

Tres Aplicaciones Importantes de la Estimación de Áreas Pequeñas en los Estados Unidos

Carolina Franco



Desagregación de Datos mediante Técnicas de Estimación de Áreas Pequeñas en México

28 de noviembre de 2023

Introducción

Estimación en áreas pequeñas (SAE)

- **Objetivo:** Estimar cantidades de interés (por ejemplo, pobreza, desempleo, etc.) para muchos dominios geográficos y/o demográficos/subconjuntos de la población
- **Problema:** A menudo, las encuestas no pueden estimar todas las cantidades de interés a través de métodos “directos” con una precisión aceptable

Estimador directo: *Basado únicamente en datos de muestra para el dominio de interés.*

Area pequeña: *Dominio donde el tamaño de muestra es demasiado pequeño para una estimación directa confiable*

- **SAE:** A través de la modelización, incorporar información de otros dominios y fuentes de datos auxiliares para “prestar fuerza”

- Registros administrativos (e.g., Erciulescu, Franco, Lahiri, 2021 habla sobre consideraciones prácticas)
- Censos
- Misma encuesta en diferentes años
- Datos comerciales, datos de satélites, datos de teléfonos celulares, etc.
- Información espacial (e.g. matrices de adyacencia)
- Otras encuestas (e.g., Franco y Maitra 2023)

- Por supuesto, se puede usar una combinación de estas
- Generalmente, cada tipo de fuente requiere supuestos de modelización distintos (temporales, bivariados, espaciales)
- La calidad de las fuentes de información es importante (como también la calidad de los datos)
- Distintos tipos de modelos tienen distintos requisitos para las fuentes de información auxiliar (e.g., modelos de nivel de unidad o modelos e nivel de área)

Posibles beneficios de SAE (“Small Area Estimation”)

Cuando se implementa **adecuadamente**, y cuando se encuentra **buena información auxiliar**, SAE puede...

Posibles beneficios de SAE (“Small Area Estimation”)

Cuando se implementa **adecuadamente**, y cuando se encuentra **buena información auxiliar**, SAE puede...

- Reducir las medidas de incertidumbre en comparación con los estimadores directos

Posibles beneficios de SAE (“Small Area Estimation”)

Cuando se implementa **adecuadamente**, y cuando se encuentra **buena información auxiliar**, SAE puede...

- Reducir las medidas de incertidumbre en comparación con los estimadores directos
- Permitir que más áreas cumplan con los umbrales de calidad establecidos por los institutos de estadística

Posibles beneficios de SAE (“Small Area Estimation”)

Cuando se implementa **adecuadamente**, y cuando se encuentra **buena información auxiliar**, SAE puede...

- Reducir las medidas de incertidumbre en comparación con los estimadores directos
- Permitir que más áreas cumplan con los umbrales de calidad establecidos por los institutos de estadística
- Permitir la publicación de estadísticas que de otra manera se suprimirían

Posibles beneficios de SAE (“Small Area Estimation”)

Cuando se implementa **adecuadamente**, y cuando se encuentra **buena información auxiliar**, SAE puede...

- Reducir las medidas de incertidumbre en comparación con los estimadores directos
- Permitir que más áreas cumplan con los umbrales de calidad establecidos por los institutos de estadística
- Permitir la publicación de estadísticas que de otra manera se suprimirían
- Proporcionar estimaciones para áreas sin muestra de la encuesta (con cautela)

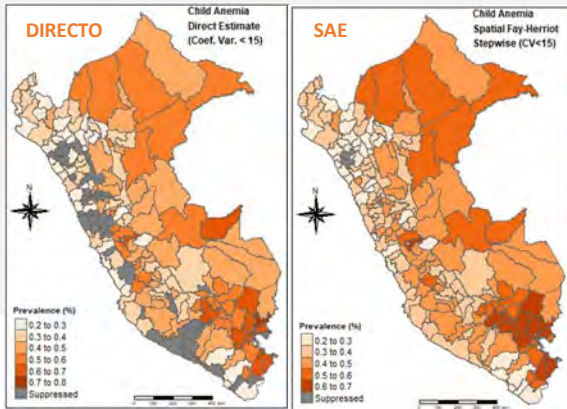
Posibles beneficios de SAE (“Small Area Estimation”)

Cuando se implementa **adecuadamente**, y cuando se encuentra **buena información auxiliar**, SAE puede...

- Reducir las medidas de incertidumbre en comparación con los estimadores directos
- Permitir que más áreas cumplan con los umbrales de calidad establecidos por los institutos de estadística
- Permitir la publicación de estadísticas que de otra manera se suprimirían
- Proporcionar estimaciones para áreas sin muestra de la encuesta (con cautela)

Muchos modelos producen estimadores con convergencia al estimador directo – propiedad importante

Ilustración: Anemia en las provincias de Perú para niños menores de 5 años



Modelo SAE:
Solo 3
estimadores se
suprimen.
**Estimadores
directos:** se
suprimen 42
estimadores

Trabajo con
Angelo
Cozzubo

1. SAIPE: Estimaciones oficiales de ingresos y pobreza en áreas pequeñas
2. VRA, Section 203: Ley de Derechos de Votación, Sección 203, estimaciones oficiales
3. Prevalencia de pérdida auditiva por raza, sexo, edad y condado de los Estados Unidos

SAIPE: Estimación oficial de pobreza para varios grupos y niveles geográficos

Estimación de la pobreza en la Oficina del Censo

- El programa SAIPE (Small Area Income and Poverty Estimates) de la Oficina del Censo de EE. UU. estima la pobreza para varios grupos de edad
- Estimaciones para estados, **condados***, y distritos escolares
- La estimación de pobreza para los niños entre 5 a 17 años de edad se usa para la distribución de fondos federales
- Para las estimaciones a nivel de condado SAIPE usa un modelo Fay-Herriot, con transformación
- En EE. UU., una **familia** y todas las personas de la familia se consideran en situación de pobreza si sus ingresos **totales** (antes de impuestos) son inferiores al umbral de pobreza para el tamaño de la familia y la composición por edad.

<https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html>

Poverty Thresholds for 2022 by Size of Family and Number of Related Children Under 18 Years
(In dollars)

Size of family unit	Related children under 18 years								
	None	One	Two	Three	Four	Five	Six	Seven	Eight or more
One person (unrelated individual):									
Under 65 years.....	15,225								
65 years and over.....	14,036								
Two people:									
Householder under 65 years.....	19,597	20,172							
Householder 65 years and over.....	17,689	20,095							
Three people.....	22,892	23,556	23,578						
Four people.....	30,186	30,679	29,678	29,782					
Five people.....	36,402	36,932	35,801	34,926	34,391				
Six people.....	41,869	42,035	41,169	40,339	39,104	38,373			
Seven people.....	48,176	48,477	47,440	46,717	45,371	43,800	42,076		
Eight people.....	53,881	54,357	53,378	52,521	51,304	49,760	48,153	47,745	
Nine people or more.....	64,815	65,129	64,263	63,536	62,342	60,699	59,213	58,845	56,578

Source: U.S. Census Bureau, 2023.

Note: The source of the weighted average thresholds is the 2023 Current Population Survey Annual Social and Economic Supplement (CPS ASEC).

Encuesta de la Comunidad Estadounidense (ACS)

- “American Community Survey”
- Encuesta principal para modelos SAIPE y para las otras 2 aplicaciones
- Aproximadamente 3.5 millones de direcciones por año
- Preguntas sobre demografía, ingresos, seguro médico, educación, discapacidades, etc.
- Encuesta compleja (estratificación, agrupación de personas dentro de hogares, submuestreo para ahogares que no responden)
- Reemplazó al “formulario largo” del censo (aproximadamente 1/6 de la población durante el censo decenal)
- Se producen estimaciones anuales de 1 año y 5 años

El modelo Fay-Herriot (1979)

- Para m áreas pequeñas:

$$y_i = Y_i + e_i \quad i = 1, \dots, m$$

$$Y_i = \mathbf{x}_i' \beta + u_i$$

- Y_i es la característica de interés de la población para el área i
- y_i es la estimación directa de la encuesta de Y_i
- e_i es el error de muestreo en y_i , generalmente asumido como $N(0, v_i)$, independiente con v_i conocido
- u_i es el efecto aleatorio del área i , usualmente asumido como *i.i.d.* $N(0, \sigma_u^2)$ e independiente de los e_i .

- Mejor predictor lineal de Y_i (β y σ_u^2 conocidos):

$$\hat{Y}_i = (1 - \gamma_i)y_i + \gamma_i\mathbf{x}'_i\beta$$

donde

$$\gamma_i = \frac{v_i}{v_i + \sigma_u^2}$$

- Combinación lineal del estimador “directo” (y_i) y el estimador “sintético” ($\mathbf{x}'_i\beta$)
- Varianzas de muestreo más pequeñas implican mayor peso de y_i
- Ajuste mediante bayes jerárquico (HB) o bayes empírico (EB)

Un modelo de Fay-Herriot:

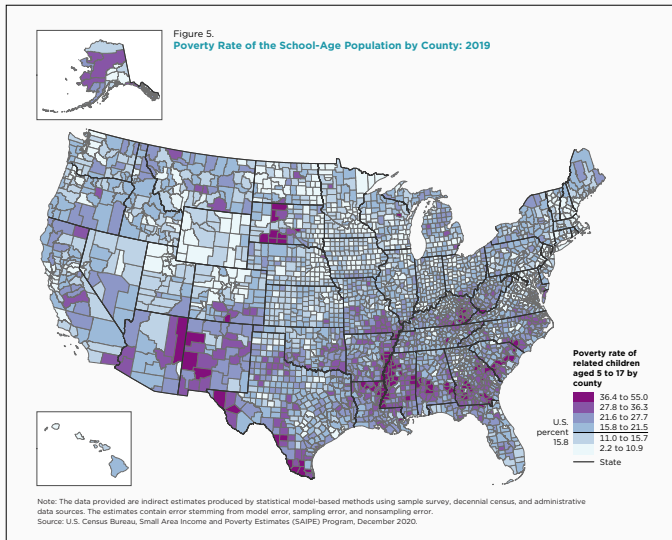
- y_i = logaritmo de la estimación ACS del número de niños de 5 a 17 años en situación de pobreza para el condado i
- Y_i = logaritmo de la cantidad verdadera correspondiente
- β y σ_u^2 se estiman mediante máxima verosimilitud
- \mathbf{x}_i es el vector de variables de regresión en escala logarítmica
- Los resultados de predicción se traducen de nuevo a la escala lineal utilizando propiedades de la distribución lognormal

Modelo de pobreza para condados SAIPE, edades 5-17- variables de regresión

El modelo incluye un intercepto. Además, en escala logarítmica, para cada condado, las covariables incluyen:

- Número de “exenciones de niños pobres” (exenciones en declaraciones de impuestos con ingresos por debajo del umbral de pobreza)
- Número de beneficiarios del Programa de Asistencia Nutricional Suplementaria (SNAP)
- Estimación del tamaño de la población de edad 0-17
- Número de exenciones de impuestos para niños
- Estimación del censo 2000 del número de niños pobres en edad escolar (de 5 a 17 años)

Mapa de pobreza de niños por condado

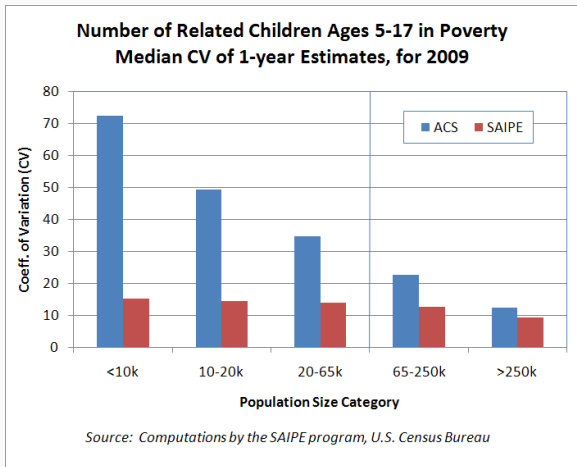


En el sitio web de SAIPE, se puede encontrar la herramienta interactiva de SAIPE:

<https://www.census.gov/data-tools/demo/saipe>

Allí, se pueden ver mapas de pobreza filtrando por grupos de edad y geografía, así como gráficos sobre la serie temporal de las tasas de pobreza

Número de niños de 5-17 años en pobreza, CV mediano de estimaciones, 2009



- Los datos son inherentemente discretos, posibles mejoras a las suposiciones de normalidad (?). Necesidad de eliminar los condados con estimaciones cero ($\approx 5\%$) debido a la transformación logarítmica;

- Los datos son inherentemente discretos, posibles mejoras a las suposiciones de normalidad (?). Necesidad de eliminar los condados con estimaciones cero ($\approx 5\%$) debido a la transformación logarítmica;

Solución potencial: Considerar otros Modelos Lineales Generalizados Mixtos (GLMM), como el modelo Binomial/Logit Normal (BLN). Ver Franco y Bell (2013, 2015, 2022); Franco (2020)

- Los datos son inherentemente discretos, posibles mejoras a las suposiciones de normalidad (?). Necesidad de eliminar los condados con estimaciones cero ($\approx 5\%$) debido a la transformación logarítmica;

Solución potencial: Considerar otros Modelos Lineales Generalizados Mixtos (GLMM), como el modelo Binomial/Logit Normal (BLN). Ver Franco y Bell (2013, 2015, 2022); Franco (2020)

- Falta de buenas estimaciones de las varianzas de muestreo

- Los datos son inherentemente discretos, posibles mejoras a las suposiciones de normalidad (?). Necesidad de eliminar los condados con estimaciones cero ($\approx 5\%$) debido a la transformación logarítmica;

Solución potencial: Considerar otros Modelos Lineales Generalizados Mixtos (GLMM), como el modelo Binomial/Logit Normal (BLN). Ver Franco y Bell (2013, 2015, 2022); Franco (2020)

- Falta de buenas estimaciones de las varianzas de muestreo

Solución potencial: Usar una Función de Varianza Generalizada para producir estimaciones de las varianzas de muestreo. Ver Franco y Bell (2013), y Franco (2020)

- Uso de la estimación del Censo 2000 como covariable
Solución potencial: Considerar el uso de información de estimaciones de años anteriores de ACS en su lugar
- Nota: el uso de una estimación directa de encuesta como covariable, sin tener en cuenta el error de muestreo, puede resultar en un error de medición
- Bell, Chung, Datta y Franco (2019) muestran que ignorar este error puede conducir a predicciones subóptimas y una estimación incorrecta de la medida de incertidumbre
- El error de muestreo se puede tener en cuenta a través de modelos temporales, extensiones bivariadas o modelos de error de medición (ver Franco y Maitra; 2023; Franco y Bel, 2013; 2015; 2022; o Arima, Datta, Bell, Franco y Liseo, 2019)

Modelo Binomial/Logit Normal (BLN)

- Sea y_i el “número de éxitos” de la muestra, n_i el tamaño de la muestra, p_i proporción verdadera

$$y_i | p_i, n_i \sim \text{Bin}(n_i, p_i) \quad i = 1, \dots, m$$

$$\text{logit}(p_i) = \mathbf{x}'_i \beta + u_i \quad (1)$$

- $\text{logit}(p_i) = \log[p_i/(1 - p_i)]$, $u_i \stackrel{i.i.d}{\sim} N(0, \sigma_u^2)$
- Puede ser más apropiado para datos discretos, maneja de manera natural estimaciones de cero y la asimetría
- Garantiza estimadores e intervalos de confianza entre (0,1)
- El muestreo complejo se puede abordar utilizando el **tamaño efectivo de la muestra** (e.g, Franco 2013), etc.
- Puede ser extendido fácilmente a bivariado (Franco y Bell 2013), temporal (Franco y Bell 2015), etc.
- Posibles ventajas para SAIPE (Franco 2020, Franco y Bell 2015)

Algunas observaciones sobre SAIPE

- SAIPE es un buen ejemplo de un programa exitoso de estimación en áreas pequeñas
- Al aprovechar otras fuentes de datos (registros, censo), SAIPE puede proporcionar estimaciones mejoradas que aprovechan contar con más información
- SAIPE hace investigación y re-evaluación constante de los modelos que usa
- Para leer más sobre el programa de SAIPE, vea Bell et al. (2015) en Analysis of Poverty Data by Small Area Estimation, o el sitio de SAIPE
www.census.gov/programs-surveys/saipe.html

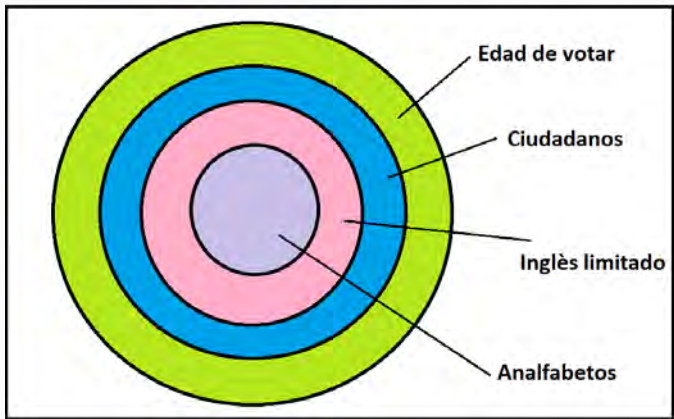
VRA Section 203: Estimación oficial de ciudadanía, conocimientos limitados del idioma inglés, y analfabetismo, para la ley de los derechos de los votantes

- Los estados y subdivisiones políticas (condados y áreas de nativos americanos) deben proporcionar materiales de votación en un idioma distinto al inglés para los miembros de Grupos Minoritarios de Idioma (LMGs) de acuerdo con reglas específicas basadas en fracciones de población
- El Director de la Oficina del Censo toma las determinaciones basadas en datos del Censo Decenal y de la ACS

Terminología principal

- Geografías: estados, Jurisdicciones (aproximadamente 8000), Áreas de Nativos Americanos (AIAs, alrededor de 570), y Corporaciones Regionales de Nativos de Alaska (ANRCs, 12)
- **73 Grupos Minoritarios de Idioma (LMGs):** (21 grupos asiáticos, 51 grupos nativo americanos/alaska nativos, hispanos)
- **VOT:** Población en edad de votar
- **CIT:** Población en edad de votar y ciudadanos
- **LEP:** Población en edad de votar, ciudadanos, inglés limitado
- **ILL:** Población en edad de votar, ciudadanos, inglés limitado, y analfabetos.

Representación visual–categorías anidadas



Cantidades

necesarias para la administracion del VRA S203

- La necesidad de desarrollar varios modelos relacionados para distintos grupos minoritarios y geografías
- La necesidad de usar modelos separados pero parecidos
- Algunos de los modelos cuentan con miles de datos (e.g., hispanos a nivel de jurisdicción), otros modelos casi no cuentan con información (varios de los grupos indígenas)
- Aun en los casos con miles de datos, hay miles de áreas pequeñas, y áreas con pocos o ningún dato
- Las cantidades de interés son anidadas

- Modelado SAE debido a la volatilidad de los estimadores directos en muchos dominios.
- Modelos ajustados por separado para diferentes LMGs
- Se utiliza un modelo Multinomial Logit Normal (MLN) para LMGs más grandes, versión “continuation ratio” (Agresti and Coull, 2000)
- Para algunos LMGs, la muestra de ACS no es suficientemente grande para el modelo MLN
- En este caso, se utilizaron modelos más simples: MLN diagonal, modelos univariados separados de Binomial Logit Normal (BLN) para las proporciones de ILL, LEP y CIT o modelos aún más simples
- Una combinación de estimaciones frequentistas y bayesianas, basadas en consideraciones de computación

Modelo para los casos con muchos datos: Modelo Logit Normal Multinomial, Continuation Ratio

MLN “Continuation Ratio” mixto (Agresti y Coull, 2000);
parametrización de probabilidades condicionales de las cantidades
anidadas

$$\begin{aligned}(\tilde{y}_{1i}, \tilde{y}_{2i}, \tilde{y}_{3i}, \tilde{y}_{4i}) &\sim \text{Multinom}(\tilde{n}_i; \underline{\omega}_i), \quad i = 1, \dots, m \\(\omega_{1,i}, \omega_{2,i}, \omega_{3,i}, \omega_{4,i}) &= \left((1 - \mu_i), \mu_i(1 - \nu_i), \mu_i\nu_i(1 - \rho_i), \mu_i\nu_i\rho_i \right) \\ \text{logit}(\mu_i) &= \mathbf{x}_{1i}\boldsymbol{\beta}_1 + u_{1i} \\ \text{logit}(\nu_i) &= \mathbf{x}_{2i}\boldsymbol{\beta}_2 + u_{2i} \\ \text{logit}(\rho_i) &= \mathbf{x}_{3i}\boldsymbol{\beta}_3 + u_{3i}\end{aligned}$$

donde μ_{ig} es la probabilidad condicional de CIT entre VOT; ν_{ig}
LEP entre CIT; y ρ_{ig} es ILL entre LEP

$\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i})$ es un vector Gaussiano tridimensional

Cuadro 1: Reducciones medias de CVs de usar el modelo MLN para ILL/LEP, casos $n \geq 5$, varianza directa > 0

CV directo	# áreas	% reducción
(0.01, 0.21]	436	9.39
(0.21, 0.32]	467	13.79
(0.32, 0.41]	422	16.77
(0.41, 0.50]	477	21.62
(0.50, 0.60]	501	25.77
(0.60, 0.71]	490	33.98
(0.71, 0.82]	441	37.22
(0.82, 0.95]	505	41.39
(0.95, 1.10]	460	45.89
(1.10,12.60]	478	61.21

- Como en otros años, se usaron covariables de la misma encuesta a mayor nivel de agregación – eso puede causar problemas de error de medición (ver Bell, Chung, Datta, Franco, 2019)
- Para investigar en el futuro: uso de covariables de registros administrativos de impuestos, seguridad social, asistencia médica para los mayores.
- Extensiones temporales del modelo “Continuation Ratio MLN,” para utilizar varios años de datos de la ACS
- Para más detalles, Slud, Franco y Hall (2024); Slud, Franco, Hall y Kang (2022)

Pérdida de la audición para grupos demográficos en los condados de Estados Unidos– trabajo con beca del Centro para el Control de las Enfermedades de los Estados Unidos

La importancia de estimar la pérdida auditiva

- Condición altamente prevalente en los Estados Unidos, a menudo no tratada adecuadamente
- Asociada con múltiples resultados de salud y calidad de vida: desarrollo del habla y lenguaje, comportamiento suicida, demencia, etc.
- Es importante tener estimaciones a niveles más bajos de agregación para la planificación de políticas e intervenciones.
- No existían estimadores a niveles de agregación de condados por grupos demográficos
- Grupos demográficos de interés—**razas**: negros, hispanos, blancos, “otros”; **sexos**: hombre y mujer; **edades**: 0-4, 5-17, 18-34, 35-64, 65-74, 75+.
- Estrategia bayesiana de combinar información de varias fuentes usando modelos SAE y calibración

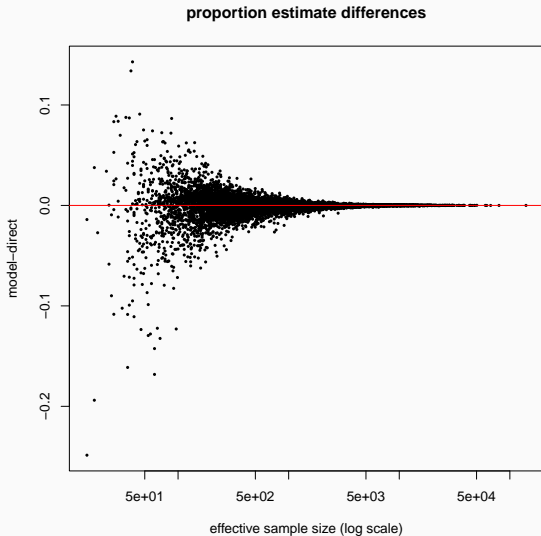
Las fuentes de datos e información auxiliar

- **Encuesta Nacional de Examen de Salud y Nutrición (NHANES)**: estimaciones de pérdida auditiva, 5000 personas por año, diferentes niveles de pérdida auditiva, **medidas por personal médico** (nos enfocamos en leve y moderada/grave)
- **Encuesta de la Comunidad Estadounidense (ACS)**: pérdida auditiva binaria, **auto-reportada**
- **Registros de Medicare** (programa de asistencia social para los mayores para obtener cuidados médicos), **y varias otras fuentes de datos administrativos**
- **Datos relacionados con el ruido**, como la proporción de personas que trabajan en industrias con exposición al ruido
- **Estimaciones de los tamaños de población para los dominios de la ACS**

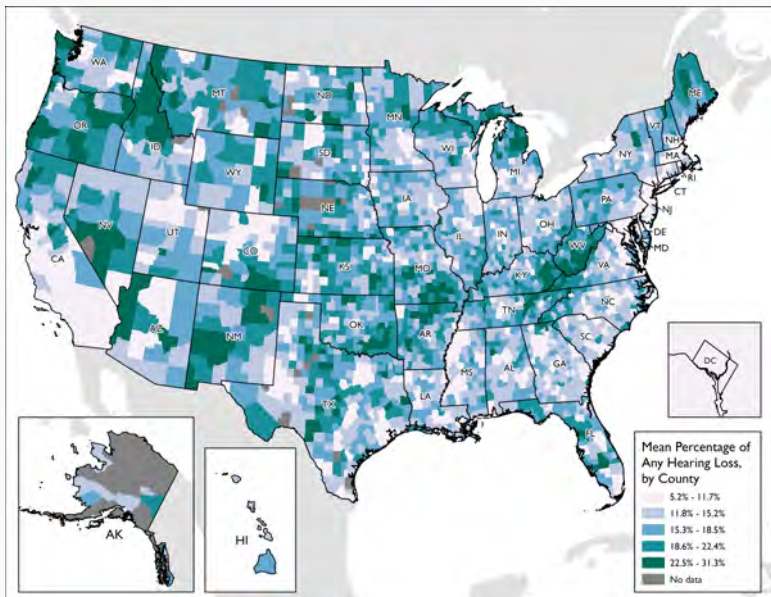
Objetivo: estimaciones para los grupos demográficos al nivel de condado calibrados a las mediciones de NHANES y para los dos niveles de pérdida auditiva

- Modelo NHANES para prevalencias en 3 categorías de pérdida auditiva por grupos demográficos: **Continuation Ratio Multinomial Logit Normal**, covariables de registros (Stan)
- Modelo ACS para pérdida auditiva, por condado por edad: **Binomial Logit Normal**, covariables de registros (JAGS)
- Calibración/raking: estimaciones del modelo ACS deben sumar a las estimaciones del modelo NHANES (para audición normal, pérdida leve, pérdida grave, y por grupo demográfico), asegurar que las probabilidades de las tres categorías sumen uno
- Para mas detalle, Rein, Franco, et al. (2023)

Predicciones del modelo ACS



Mapa a nivel de condados



Discusión de modelos de la pérdida de audición

- Nuestra estrategia bayesiana resulta en medidas de incertidumbre que capturaron aproximadamente todas las fuentes cuantificables de error.
- Nuestra metodología nos permitió aproximar tres grados de gravedad de la pérdida auditiva por grupos demográficos a nivel de condado, previamente no disponibles
- Algunas observaciones interesantes sobre los resultados fueron las tendencias de mayor pérdida auditiva en áreas rurales y una mayor prevalencia de pérdida auditiva en los blancos y los hombres
- Cuando se desarrollan estrategias de SAE, es importante consultar con expertos del tema

Reflexiones

Discusión de las tres aplicaciones

- Vimos tres ejemplos de la utilidad de SAE en aplicaciones reales e importantes. Cada una muestra grandes beneficios en el uso de SAE
- Cada aplicación presenta desafíos únicos que requieren metodologías a la altura
- SAE es una excelente herramienta para entender las desigualdades en la sociedad, pero hay que tener mucho cuidado con las suposiciones del modelo y con las características de las fuentes de datos y de la información auxiliar
- Aunque no hablé mucho de verificación, comparación y diagnósticos de modelos estos fueron parte de las tres aplicaciones, y son muy importantes

- La necesidad de estadísticas a niveles inferiores de agregación está aumentando en todo el mundo.
- SAE puede ayudar a informar políticas basadas en datos, contribuir a la asignación eficiente de recursos
- Descubrimiento global del potencial de SAE
- El estudio de SAE está mas activo que nunca (¡Conferencia de 2024 es en Latino America!
<https://sae2023.pucp.edu.pe/>)
- Época de “proliferación de información”
- Oportunidad para una mayor aplicación de SAE en las estadísticas oficiales!

Preguntas?

Franco-Carolina@norc.org

