Desagregación de datos: Posibilidades y limitaciones de las encuestas de hogares y métodos de estimación en áreas pequeñas

Xavier Mancero
División de Estadísticas
CEPAL

NACIONES UNIDAS
CEPAL

Proyecto “Enfoques innovadores para examinar la desigualdad mediante la integración de diferentes fuentes de datos en América Latina y el Caribe”

Motivación

- La implementación de los ODS pone énfasis en "no dejar a nadie atrás“, para superar las desigualdades que caracterizan a la región.
  - Para ello, las estadísticas deben estar desagregadas por características relevantes de la población.
  - "Los indicadores de los Objetivos de Desarrollo Sostenible deben desglosarse, cuando corresponda, por ingreso, género, edad, raza, etnia, estatus migratorio, discapacidad y ubicación geográfica; u otras características, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales".
- El conocimiento sobre las desigualdades está limitado por la disponibilidad de información:
  - Encuestas de hogares subestiman la desigualdad en la distribución de los ingresos, ya que no captan adecuadamente a los hogares más ricos.
  - Capacidad limitada para proporcionar información desagregada para grupos de población y áreas geográficas específicas.
  - Aprovechamiento insuficiente de datos satelitales e información geográfica para cuantificar y visibilizar las desigualdades.

Posibilidades del uso combinado de fuentes de información

- Desarrollos metodológicos recientes para combinar información de diferentes fuentes de datos, tales como encuestas de hogares, censos de población, registros administrativos o Cuentas Nacionales.
- Uso de datos satelitales e información geográfica para producir estadísticas y representación de la información a través de mapas.
- La disponibilidad de mayor y mejor información sobre las desigualdades es un insumo valioso para las políticas públicas que apuntan a corregirlas.

Componentes del proyecto

Medición de desigualdad del ingreso con datos de EH, registros tributarios y Cuentas Nacionales

Integración de información estadística y geoespacial

Desagregación de información mediante metodologías de “estimación de áreas pequeñas”

Actividades del proyecto

Posibilidades y limitaciones de las encuestas de hogares para la desagregación de datos

Las encuestas de hogares como fuente de información

- Una de las principales fuentes de información sobre las condiciones de vida.
- Útil para producir desagregaciones para diferentes grupos de población.
- Una encuesta se planea para generar información para dominios de estudio predefinidos.
  - Ej.: Tasa de desempleo, por área urbana y rural
- Sin embargo, podríamos estar interesados en estimaciones para subgrupos de población que no se abordaron en el diseño.
  - Ej.: Asistencia escolar de los niños (6-12 años), por quintil de ingreso
- En estos casos, la desagregación podría no ser viable:
  - Falta de información (sin casos observados)
  - Baja precisión

Precisión de las estimaciones

- Una encuesta es una colección de datos para un subconjunto, o una muestra, de una población finita.
- Error de muestreo: diferencia entre la estimación de la encuesta y el verdadero parámetro de la población.
  - Cuanto mayor sea el error de muestreo, menor será la precisión.

Intervalo de confianza

- El tamaño del error de muestreo se puede representar mediante el intervalo de confianza.
- Indica el posible rango de valores en el que es más probable que se encuentre el valor real del parámetro
  - "más probable" suele significar una probabilidad del 90%, 95% o 99%
- Un intervalo de confianza del 95% para un parámetro de interés (θ) viene dado por la siguiente expresión:

$$\left(\hat{\theta} - t_{0.975,gl} \times se(\hat{\theta}) \quad , \quad \hat{\theta} + t_{0.975,gl} \times se(\hat{\theta})\right)$$

- Donde θ es un estimador para el parámetro de interés, $t0.975,gl$ es el percentil 0,975 de una distribución t-student con gl grados de libertad (UPM – estratos) y $se(\theta)$ es el error estándar.
- Los intervalos de confianza nos permiten inferir la precisión de un estimador

Coeficiente de variación

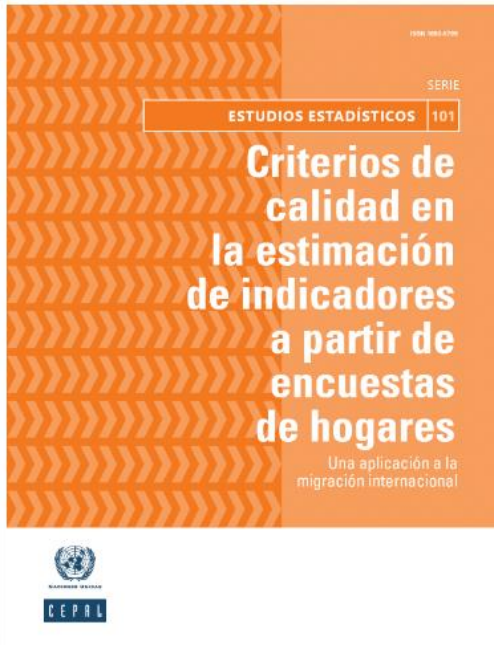
- Otro indicador útil de la precisión de una estimación es el coeficiente de variación.
- El coeficiente de variación es una medida de error relativa a un estimador, definida como:

$$cve(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}}$$

- Muchas Oficinas Nacionales de Estadística utilizan el CV como criterio para evaluar la precisión
  - CV > x, donde x toma valores típicamente entre 10 % y 30 %.

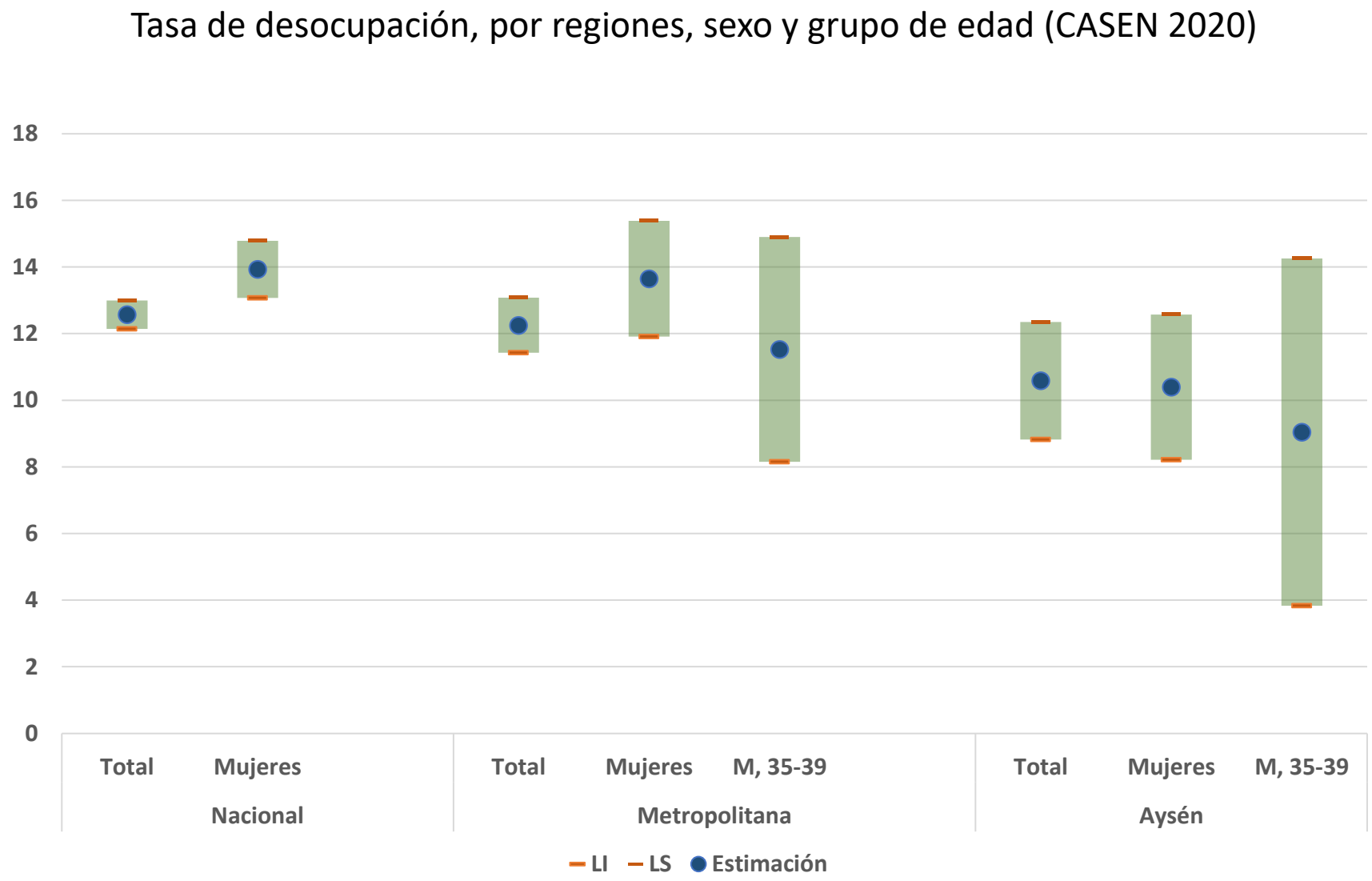
Criterios adicionales para evaluar la precisión

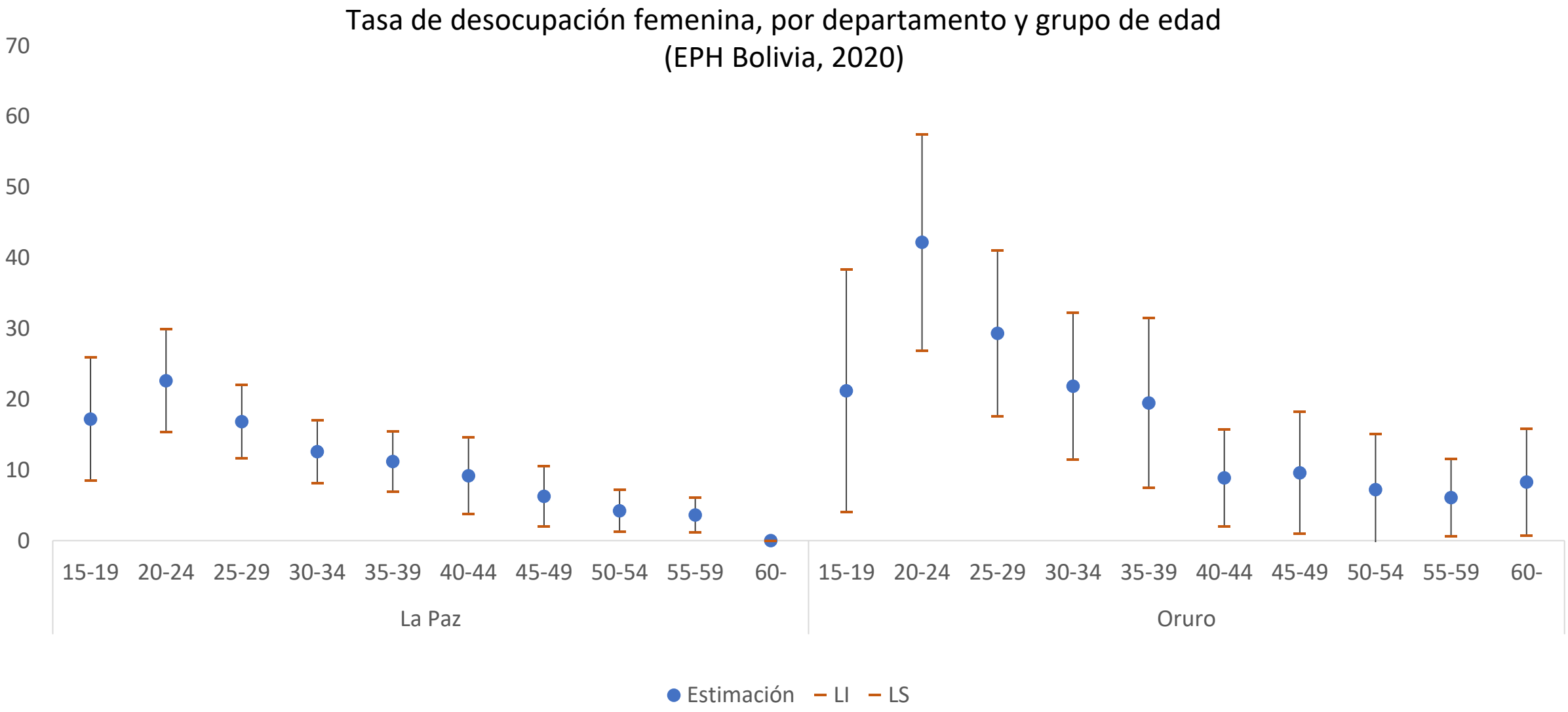
- Para evaluar si una estimación tiene la calidad suficiente para ser publicada, es conveniente tomar en consideración criterios adicionales:
  - Tamaño de muestra
  - Tamaño de muestra efectivo
  - Grados de libertad
  - Coeficiente de variación logarítmico
  - Conteo de casos no ponderado
- El INE Chile ha implementado la librería “calidad” en R para evaluar la precisión de las estimaciones (https://cran.r-project.org/web/packages/calidad/index.html)



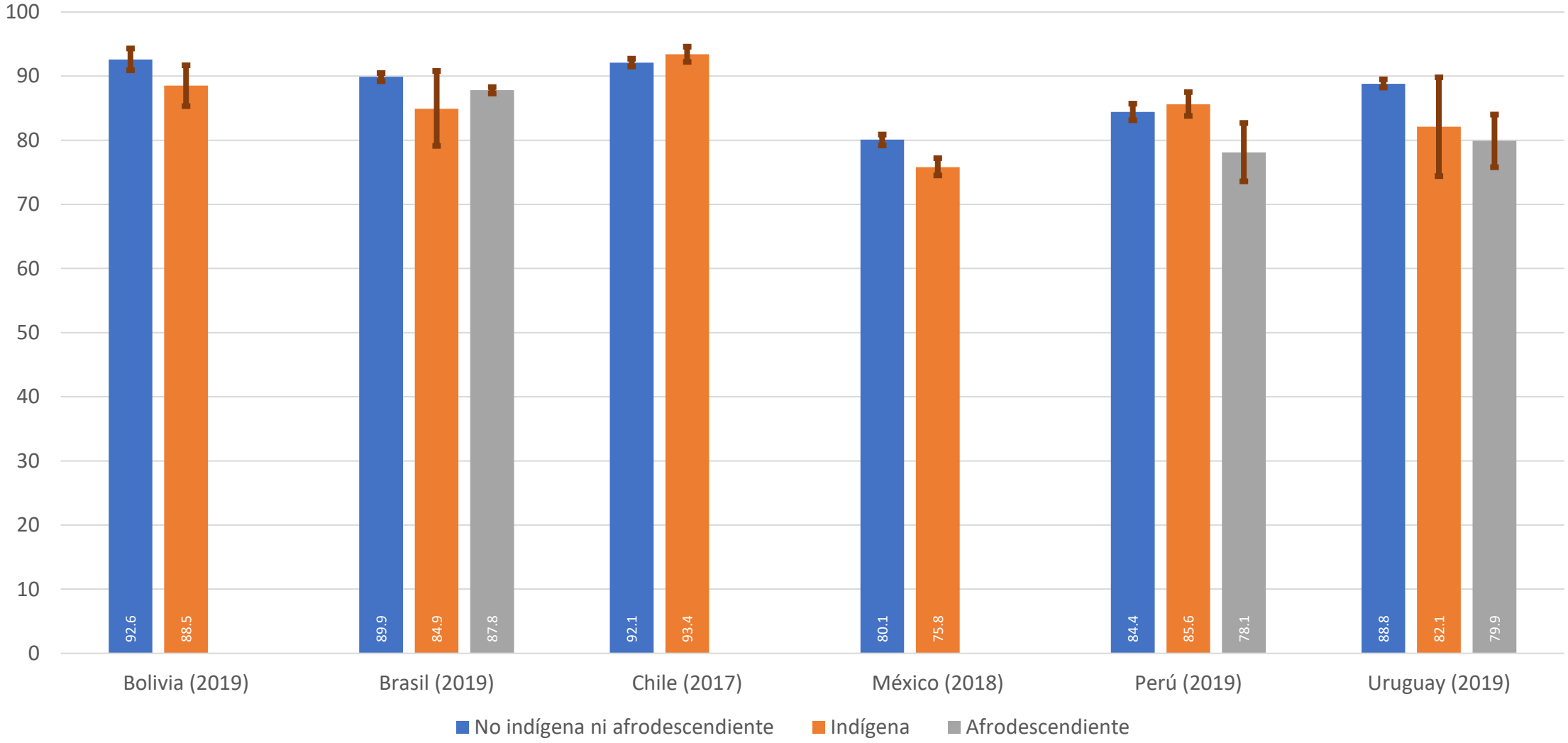
https://repositorio.cepal.org/handle/11362/45681

Algunos ejemplos




Tasa de desocupación femenina, por departamento y grupo de edad
(EPH Bolivia, 2020)
70
60
50
40
30
20
10
0
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-
La Paz
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-
Oruro
Estimación
LI
LS

Tasa de asistencia a la educación secundaria, población de 12 a 18 años de edad, por grupo étnico o raza



Estimaciones de áreas pequeñas (SAE) para la desagregación de datos

- Modelos estadísticos que permiten aprovechar la capacidad de desagregación de otras fuentes de datos:
  - Censos de población
  - Registros administrativos
- Un área pequeña es un dominio para el cual el tamaño de muestra específico no es lo suficientemente grande para obtener estimaciones confiables.
- Suelen ser dominios de estudio no planificados, con un tamaño de muestra esperado aleatorio.
- La subpoblación de interés puede ser un área geográfica o un subgrupo socioeconómico.
  - Geográfico: Provincia, municipio, etc.
  - Subgrupos: Desagregación por edad x género x grupo étnico dentro de un área.

¿Cómo funciona SAE?

- 1. Identificar las variables auxiliares ***x*** que están disponibles en la encuesta y en la fuente de datos complementaria.
- 2. Estimar un modelo para predecir la variable de interés ***y***, utilizando las variables auxiliares identificadas ***x***, con la encuesta.
- 3. Aplicar los parámetros estimados a la fuente complementaria, para predecir la variable de interés ***y*** al nivel de desagregación deseado.
- 4. Medir el error de estimación y evaluar la confiabilidad de los resultados.

Modelos de área y modelos de unidad

- Modelos a nivel de área: predicen la variable de interés ***y*** a partir de variables auxiliares ***x*** agregadas a nivel de área (ej. comuna).
- Modelos a nivel de unidad: predicen la variable ***y*** usando variables auxiliares ***x*** para cada individuo y luego se agregan al nivel de área deseado.
- Los modelos a nivel de área también se pueden usar con datos a nivel de unidad, si la información se resume en el nivel de área apropiado.
- Los modelos a nivel de unidad pueden incorporar variables a nivel de área.

Comentarios finales

- La demanda por información desagregada suele exceder la capacidad de las encuestas para producir estimaciones precisas.
- Los métodos SAE ofrecen una forma práctica de producir desagregaciones que no son alcanzables mediante una estimación directa.
- Todos los métodos SAE requieren datos auxiliares al nivel del área pequeña de interés.
- Algunos aspectos a tener en consideración:
  - Obtener variables auxiliares con alto grado de asociación con la variable de interés.
  - Es importante estimar las medidas de error para SAE. Pero también tener en consideración que pueden existir otras fuentes de error.
  - Puede no ser posible usar las mismas fuentes de datos y modelo estadístico a lo largo del tiempo.