Hacia una mayor aplicación de SAE en indicadores sociales

Carolina Franco

NORC at the University of Chicago

Taller regional sobre metodología de estimación en áreas pequeñas y desagregación de datos,

6 de junio, 2023

Introducción

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 las encuestas
- Permitir la publicación de estadísticas que de otra manera se suprimirían
- Proporcionar estimaciones para áreas sin muestra de la encuesta (con cautela)

Muchos estimadores SAE tienen la propiedad de convergencia al estimador directo – propiedad importante

Ejemplos de fuentes de información auxiliar

- 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 (articulo de introducción Franco y Maitra 2023)

Comentarios sobre laa fuentes de información auxiliar

- Por supuesto, se puede usar una combinación de estas
- Generalmente, cada tipo de fuente requiere supuestos de modelización distintos (temporales, bivariadas, espaciales)
- Distintos tipos de modelos tienen distintos requisitos para las fuentes auxiliares (modelos de unidad vs. de área)
- La calidad de las fuentes de información es importante

Tres aplicaciones importantes en SAE en los Estados Unidos

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 de edad entre 5 a 17 años se usa para la distribución de fondos federales
- Pare 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.

Encuesta de la Comunidad Estadounidense (ACS)

- 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:

$$y_i = Y_i + e_i \qquad i = 1, \ldots, m$$
$$Y_i = \mathbf{x}_i'\beta + u_i$$

- 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.

Modelo de pobreza para condados SAIPE, edades 5-17

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

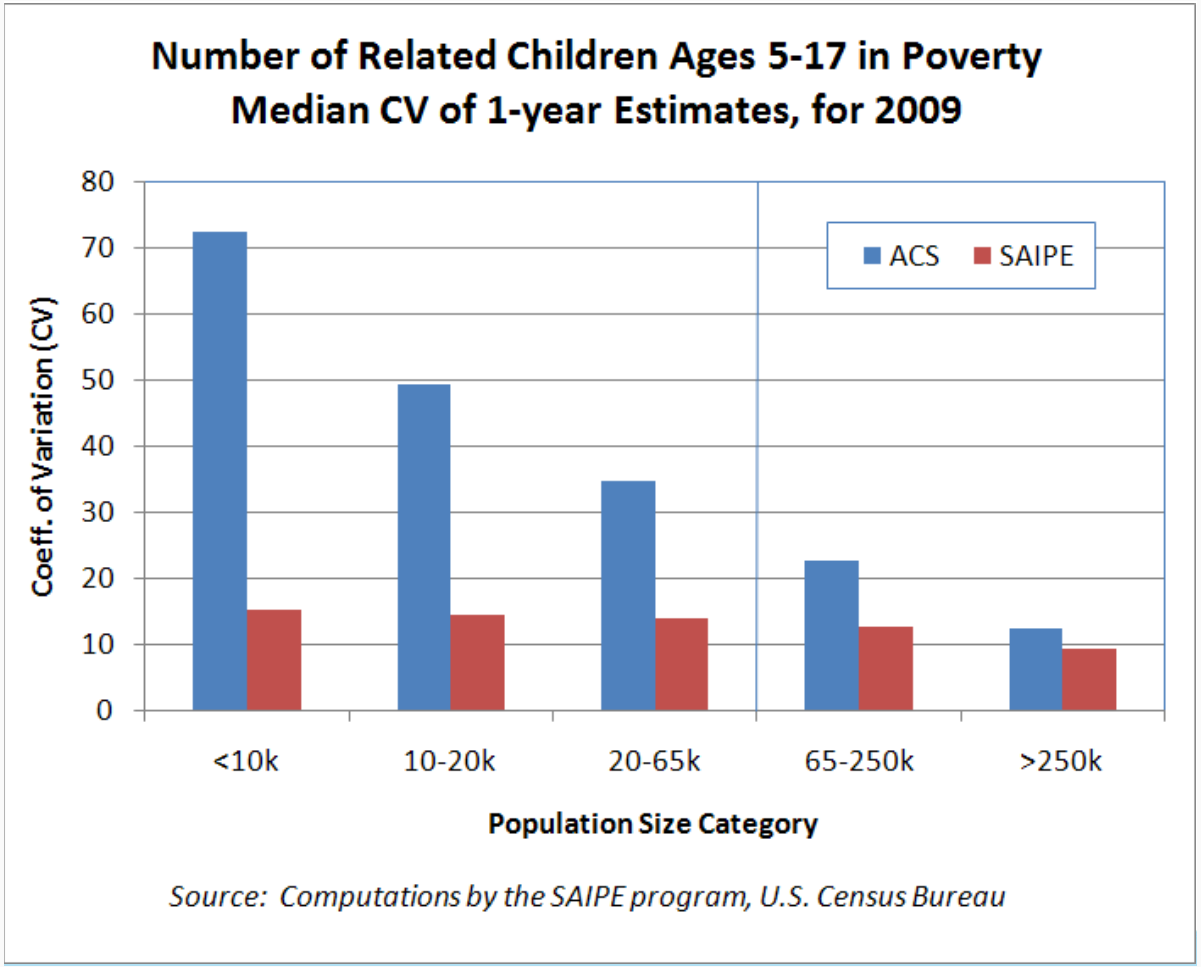
En escala logarítmica, para cada condado:

- 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)

El mapa interactivo de pobreza de SAIPE

En el sitio web de SAIPE, se puede encontrar la herramienta interactiva de SAIPE:
`https://www.census.gov/data-tools/demo/saipe`
Allí, se puede 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



Desafíos/investigaciones relacionadas con SAIPE

- Necesidad de eliminar los condados con estimaciones cero ($\approx 5\%$) debido a la transformación logarítmica; 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. Modelar las tasas directamente en lugar de los conteos logarítmicos. Ver Franco y Bell (2013), y Franco (2020)

- Los datos son inherentemente discretos/posibles mejoras a las suposiciones de normalidad (?)

  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)

Desafíos/investigaciones relacionadas con SAIPE continuación

- 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

$$y_i|p_i, n_i \sim \text{Bin}(n_i, p_i) \qquad i = 1, \ldots, m$$
$$\text{logit}(p_i) = \mathbf{x}_i'\beta + ui \tag{1}$$

- $\text{logit}(p_i) = \log[p_i/(1-p_i)]$, $u_i \overset{i.i.d}{\sim} N(0, \sigma_u^2)$
- 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)
- Puede ser extendido fácilmente a bivariado (Franco y Bell 2013), temporal (Franco y Bell 2015)
- 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, ver 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

Visión general de la Sección 203 de VRA

- 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 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.

Reglas para las determinaciones

Estados - dentro de LMG: LEP $>5\%$ CIT y ILL/LEP $>$ tasa nacional de analfabetismo

Jurisdicción (condado o MCD) - dentro de LMG: :
{numero con LEP $>10,000$ o (Jurisd, LMG) LEP $>5\%$ Jurisd CIT}
y ILL/LEP $>$ tasa nacional de analfabetismo

AIA – dentro de LMG: (AIA,LMG) LEP $>5\%$ (AIA AIAN CIT)
y (AIA,LMG) ILL/LEP $>$ tasa nacional de analfabetismo

El reto del VRA Sección 203

- 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 datos (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

Estrategia de modelado

- 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 II - MLN2

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

$$\left(\tilde{y}_{1i}, \tilde{y}_{2i}, \tilde{y}_{3i}, \tilde{y}_{4i}\right) \sim \text{Multinom}(\tilde{n}_i\,;\, \underline{\omega}_i), \quad i = 1, \ldots, m$$

$$\left(\omega_{1,i}, \omega_{2,i}, \omega_{3,i}, \omega_{4,i}\right) = \left((1-\mu_i), \mu_i(1-\nu_i), \mu_i\nu_i(1-\rho_i), \mu_i\nu_i\rho_i\right)$$

$$\text{logit}(\mu_i) = \mathbf{x}_{1i}\boldsymbol{\beta}_1 + u_{1i}$$

$$\text{logit}(\nu_i) = \mathbf{x}_{2i}\boldsymbol{\beta}_2 + u_{2i}$$

$$\text{logit}(\rho_i) = \mathbf{x}_{3i}\boldsymbol{\beta}_3 + u_{3i}$$

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

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

Reduccion mediana de CV, jurisdicciones, ILL/LEP

Table 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

Desafíos/investigaciones futuras relacionadas con VRA

- Como en otros anõs, 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 ACS
- Para más detalles, Slud, Franco y Hall (2023, en revisión), Slud, Franco, Hall y Kang (2022)

Pérdida de la audición para grupos demográficos en los condados de Estados Unidos– trabajo con David Rein y otros colegas en NORC (articulo en revisión), con beca del Centro para el Control de las Enfermedades

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 cruzados con 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)
- Datos sobre el número de audiologistas por condado
- Otros registros: Archivos de Recursos de Salud del área
- Estimaciones de población 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**

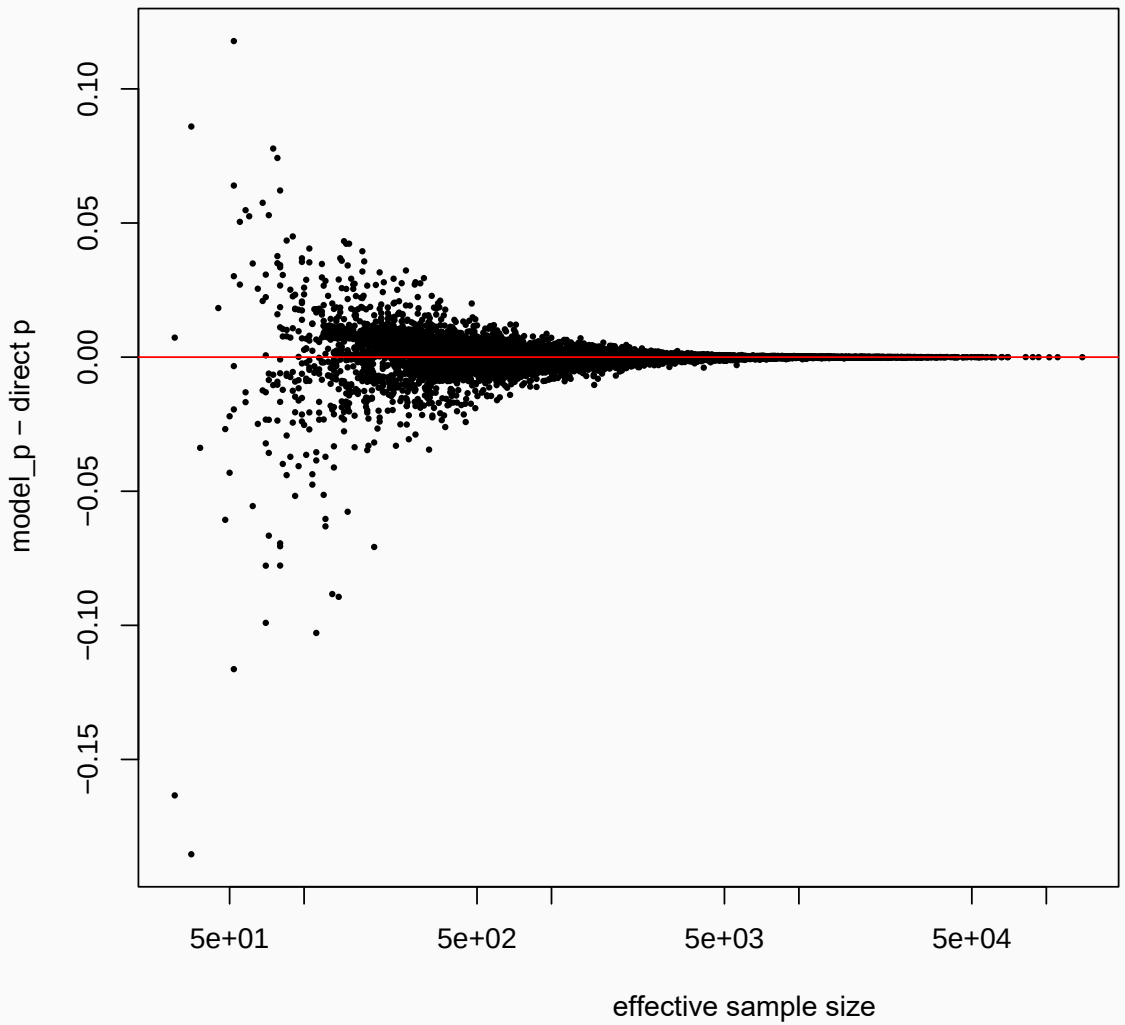
La estrategia de estimación

- 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)
- Obtener totales de población de perdida de audición en "moneda" de ambas encuestas mediante el uso de estimaciones de población directa de ACS y los modelos SAE
- Para las estimaciones del modelo ACS: suposición inicial de prevalencia constante para un grupo de edad y condado

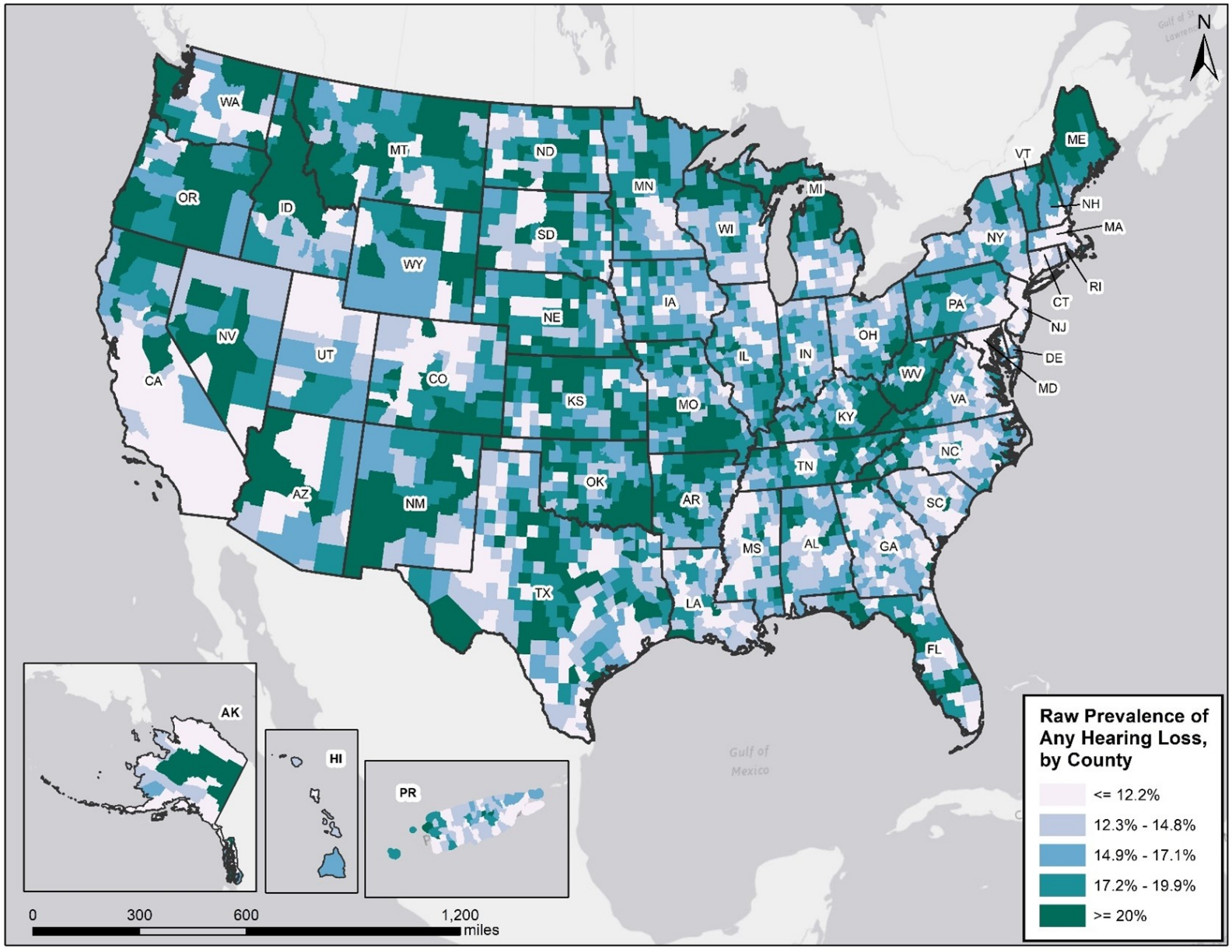
La estrategia de estimación en general–continuación

- 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
- Realizar calibración en cada réplica de cadena de Markov Chain Monte Carlo para obtener una distribución posterior aproximada de las estimaciones calibradas
- Para capturar la incertidumbre en las estimaciones de población de la ACS, simulamos un modelo lognormal con la varianza de muestreo y la media del estimador de muestreo
- Consideramos hacer un "modelo de conexión" formal (e.g., Erciulescu Opsomer y Breidt 2021), pero temas de convergencia lo dificultaron

Predicciones del modelo ACS



Mapa a nivel de condados



Discusión de modelos de la pérdida de audición

- Nuestras medidas de incertidumbre capturaron aproximadamente todas las fuentes cuantificables de error.
- Nuestra metolodogía 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
- Quedan varios temas de investigación pendientes: modelos "de conexión" al nivel del estado; análisis mas cuidadoso de nuestras medidas de incertidumbre, etc.

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
- La buena información auxiliar es clave

Discusión de las tres aplicaciones

- Los supuestos de normalidad no siempre son adecuados
- Cuando se trata de entender las desigualdades en la sociedad usando SAE, 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

Elementos de un programa exitoso de SAE

- Normalmente, una encuesta que mida la cantidad de interés aproximadamente sin sesgo, pero con un tamaño de muestra limitado
- Buena información auxiliar y experiencia en ella
- Experiencia en estimación de áreas pequeñas o acceso a gente con experiencia
- Experiencia en el tema de aplicación en cuestión o acceso a gente con experiencia
- Por supuesto, estimaciones de calidad requieren fuentes de datos de calidad
- Hemos visto muchísimos ejemplos de los beneficios del uso de SAE

Hacia una mayor aplicación de SAE en indicadores sociales

- SAE abre las puertas a producir estadísticas con mayor desagregación
- Relevante para entender las desigualdades en nuestras sociedades
- SAE puede informar intervenciones locales – basadas en datos reales – para problemas sociales
- Descubrimiento global del potencial de SAE
- El estudio de SAE está mas activo que nunca
- Estamos en un época de "proliferación de información", lo cual representa una oportunidad para implementar SAE
- Espero que estas charlas tan interesantes los inspire a considerar una mayor aplicación de SAE

Gracias a CEPAL y cetic.br

Preguntas?

Franco-Carolina@norc.og

NORC at the University of Chicago