estimaciones realizadas).

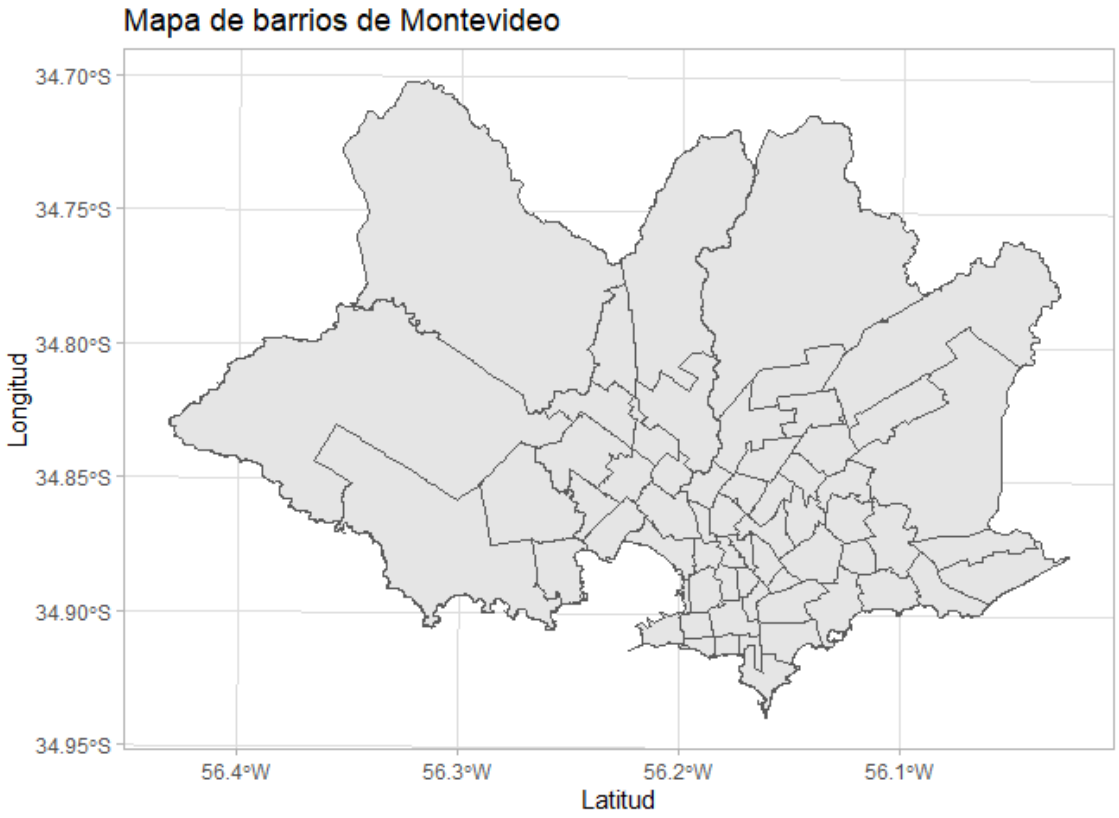


Figura 7.1: Barrios de Montevideo

# 8. Resultados

En este capítulo se presentan los resultados de aplicar las técnicas estudiadas siguiendo el esquema de trabajo planteado en el capítulo anterior, comenzando con el estudio de una posible autocorrelación espacial, siguiendo por los resultados de los estimadores **EBLUP** y **SEBLUP** en conjunto con sus respectivas medidas de calidad (y gráficos para comparar los rendimientos), se buscará analizar cual de los estimadores se adapta mejor al problema.

## 8.1. Estimaciones directas

Se realiza un primer mapa de pobreza haciendo uso de las estimaciones directas de la cantidad de personas por debajo de la línea de pobreza a nivel de barrio.

El resultado fue el siguiente:

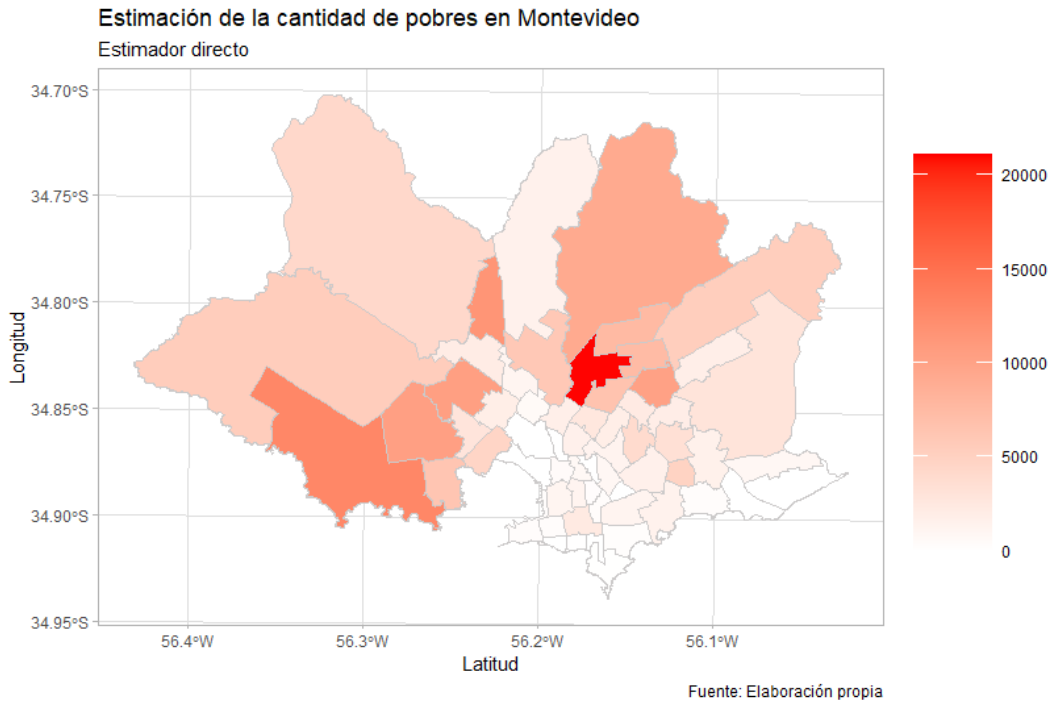


Figura 8.1: Estimación directa de la cantidad de personas por debajo de la línea de pobreza en Montevideo

Un aspecto relevante a destacar, es que en los mapas construidos para visualizar la cantidad estimada de personas por debajo de la línea de de pobreza, **se fijó el margen**

**de escala para valores de 0 a 21100**. De esta forma se logra que los mapas sean realmente contrastables entre ellos. De no hacer esto, R adapta los colores al conjunto de datos y puede dar la sensación de que las estimaciones son idénticas entre mapas (cuando en realidad lo único que comparten es el orden en cuanto a cantidad de individuos con ingresos por debajo de la línea de pobreza).

Puede verse como en los barrios periféricos la cantidad de personas por debajo de la línea de pobreza tienden a ser mayor, mientras que para los barrios costeros se tiene en la mayoría de los casos, cero personas por debajo del umbral de pobreza en la muestra. Apreciando el gráfico, se esperaría rechazar la hipótesis nula de ausencia de autocorrelación espacial tras aplicar el índice de Moran.

Acompañando a esta figura, se grafica el $\mathbb{CV}$ de las estimaciones directas por barrio.

Es claro como la calidad de las estimaciones en la mayor parte de los barrios en ineficiente (de un total de 62 barrios, tan solo para 9 se obtienen estimaciones apenas confiables).

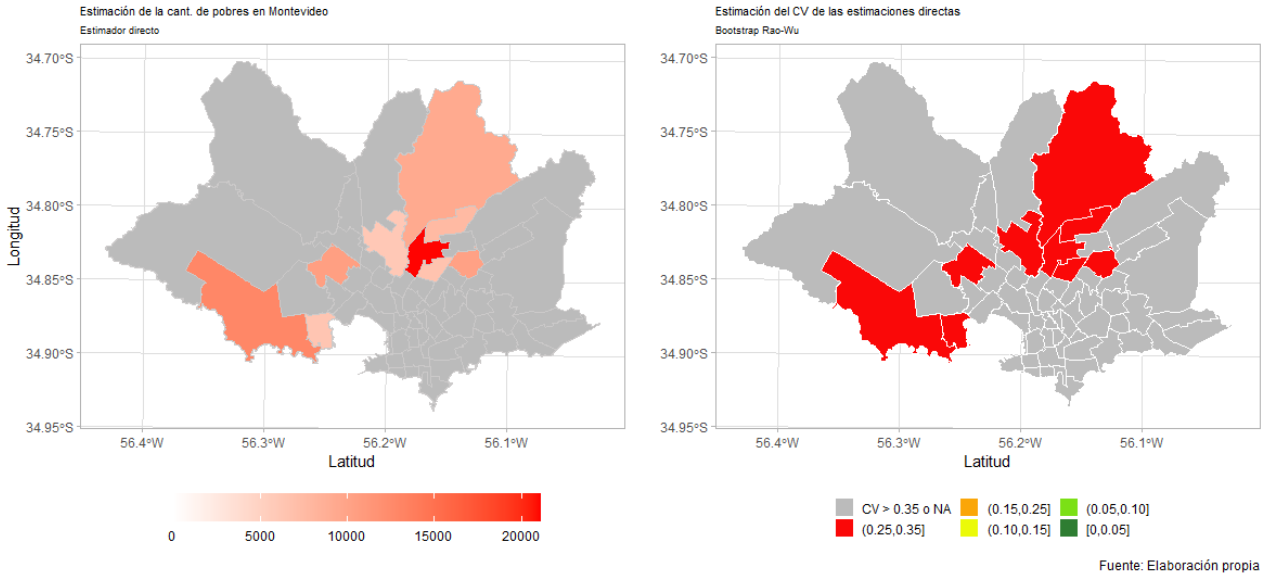


Figura 8.2: Mapa de estimaciones directas y calidad de las mismas

En el transcurso de este capítulo se analizará qué tanto los estimadores **EBLUP** y **SEBLUP** logran mejorar la calidad de las estimaciones directas.

(Statistics Canada, 2013) menciona que estimaciones con un $\mathbb{CV}$ por debajo del 35 % son publicables, aunque destaca que aquellas con un $\mathbb{CV}$ entre 25 % y 35 % deben ser tratadas con precaución.

## 8.2. Estimaciones EBLUP

Comenzando con la aplicación de técnicas **SAE** se presentan los resultados del estimador empírico del **FH**.

Vale la pena resaltar que en el caso de los barrios en donde no se haya registrado ninguna persona por debajo del umbral de pobreza, el trabajo se apoya en la metodología propuesta por (Molina, 2019), que consiste en darle todo el peso al estimador sintético. De no hacer esto, sucedería lo contrario. Todo el peso sería dado al estimador directo (cuya estimación será de pobreza cero y varianza nula), esto no genera una estimación confiable ya que de cierta manera afirmaría sin duda alguna que en esos barrios no hay pobres (a partir de una muestra jamás se tendrá certeza total).

Para la estimación del $\mathbb{ECM}$ en esos barrios se aplicará la fórmula para estimar el $\mathbb{ECM}$ en dominios no muestreados propuesto por (Rao y Molina, 2014).

A simple vista, resalta como a nivel general las estimaciones del **EBLUP** se encuentran por debajo de las estimaciones directas.

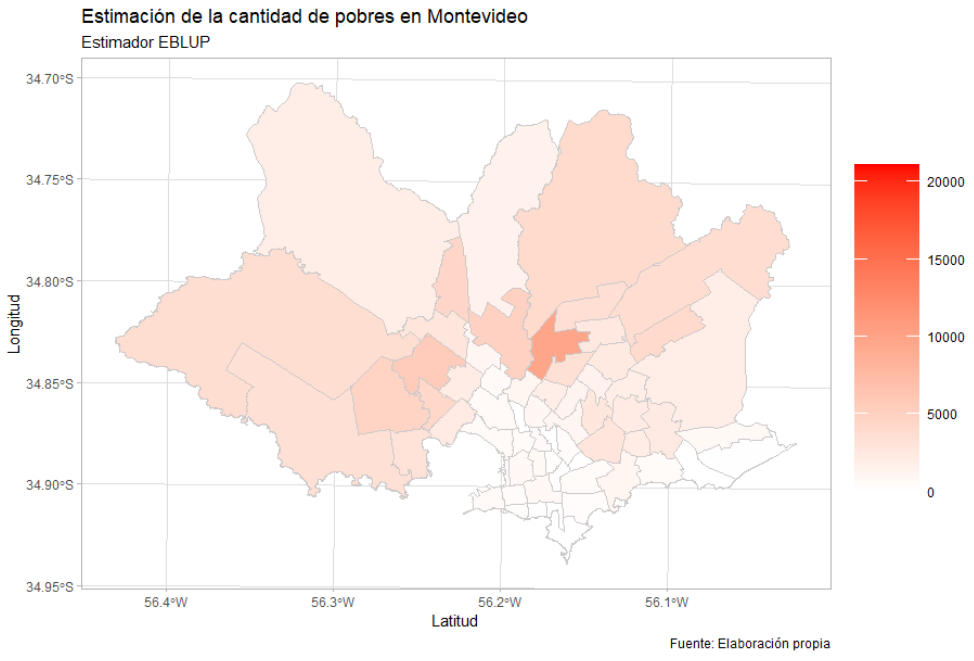


Figura 8.3: Mapa estimaciones EBLUP

La presencia de autocorrelación espacial positiva sigue siendo evidente. Los barrios ubicados al sureste tienden a tener una cantidad de personas por debajo del umbral de pobreza estimada ínfima, y cuanto más al norte nos ubiquemos, mayor. En la sección si-

guiente, cuando se estudien los estimadores **SEBLUP**, la estimación de $\rho$ fundamentará más formalmente la idea aquí planteada.

Las estimaciones de los parámetros del **EBLUP** fueron las siguientes:

| Parámetros estimados del EBLUP | | | |
|---|---|---|---|
| Parámetro | Valor | $\hat{\sigma}$ | P-valor |
| $\hat{\beta}_0$ | -231.703 | 140.748 | 0.100 |
| $\hat{\beta}_{NUOS}$ | -0.162 | 0.041 | <0.001 |
| $\hat{\beta}_{DOI}$ | 0.275 | 0.067 | <0.001 |
| $\hat{\beta}_{VNBI}$ | 6.238 | 1.401 | <0.001 |
| $\hat{\sigma}_u^2$ | 110597.9 | - | - |

El valor del intercepto ($\beta_0$) es incoherente en el contexto en el que trabajamos, el mismo expresa que dado un barrio sin universitarios, desocupados, inactivos y sin personas con más de 3 necesidades básicas insatisfechas, se espera que la cantidad de personas por debajo del umbral de pobreza sea negativa. Si bien es posible eliminarlo del modelo, su inclusión (en este caso) mejoró la calidad global de los estimadores.

De todos modos, no se debe olvidar que el estimador de **FH** es una combinación convexa de un estimador sintético (el afectado por este problema con $\hat{\beta}_0$) y uno directo, por lo que una estimación negativa del estimador sintético no necesariamente implica una estimación negativa del estimador **FH**.

El resto de las estimaciones coincide con lo esperado. A mayor cantidad de universitarios, menor será la cantidad de personas pobres estimada. Al contrario, cuanto más desocupados y hogares con 3 o más necesidades básicas insatisfechas, mayor será la cantidad de pobres estimada. El valor de $\hat{\beta}_{VNBI}$ expresa que por cada hogar con 3 o más NBI se espera hayan 6 personas por debajo del umbral de pobreza (según el modelo sintético)

Por otra parte, la estimación de los parámetros de varianza convergió rápidamente. Solo tomó 9 iteraciones. El modelo estima un gran nivel de heterogeneidad entre los dominios, el valor de $\hat{\sigma}_u^2$ es 110597,9.

Con respecto a la calidad de las estimaciones, el estimador de **FH** mostró una gran mejoría de la calidad de las estimaciones a nivel global. De 9 barrios con estimaciones apenas aceptables en el estimador directo, se pasó a 35 barrios con estimaciones aceptables. Puede verse como incluso existen una gran cantidad de barrios cuyo $\mathbb{CV}$ se

encuentra por debajo del 25 % e incluso por debajo del 15 %.

Por lo que no solo aumentó a más del doble la cantidad de barrios con estimaciones aceptables, sino que gran parte de los que lo hicieron alcanzaron un nivel de calidad medianamente bueno.

Lo explicado se resume en el siguiente gráfico:

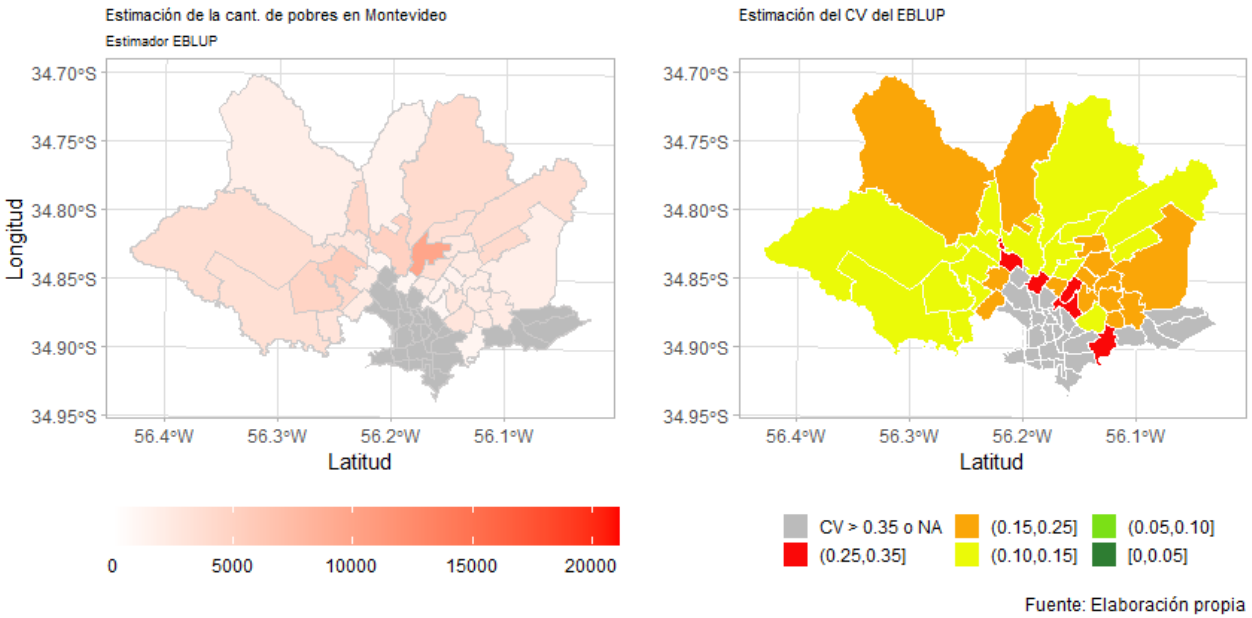


Figura 8.4: Mapa estimaciones EBLUP y calidad

## 8.3. Correlación espacial

Con el objetivo es estudiar la autocorrelación espacial de la pobreza en los distintos barrios de Montevideo, se comienza definiendo las distintas matrices de vecindad a ser utilizadas:

- $W_1$: matriz de conectividad binaria.
- $W_2$: matriz de 2 vecinos más cercanos.
- $W_3$: matriz de vecindad definida a partir de distancia menor a 5km.

### 8.3.1. Matriz W1

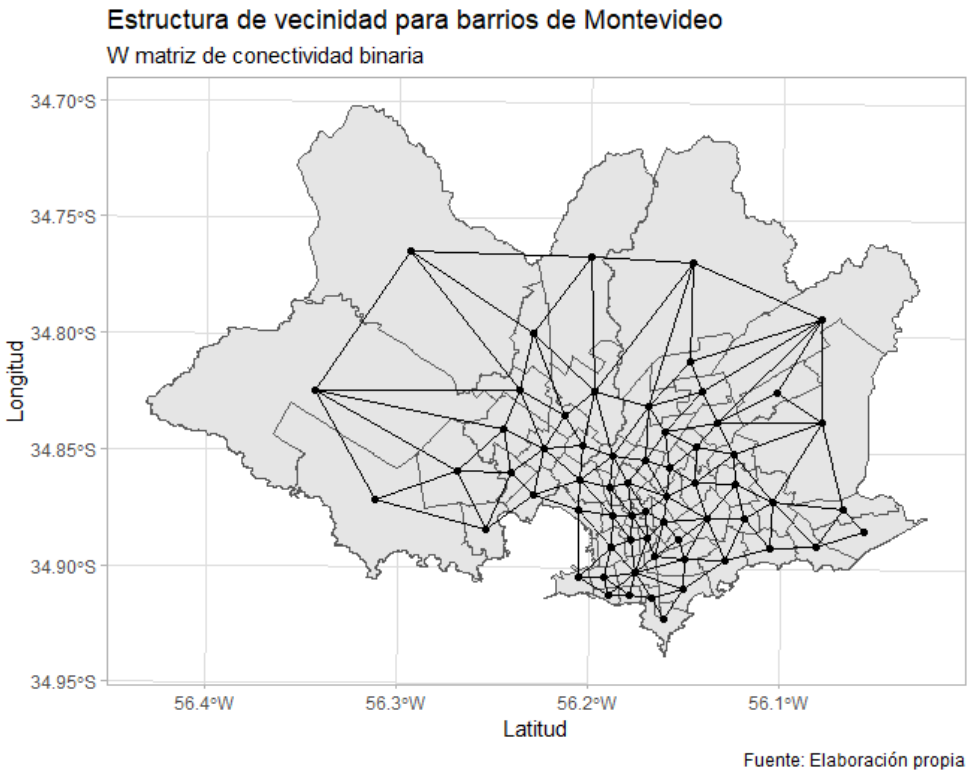


Figura 8.5: Representación de $W_1$ para los barrios de Montevideo

### 8.3.2. Matriz W2

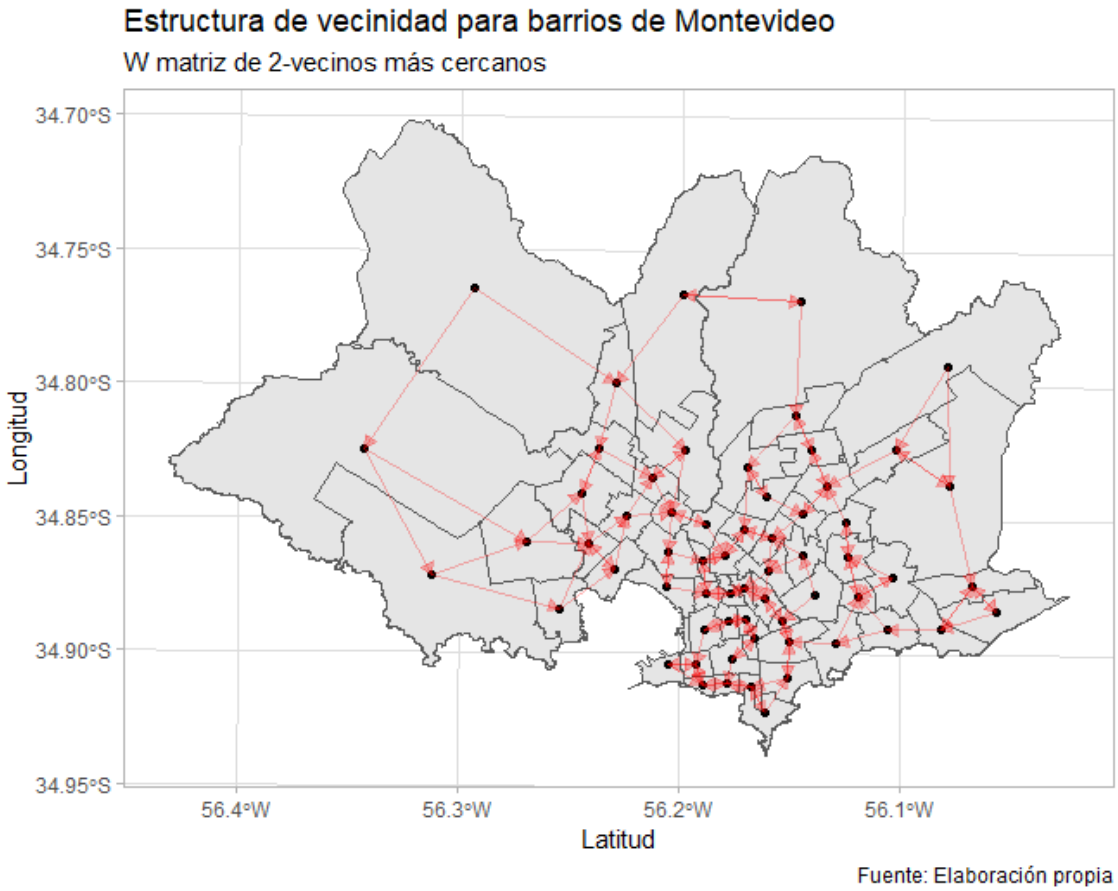


Figura 8.6: Representación de $W_2$ para los barrios de Montevideo

$A \to B$ si $B$ es vecino de $A$ (y no necesariamente A vecino de B).

### 8.3.3. Matriz W3

En este caso puede verse que a nivel general, los barrios periféricos tienden a tener menos vecinos. Son barrios grandes y sus centroides son lejanos a los del resto. En esos casos la matriz de vecindad será nula para sus respectivas filas, por lo que la expresión algebráica del estimador de FH espacial será equivalente al común y corriente.

Los barrios céntricos tienen muchos vecinos y se ven influenciados por todos ellos.

Cabe destacar que en el gráfico se presentan 2 ejemplos, uno asociado a la matriz de peso cuando se toma 1km como distancia de referencia y el otro cuando se toma 5km.

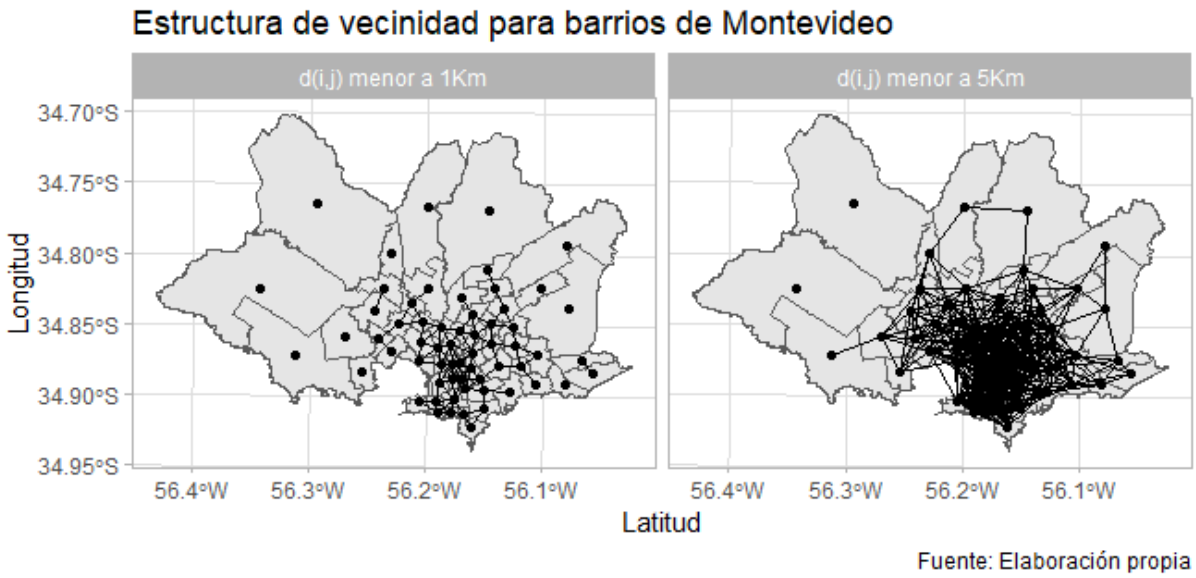


Figura 8.7: Representación de $W_3$ para los barrios de Montevideo

En este trabajo se optó por trabajar con una distancia de 5km.

### 8.3.4. Testeo de la autocorrelación espacial

Es de interés testear la existencia de autocorrelación espacial en los 3 casos para estudiar la coherencia de introducír estructura espacial a los efectos aleatorios (es decir, ¿tiene sentido aplicar el estimador de **FH-Espacial**?).

Construidas las matrices y aplicadas las funciones correspondientes se obtuvieron los sigueintes resultados:

| Resultados del Test de Moran | | | |
|---|---|---|---|
| W | Índice de Moran | Varianza | P-valor |
| $W_1$ | 0.201 | 0.005 | 0.001 |
| $W_2$ | 0.351 | 0.011 | <0.001 |
| $W_3$ | 0.195 | 0.002 | <0.001 |

Como puede verse, en cada uno de los casos se rechaza la hipótesis nula de ausencia de autocorrelación espacial. Los P-valores tienden a 0.

Visto esto y con el objetivo de estimar la cantidad de personas por debajo del umbral de pobreza a nivel de barrios, será de interés aplicar el **SEBLUP** como alternativa al **EBLUP** (y luego evaluar diferencias en el rendimiento).

**Observación:**

En esta sección se estudió la autocorrelación espacial de las estimaciones directas, estimados los modelos **SEBLUP** en R, la estimación de $\rho$ se corresponderá con la autocorrelación espacial existente entre los efectos aleatorios. Claramente esta se relaciona con la autocorrelación de las estimaciones directas pero no tienen por qué ser iguales.

## 8.4. Estimaciones SEBLUP

En este apartado se estima la cantidad de personas por debajo del umbral de pobreza haciendo uso del estimador de Fay-Herriot espacial. Para ello se desplegará en cada subsección los resultados según cada una de las matrices de vecindad construidas acompañados de sus respectivas medidas de calidad. Se hace uso del paquete **SAE**.

### 8.4.1. Estimaciones SEBLUP CON $W_1$

Realizadas las estimaciones del **SEBLUP-W1** puede verse que la distribución de las estimaciones es muy parecida a la del **EBLUP**.

A nivel general y frente a las estimaciones directas existe un descenso generalizado (con excepción de los barrios con pobreza cero en la muestra) de la cantidad de personas con ingresos por debajo de la línea de pobreza.

Al igual que en el caso anterior, se presenta a continuación el mapa de las estimaciones obtenidas:

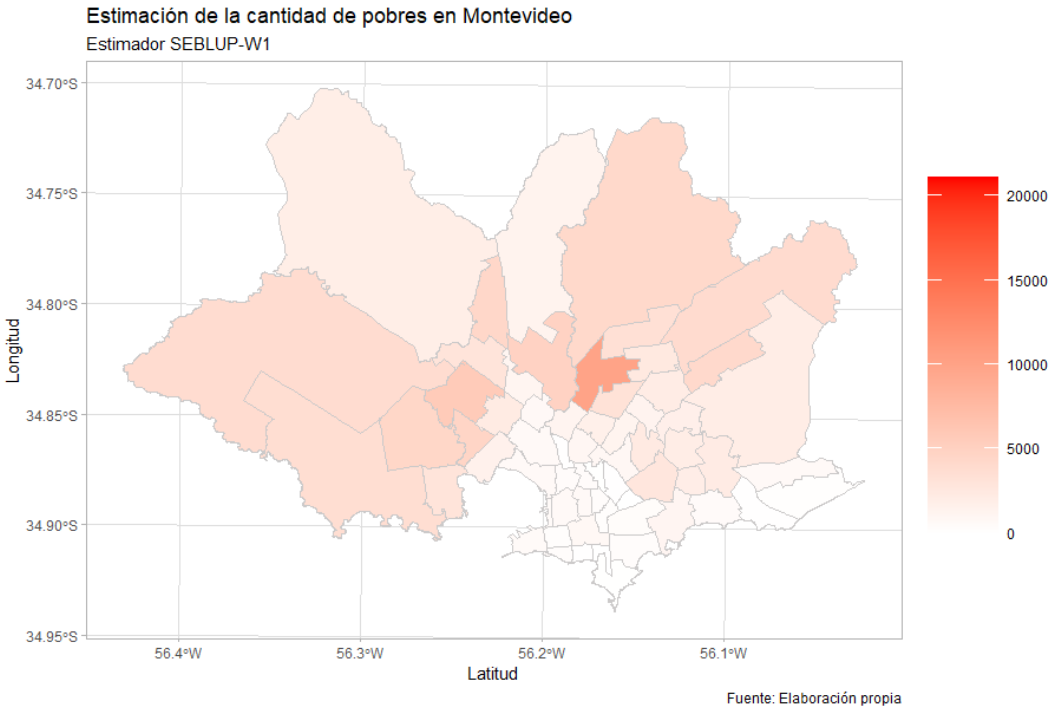


Figura 8.8: Mapa estimaciones SEBLUP-W1

La diferencia mayor se encuentra en la calidad de las estimaciones. De las estimacio-

nes **SEBLUP-W1**, 42 alcanzan los criterios de calidad establecidos. Esta es una gran mejoría frente a la situación inicial en donde tan solo 9 barrios poseían estimaciones de calidad.

A su vez, se resalta un patrón que se verá en todos los estimadores presentados y por presentar: **la mayor parte de barrios céntricos y costeros no obtuvieron estimaciones de calidad**. Estos barrios se caracterizan por haber tenido poca cantidad de personas por debajo de la línea de pobreza en la muestra, lo que hace que la varianza del estimador sea elevada. Esta última afirmación se verá mejor justificada en las tablas anexas finales.

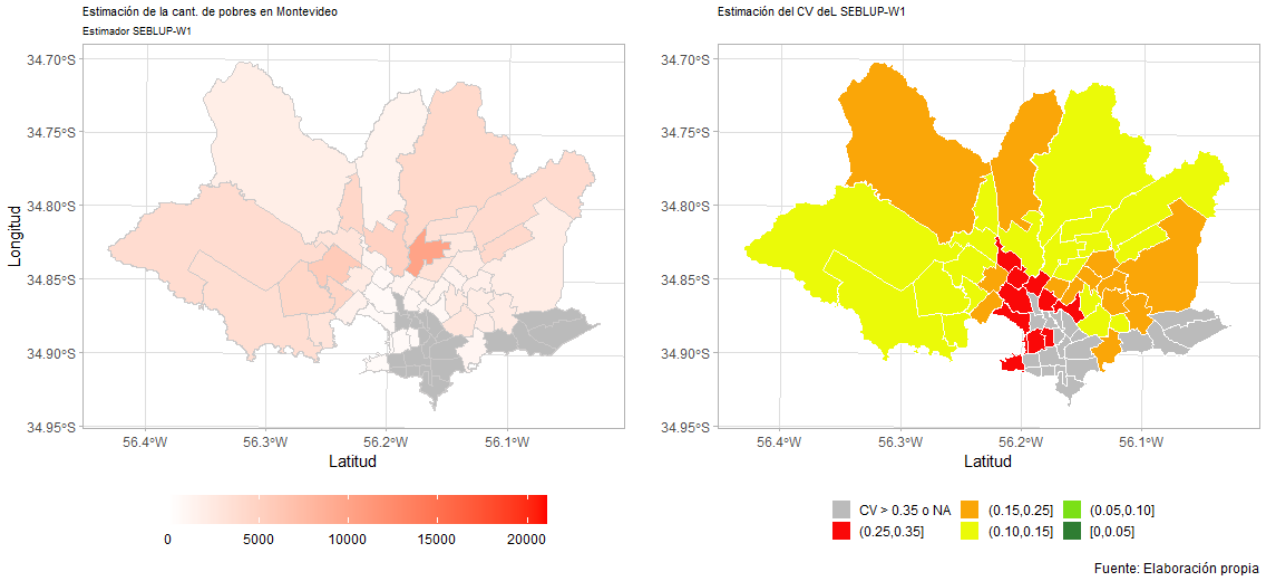


Figura 8.9: Mapa estimaciones SEBLUP-W1 y calidad

En cuanto al análisis de los parámetros del modelo, se tiene:

| Parámetros estimados del SEBLUP-W1 | | | |
|---|---|---|---|
| Parámetro | Valor | $\hat{\sigma}$ | P-valor |
| $\hat{\beta}_0$ | -77.495 | 268.931 | 0.773 |
| $\hat{\beta}_{NUOS}$ | -0.117 | 0.040 | 0.003 |
| $\hat{\beta}_{DOI}$ | 0.190 | 0.064 | 0.003 |
| $\hat{\beta}_{VNBI}$ | 7.153 | 1.270 | <0.001 |
| $\hat{\rho}$ | 0.894 | - | - |
| $\hat{\sigma}_u^2$ | 33463.7 | - | - |

Los resultados siguen el esquema del modelo anterior en cuanto a magnitudes. Es claro que todos los parámetros (con excepción del intercepto) son estadísticamente significativos.

A diferencia del modelo anterior, en este se suma el parámetro $\rho$ cuya estimación tiene el valor de 0.8941. Esto indica un altísimo nivel de autocorrelación espacial de los efectos aleatorios (y por consiguiente del valor del estimador en los distintos dominios). Si los barrios vecinos tienen una gran cantidad de personas por debajo del umbral de pobreza, muy probablemente ese barrio también los tenga.

Otro aspecto relevante es que la varianza estimada de los efectos aleatorios toma el valor de 33463.69, 3 veces menor que en el modelo de **FH** clásico.

### 8.4.2. Estimaciones SEBLUP CON $W_2$

Se procede ahora con el análisis del **EBLUP** espacial con matriz de vecindad $W_2$.

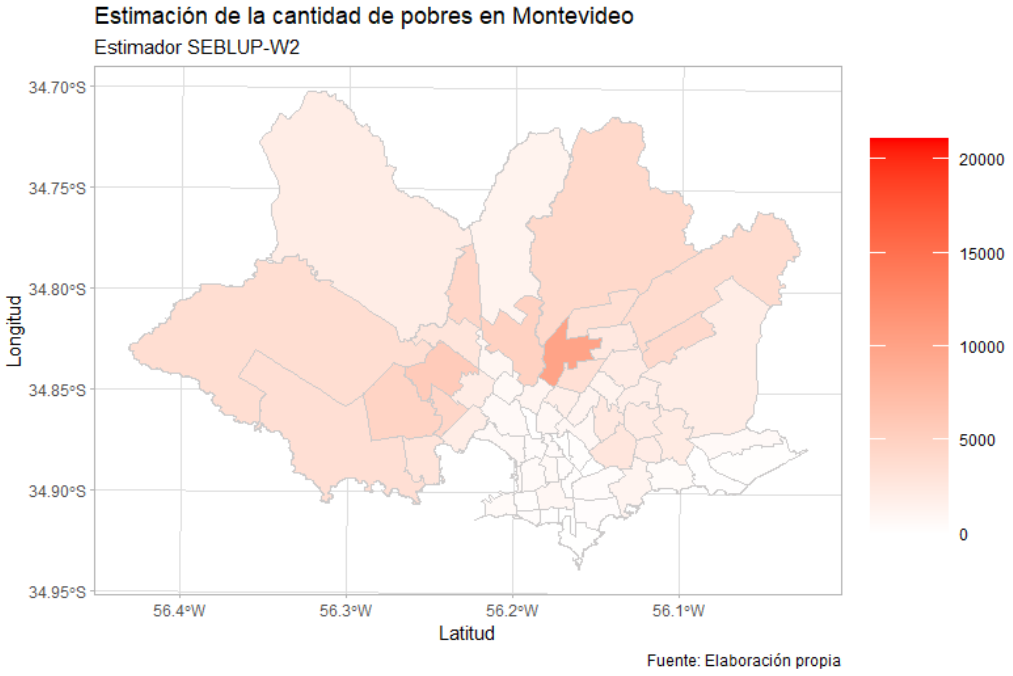


Figura 8.10: Mapa estimaciones SEBLUP-W2

Los resultados de la estimación del **SEBLUP-W2** son muy similares a las del **SEBLUP-W1** y como veremos posteriormente a las del **SEBLUP-W3**, por lo que independientemente de la matriz de vecindad con la que se opte trabajar, en este problema, los resultados serán similares. A pesar de ello, se verá que la calidad de los estimadores no

necesariamente es igual.

Con respecto a la calidad de las mismas, existe un descenso en la cantidad de barrios con estimaciones de calidad aceptable. Tan solo 34 barrios se atienen a los criterios de calidad. Esto puede deberse al aumento de la varianza de los efectos aleatorios con respecto al **SEBLUP-W1**, recordemos que la combinación convexa queda definida a partir del parámetro $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2+\psi_d}$ por lo que ante un aumento de la varianza de $\sigma_u^2$, el estimador asigna un mayor peso al estimador directo (que en la mayoría de los barrios tiene un rendimiento deficiente).

Se presenta a continuación el mapa censurado acompañado del mapa de calidad, la presencia de una mayor cantidad de zonas grises apoya lo explicado.

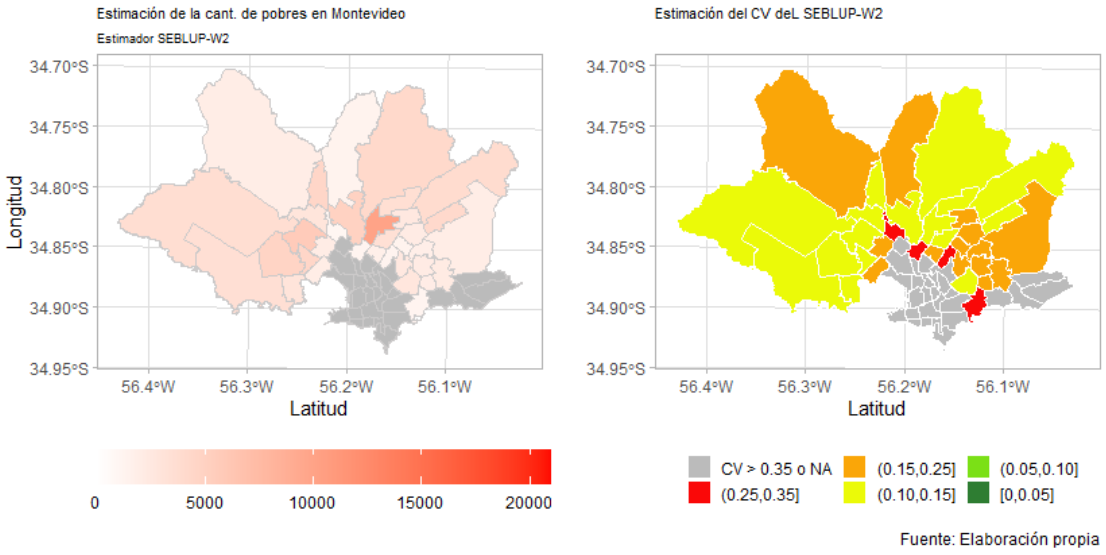


Figura 8.11: Mapa estimaciones SEBLUP-W2 y calidad

En cuanto a los parámetros estimados se tiene:

| Parámetros estimados del SEBLUP-W2 | | | |
|---|---|---|---|
| Parámetro | Valor | $\hat{\sigma}$ | P-valor |
| $\hat{\beta}_0$ | -150.240 | 165.519 | 0.364 |
| $\hat{\beta}_{NUOS}$ | -0.138 | 0.045 | 0.002 |
| $\hat{\beta}_{DOI}$ | 0.229 | 0.073 | 0.002 |
| $\hat{\beta}_{VNBI}$ | 6.855 | 1.492 | <0.001 |
| $\hat{\rho}$ | 0.457 | - | - |
| $\hat{\sigma}_u^2$ | 86309.1 | - | - |

Estos tienen el mismo comportamiento que se vienen describiendo desde el **EBLUP**.

Existe también un descenso en el nivel de autocorrelación espacial estimado. El valor de $\hat{\rho}$ es de 0.4569989 (correlación positiva más leve que en el caso anterior). A su vez, $\hat{\sigma}_u^2 = 86309.13$ (más del doble del estimador anterior).

La tendencia muestra una relación negativa entre la correlación espacial y la varianza de los efectos aleatorios (esto es obvio, cuanto mayor correlación positiva se espera un mayor nivel de homogeneidad entre los efectos).

### 8.4.3. Estimaciones SEBLUP CON $W_3$

Finalmente, se presetan los resultados obtenidos tras aplicar el **SEBLUP-W3**.

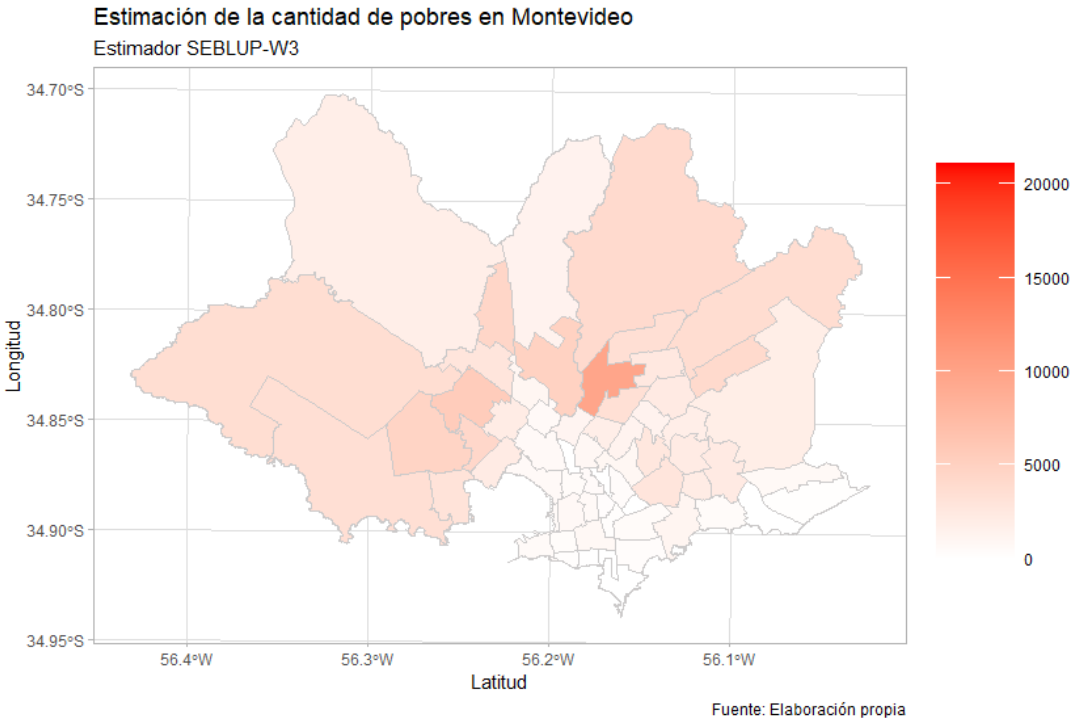


Figura 8.12: Mapa estimaciones SEBLUP-W3

Nuevamente, la diferencia con el resto de los **SEBLUP** es imperceptible.

En cambio, si bien la cantidad de barrios cuya calidad en las estimaciones es aceptable es de 34 (menos que en el caso del **EBLUP** y el **SEBLUP-W1**) este es el único estimador que alcanzó margenes de calidad muy buenos en 3 barrios (en el mapa de CV se ven por primera vez barrios de color verde).

El nivel de correlación espacial en este caso es positivo pero leve (0.362738). Esto puede deberse a que en el caso de los barrios que no limitan con Canelones, el radio de vecindad es tan grande que hace que todos sean vecinos entre sí. Con esto se tiene que si bien parte de los vecinos son parecidos, otros no tanto.

Esto trajo consigo un aumento de la varianza de los efectos aleatorios $\hat{\sigma}_u^2 = 109304.4$, valor que se acerca al del **EBLUP**).

En la siguiente página se muestra el mapa mencionado y la tabla con los resultados de las estimaciones realizadas.

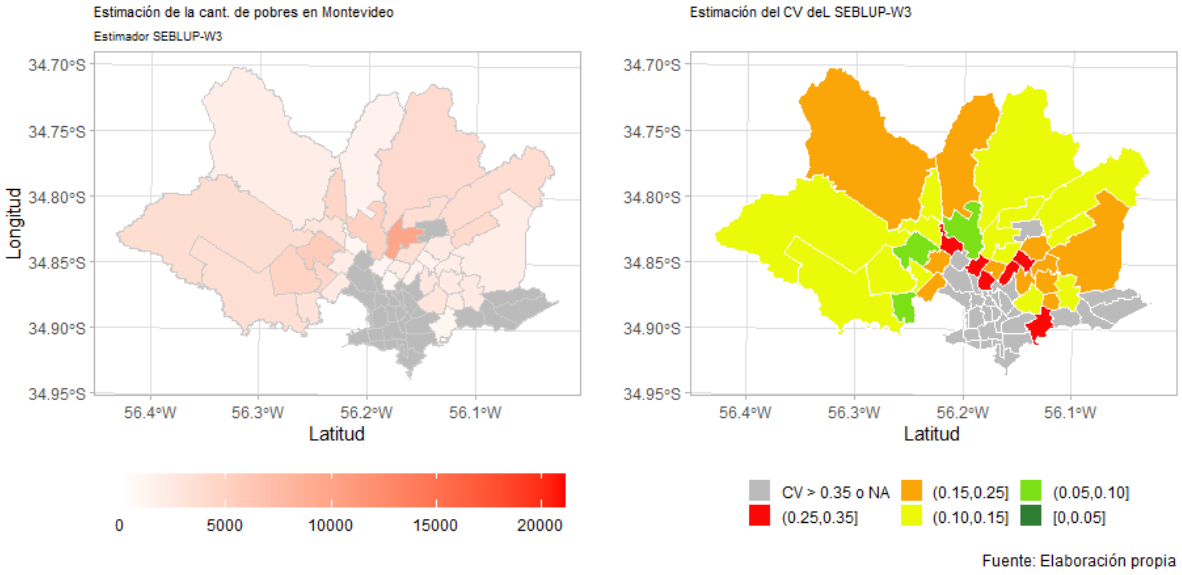


Figura 8.13: Mapa estimaciones SEBLUP-W3 y calidad

| Parámetros estimados del SEBLUP-W3 | | | |
|---|---|---|---|
| Parámetro | Valor | $\hat{\sigma}$ | P-valor |
| $\hat{\beta}_0$ | -199.801 | 153.715 | 0.194 |
| $\hat{\beta}_{NUOS}$ | -0.156 | 0.042 | <0.001 |
| $\hat{\beta}_{DOI}$ | 0.263 | 0.068 | <0.001 |
| $\hat{\beta}_{VNBI}$ | 6.319 | 1.407 | <0.001 |
| $\hat{\rho}$ | 0.363 | - | - |
| $\hat{\sigma_u}^2$ | 109304.4 | - | - |

## 8.5. Comparación de resultados

En esta subsección se busca analizar detenidamente las diferencias entre los estimadores, comenzando con los valores propiamente predichos y finalizando con criterios de calidad.

A nivel general, puede verse como las estimaciones directas tienden a ser mayores que el resto, con la clara excepción de aquellos barrios en los cuales no hubo personas pobres en la muestra.

Otro aspecto a destacar es que las diferencias existentes entre los estimadores **SAE** y particularmente en los **SEBLUP** es mínima. Esto ya se veía reflejado en los mapas presentados.

Por lo que la elección de la matriz de vecindad pareciera no ser un aspecto determinante al momento de estimar la cantidad de pobres en un barrio.

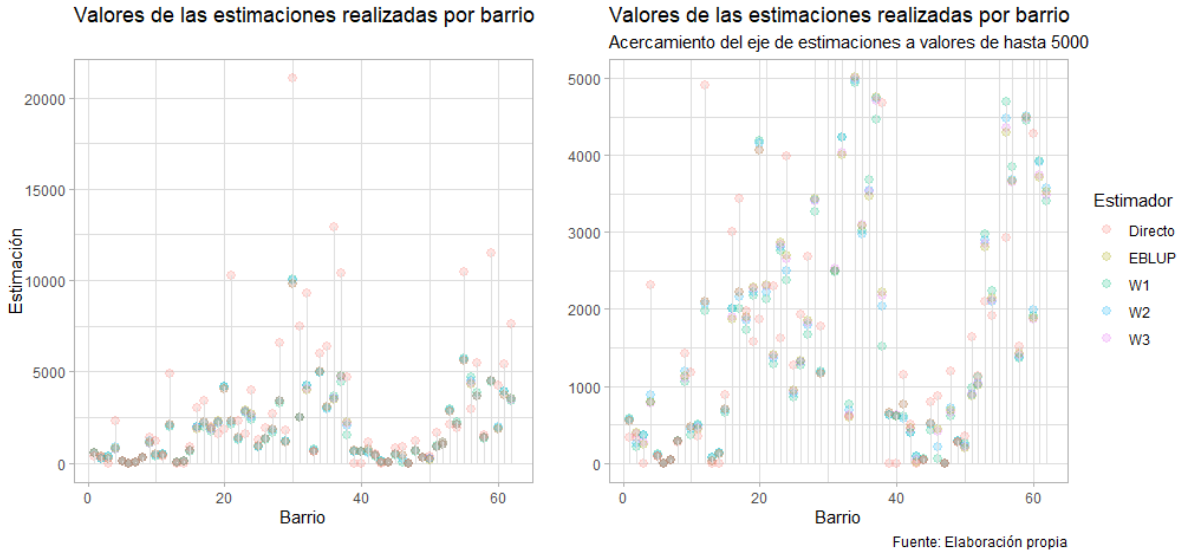


Figura 8.14: Gráfico burbujas de las estimaciones por barrio

A pesar de lo anterior, las estimaciones de los $\mathbb{CV}$ suelen diferir.

En el gráfico se observa como a nivel general el **SEBLUP-W1** obtuvo estimaciones de mejor calidad. Se destaca como en barrios particulares (Cerro, Nuevo París y Lavalleja) el estimador **SEBLUP-W3** obtuvo estimaciones de nivel excepcional en comparación al resto de los estimadores. En partícular, la estimación para el barrio Cerro tuvo un $\mathbb{CV}$ estimado de 5,02 %.

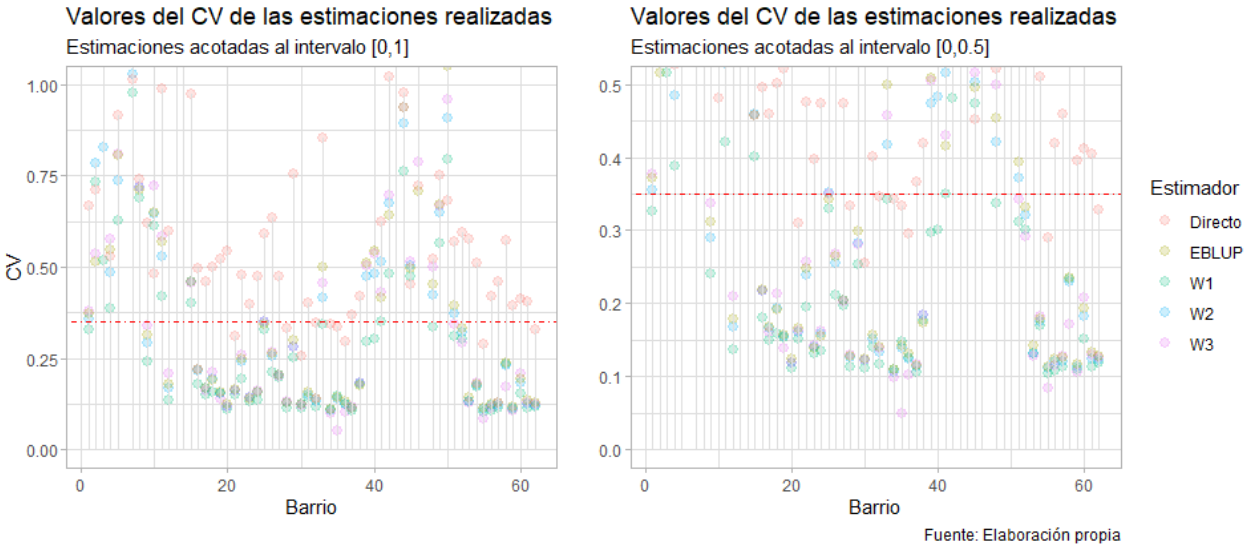


Figura 8.15: Gráfico burbujas de la calidad de las estimaciones por barrio

El siguiente Boxplot resume la dispersión del $\mathbb{CV}$ de los distintos estimadores:

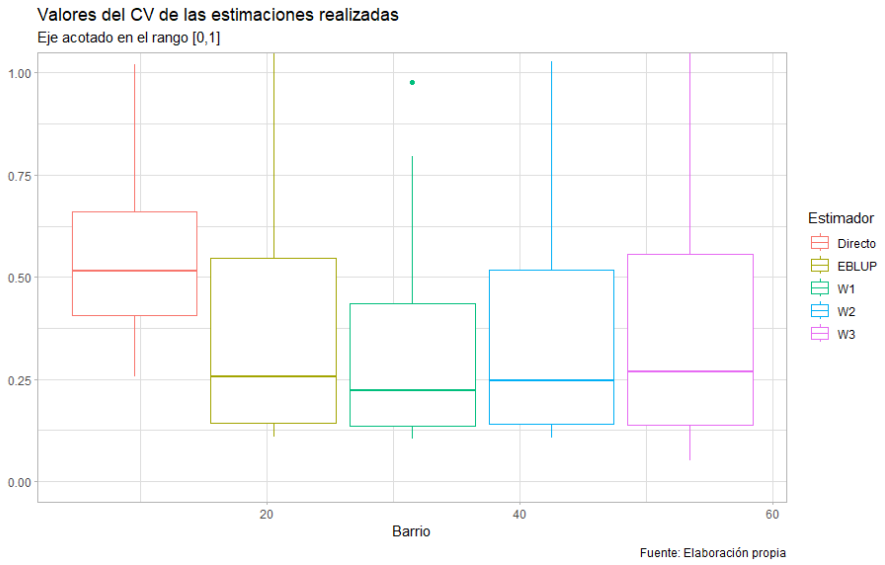


Figura 8.16: Boxplot de la distribución del CV de las estimaciones por barrio

Se debe prestar atención a que el gráfico fue acotado al intervalo $[0, 1]$ ya que outliers muy grandes "achataban" en exceso las cajas haciendo imposible diferenciar el comportamiento del $\mathbb{CV}$ entre estimadores.

Es claro como el estimador directo tiende a obtener estimaciones de peor calidad, seguido del **EBLUP** y finalmente los **SEBLUP's**. Lo que indica que considerar la autocorrelación espacial en el modelo no fue en vano.

# 9. Conclusiones

Con el objetivo de construir estimaciones de calidad que reflejen los niveles de pobreza en los distintos barrios de Montevideo, se recurrió a la inferencia basada en modelos como una alternativa al paradigma predominante de la inferencia basada en el diseño. Más en particular, se exploraron técnicas de Estimación en Áreas Pequeñas, conjunto de técnicas que como se mencionó previamente ha ganado relevancia en los últimos años.

Lejos de decepcionar, los resultados obtenidos fueron concordantes con la bibliografía consultada: *los estimadores reflejan una mejoría generalizada en la calidad de las estimaciones con respecto a las estimaciones directas.*

Esta mejoría fue independiente de si se trabajaba con **EBLUP** o **SEBLUP** (en cualquiera de sus variedades). Aunque sí destaca como ante la presencia de una alta correlación espacial (definida la matriz de vecindad) la calidad de las estimaciones tienden a mejorar de sobremanera.

Se debe tener en consideración que **SAE** no es una solución garantizada para todo barrio, el punto de arranque influye en el resultado de las estimaciones. A nivel general, barrios con estimaciones directas de calidad paupérrima no alcanzaron tampoco estimaciones **SAE** de calidad, aunque sí se logró reducir el error cuadrático medio de los estimadores.

Otro aspecto a tener en cuenta es que el rendimiento de los estimadores se ve condicionado a la capacidad explicativa del conjunto de variables auxiliares a disposición. No debe olvidarse que aquí fueron tres las utilizadas, casos como los de **CEPAL** en Chile (MDSF-CEPAL, 2022) muestran un uso exhaustivo de un conjunto mucho más numeroso de variables auxiliares. Sería interesante evalúar el rendimiento de las estimaciones aquí planteadas haciendo uso de un conjunto de covariables reforzado, proveniente de distintas fuentes gubernamentales y administrativas, no disponibles al momento de realizar esta investigación. El uso de información proveniente de la teledetección puede tener un aporte relevante, índices de textura y color podrían ayudar a distinguir regiones de mayores o menores ingresos.

A su vez, el trabajo buscó construir estimaciones a nivel de barrios en Montevideo. Es de interés personal y se plantea como potencial línea de investigación, la construcción de mapas de pobreza para otras localidades del país. La implementación de técnicas que permitan estimaciones a nivel más desagregado es también relevante. Transcurrido el Censo 2023, información actualizada y de calidad estará disponible en unos meses. Esta

información se colecta a nivel de individuo, por lo que nuevas herramientas metodológicas quedarán disponibles.

Desagregar las estimaciones de pobreza a nivel de etnia, sexo y otras variables es muy importante para dimensionar la pobreza. Extender la metodología utilizada para obtener resultados en estas desagregaciones implica revisar la misma, y la necesidad de mejorar la información auxiliar, dado que los tamaños de muestra se verían aun más reducidos.

En este trabajo se estimó el total de personas por debajo de la línea de pobreza, sin embargo, existen enfoques alternativos para estudiar el fenómeno de la pobreza. Entre ellos:

- Tasa de pobreza: para este caso debería trabajarse idealmente con estimadores que traten la no linealidad, aunque es posible aplicar transformaciones (transformación arcoseno) al conjunto de variables y utilizar las técnicas planteadas en este trabajo.
- Índices de pobreza/privación multidimensionales.

Enfoques no tan tradicionales están siendo desarrollados, cercano en el tiempo, pueden encontrarse artículos que estudian técnicas de **Machine Learning** aplicadas a **SAE**. (Viljanen, Meijerink, Zwakhals, y Van de Kassteele, 2022) muestra el resultado de la aplicación de este enfoque, en donde pueden verse predicciones a nivel muy desagregado. Resulta interesante la aplicación de este enfoque a nivel de individuo una vez transcurrido el censo de poblaciones (o en un futuro, utilizando registros administrativos).

Para futuros trabajos, es relevante la aplicación de técnicas de benchmarking con motivo de que la suma de las estimaciones **SAE** a nivel de barrio coincida con el valor del estimador directo de la cantidad de personas por debajo del umbral de pobreza en Montevideo.

Con el crecimiento acelerado de las fuentes de información, y de la demanda de estimaciones a nivel cada vez más desagregado, el desarrollo de técnicas que puedan integrar fuentes de información diversas y complejas que modelen correctamente su relación con la variable de interés, es vital. Como se analizó, **SAE** es un conjunto de técnicas cuya aplicación puede lograr este objetivo, brindando un marco metodológico flexible adaptable a diferentes problemáticas y coyunturas.

# 10. Anexo

| N° | Barrio | $\hat{\delta}^{\text{DIR}}$ | CV | $\hat{\delta}^{\text{EBLUP}}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W1}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W2}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W3}$ | CV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ciudad Vieja | 338 | 0.67 | 548 | 0.37 | 577 | 0.33 | 571 | 0.36 | 551 | 0.38 |
| 2 | Centro | 328 | 0.71 | 390 | 0.52 | 213 | 0.73 | 251 | 0.78 | 386 | 0.54 |
| 3 | Barrio Sur | 0 | NA | 244 | 1.37 | 358 | 0.52 | 358 | 0.83 | 266 | 1.25 |
| 4 | Cordon | 2320 | 0.53 | 793 | 0.55 | 788 | 0.39 | 888 | 0.48 | 779 | 0.58 |
| 5 | Palermo | 81 | 0.91 | 91 | 0.81 | 114 | 0.63 | 100 | 0.74 | 91 | 0.81 |
| 6 | Parque Rodó | 0 | NA | 0 | NA | 0 | NA | 0 | NA | 0 | NA |
| 7 | Punta Carretas | 44 | 1.01 | 40 | 1.10 | 45 | 0.98 | 43 | 1.03 | 39 | 1.13 |
| 8 | Pocitos | 274 | 0.74 | 282 | 0.71 | 289 | 0.69 | 280 | 0.72 | 284 | 0.72 |
| 9 | Buceo | 1428 | 0.62 | 1136 | 0.31 | 1058 | 0.24 | 1190 | 0.29 | 1103 | 0.34 |
| 10 | Pque. Batlle, V. Dolores | 1184 | 0.48 | 470 | 0.65 | 366 | 0.61 | 440 | 0.65 | 466 | 0.72 |
| 11 | Malvín | 349 | 0.99 | 448 | 0.57 | 503 | 0.42 | 480 | 0.53 | 451 | 0.58 |
| 12 | Malvín Norte | 4905 | 0.60 | 2105 | 0.18 | 1978 | 0.14 | 2064 | 0.17 | 2093 | 0.21 |
| 13 | Punta Gorda | 0 | NA | 18 | 18.49 | 66 | 2.81 | 66 | 4.48 | 32 | 10.39 |
| 14 | Carrasco | 0 | NA | 111 | 3.04 | 134 | 1.39 | 134 | 2.22 | 121 | 2.76 |
| 15 | Carrasco Norte | 888 | 0.97 | 709 | 0.46 | 648 | 0.40 | 680 | 0.46 | 702 | 0.46 |
| 16 | Bañados de Carrasco | 3004 | 0.50 | 1872 | 0.22 | 2001 | 0.18 | 2000 | 0.22 | 1904 | 0.22 |
| 17 | Maroñas, Parque Guaraní | 3437 | 0.46 | 2219 | 0.17 | 2014 | 0.15 | 2160 | 0.17 | 2215 | 0.16 |
| 18 | Flor de Maroñas | 1975 | 0.50 | 1907 | 0.19 | 1736 | 0.16 | 1847 | 0.19 | 1888 | 0.21 |

| N° | Barrio | $\hat{\delta}^{\text{DIR}}$ | CV | $\hat{\delta}^{\text{EBLUP}}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W1}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W2}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W3}$ | CV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Las Canteras | 1580 | 0.52 | 2281 | 0.16 | 2173 | 0.15 | 2229 | 0.15 | 2276 | 0.14 |
| 20 | Pta. Rieles, Bella Italia | 1874 | 0.54 | 4059 | 0.12 | 4193 | 0.11 | 4164 | 0.12 | 4059 | 0.12 |
| 21 | Jardines del Hipódromo | 10250 | 0.31 | 2312 | 0.17 | 2138 | 0.15 | 2220 | 0.16 | 2305 | 0.16 |
| 22 | Ituzaingo | 2294 | 0.48 | 1405 | 0.25 | 1288 | 0.19 | 1355 | 0.24 | 1397 | 0.26 |
| 23 | Unión | 1622 | 0.40 | 2868 | 0.14 | 2755 | 0.13 | 2801 | 0.14 | 2843 | 0.14 |
| 24 | Villa Española | 3982 | 0.47 | 2695 | 0.16 | 2378 | 0.14 | 2495 | 0.16 | 2656 | 0.16 |
| 25 | Mcdo. Modelo, Bolivar | 1275 | 0.59 | 945 | 0.34 | 855 | 0.33 | 895 | 0.35 | 930 | 0.35 |
| 26 | Castro, P. Castellanos | 1929 | 0.63 | 1325 | 0.26 | 1272 | 0.21 | 1314 | 0.25 | 1318 | 0.27 |
| 27 | Cerrito | 2687 | 0.47 | 1847 | 0.20 | 1667 | 0.20 | 1791 | 0.20 | 1828 | 0.20 |
| 28 | Las Acacias | 6585 | 0.33 | 3435 | 0.13 | 3263 | 0.11 | 3425 | 0.13 | 3399 | 0.13 |
| 29 | Aires Puros | 1784 | 0.75 | 1155 | 0.30 | 1191 | 0.25 | 1178 | 0.28 | 1169 | 0.28 |
| 30 | Casavalle | 21060 | 0.26 | 9826 | 0.12 | 10058 | 0.11 | 10015 | 0.12 | 9808 | 0.12 |
| 31 | Piedras Blancas | 7488 | 0.40 | 2501 | 0.16 | 2484 | 0.14 | 2496 | 0.15 | 2523 | NA |
| 32 | Manga, Toledo Chico | 9325 | 0.35 | 4006 | 0.14 | 4240 | 0.12 | 4242 | 0.13 | 4037 | 0.14 |
| 33 | Paso de las Duranas | 604 | 0.85 | 598 | 0.50 | 764 | 0.34 | 693 | 0.42 | 631 | 0.46 |
| 34 | Peñarol, Lavalleja | 6010 | 0.34 | 5024 | 0.11 | 4944 | 0.11 | 4977 | 0.11 | 5010 | 0.10 |
| 35 | Cerro | 6355 | 0.33 | 3076 | 0.14 | 3022 | 0.15 | 2979 | 0.14 | 3096 | 0.05 |
| 36 | Casabó, Pajas B. | 12914 | 0.30 | 3471 | 0.13 | 3676 | 0.13 | 3537 | 0.12 | 3531 | 0.10 |
| 37 | La Paloma, Tomkinson | 10420 | 0.37 | 4756 | 0.11 | 4462 | 0.11 | 4743 | 0.11 | 4710 | 0.12 |
| 38 | La Teja | 4676 | 0.42 | 2228 | 0.17 | 1514 | 0.18 | 2031 | 0.18 | 2174 | 0.18 |
| 39 | Prado, Nueva Savona | 0 | NA | 660 | 0.51 | 628 | 0.30 | 628 | 0.47 | 662 | 0.51 |
| 40 | Capurro, Bella Vista | 0 | NA | 614 | 0.55 | 613 | 0.30 | 613 | 0.48 | 620 | 0.54 |
| 41 | Aguada | 1141 | 0.62 | 763 | 0.42 | 614 | 0.35 | 579 | 0.52 | 762 | 0.43 |

| N° | Barrio | $\hat{\delta}^{\text{DIR}}$ | CV | $\hat{\delta}^{\text{EBLUP}}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W1}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W2}$ | CV | $\hat{\delta}^{\text{SEBLUP}}_{W3}$ | CV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 42 | Reducto | 501 | 1.02 | 461 | 0.64 | 392 | 0.48 | 402 | 0.67 | 444 | 0.70 |
| 43 | Atahualpa | 0 | NA | 16 | 21.47 | 93 | 1.96 | 93 | 3.15 | 35 | 9.37 |
| 44 | Jacinto Vera | 38 | 0.98 | 40 | 0.94 | 49 | 0.76 | 42 | 0.89 | 40 | 0.94 |
| 45 | La Figurita | 794 | 0.45 | 522 | 0.50 | 427 | 0.47 | 503 | 0.50 | 520 | 0.52 |
| 46 | Larra1aga | 864 | 0.72 | 441 | 0.71 | 56 | 3.22 | 208 | 1.34 | 417 | 0.79 |
| 47 | La Blanqueada | 0 | NA | 0 | NA | 0 | NA | 0 | NA | 0 | NA |
| 48 | Villa Muñoz, Retiro | 1196 | 0.52 | 686 | 0.45 | 614 | 0.34 | 720 | 0.42 | 659 | 0.50 |
| 49 | La Comercial | 278 | 0.75 | 274 | 0.67 | 289 | 0.57 | 284 | 0.65 | 278 | 0.67 |
| 50 | Tres Cruces | 346 | 0.68 | 192 | 1.05 | 261 | 0.79 | 221 | 0.91 | 207 | 0.96 |
| 51 | Brazo Oriental | 1640 | 0.57 | 871 | 0.39 | 979 | 0.31 | 884 | 0.37 | 901 | 0.34 |
| 52 | Sayago | 1128 | 0.59 | 1006 | 0.33 | 1120 | 0.30 | 1017 | 0.32 | 1035 | 0.29 |
| 53 | Conciliación | 2094 | 0.57 | 2800 | 0.14 | 2973 | 0.13 | 2906 | 0.13 | 2846 | 0.13 |
| 54 | Belvedere | 1909 | 0.51 | 2141 | 0.18 | 2245 | 0.17 | 2101 | 0.17 | 2118 | 0.18 |
| 55 | Nuevo París | 10461 | 0.29 | 5623 | 0.11 | 5744 | 0.10 | 5685 | 0.11 | 5621 | 0.08 |
| 56 | Tres Ombúes, Victoria | 2936 | 0.42 | 4302 | 0.12 | 4690 | 0.11 | 4482 | 0.12 | 4365 | 0.11 |
| 57 | Paso de la Arena | 5494 | 0.46 | 3668 | 0.13 | 3856 | 0.11 | 3682 | 0.12 | 3656 | 0.13 |
| 58 | Colon Sureste, Abayuba | 1510 | 0.57 | 1430 | 0.24 | 1361 | 0.23 | 1381 | 0.23 | 1410 | 0.17 |
| 59 | Colón Centro y Noroeste | 11516 | 0.39 | 4492 | 0.12 | 4452 | 0.11 | 4507 | 0.11 | 4481 | 0.11 |
| 60 | Lezica, Melilla | 4274 | 0.41 | 1887 | 0.19 | 1912 | 0.15 | 1987 | 0.18 | 1868 | 0.21 |
| 61 | Villa Garcia, Manga Rur. | 5403 | 0.40 | 3719 | 0.13 | 3929 | 0.11 | 3907 | 0.13 | 3750 | 0.13 |
| 62 | Manga | 7627 | 0.33 | 3522 | 0.13 | 3398 | 0.12 | 3567 | 0.12 | 3485 | 0.13 |

# Referencias

Bedi, T., Coudouel, A., y Simler, K. (2007). More than a pretty picture: Using poverty maps to design better policies and interventions. World Bank Publications. https://doi.org/10.1596/978-0-8213-6931-9

Bivand, R. S., Pebesma, E. J., y Gomez-Rubio, V. (2013). Applied spatial data analysis with r (2a ed.). Springer. https://doi.org/10.1007/978-1-4614-7618-4

Chambers, R., R. Clark. (2012). An introduction to model-based survey sampling with applications. Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780198566625.001.0001

Cliff, A. D., y Ord, J. K. (1973). Spatial autocorrelation. Progress in Human Geography, 85(409). https://doi.org/10.1177/030913259501900205

Drew, D., Singh, M., y Choudhry, G. (1982). Evaluation of small area estimation techniques for the canadian labour force survey. , 8(1-2), 17–47. Descargado de `https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198200114328`

Fay III, E., R., y Herriot, A., R. (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. https://doi.org/10.2307/2286322

Ferreira, J. (2022). Metodología de la encuesta continua de hogares(ech). , Instituto Nacional de Estadística (INE), Uruguay. Descargado de `https://www.gub.uy/instituto-nacional-estadistica/comunicacion/publicaciones/metodologia-encuesta-continua-hogares-ech-2021`

Fuster, T., Glejberman, D., y Vernengo, A. (2006). Líneas de pobreza e indigencia 2006 uruguay - metodología y resultados. Descargado de `https://www.gub.uy/instituto-nacional-estadistica/comunicacion/publicaciones/lineas-pobreza-indigencia-2006-metodologia-resultados`

González-Manteiga, W., Lombarda, M., Molina, I., Morales, D., y Santamaría, L. (2010). Small area estimation under fay–herriot models with non-parametric estimation of heteroscedasticity. Statistical Modelling: An International Journal, 10(2). https://doi.org/10.1177/1471082X0801000206

Horvitz, D. G., y Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685. https://doi.org/doi.org/10.2307/2280784

Kackar, R. N., y Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the American

Statistical Association, 79(388), 853–862. https://doi.org/10.2307/2288715

Lahiri, P., y Pramanik, S. (2018). Evaluation of synthetic small-area estimators using design-based methods. Austrian Journal of Statistics, 48, 48–48. https://doi.org/10.17713/ajs.v48i4.790

MDSF-CEPAL. (2022). Estimaciones comunales de pobreza por ingresos en chile mediante métodos de estimación en Áreas pequeñas - informe metodológico. Santiago de Chile. Descargado de http://observatorio.ministeriodesarrollosocial.gob.cl/pobreza-comunal-2020

Molina, I. (2019). Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas. Comisión Económica para América Latina y el Caribe, (CEPAL). Descargado de https://www.cepal.org/es/publicaciones/44214-desagregacion-datos-encuestas-hogares-metodologias-estimacion-areas-pequenas

Molina, I., Salvati, N., y Pratesi, M. (2008). Bootstrap for estimating the mse of the spatial eblup. Computational Statistics, 24(3), 441-458. https://doi.org/10.1007/s00180-008-0138-4

Morales, D., Esteban, M. D., Pérez, A., y Hobza, T. (2021). A course on small area estimation and mixed models: Methods, theory and applications in r. Springer. https://doi.org/10.1007/978-3-030-63757-6

ONU. (2016). Transforming our world: The 2030 agenda for sustainable development. Descargado de https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981

Prasad, N. G. N., y Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, 85(409), 163—171. https://doi.org/10.2307/2289539

Rao, J. N. K., y Molina, I. (2014). Small area estimation. John Wiley Sons. https://doi.org/10.1002/9781118735855

Rao, J. N. K., y Wu, C. F. J. (1988). Resampling inference with complex survey data. Journal of the American Statistical Association, 83(401), 231–241. https://doi.org/10.2307/2288945

Rencher, B., A.C. Schaalje. (2008). Linear models in statistics. John Wiley Sons, Inc. https://doi.org/10.1002/9780470192610

Statistics Canada, L. S. D. (2013). Guide to the survey of employment, payroll and hours. Descargado de https://www150.statcan.gc.ca/n1/pub/72-203-g/72-203-g2021001-eng.htm

Särndal, C., Swensson, B., y Wretman, J. (1992). Model assisted survey sampling. Springer. https://doi.org/10.1007/978-1-4612-4378-6

Viljanen, M., Meijerink, L., Zwakhals, L., y Van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of netherlands. International Journal of Health Geographics, 21(4). https://doi.org/10.1186/s12942-022-00304-5

West, B. T., Welch, K. B., y Galecki, A. T. (2007). Linear mixed models: A practical guide using statistical software (3a ed.). Chapman Hall/CRC. https://doi.org/10.1201/9781003181064