

Estimación en áreas pequeñas ante la falta de datos censales actualizados

Alejandra Arias-Salazar

Escuela de Estadística, Universidad de Costa Rica

06.09.2023

- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 Proyectos en desarrollo

Las Oficinas Nacionales de Estadística producen:

→ Estadísticas oficiales

e.g., indicadores socio-económicos

→ Microdatos

e.g., datos de encuestas

Las Oficinas Nacionales de Estadística producen:

→ Estadísticas oficiales
e.g., indicadores socio-económicos

→ Microdatos
e.g., datos de encuestas



Características de calidad:

- Confiables
- Desagregados
- Actualizados

Las Oficinas Nacionales de Estadística producen:

→ Estadísticas oficiales
e.g., indicadores socio-económicos



Características de calidad:

- Confiables
- Desagregados
- Actualizados

→ Microdatos
e.g., datos de encuestas



Balance entre:

- Utilidad
- Privacidad



Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo.





Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo.

1 NO
POVERTY



(a) Monetaria



(b) Multidimensional



Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo.

1 NO POVERTY



(a) Monetaria



(b) Multidimensional



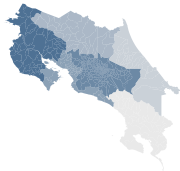
Ejemplo: Desagregación administrativa de Costa Rica, 2011



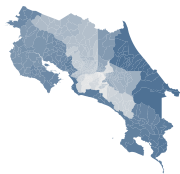
(a) Nacional



(b) Dos zonas



(c) Seis regiones



(d) 81 Cantones



(e) 473 Distritos

Ejemplo: Desagregación administrativa de Costa Rica, 2011



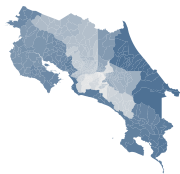
(a) Nacional



(b) Dos zonas



(c) Seis regiones



(d) 81 Cantones



(e) 473 Distritos

} Pobreza monetaria
Nivel de publicación

- ✓ Confiable
- ✗ Desagregado
- ✓ Actualizado

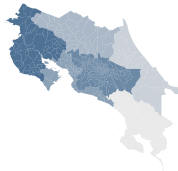
Ejemplo: Desagregación administrativa de Costa Rica, 2011



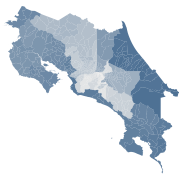
(a) Nacional



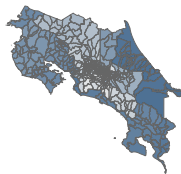
(b) Dos zonas



(c) Seis regiones



(d) 81 Cantones



(e) 473 Distritos

} Pobreza monetaria
Nivel de publicación

- ✓ Confiable
- ✗ Desagregado
- ✓ Actualizado



Estimación en áreas
pequeñas (SAE)

Fuente	Y	X	Dominios de interés
Encuesta	✓	✓	✗
Censo	✗	✓	✓

$$\bar{Y}_a = N_a^{-1} \left(\sum_{i \in s_a} Y_{ai} + \sum_{i \in r_a} Y_{ai} \right)$$

$a = 1, \dots, A$ dominios

$i = 1, \dots, N_a$ individuos

s = elementos muestreados

r = elementos no muestreados

Fuente	Y	X	Dominios de interés
Encuesta	✓	✓	✗
Censo	✗	✓	✓

$$\bar{Y}_a = N_a^{-1} \left(\sum_{i \in S_a} Y_{ai} + \sum_{i \in r_a} Y_{ai} \right)$$

$a = 1, \dots, A$ dominios
 $i = 1, \dots, N_a$ individuos
 $s =$ elementos muestreados
 $r =$ elementos no muestreados

Especificación

Ajuste

Predicción

$$y_{ai} = \mathbf{x}_{ai}^T \boldsymbol{\beta} + u_a + e_{ai}$$

Donde,

y_{ai} variable de interés
 \mathbf{x}_{ai} conjunto de predictores

u_a efectos aleatorios del área (dominio)

e_{ai} error a nivel de individuo

$u_a \stackrel{iid}{\sim} N(0, \sigma_u^2)$, $e_{ai} \stackrel{iid}{\sim} N(0, \sigma_e^2)$.

Utilizamos datos de encuesta

Obtenemos los estimadores de los parámetros del modelo:

$$\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$$

Utilizamos datos de censo

Obtenemos estimaciones para todos los dominios

Métodos SAE

Directos

Indirectos

Básicos

Basados en modelos

- Área (Fay-Herriot)
- Unidad (EBP)
- GLMM

Ejemplo: Pobreza monetaria para dominios pequeños



(d) 81 Cantones



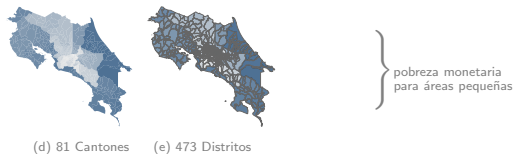
(e) 473 Distritos

} pobreza monetaria
para áreas pequeñas

Small area estimation (SAE)

Censo + Encuesta

Ejemplo: Pobreza monetaria para dominios pequeños



Small area estimation (SAE)

Censo + Encuesta
2011 2017

- ✓ Confiable
- ✓ Desagregada
- ✗ Actualizada

Ejemplo: Pobreza monetaria para dominios pequeños



(d) 81 Cantones



(e) 473 Distritos

} pobreza monetaria
para áreas pequeñas

Small area estimation (SAE)

Censo + Encuesta
2011 2017

Structure PREserving Estimation (SPREE)

- ✓ Confiable
- ✓ Desagregada
- ✓ **Actualizada**



- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 Proyectos en desarrollo

Structure PREserving Estimation (SPREE)

Census Z_{a_j, t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a, t_1} and \hat{Y}_{j, t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{a_j, t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

- Técnica para actualizar conteos poblacionales

Structure PREserving Estimation (SPREE)

Census Z_{aj,t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{aj,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

- Técnica para actualizar conteos poblacionales
- Proceso de ajuste a través del algoritmo “ajuste proporcional iterativo”- Iterative proportional fitting (IPF) (Deming and Stephan (1940)):

$$\hat{Y}_{aj,t_1}^S = \text{IPF}(Z_{aj,t_0}, \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1}).$$

Structure PREserving Estimation (SPREE)

Census Z_{aj,t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{aj,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

- Técnica para actualizar conteos poblacionales
- Proceso de ajuste a través del algoritmo “ajuste proporcional iterativo”- Iterative proportional fitting (IPF) (Deming and Stephan (1940)):

$$\hat{Y}_{aj,t_1}^S = \text{IPF}(Z_{aj,t_0}, \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1}).$$

La variable aleatoria Y_{aj,t_1} organizada en una tabla de contingencia puede representarse como:

$$\log Y_{aj,t_1} = \alpha_{0,t_1}^Y + \alpha_{a,t_1}^Y + \alpha_{j,t_1}^Y + \alpha_{aj,t_1}^Y.$$

La estructura de *asignación* es brindada por los márgenes y la estructura de *asociación* representa las interacciones .

Structure PREserving Estimation (SPREE)

Census Z_{aj,t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{aj,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

- Técnica para actualizar conteos poblacionales
- Proceso de ajuste a través del algoritmo “ajuste proporcional iterativo”- Iterative proportional fitting (IPF) (Deming and Stephan (1940)):

$$\hat{Y}_{aj,t_1}^S = \text{IPF}(Z_{aj,t_0}, \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1}).$$

La variable aleatoria Y_{aj,t_1} organizada en una tabla de contingencia puede representarse como:

$$\log Y_{aj,t_1} = \alpha_{0,t_1}^Y + \alpha_{a,t_1}^Y + \alpha_{j,t_1}^Y + \alpha_{aj,t_1}^Y.$$

La estructura de *asignación* es brindada por los márgenes y la estructura de *asociación* representa las interacciones .

Structure PREserving Estimation (SPREE)

Census Z_{aj,t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{aj,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

- Técnica para actualizar conteos poblacionales
- Proceso de ajuste a través del algoritmo “ajuste proporcional iterativo”- Iterative proportional fitting (IPF) (Deming and Stephan (1940)):

$$\hat{Y}_{aj,t_1}^S = \text{IPF}(Z_{aj,t_0}, \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1}).$$

La variable aleatoria Y_{aj,t_1} organizada en una tabla de contingencia puede representarse como:

$$\log Y_{aj,t_1} = \alpha_{0,t_1}^Y + \alpha_{a,t_1}^Y + \alpha_{j,t_1}^Y + \alpha_{aj,t_1}^Y.$$

La estructura de *asignación* es brindada por los márgenes y la estructura de *asociación* representa las interacciones .

Método	Supuesto	Fuentes
SPREE (Purcell and Kish, 1980)	$\alpha_{aj, t_1}^Y = \alpha_{aj, t_0}^Z$	Datos de censo + Totales de encuesta

Método	Supuesto	Fuentes
SPREE (Purcell and Kish, 1980)	$\alpha_{aj,t_1}^Y = \alpha_{aj,t_0}^Z$	Datos de censo + Totales de encuesta
Generalized SPREE (Zhang and Chambers, 2004)	$\alpha_{aj,t_1}^Y = \beta \alpha_{aj,t_0}^Z$	Datos de censo + Totales de encuesta + Datos de encuesta

Método	Supuesto	Fuentes
SPREE (Purcell and Kish, 1980)	$\alpha_{aj,t_1}^Y = \alpha_{aj,t_0}^Z$	Datos de censo + Totales de encuesta
Generalized SPREE (Zhang and Chambers, 2004)	$\alpha_{aj,t_1}^Y = \beta \alpha_{aj,t_0}^Z$	Datos de censo + Totales de encuesta + Datos de encuesta
Multivariate SPREE (Luna-Hernandez, 2016)	$\begin{bmatrix} \alpha_{a1,t_1}^Y \\ \vdots \\ \alpha_{aJ,t_1}^Y \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1J} \\ \vdots & \ddots & \vdots \\ \beta_{J1} & \dots & \beta_{JJ} \end{bmatrix} \begin{bmatrix} \alpha_{a1,t_0}^Z \\ \vdots \\ \alpha_{aJ,t_0}^Z \end{bmatrix}$	Datos de censo + Totales de encuesta + Datos de encuesta

Requerimientos para aplicar SPREE

- Datos organizados en tablas de contingencia.
- Los márgenes (totales) del censo y de la encuesta deben ser del mismo tamaño.
- Los totales por filas y por columnas deben ser iguales.

Census Z_{a_j,t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{a_j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Requerimientos para aplicar SPREE

- Datos organizados en tablas de contingencia.
- Los márgenes (totales) del censo y de la encuesta deben ser del mismo tamaño.
- Los totales por filas y por columnas deben ser iguales.

Census Z_{aj,t_0}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

Target Y_{aj,t_1}

A/J	j_1	j_2	j_3	Total
a_1				
a_2				
a_3				
Total				

- Estructura de asignación (Totales de encuesta) son confiables. ?

Requerimientos para aplicar SPREE

- Datos organizados en tablas de contingencia.
- Los márgenes (totales) del censo y de la encuesta deben ser del mismo tamaño.
- Los totales por filas y por columnas deben ser iguales.

Census Z_{aj,t_0}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

Target Y_{aj,t_1}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

- Estructura de asignación (Totales de encuesta) son confiables. ?
- Estructura de asociación (Censo) contiene el indicador de interés. ?

- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 Proyectos en desarrollo

1. ¿Cómo cumplir con el requisito de que la estructura de asignación (totales por área) es confiable?

1. ¿Cómo cumplir con el requisito de que la estructura de asignación (totales por área) es confiable?
2. ¿Cómo hacer uso de los métodos SPREE si el censo no contiene la variable de interés?

- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 Proyectos en desarrollo

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

$$\hat{Y}_{aj,t_1}^S = \text{IPF}(Z_{aj,t_0}, \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1})$$

Métodos tradicionales para tener tamaños poblacionales en áreas pequeñas:

- Datos de encuesta
- Proyecciones poblacionales

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

$$\hat{Y}_{aj,t_1}^S = \text{IPF}(Z_{aj,t_0}, \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1})$$

Métodos tradicionales para tener tamaños poblacionales en áreas pequeñas:

- Datos de encuesta
- Proyecciones poblacionales
- **Enfoque “fijo”**

- **Enfoque fijo:** proporciones poblacionales p_{a,t_0} son utilizadas para distribuir conteos de áreas “grandes” K en áreas pequeñas A

$$\hat{Y}_{a,t_1} = \hat{Y}_{k,t_1} \cdot p_{a,t_0} \quad \text{with} \quad \sum_{a=1}^A p_{a,t_0} = \sum_{a=1}^A \frac{Y_{a,t_0}}{Y_{k,t_0}} = 1 \quad \forall \quad a \subseteq k$$

- **Enfoque fijo:** proporciones poblacionales p_{a,t_0} son utilizadas para distribuir conteos de áreas “grandes” K en áreas pequeñas A

$$\hat{Y}_{a,t_1} = \hat{Y}_{k,t_1} \cdot p_{a,t_0} \quad \text{with} \quad \sum_{a=1}^A p_{a,t_0} = \sum_{a=1}^A \frac{Y_{a,t_0}}{Y_{k,t_0}} = 1 \quad \forall \quad a \subseteq k$$

- **Enfoque dinámico:** la información auxiliar actualizada se puede utilizar para tomar en cuenta cambios poblacionales en áreas pequeñas a través del tiempo

$$\hat{Y}_{a,t_1} = \hat{Y}_{k,t_1} \cdot \hat{p}_{a,t_1} \quad \text{with} \quad \sum_{a=1}^A \hat{p}_{a,t_1} = \sum_{a=1}^A \frac{\hat{Y}_{a,t_1}^{Aux}}{\hat{Y}_{k,t_1}^{Aux}} = 1 \quad \forall \quad a \subseteq k$$

Mejorando la estructura de asignación con información auxiliar

- **Enfoque fijo:** proporciones poblacionales p_{a,t_0} son utilizadas para distribuir conteos de áreas “grandes” K en áreas pequeñas A

$$\hat{Y}_{a,t_1} = \hat{Y}_{k,t_1} \cdot p_{a,t_0} \quad \text{with} \quad \sum_{a=1}^A p_{a,t_0} = \sum_{a=1}^A \frac{Y_{a,t_0}}{Y_{k,t_0}} = 1 \quad \forall \quad a \subseteq k$$

- **Enfoque dinámico:** la información auxiliar actualizada se puede utilizar para tomar en cuenta cambios poblacionales en áreas pequeñas a través del tiempo

$$\hat{Y}_{a,t_1} = \hat{Y}_{k,t_1} \cdot \hat{p}_{a,t_1} \quad \text{with} \quad \sum_{a=1}^A \hat{p}_{a,t_1} = \sum_{a=1}^A \frac{\hat{Y}_{a,t_1}^{Aux}}{\hat{Y}_{k,t_1}^{Aux}} = 1 \quad \forall \quad a \subseteq k$$

Census Z_{aj,t_0}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

Survey margins \hat{Y}_{a,t_1} and \hat{Y}_{j,t_1}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

Target Y_{aj,t_1}

A/J	j ₁	j ₂	j ₃	Total
a ₁				
a ₂				
a ₃				
Total				

- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 Proyectos en desarrollo

Ejemplo: pobreza monetaria para áreas pequeñas



(d) 81 Cantones



(e) 473 Distritos

} Pobreza monetaria
para áreas pequeñas

Estimación en áreas pequeñas (SAE)

Censo + Encuesta
2011 2017

Structure PREserving Estimation (SPREE)

- ✓ Confiable
- ✓ desagregado
- ✓ **Actualizado**



Ejemplo: pobreza monetaria para áreas pequeñas



(d) 81 Cantones



(e) 473 Distritos

} Pobreza monetaria
para áreas pequeñas

Estimación en áreas pequeñas (SAE)

Censo + Encuesta
2011 2017

Structure PREserving Estimation (SPREE)

- ✓ Confiable
- ✓ desagregado
- ✓ **Actualizado**



x La información del ingreso **no** está en el censo!

Propuesta para **estimar** y **actualizar** conteos poblacionales en áreas pequeñas

Métodos SAE

Variable de interés en encuesta
pero no en censo. Censo
y encuesta del mismo año

ELL
Elbers et
al. (2003)

EBP
Molina &
Rao (2010)

...

Métodos de Actualización

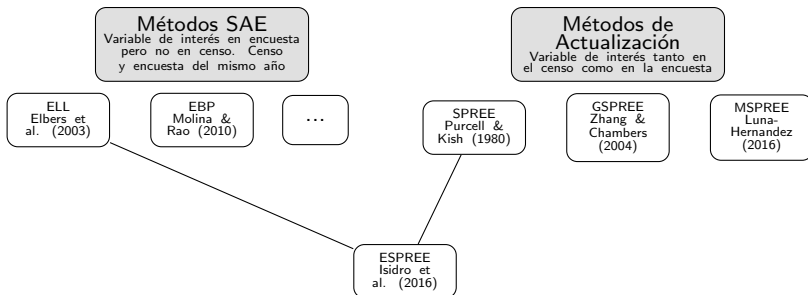
Variable de interés tanto en
el censo como en la encuesta

SPREE
Purcell &
Kish (1980)

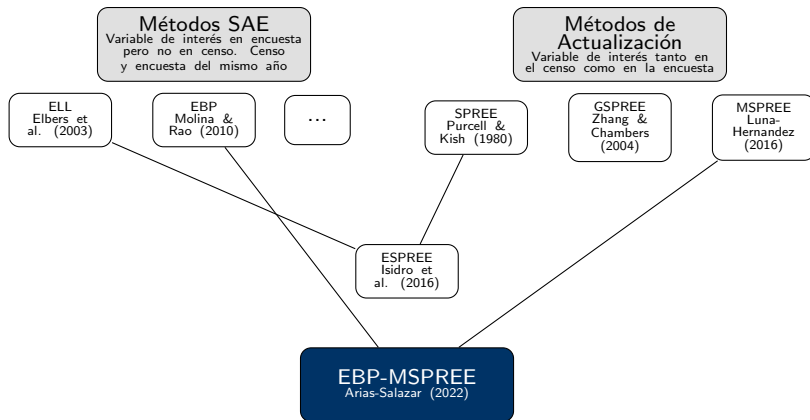
GSPREE
Zhang &
Chambers
(2004)

MSPREE
Luna-
Hernandez
(2016)

Propuesta para **estimar** y **actualizar** conteos poblacionales en áreas pequeñas



Propuesta para **estimar** y **actualizar** conteos poblacionales en áreas pequeñas



Propuesta: procedimiento en dos pasos

Census 2011

Cantons	Poor	Extreme	Not poor	Total
a ₁				
a _{..}				
a ₈₁				
Total				



Survey margins 2017

Cantons	Poor	Extreme	Not poor	Total
a ₁				
a _{..}				
a ₈₁				
Total				



Target 2017

Cantons	Poor	Extreme	Not poor	Total
a ₁				
a _{..}				
a ₈₁				
Total				

1. Estimación con EBP
(Empirical Best Predictor)

2. Actualización con MSPREE

Modelo de regresión de errores anidados (Battese et. al, 1988):

$$y_{ai} = \mathbf{x}_{ai}^T \boldsymbol{\beta} + u_a + e_{ai},$$

donde y_{ai} es la variable de interés, \mathbf{x}_{ai} el conjunto de predictores para el individuo i ésimo en el área a , u_a es el efecto aleatorio de área, e_{ai} , el error a nivel de individuo, $u_a \stackrel{iid}{\sim} N(0, \sigma_u^2)$ y $e_{ai} \stackrel{iid}{\sim} N(0, \sigma_e^2)$.

Modelo de regresión de errores anidados (Battese et. al, 1988):

$$y_{ai} = \mathbf{x}_{ai}^T \boldsymbol{\beta} + u_a + e_{ai},$$

donde y_{ai} es la variable de interés, \mathbf{x}_{ai} el conjunto de predictores para el individuo i ésimo en el área a , u_a es el efecto aleatorio de área, e_{ai} , el error a nivel de individuo, $u_a \stackrel{iid}{\sim} N(0, \sigma_u^2)$ y $e_{ai} \stackrel{iid}{\sim} N(0, \sigma_e^2)$.

- 1 Con los datos de encuesta obtener: $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$
- 2 Con los datos censales: simular $l = 1, \dots, L$ poblaciones sintéticas.
- 3 Calcular la media sobre las L réplicas para obtener el indicador final.

- 1 Tomar Y_{aj} de los datos de encuesta (2012-2017), Z_{aj} de las estimaciones via EBP (2011) como datos censales, para obtener α_{a,t_0}^Z
- 2 Obtener β via Máxima Verosimilitud (ML) o mínimos cuadrados ponderados iterativos (IWLS)
- 3 Entonces $\alpha_{a,t_1}^Z = \beta \alpha_{a,t_0}^Z$
- 4 Tomar en consideración todos estos elementos para obtener:
 $\hat{Y}_{aj,t_1}^{EM} = \text{IPF}[\exp(\hat{\alpha}_{a,t_1}^Y), \hat{Y}_{a,t_1}, \hat{Y}_{j,t_1}]$ para 2012-2017.

$$\widehat{\text{MSE}}(\hat{Y}_{aj,t_1}^{EM}) = \frac{1}{B} \sum_{b=1}^B (\hat{Y}_{aj,t_1}^{EM,b} - \hat{Y}_{aj,t_1}^{EM})^2.$$

1 Incertidumbre que proviene de la estructura de asociación (censo)

- Del procedimiento EBP, tomar las L-pseudo poblaciones para crear $\hat{\alpha}_{a,t_0}^{Z,b}$ con $L=B$ (Isidro, et al. 2016).

2 Incertidumbre que proviene de la estructura de asignación (totales de encuesta)

- De la estimación puntual, generar B pares de márgenes asumiendo e.g.,

$$Y_{a,t_1}^b \stackrel{\text{ind}}{\sim} \text{Multinomial}(\tilde{Y}, \tilde{\pi}_a), \text{ donde } \tilde{\pi}_a = \frac{\hat{Y}_{a,t_1}}{\tilde{Y}} \text{ y } \tilde{Y} = \sum_{a=1}^A \hat{Y}_{a,t_1}.$$

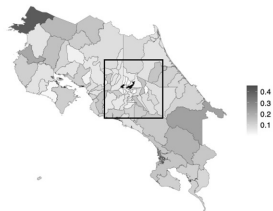
3 Incertidumbre a causa de la estimación de β .

- Obtener B-veces estimaciones EBP-MSPREE $\hat{Y}_{aj,t_1}^{EM,b} = \text{IPF}[\exp(\hat{\alpha}_{a,t_1}^{Y,b}), \hat{Y}_{a,t_1}^b, \hat{Y}_{j,t_1}^b]$

- Objetivo: Obtener estimaciones de pobreza monetaria a nivel cantonal (2011-2017).
- Datos disponibles: ENAHO (2011-2017) y Censo 2011.
- Paso 1: estimación vía EBP.

Estudio de caso: Pobreza en Costa Rica

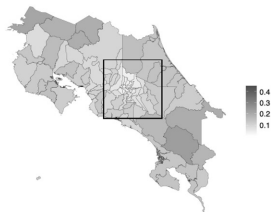
- Objetivo: Obtener estimaciones de pobreza monetaria a nivel cantonal (2011-2017).
- Datos disponibles: ENAHO (2011-2017) y Censo 2011.
- Paso 1: estimación vía EBP.



(a) Direct estimates total



(b) Direct estimates Metropolitan Area



(c) Model- based estimates total



(d) Model- based estimates Metropolitan Area

Tabla 1: Coeficientes de variación de las estimaciones directas y las basadas en el modelo, 2011

CV	Direct			EBP		
	Extreme	Poor	Not Poor	Extreme	Poor	Not Poor
Min	0.068	0.071	0.033	0.085	0.045	0.012
1st Q.	0.232	0.164	0.086	0.127	0.065	0.023
Median	0.332	0.323	0.110	0.155	0.085	0.033
Mean	0.412	0.271	0.123	0.160	0.089	0.033
3rd Q.	0.524	0.234	0.145	0.181	0.106	0.040
Max	1.000	1.000	0.281	0.299	0.182	0.062

¿Cuáles cantones han tenido la mayor reducción de pobreza extrema entre 2011-2017?

En el 2017, ¿Cuáles cantones tuvieron mayor incidencia de pobreza extrema?

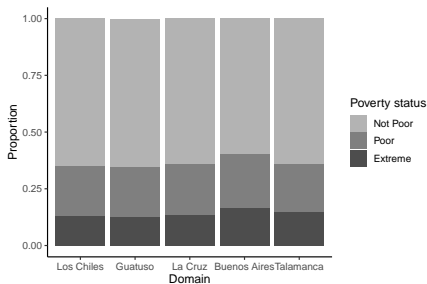
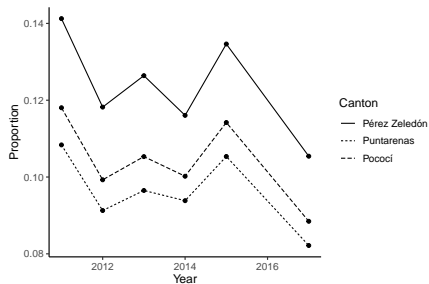


Gráfico 1: Cantones con mayor reducción en incidencia de pobreza extrema, 2011-2017

Gráfico 2: Cantones con la mayor incidencia en pobreza extrema, 2017

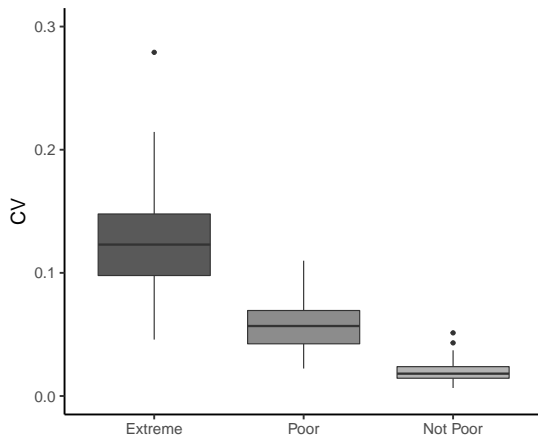
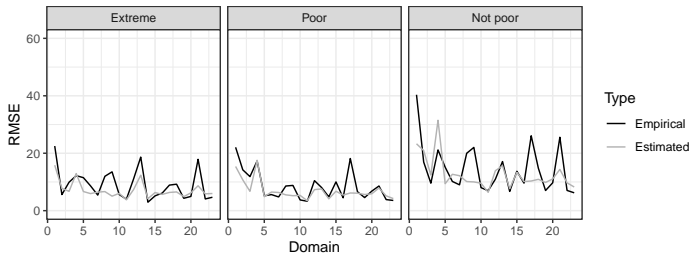


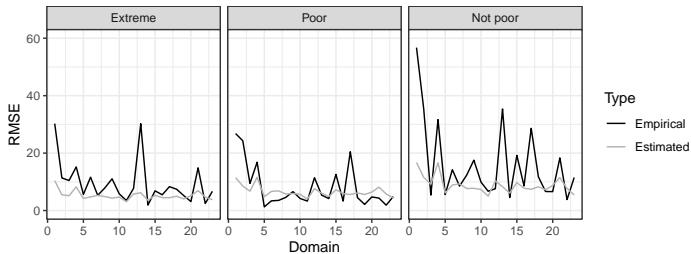
Gráfico 3: Coeficientes de variación, 2017

Estudio de caso: Pobreza en Costa Rica

MSPREE



SPREE



- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 Proyectos en desarrollo

Institutos Nacionales de Estadística

→ Estadísticas oficiales
e.g., indicadores socio-económicos



Características de calidad:

- Confiables
- Desagregados
- Actualizados

→ Microdata
e.g., Datos de encuesta



Balance entre:

- Utilidad
- Privacidad

Institutos Nacionales de Estadística

→ Estadísticas oficiales
e.g., indicadores socio-económicos



Características de calidad:

- Confiables
- Desagregados
- Actualizados

Métodos SPREE!

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?
 - Uso de información auxiliar (e.g., imágenes satelitales) para producir proporciones de subpoblaciones anualmente.

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?

– Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1-27. <https://doi.org/10.1111/rssa.12802>

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?

– Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1-27. <https://doi.org/10.1111/rssa.12802>

2. ¿Cómo aplicar los métodos SPREE si el censo no contiene la variable de interés?

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?

– Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1-27. <https://doi.org/10.1111/rssa.12802>

2. ¿Cómo aplicar los métodos SPREE si el censo no contiene la variable de interés?

Procedimiento en dos pasos: EBP-MSPREE.

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?

– Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1-27. <https://doi.org/10.1111/rssa.12802>

2. ¿Cómo aplicar los métodos SPREE si el censo no contiene la variable de interés?

– Arias-Salazar, A. (2022). **Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach**, *Journal of Official Statistics*, Accepted.

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?
 - Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1-27. <https://doi.org/10.1111/rssa.12802>
2. ¿Cómo aplicar los métodos SPREE si el censo no contiene la variable de interés?
 - Arias-Salazar, A. (2022). **Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach**, *Journal of Official Statistics*, Accepted.
3. ¿Cómo implementar fácilmente los métodos básicos de SPREE?

1. ¿Cómo cumplir el requerimiento de que la estructura de asignación (totales por área) es confiable?

– Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1-27. <https://doi.org/10.1111/rssa.12802>

2. ¿Cómo aplicar los métodos SPREE si el censo no contiene la variable de interés?

– Arias-Salazar, A. (2022). **Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach**, *Journal of Official Statistics*, Accepted.

3. ¿Cómo implementar fácilmente los métodos básicos de SPREE?

– Arias-Salazar, A. (2022). **R package spree**. *Working paper*, https://github.com/AlejandraAriasSalazar/spree_pkg

Institutos Nacionales de Estadística

→ Microdatos



Balance entre:

- Utilidad
- Privacidad

Institutos Nacionales de Estadística

→ Microdatos



Balance entre:

- Utilidad
- Privacidad

4. ¿Cómo mejorar la relación privacidad-utilidad al publicar microdatos de encuestas?

Institutos Nacionales de Estadística

→ Microdatos



Balance entre:

- Utilidad
- Privacidad

4. ¿Cómo mejorar la relación privacidad-utilidad al publicar microdatos de encuestas?

Koebe, T., Arias-Salazar, A. & Schmid, T. **Releasing survey microdata with exact cluster locations and additional privacy safeguards.** *Humanit Soc Sci Commun* 10, 220 (2023). <https://doi.org/10.1057/s41599-023-01694-y>

- 1 Motivación
- 2 Structure PREserving Estimation (SPREE)
- 3 Nuevos desarrollos en SPREE: **Llenando vacíos**
 - Estructura de asignación no confiable
 - Estructura de asociación no disponible
- 4 Resumen
- 5 **Proyectos en desarrollo**

Objetivo 1: Erradicar la pobreza en todas sus formas en todo el mundo.



(a) Monetaria



(b) Multidimensional



- Arias-Salazar A., Gutiérrez, A., Guerrero-Gómez, S., Rojas-Perilla, N., Mancero, X., and Zhang, H. **Small area estimation for composite indicators: the case of multidimensional poverty incidence.** preprint arXiv: 2304.03901. <https://doi.org/10.48550/arXiv.2304.03901>.

Objetivo 1: Erradicar la pobreza en todas sus formas en todo el mundo.



(a) Monetaria



(b) Multidimensional



- Arias-Salazar A., Gutiérrez, A., Guerrero-Gómez, S., Rojas-Perilla, N., Mancero, X., and Zhang, H. **Small area estimation for composite indicators: the case of multidimensional poverty incidence.** preprint arXiv: 2304.03901. <https://doi.org/10.48550/arXiv.2304.03901>.
- **Comparing data integration approaches to deal with missingness in composite indicators.** Con Moretti, A.

Objetivo 1: Erradicar la pobreza en todas sus formas en todo el mundo.



(a) Monetaria



(b) Multidimensional



- Arias-Salazar A., Gutiérrez, A., Guerrero-Gómez, S., Rojas-Perilla, N., Mancero, X., and Zhang, H. **Small area estimation for composite indicators: the case of multidimensional poverty incidence.** preprint arXiv: 2304.03901. <https://doi.org/10.48550/arXiv.2304.039017>.
- **Comparing data integration approaches to deal with missingness in composite indicators.** Con Moretti, A.
- Multidimensional poverty index for small domains: a Bayesian approach. Con CEPAL.

Objetivo 1: Erradicar la pobreza en todas sus formas en todo el mundo.



(a) Monetaria



(b) Multidimensional



- Arias-Salazar A., Gutiérrez, A., Guerrero-Gómez, S., Rojas-Perilla, N., Mancero, X., and Zhang, H. **Small area estimation for composite indicators: the case of multidimensional poverty incidence.** preprint arXiv: 2304.03901. <https://doi.org/10.48550/arXiv.2304.039017>.
- **Comparing data integration approaches to deal with missingness in composite indicators.** Con Moretti, A.
- Multidimensional poverty index for small domains: a Bayesian approach. Con CEPAL.
- Metodologías para la estimación de pobreza en áreas pequeñas en Costa Rica. Proyecto UCR.

Gracias por su atención!

Alejandra Arias-Salazar (alejandra.ariassalazar@ucr.ac.cr)

- Deming, E. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*. 11(4), 427–444.
- Isidro, M., S. Haslett, and G. Jones (2016). Extended structure preserving estimation (ESPREE) for updating small area estimates of poverty. *The Annals of Applied Statistics*. 10(1), 451–476.
- Luna-Hernández, A. (2016). Multivariate structure preserving estimation for population compositions. Ph. D. thesis, University of Southampton.
- Molina, I. and J. Rao (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*. 38(3), 369–385.
- Purcell, N. J. and L. Kish (1980). Postcensal estimates for local areas (or domains). *International Statistical Review* 48(1), 3–18.

Table of appendix

- Uncertainty EBP-MSE [Details](#)
- IPF [Details](#)
- Beta Estimation [Details](#)
- Z scores [Details](#)
- Complex sampling design [Details](#)
- Constraints [Details](#)

A two-way contingency table for the Censo year t_0 can be represented as a saturated log-linear model.

$$\log Z_{aj,t_0} = \alpha_{0,t_0}^Z + \alpha_{a,t_0}^Z + \alpha_{j,t_0}^Z + \alpha_{aj,t_0}^Z.$$

These terms can be defined using a centred-constraint parametrization:

$$\alpha_{0,t_0}^Z = \frac{1}{AJ} \sum_{a=1}^A \sum_{j=1}^J \log Z_{aj,t_0},$$

$$\alpha_{a,t_0}^Z = \frac{1}{J} \sum_{j=1}^J \log Z_{aj,t_0} - \alpha_{0,t_0}^Z,$$

$$\alpha_{j,t_0}^Z = \frac{1}{A} \sum_{a=1}^A \log Z_{aj,t_0} - \alpha_{0,t_0}^Z,$$

$$\alpha_{aj,t_0}^Z = \log Z_{aj,t_0} - \alpha_{a,t_0}^Z - \alpha_{j,t_0}^Z - \alpha_{0,t_0}^Z,$$

and the constraints:

$$\sum_{a=1}^A \alpha_{a,t_0}^Z = \sum_{j=1}^J \alpha_{j,t_0}^Z = \sum_{a=1}^A \alpha_{aj,t_0}^Z = \sum_{j=1}^J \alpha_{aj,t_0}^Z = 0. \quad (1)$$

Using the Censo Z_{aj,t_0} as starting point, the first cycle of the algorithm is:

- 1 Rescale to the first row margins of Z_{aj,t_0} :

$$\hat{Y}_{aj,t_1}^{(1)} = Z_{aj,t_0} \frac{Y_{j,t_1}}{Z_{j,t_0}}.$$

- 2 The cells rescaled in the previous step are rescaled again, now with the column margins of $\hat{Y}_{aj,t_1}^{(1)}$:

$$\hat{Y}_{aj,t_1}^{(2)} = \hat{Y}_{aj,t_1}^{(1)} \frac{Y_{a,t_1}}{\hat{Y}_{a,t_1}^{(1)}}.$$

- 3 Cells are rescaled again, with row margins of $\hat{Y}_{aj,t_1}^{(2)}$:

$$\hat{Y}_{aj,t_1}^{(3)} = \hat{Y}_{aj,t_1}^{(2)} \frac{\hat{Y}_{j,t_1}^{(2)}}{Y_{j,t_1}}.$$

In this way, the IPF follows an iterative process, repeating the last two final steps until reaching convergence. Convergence occurs when the maximum deviation between observed and fitted margins is less than a specific value (e.g., 0.00001). [Overview](#)

- 1 **Uncertainty from the association structure.** From the EBP model, L Monte Carlo pseudo populations in t_0 are used to create $\alpha_{aj,t_0}^{Z,b}$ (with $L = B$)
- 2 **Uncertainty from the allocation structure.** Generate B pairs of margins Y_{a,t_1}^b, Y_{j,t_1}^b from the point estimate \hat{Y}_{aj,t_1}^{EM} assuming a multinomial distribution, i.e., $Y_{a,t_1}^b \stackrel{\text{ind}}{\sim} \text{Multinomial}(\tilde{Y}, \hat{\pi}_a)$, where $\hat{\pi}_a = \frac{\hat{Y}_{a,t_1}}{\tilde{Y}}$ and $\tilde{Y} = \sum_{a=1}^A \hat{Y}_{a,t_1}$.
- 3 **Uncertainty due to the estimation of β .** From the point estimate \hat{Y}_{aj,t_1}^{EM} calculate the within-area proportions $\hat{\pi}_{a,t_1} = \frac{\hat{Y}_{aj,t_1}^M}{Y_{a,t_1}}$ for each area $a = 1, \dots, A$.
For $b = 1, \dots, B$:
 - Generate B bootstrap populations Y_{aj,t_1}^{*b} assuming $\mathbf{Y}_a | \alpha_{a,t_0}^Z \stackrel{\text{ind}}{\sim} \text{Multinomial}(Y_{a,t_1}, \hat{\pi}_{a,t_1})$.
 - draw a sample $y_{aj,t_1}^{EM,b}$.
- 4 With the output of Step 1 and 3, specify the MSPREE structural assumption $\alpha_a^Y = \beta \alpha_a^Z$.
- 5 With $L = B$ and the output of Step 2 and 4, compute B EBP-MSPREE estimates

$$\hat{Y}_{aj,t_1}^{EM,b} = \text{IPF}[\exp(\hat{\alpha}_{a,t_1}^{Y,b}), Y_{a,t_1}^b, Y_{j,t_1}^b].$$

- 6 Finally, estimate the MSE:

$$\widehat{\text{MSE}}(\hat{Y}_{aj,t_1}^{EM}) = \frac{1}{B} \sum_{b=1}^B (\hat{Y}_{aj,t_1}^{EM,b} - Y_{aj,t_1}^{*b})^2. \quad (2)$$

Encuesta-based direct estimates are required to compute β , and several approaches can be implemented: via maximum likelihood (ML) or iterative weighted least squares (IWLS) Jiang2007, in both cases by using a log or a logit link. ML estimates of β can be obtained by assuming

$$\hat{Y}_{aj,t_1}^{Dir} | \alpha_{a,t_0}^Z \stackrel{\text{ind}}{\sim} \text{Poisson}(\mu_{aj,t_1})$$

or

$$\hat{Y}_{a,t_1}^{Dir} | \alpha_{a,t_0}^Z \stackrel{\text{ind}}{\sim} \text{Multinomial}\left(\sum_{j=1}^J \hat{Y}_{aj,t_1}^{Dir}, \pi_{a,t_1}\right),$$

where μ_{aj,t_1} are Poisson expected frequencies and π_{a,t_1} the cell probabilities used under multinomial sampling. [Overview](#)

Since this matrix is usually not available, Luna-Hernández (2016) proposes a solution by multiplying a design effect with the variance that corresponds to a simple random sampling design without replacement. The estimate of this matrix is defined as

$$\hat{V}_{aj,t_1} = \frac{\text{deff}_{j,t_1} \hat{\pi}_{aj,t_1} (1 - \hat{\pi}_{aj,t_1})}{n_{a,t_1}}, \quad (3)$$

with $\hat{\pi}_{aj,t_1} = \frac{\hat{Y}_{aj,t_1}^M}{Y_{a,t_1}}$ and n_{a,t_1} the area sample sizes. The term deff_{j,t_1} is a vector of size J , containing a design effect for each category.

For the intercensal years, estimates on incidence of extreme poverty are obtained with EBP-MSPREE. To analyse major changes in this indicator between 2011 and 2017 Z-scores are used:

$$Z = \frac{\text{Estimate}_{2011} - \text{Estimate}_{2017}}{\sqrt{(\text{Standard error}_{2011})^2 + (\text{Standard error}_{2017})^2}}.$$

This measure represents the standardised distance between the estimates in both years (Isidro, M., 2010).

Overview