Estimación de Área Pequeña

Equipo Investigación, CIT

2023-07-01

Table of contents

Prefacio

El presente documento presenta la documentación general de los procesos de estimación en áreas pequeñas o desagregación de información.

Rapido desarrollo en la actualidad

Papers clasicos

1 Introducción

Definicion SAE: conjunto de métodos usadaos para producir estimadores basados en encuestas para areas geográficas o dominios de estudio en los cuales los tamaños muestrales son demasiado pequeños, o incluso ausentes, y así entregar estimaciones válidas.

Para producir estos estimadores válidos, en general es necesario incluir bases de datos adicionales, mediante un proceso de modelado estadístico.

Dolores: necesidad de datos para diseño de politica publica

Censos se demoran, encuestas son caras

Tradicion de investigacion en el area, desde libros, hasta la producción de guías de trabajo que entregan directrices para que estos procedimientos puedan ser utilizados por agencias estadísticias nacionales e investigadores interesados en el tema.

A lo largo del libro se tratarán los siguientes temas

- marco de trabajo
- revisión teórica de modelos
- Ejemplos de aplicaciones
- Demostración

Producir informacion desagregada para distintos grupos de poblacion

Insumo para la definición de políticas públicas

Integración de datos: registros administrativos, web scrapping, datos

2 Resumen

Marco de trabajo

Técnicas de estimación

Software estadístico y paquetes

Utilidad y ejemplos

Desagregación de información

"Tomar fuerza"

3 Marco de Trabajo

El trabajo de Tzavidis et al. (2018) entrega un marco de trabajo basado en la interacción con usuarios y con la parsimonia como principio rector. Este luego se encuentra simplificado en la web de la ONU para la producción de estimadiones SAE SAE4SDG

Relación con Stake holders

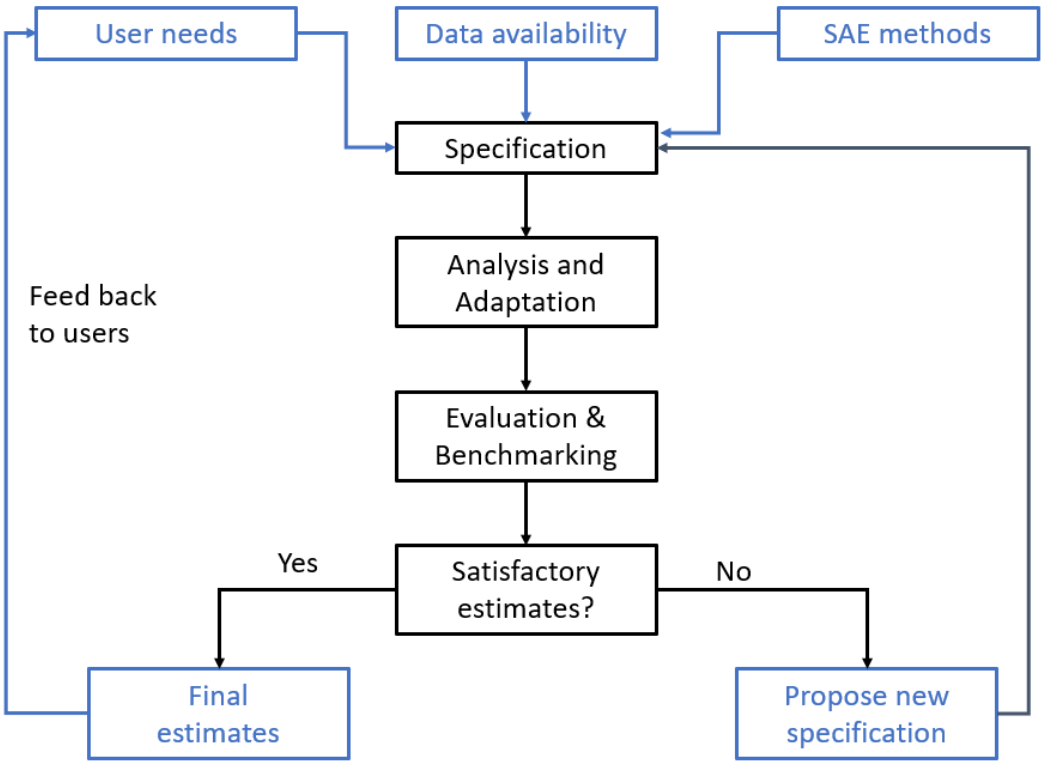

Figure 3.1: Flujo de Trabajo SAE

3.1 Especificación

Entrada:

- Necesidades de usuarios
- Disponibilidad de Datos
- Métodos SAE

Flujo de decision

Salida:

- Objetivos de estimación y geografías
- Elección inicial de métodos y software

3.2 Análisis y adaptación

- Estimaciones preliminares

3.3 Evaluación

- Responder la pregunta ¿son satisfactorios los resultados?

Si la respuesta es negativa se vuelve a la etapa de especificación

Si la respuesta es positiva se generan estimaciones finales y se calculan medidas de incertidumbre

4 Marco Teórico

El término area pequeña se relaciona no solo a un espacio geográfico sino tambien puede hacer referencia a un subgrupo específico de la población. El objetivo es entonces estimar las características de esta area o subgrupo, y el problema viene de que la información disponible para realizar una estimación directa entrega resultados muy variables o poco confiables. Esto último debido a un tamaño muestral pequeño (pocas observaciones).

Para solucionar esto y mejorar la precisión de las estimaciones directas de las encuestas, se usa información suplementaria relevante como por ejemplo datos de areas relacionadas y covariantes de otras fuentes. Suele usarse el término "prestar fuerza" en este contexto.

Respecto de los modelos usados destacan los modelos multinivel o lineales mixtos, destacando el trabajo de Sugasawa and Kubokawa (2020) por entregar una revisión y resumen de aplicaciones en contexto de estimaciones de areas pequeñas. Los resultados que entregan son basados en modelos y son llamados BLUP por ser los mejores predictores lineales no sesgados, según las siglas en inglés (*Best Linear Unbiased Predictors*). Siendo la ventaja de este enfoque el permitir acotar (disminuir la variabilidad de) los resultados en areas pequeñas hacia una cantidad estable construída mediante la combinación de datos.

Esto último deriva principalmente por la estructura de los modelos multinivel donde la observacion se explica tanto por parametros comunes, efectos aleatorios y errores residuales. Es así como el efecto de acotar viene por el modelamiento del efecto aleatorio, y la combinación de información se expresa en los parámetros comunes. Mayor detalle de la estimación de estos modelos y su implementación en R puede encontrarse en la web mediante una simple búsqueda en google, sin embargo destaca el siguiente tutorial por su parsimonia y profundidad.

A continuación se describen la aplicación de modelos básicos basados en modelos multinivel, el modelo de area de Fay-Herriot y el modelo de unidad de Errores Anidados, además de revisar la forma de cálculo de las medidas de incertidumbre, siguiendo el trabajo de Sugasawa and Kubokawa (2020).

4.1 Modelos de Área

Referencia Wiki

La mayoría de los datos públicos se reportan en datos agregados o promedios para ciudades o regiones. El modelo de Fay-Herriot (FH) es un modelo multinivel para estimar las medias reales

de area $\theta_1, ..., \theta_m$ basado en estadísticas promedios a nivel de area, denotadas por $y_1, ..., y_m$ donde y_i es un estimador directo de θ_i para $i = 1, .., m$. Notar que y_i es un estimador crudo de alta varianza. Debido a que el tamaño muestral para calcular y_i en la pratica es pequeño, se usa información adicional. Siendo x_i un vector de características conocidas con un término de intercepto, el modelo FH viene dado por:

$$y_i = \theta_i + \epsilon_i \quad \theta_i = x_i^t \beta + \nu_i, \quad i = 1, ..., m$$

Con β un vector de coeficientes de regresion, ϵ_i y ν_i son repecticamente errores de muestreo y efectos aleatorios, los cuales se distribuyen de forma independiente como $\epsilon \sim N(0, D_i)$ y $\nu_i \sim N(0, A)$. Donde D_i es la varianza de y_i dado θ_i, la cual se asume conocida, y A es un parametro de varianza desconocido. El supuesto de D_i conocido parece restrictivo pero puede estimarse con data a priori.

El mejor predictor de θ_i bajo perdida cuadratica es la expectativa condicional:

$$E[\theta_i | y_i] = \gamma_i y_i + (1 - \gamma_i) x_i^t \beta$$

Donde $gamma_i = A/(A + D_i)$ es conocido como un coeficiente de acotamiento. Es decir que genera un equilibrio basado en las varianzas respectivas de la estimación directa y la varianza de los datos auxiliares.

Siguiendo la formulación propuesta, β puede estimarse por minimos cuadrados generalizados (GLS):

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{m} \frac{x_i x_i^t}{A + D_i} \right)^{-1} \left(\sum_{i=1}^{m} \frac{x_i y_i}{A + D_i} \right)$$

Al reemplazar β con $\hat{\beta}$ se obtiene el mejor predictor lineal no sesgado (BLUP):

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) x_i^t \hat{\beta}_{GLS}$$

Ya que $\hat{\beta}_{GLS}$ se construye a partir de todos los datos, el estimador de regresion $x_i^t \hat{\beta}_{GLS}$ es mucho más estable que los estimadores directos y_i.

En la práctica, la varianza de efectos aleatorios A es desconocida y debe ser reemplazada en γ_i y $\hat{\beta}_{GLS}$ por un estimador basado en la muestra, lo que genera el mejor predictor lineal no sesgado empirico (EBLUP).

4.2 Modelos de Unidad

Referencia wiki

Cuando existen datos disponibles a nivel de unidad (por ejemplo a nivel de hogares) se puede usar un analisis más profundo. Sea $y_{i1}, ..., y_{yn_i}$ una muestra a nivel de unidad de la area i-esima para $i = 1, .., m$ y seav $x_{i1}, ..., x_{yn_i}$ los vectores fijos de covariantes con o sin el intercepto, el modelo de error anidado se describe como:

$$ y_{ij}=x_{ij}^t + i+ {ij}, \quad j=1,...n_i, \quad i=1,...,m,

$$

Donde ν_i y ϵ_{ij} son efectos aleatorios y el terminos de error y son independientes y distribuidos como $\nu_i \sim N(0, \tau^2)$ y $\epsilon_{ij} \sim N(0, \sigma^2)$, β es un vector de coeficientes de regresion desconocidos, y τ^2 y σ^2 son parametros de varianza desconocidos.

Se nota que ν_i es un efecto aleatorio que depende del area i-esima y es comun a las observaciones en la mismas area_. Esto induce correlaciones entre las observaciones y_{ij} las cuales se expresan como $Cov(y_{ij}, y_{ij'}) = \tau^2$ para $j \neq j'$, notando que las observaciones en diferentes areas son independientes. Por tanto estas se llaman varianzas *within* y *between* (dentro y entre).

Este modelo se usa tipicamente en un marco de trabajo de modelos de pobalcion finita. Asumiento que el area i contiene N_i unidades en total, pero solo n_i son muestreadas. Por simplicidad, se asume un mecanismo de muestreo aleatorio simple (por lo que no se consideran factores de expansion). Para todas las unidades se asume un modelo de población:

$$Y_{ij} = x_{ij}^t\beta + \nu_i + \epsilon_{ij}, \quad j = i, ..., N_i, \quad i = 1, ..., m$$

Donde Y_{ij} son las características de la unidad j en el area i. Sin pérdida de generalidad, se asumen que se observan las primeras n_i caracteristicas y el resto no son observadas. Bajo esta configuracion el promedio real de area se define como:

$$\frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij} = \frac{1}{N_i}\sum_{j=1}^{N_i}(x_{ij}^t\beta + \nu_i + \epsilon_{ij}) = \bar{X}_i^t\beta + \nu_i + \frac{1}{N_i}\sum_{j=1}^{N_i}\epsilon_{ij}$$

En la práctica, el numero total de unidades N_i es grande aunque el numero total de unidades muestradas n_i no es grande. Por tanto el ultimo termino puede ser muy pequeño, por tanto se puede definir el parametro de la media como $\theta_i = \bar{X}_i^t\beta + \nu_i$. Y estimarlo al conocer el vector de información auiliar $\bar{X}_i^t$, lo cual es comun e la práctica.

El mejor de predictr de ν_i viene dado por:

$$\tilde{\nu_i} = \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2}(\bar{y_i} - \bar{x_i}^t \beta)$$

Donde $\bar{y_i} = n_i{-}1 \sum_{j=i}^{n_i} y_{ij}$ y $\bar{x_i} = n_i{-}1 \sum_{j=i}^{n_i} x_{ij}$.

De forma similar al modelo FH, se puede estimar β por el estimador general de minimos cuadrados (GLS) basados en los datos muestrales. Los parámetros de varianza también se estiman de esta forma. Es así que el mejor predictor EBLUP viene dado por $\hat{\theta}_i = \bar{X}_i^t \hat{\beta}_{GLS} + \hat{\nu_i}$

4.3 Estimación de MSE e intervalos de Confianza

Una parte importante de la estimación de area pequeña es la evaluación de la confianza y precisión de los resultados. Para esto se usan los errores cuadraticos medios (MSE) y los intervalos de confianza.

4.3.1 Estimacion de MSE

Considerando una situación donde el parámetro es θ_i y $\tilde{\theta}_i$ representa la expectativa condicional de θ_i dado y_i el cual depende del parámetro desconocido Ψ. Siendo $\hat{\theta}_i$ el mejor predictor empirico de $theta_i$, se define el MSE de $\hat{\theta}_i$ como:

$$MSE_i(\Psi) = E[(\hat{\theta}_i - \theta_i)^2]$$

Usando el hecho de que $\tilde{\theta}_i$ es la expectativa condiciona de θ_i dado y_i tenemos que:

$$MSE_i(\Psi) = E[(\tilde{\theta}_i - \theta_i)^2] + E[(\hat{\theta}_i - \tilde{\theta}_i)^2] = g_{i1}(\Psi) + g_{i2}(\Psi)$$

Con $g_{i1}(\Psi)$ representando la variablidad del mejor predictor dado Ψ y tipicamente de orden $O(1)$, mientras que $g_{i2}(\Psi)$ mide la variabilidad adicional derivada de la estimación de Ψ, tipicamente de orde $O(m_{-1})$, en la mayoria de los casos, en general se deriva una formula de aproximacion hasta el segundo orden.

Para estimar estos elementos existen derivaciones analíticas, método de bootstrap y método de jackknife. Estos pueden profundizarse aparte. Sin embargo el método de boostrap cuenta entre sus variantes el método de bootrap hibrido, el cual separa los estimadodes de g_{i1} y g_{i2} mediante el bootrap paramétrico, cuyo enfoque a grandes razgos consiste en estimar Ψ mediante la simulación de muestras. Esto último implica generar realizaciones de los errores de muestreo y efectos aleatorios derivados de la estimación inicial.

4.3.2 Estimación de intervalos de confianza

Respecto de los intervalos de confianza, existen dos enfoques generales, el método analítico basado en expansion de series de Tayler y el metodo de bootstrap paramétrico.

Por simplicidad se asume que $\theta_i|y_i \sim N(\tilde{\theta}_i(y_i, \Psi), s_i(\Psi)^2)$, donde $\tilde{\theta}_i$ y s_i^2 son expectativas condicionales y varianza de θ_i.

El método de bootstrap oarpanetruci se basa en generar un estadístico pivote. Se define $U_i(\Psi) = (\theta_i - \tilde{\theta}_i)/s_i(\Psi)$, luego $U_i(\Psi) \sim N(0,1)$ con Ψ el parámetro verdadero. Se aproxima la distribución de $U_i(\hat{\Psi})$ mediante bootrsap parametrico, es deceir generar muestras bootrap desde le modelo estimado, y se computanel estimador de bootstrap $\hat{\Psi}^*_{(b)}$ para $b = 1, ..., B$. Luego la distribución de $U_i(\hat{\Psi})$ puede aproximarse por B realización bootrap.

Siendo $z^*_{iu}(\alpha)$ y $z^*_{il}(\alpha)$ los $100\alpha\%$ cuantiles empiricos de la distribucion simulada, el intervalo de confianza calibrado viene dado por:

$$(\hat{\theta}_i + z^*_{il}(\alpha))s_i(\Psi), \hat{\theta}_i + z^*_{iu}(\alpha))s_i(\Psi))$$

5 Aplicaciones

Estimaciones de parámetros no lineales de la población en subareas

Desagregación de información

5.1 Pobreza Comunal

5.2 Uso de imágenes satelitales

5.3 Desigualdad comunal en Chile

6 Demostración Consumo Energético

6.1 Objetivos del Análisis

6.2 Convalidar variables de fuentes de datos

6.3 Especificaciones

6.4 Comparación de Modelos

6.5 Resultados

7 Recursos

Esta sección detalla algunos recursos relevantes para profundizar en el tema de estimación de areas pequeñas.

7.1 Guias, Manuales y Seminarios

- SAE4SDG: Página en formato Wiki que incluye guías para el desarrollo de estimaciones de area pequeña desarrollada por el departamento de estadísticas de la ONU en el contexto de generar herramientas para el monitoreo de objetivos de desarrollo sostenible. Es un excelente punto de partida para tener una vision global del tema, además de incluir recursos, ejemplos, bases de datos y referencias a otros recursos de educación relevante.
- Seminario SAE Chile 2022 Cepal: Repositorio con presentaciones de aplicaciones destacadas de SAE en Chile.

7.2 Blogs y Presentaciones

- Encuentro SAE Chile cepal

7.3 Software estadístico

- emdi: "Estimating and Mapping Disagrgregated Indicators" Paquete de R que destaca por su flexibilidad y por ser usado como punto de partida para generar nuevas implementaciones.
- stata

...

7.4 videos

- Seminario CEPAL SAE 2023: Este seminario realizado el 2023 contiene un conjunto de presentaciones y referencias a desarrollos actuales metodológicos sobre la estimaciones de areas pequeñas, destacando la actualidad de estos además del contexto asociado a sudamérica y países en desarrollo.
- Paula Moraga

7.5 libros

- Molina y Rao
- Multilevel SAE

7.6 Papers

- Molina y Rao 2010 (Molina and Rao 2010)
- Sugasawa y Kobokawa 2021 (Sugasawa and Kubokawa 2020)
- Molina 2019 (Molina 2019)
- Newhouse et al 2022 (Newhouse et al. 2022)

References

Molina, Isabel. 2019. "Desagregación de Datos En Encuestas de Hogares: Metodologías de Estimación En Áreas Pequeñas."

Molina, Isabel, and J. N. K. Rao. 2010. "Small Area Estimation of Poverty Indicators." *Canadian Journal of Statistics* 38 (3): 369–85. https://doi.org/10.1002/cjs.10051.

Newhouse, David Locke, Joshua D. Merfeld, Anusha Ramakrishnan, Tom Swartz, and Partha Lahiri. 2022. "Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning." SSRN Scholarly Paper. Rochester, NY. October 3, 2022. https://doi.org/10.2139/ssrn.4235976.

Sugasawa, Shonosuke, and Tatsuya Kubokawa. 2020. "Small Area Estimation with Mixed Models: A Review." *Japanese Journal of Statistics and Data Science* 3 (2): 693–720. https://doi.org/10.1007/s42081-020-00076-x.

Tzavidis, Nikos, Li-Chun Zhang, Angela Luna, Timo Schmid, and Natalia Rojas-Perilla. 2018. "From Start to Finish: A Framework for the Production of Small Area Official Statistics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181 (4): 927–79. https://doi.org/10.1111/rssa.12364.