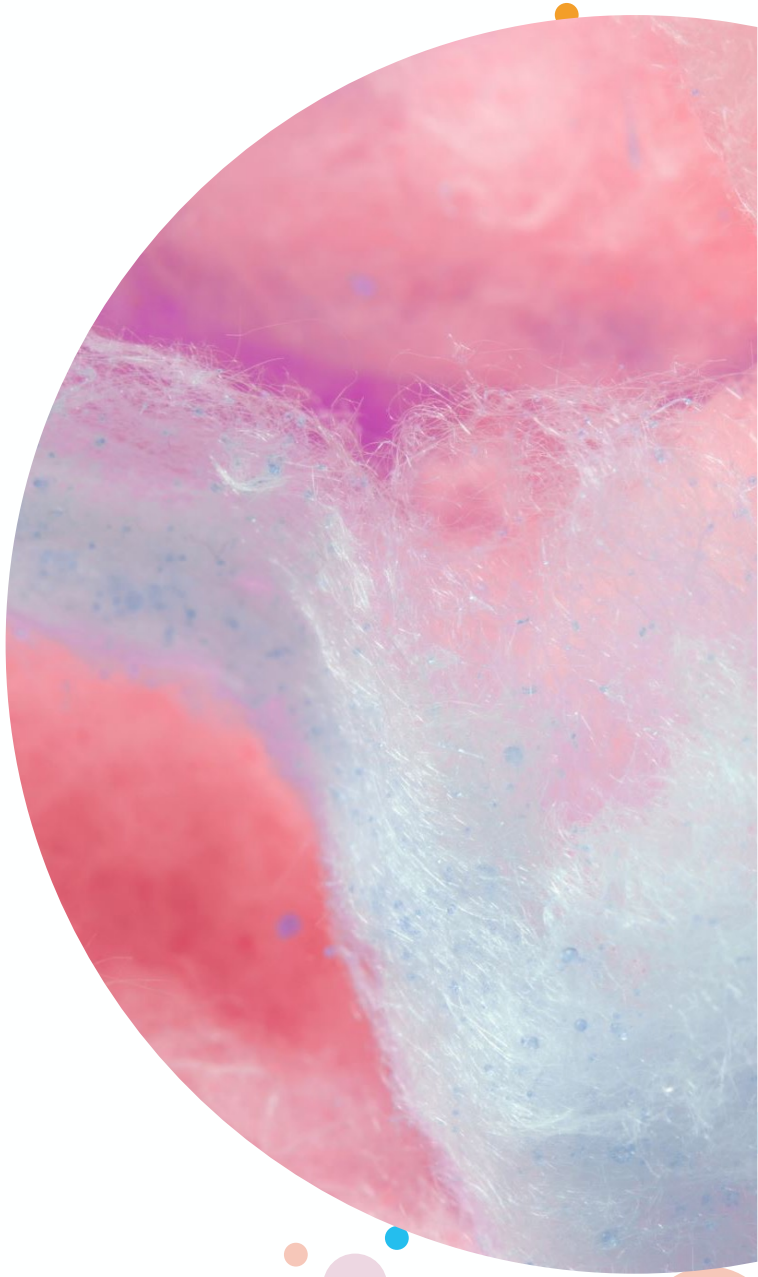
Escuela de Salud Pública
DR. SALVADOR ALLENDE
UNIVERSIDAD DE CHILE

Estimación en Áreas Pequeñas

Basado en el trabajo de Azizur Rahman: "Estimating small area health-related characteristics of populations: a methodological review"

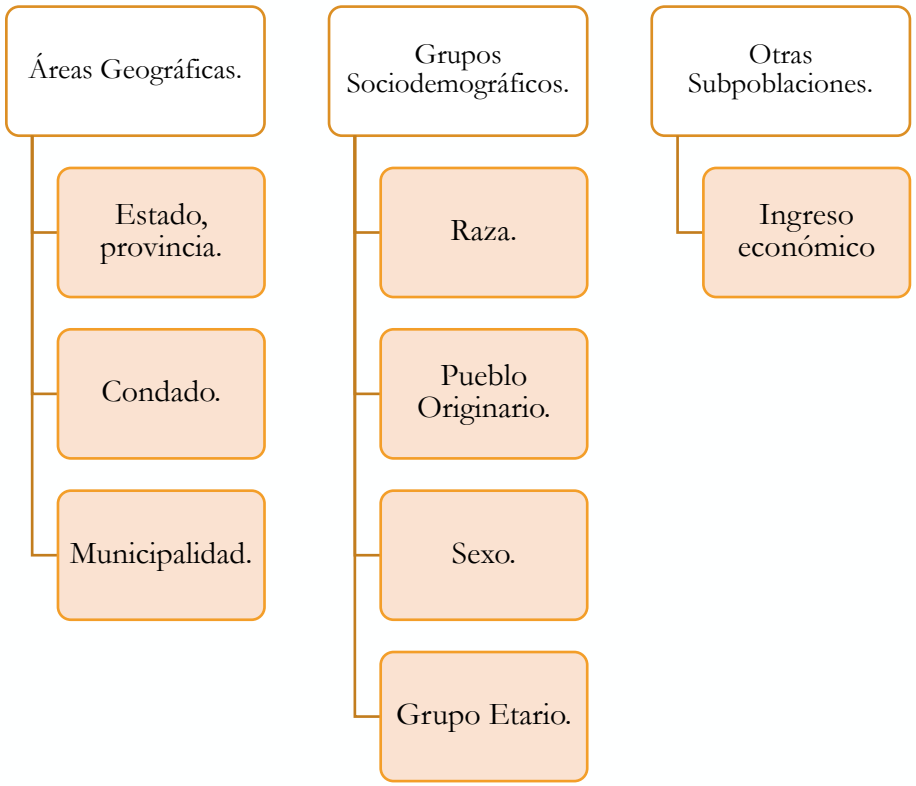
Inferencia Bayesiana 2022
Magister en Bioestadística

Alumna: Fernanda Vallejos L.



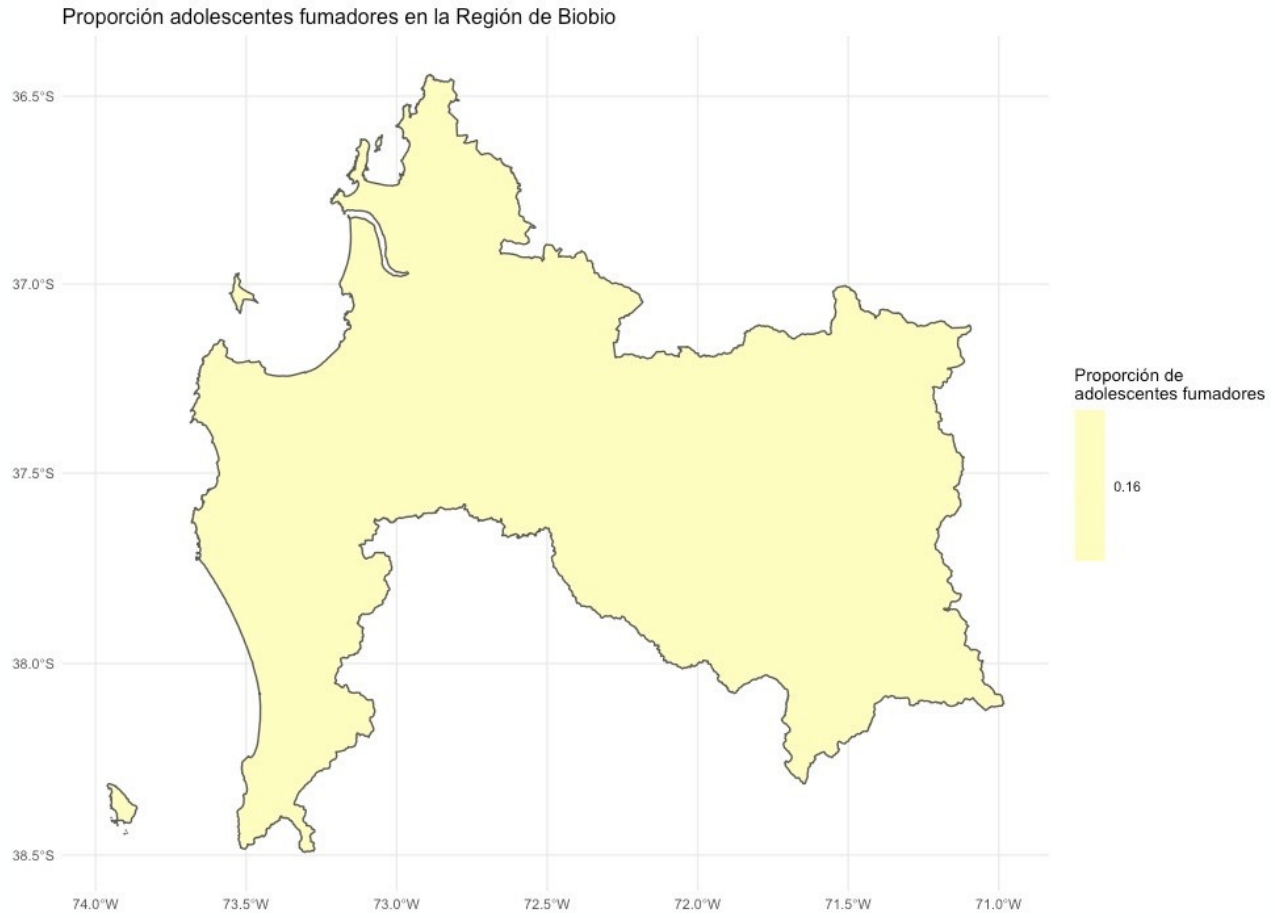
Introducción

- Llamamos “Estimación en área pequeña” a un conjunto de técnicas estadísticas utilizadas para realizar inferencia en un dominio para la cual los datos muestrales recogidos atienden a un diseño para el cual dichas áreas no recibieron consideración especifica.



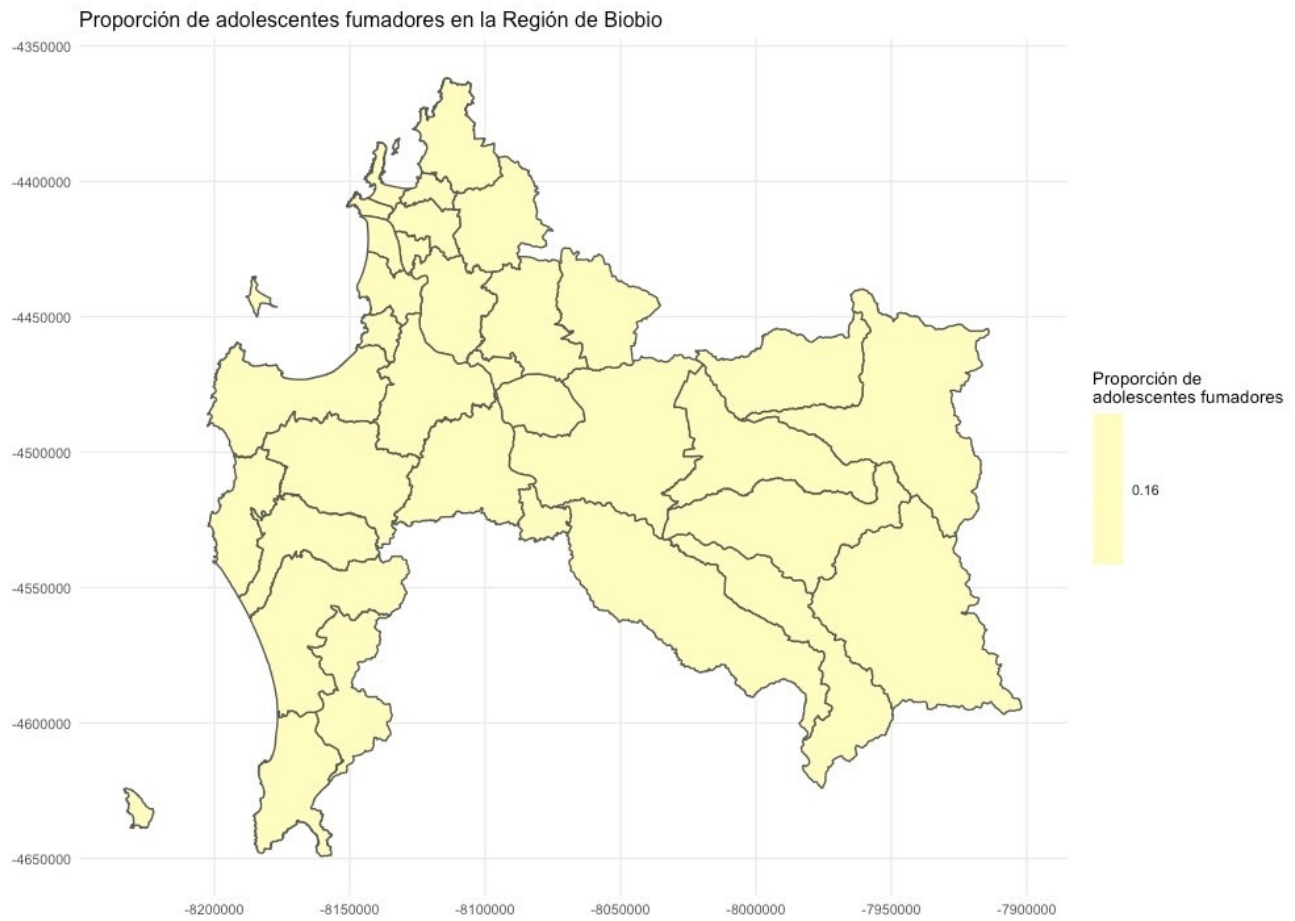
Un dominio o subpoblación es llamado “área pequeña” si su tamaño muestral para ese mismo dominio es pequeño.

Estimación en Región del Biobío

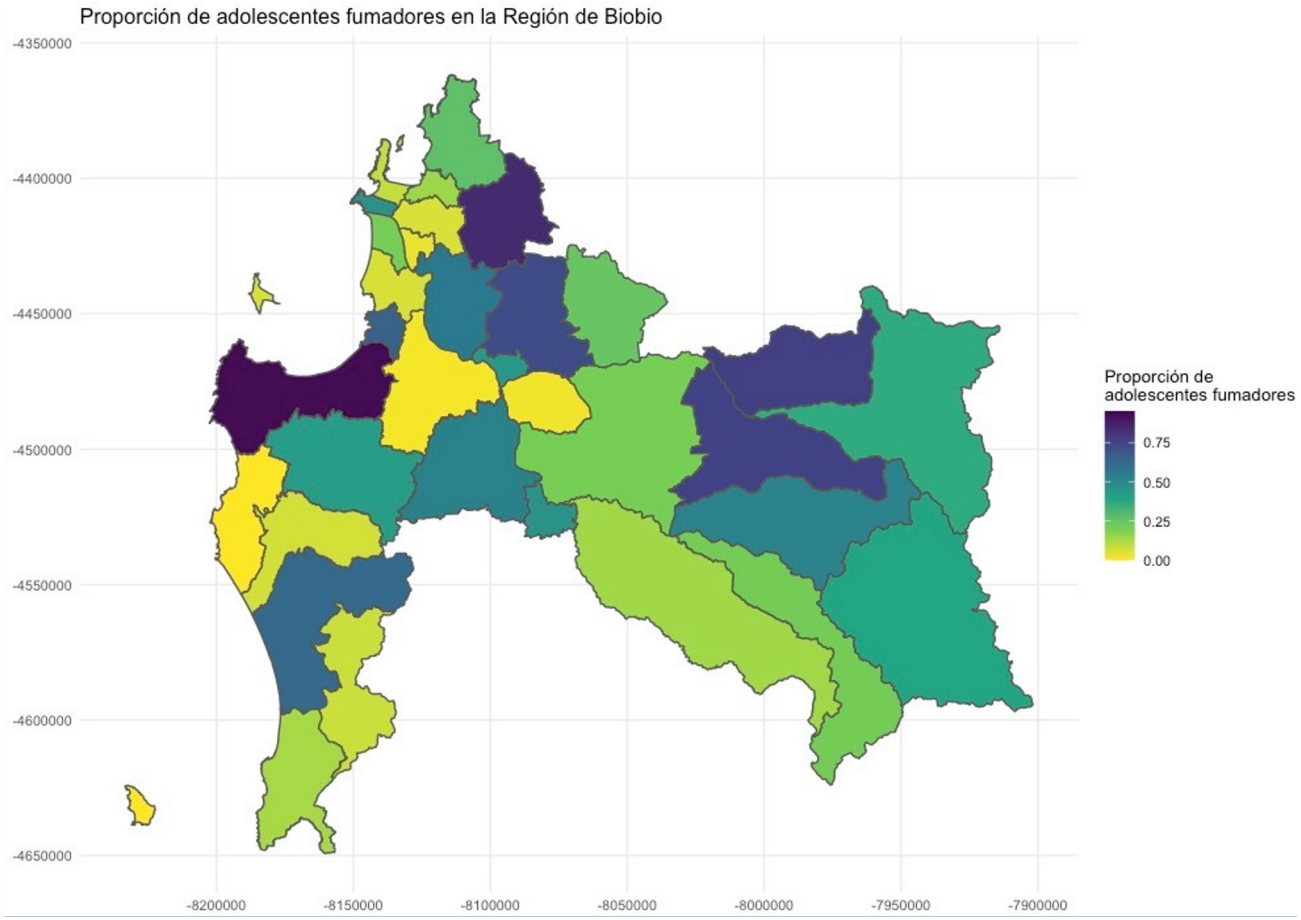


Encuesta Mundial del Tabaquismo en adolescentes (EMTA) Chile, 2008, Departamento de Epidemiología Ministerio de Salud.

Estimación en Región del Biobío

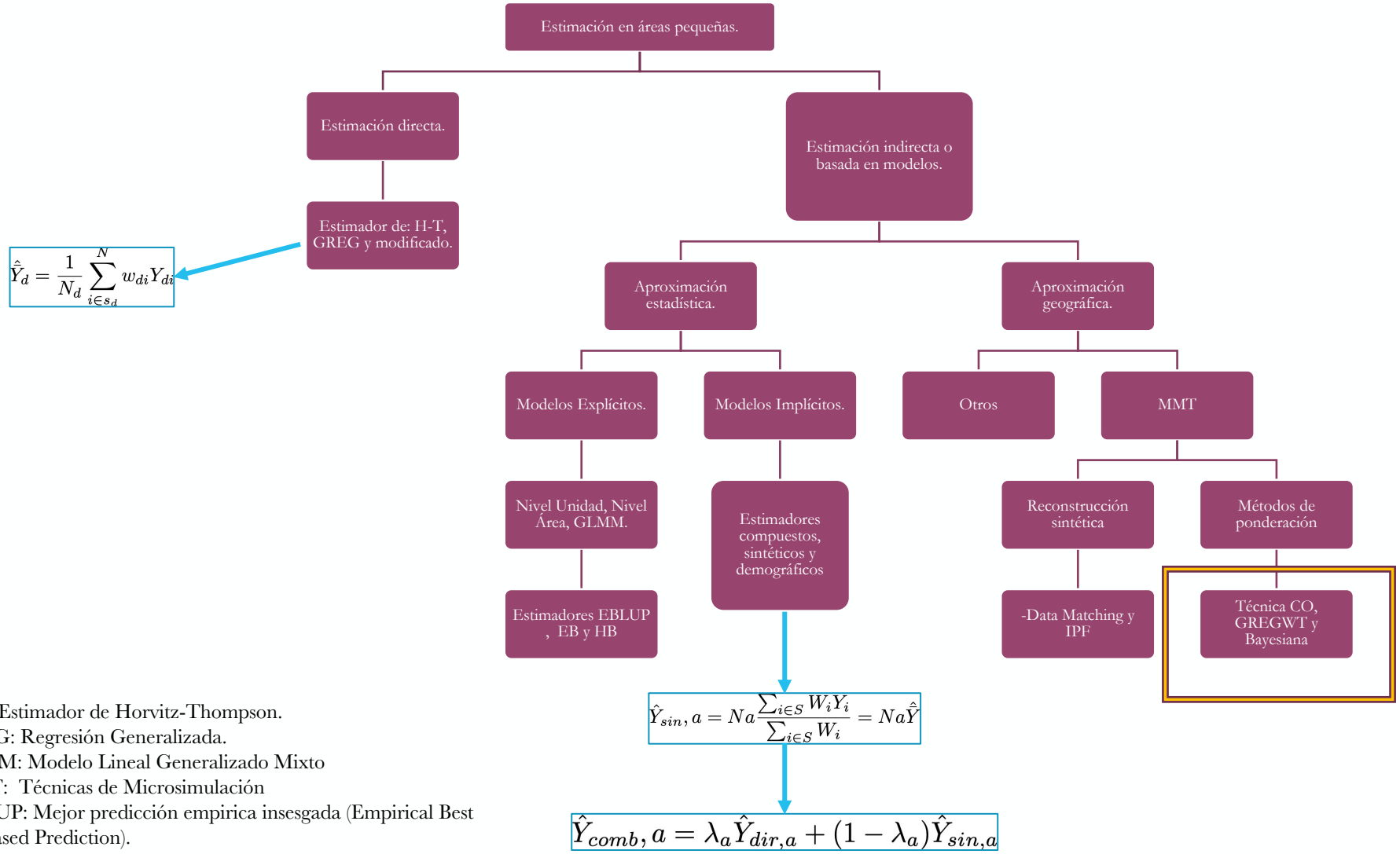


Estimación en Región del Biobío



Datos simulados

Resumen Técnicas SAE



- H-T: Estimador de Horvitz-Thompson.
- GREG: Regresión Generalizada.
- GLMM: Modelo Lineal Generalizado Mixto
- MMT: Técnicas de Microsimulación
- E-BLUP: Mejor predicción empirica insesgada (Empirical Best Unbiased Prediction).
- EB: Empirical Bayes.
- HB: Bayes Jerárquico (Hierarchical Bayes).
- CO: optimización combinatoria (combinatorial optimization)
- GREGWT: Regresión de ponderación generalizada.

Técnicas de Microsimulación

- ¿Ventajas sobre estimación directa?
- Estimación directa…
    - No siempre es posible.
    - Estimaciones no fidedignas a nivel de dominio.
    - Grandes errores estándar:
        - tamaño muestral insuficiente
        - mala especificación de los modelos.
        - inconsistencia asintótica del diseño.

- Estimación indirecta por microsimulación…
    - Popular y costo-efectivo los últimos 20 años.
    - Permite generar una micropoblación y realizar estimaciones en esta a nivel de unidad y dominio.

Insumos para estimación por Microsimulación.

- Archivo con microdatos (censo)
  - Información sobre características de cada individuo del dominio.

- Datos muestrales (encuesta)
  - Información sobre variables de interés,
  - Puede **no** contener indicadores de grupo apropiados
  - Puede contener pocas unidades muestrales en los mismos grupos, lo que puede afectar la habilidad e estimar diferentes efectos a nivel de dominio.

- Para los ajustes de modelar por microsimulación, se necesita **crear** un **conjunto de microdatos** de una **población sintética** usando información de otras fuentes (por ejemplo de datos agregados), de este modo se asegura que refleje tan completo como sea posible la población que esta siendo modelada.

Modelos de Microsimulación

- Son representados los individuos de la población
- Permite poder anticipar tendencias en la poblaciónde modo que resulta posible predecir consecuencias a corto y largo plazo en la implementación de medidas de salud, económicas, políticas, etc.

En el caso de la generación de datos espaciales (microsimulación espacial) , las metodologías están clasificadas principalmente en 2:

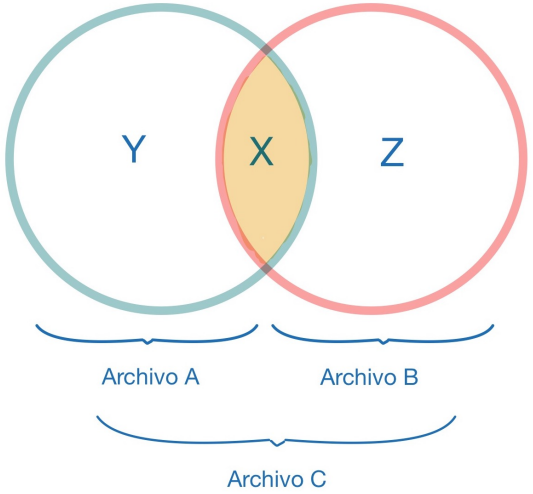
- Reconstrucción sintética.
- Métodos de reponderación.

La reconstrucción sintetica intenta construir micropoblacones sintéticas a nivel de área pequeña de la forma que todas las restricciones conocidas a nivel de área sin reproducidas. Se puede realizar a través de coincidencia de datos o fusión de datos, y a través de ajuste proporcional iterativo (IPF) como parte de un muestreo a través de MCMC.

Los métodos de repoderación son mas nuveos y consisten en calibrar los pesos de diseño de muestreo a un conjunto de nuevos pesos basados en una medida de distancia, esta aproximación incluye un proceso de optimización combinatoria (CO) y la regresión generalizada (en este contexto llamada GREGWT).

Reconstrucción sintética

- Atributos de cada unidad son estimadas por muestreo aleatorio usando un marco de trabajo probabilistico condicional.
- Basada en un proceso secuencial, paso a paso.
- El orden en que se generan los datos es importante. (el orden es fijo)
- Método complejo y consume mucho tiempo.
- Los efectos de inconsistencia entre los datos de las tablas de restricción podría ser significativo para este abordaje.



Data Matching.
Variables compartidas y codificadas del mismo modo permiten unir set de datos y crear un nuevo conjunto que permitirá reconstruir lo sucedido a nivel de dominios.

$$p_{ij(k+1)} = \frac{p_{ij(k)}}{\sum_j p_{ij(k)}} \times Q_i$$

$$p_{ij(k+2)} = \frac{p_{ij(k+1)}}{\sum_i p_{ij(k+1)}} \times Q_j$$

IPF (iterative proportional fitting).
A través de tablas de contingencia basadas en individuos de datos muestrales se identifican la esperanza de los totales marginales.
Este método ha sido usado como parte del muestreo MCMC para crear microdatos desde una variedad de fuentes de datos agregados.

$p_{ij(k)}$ es el elemento en la matriz de la fila i, columna j, iteración k
Q_i y Q_j son las sumas predefinidas de las filas y las columnas

- A pesar de la existencia de una gran cantidad de técnicas para generar microdatos espaciales, ninguno de estos métodos puede considerar un escenario de una completa micropoblacion a nivel de dominio.
- Como resultado a esto, las estimaciones derivadas de esos métodos, llevan a estimaciones imprecisas para muchos dominios.
- Además, la validación de esos outputs generados de un modelo de microsimulación basado en microdatos sintéticos es también difícil.
- Actualmente no existe un método robusto matemático o estadístico que nos lleve instantáneamente a sortear este tipo de problemas.

Otras Metodologías de Simulación

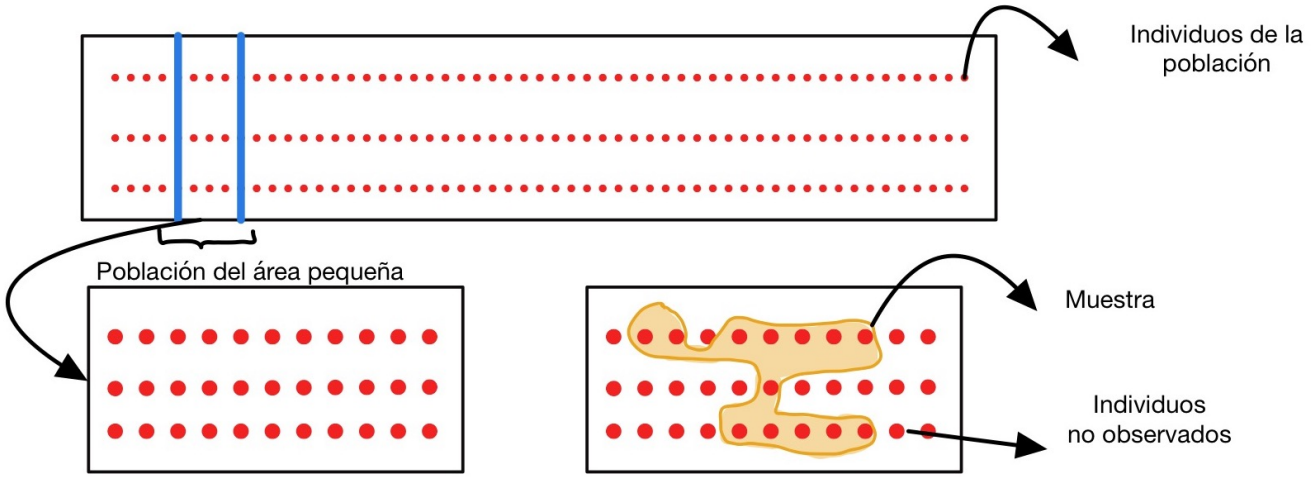
El objetivo de estos métodos consiste en generar set de datos con micropoblaciones al nivel de pequeñas áreas para las cuales no está medida la variable de interés.
Tienen la dificultad de la validación de los microdatos simulados.
Los nuevos modelos de predicción Bayesiana logran superar esta dificultad simulando un escenario completo de micropoblaciones y luego produciendo estadísticas confiables. Otras metodologías incluyen:

- GREGWT: Es un algoritmo de regresión generalizada iterativo escrito en SAS para calibrar las estimaciones de las encuestas a marcos de referencia. Por lo general tiene como foco simular micro-datos a nivel de pequeñas áreas y la agregación es posible en dominios mayores, utiliza la iteración de Newton-Raphson basado en una función de distancia restringida (función χ^2 truncada). A veces tiene problema de convergencia.

- CO: Ofrece flexibilidad y coherencia colectiva de microdatos por lo cual el análisis será posible a cualquier nivel, utiliza un abordaje de iteración basado en una combinación de hogares. No tiene problemas de convergencia.

Predicción bayesiana basada en simulación de microdatos

- A diferencia de los otros métodos de microsimulación de datos, este toma en consideración escenarios completos de unidades de micropoblacion a nivel de dominio. <- se utilza para cálculos a nivel de unidad.
- El mayor desafío del método es poder vincular los datos observados de los no observados, recordar:



Para una variable de interés y_{ij} en el i-esimo dominio siempre se tendrá $t_{yi} = \sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} y_{ij}$

Pasos básicos para simulación de microdatos

1. Obtener una distribución a priori de evento en estudio Y_i en el dominio i-esimo, esto sería $p(E_i)$ para $\forall_i$.

2. Encontrar la distribución condicional de las unidades muestrales no observadas, dado las unidades observadas esto sería $p(y_{ij} : j \in \bar{s}_i | y_i j : j \in s_i$ para $\forall_i$.

3. Derivar la distribución posterior usando el teorema de Bayes $p(\theta|s, X)$; $E_i \subseteq \theta$ donde θ es el vector de parámetros del modelo y X es el vector de información auxiliar.

4. Obtener copias de la población completa desde esta distribución posterior utilizando la técnica de simulación por MCMC.

Teoría de predicción Bayesiana

Basada en la distribución posterior de los parámetros desconocidos, Sea y un conjunto de unidades muestrales observadas de un modelo con una densidad de probabilidad conjunta $p(y|\theta)$ en la cual θ es un es $g(\theta)$ la distribución posterior de θ para y

$$p(\theta|y) \propto p(y|\theta)g(\theta).$$

Ahora si $\bar{y}$ es el conjunto de unidades no observadas de una población finita, entonces a través de la metodología bayesiana la distribución de predicción puede ser obtenida resolviendo la integral:

$$p(\bar{y}|y) \propto \int_{\theta} p(\theta|y)p(\bar{y}|\theta)d\theta,$$

en donde $p(\bar{y}|\theta)$ es la pdf de las unidades no observadas en la población finita

- Es decir… A través de un modelo de enlace, se logra conectar

$$Y_i = X_i\beta + E_i$$

$$\bar{Y}_i = \bar{X}_i\beta + \bar{E}_i$$

- Luego de una larga derivación matemática [1] , la distribución posterior conjunta de los parámetros para las unidades observadas y no observadas puede determinarse como:

$$f(\beta, \Sigma | Y_i, \bar{Y}_i) \cong |\Sigma|^{-\frac{N_i+p+1}{2}} |I_p + \Sigma^{-1} Q|^{-\frac{\nu+p+N_i-1}{2}} f\left(\bar{Y}_i | Y_i\right)$$

Donde $Q = (Y - X_i\beta)'(Y_i - X_i\beta) + (\bar{Y}_i - \bar{X}_i\beta)'(\bar{Y}_i - \bar{X}_i\beta)$

Ahora aplicando el método de simulación MCMC a la ecuación, se pueden obtener copias de los microdatos de la población para el dominio i-ésimo.

[1] Rahman, A., & Harding, A. (2016). Small area estimation and microsimulation modeling. Cap.5, pags:104-116, Chapman and Hall/CRC.

- A diferencia de los otros métodos de CO y de GREGWT este es un método probabilistico y no determinístico, pero de todos modos puede tomar el algoritmo utilizado para de GREGWT para enlazar las unidades observadas en la muestra con las no observadas.
- En contraste al punto de vista utilizado para la CO, este método utiliza la simulación por MCMC con un algoritmo iterativo basado en la distribución posterior.
- Como las probabilidades posteriores conjuntas de los parámetros para la unidades observadas y no observadas son estimadas a través de MCMC, la metodología de simulación propuesta esta enlazada con un muestreo de Cadenas de Markov. Y es muy distinta a la técnica de imputación múltiple.

El proceso básico de este método esta basado en una distribución de predicción de las unidades no observadas, dado las unidades observadas.

Conclusiones

- Existe una gran cantidad de técnicas mejorar la estimación en áreas pequeñas cuando la estimación directa no es posible, acá se presentan algunas pero existen muchas más.
- La forma de abordar el problema debe considerar la calidad de la información auxiliar disponible a nivel agregado y desagregado. Si su calidad no es buena, las estimaciones tampoco lo serán.
- Los métodos de microsimulación son los que entregan estimaciones con mayor precisión, sin embargo la validación de los microdatos simulados sigue siento una barrera para su mayor uso.
- Dentro de los métodos de microsimulación el enfoque Bayesiano es el que permite realizar predicciones tanto puntuales como por intervalos, dando así un grado mayor de confiabildad en la predicción.

Bibliografía

- Rahman, A. (2017). Estimating small area health-related characteristics of populations: a methodological review. *Geospatial Health*, *12*(1).
- Rahman, A., & Harding, A. (2016). *Small area estimation and microsimulation modeling*. Chapman and Hall/CRC.
- Rao, J. N., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons.
- Nieto Barajas,Luis (2003). *Enfoque bayesiano en la estimación de área pequeña*. Revista Nacional de Estadística y Geografía.

Propiedad	Estandarización indirecta y modelos a nivel de unidad	Modelos Multinivel	Modelos vía micro simulación
Comentario	Modelos basados en covariables a nivel individual.	Modelos basados en variables multinivel.	Modelos basados en creación sintética de micropoblaciones.
Ventajas	-Fácil. -Insesgado para muestras grandes.	-Fácil de aplicar y modelo mas explicativos pueden proveer CI de las estimaciones. -Flexible para permitir efectos en cualquier nivel	Mas sofisticado. Estado del arte. -Puede generar mediciones de confiabilidad estad´ıstica. -Abordaje robusto en términos de agregación o desagregación en distintas escalas. -Posible utilizar microdatos sintéticos a nivel de área pequeña para futuros análisis y actualización de las estimaciones.
Limitaciones	-Considera que las prevalencias a nivel agregado aplican uniformemente en nivel desagregado. -La elección de las covariables del modelo es restringida por el requerimiento de tener esta información en todos los dominios. -Muchas veces las estimaciones producidas por este método no son confiables debido a inconsistencia de las variables auxiliares.	-Estos métodos imponen la restricción de que es exigente en términos exista una congruencia entre los datos entregados por censo y las covariables usadas en el modelo. -Predictores importantes pueden ser eliminados del modelo simplemente por no conocer su distribución a nivel de área pequeña. -Estimar SE para estimadores que utilizan covariables a nivel individual y de área resulta mas complejo que la técnica de modelamiento a nivel individual	Intensivo computacionalmente. -Depende solamente de una buena generación de la micropoblación. -Existen muchas formas de realizarlo, GREGWT esta en desarrollo -Cuando existen pocas observaciones en un estrato muestral, los SE de los estimadores se vuelven poco confiables -Validaciones para algunas estimaciones SAE son desafiantes.
Aplicaciones	-Mayormente usado para tamaños muestrales grandes.	Ampliamente utilizado para datos multivariantes y con medidas repetidas, sets con observaciones perdidas o con conglomerados	Comienza a volverse una metodología popular en países desarrollados y frecuentemente utilizados para análisis de políticas sociales a nivel de área pequeña