Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

Objetivos de Desarrollo Sostenible y limitaciones de las encuestas

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

Objetivos de Desarrollo Sostenible

Agenda 2030

Figura1: Los 17 ODS

Algunas metas del ODS1 (Poner fin a la pobreza)

- De aquí a 2030, erradicar para todas las personas y en todo el mundo la pobreza extrema (actualmente se considera que sufren pobreza extrema las personas que viven con menos de 1,25 dólares de los Estados Unidos al día).
- De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales.

Algunas metas del ODS2 (Hambre cero)

- De aquí a 2030, poner fin al hambre y asegurar el acceso de todas las personas, en particular los pobres y las personas en situaciones de vulnerabilidad, incluidos los niños menores de 1 año, a una alimentación sana, nutritiva y suficiente durante todo el año.
  - Prevalencia de la subalimentación.
  - Prevalencia de la inseguridad alimentaria moderada o grave en la población, según la Escala de Experiencia de Inseguridad Alimentaria.

Algunas metas del ODS8 (Empleo decente)

- Promover políticas orientadas al desarrollo que apoyen las actividades productivas, la creación de puestos de trabajo decentes, el emprendimiento, la creatividad y la innovación, y fomentar la formalización y el crecimiento de las microempresas y las pequeñas y medianas empresas, incluso mediante el acceso a servicios financieros.
  - Proporción del empleo informal en el empleo no agrícola, desglosada por sexo.

Algunas metas del ODS8 (Empleo decente)

- De aquí a 2030, lograr el empleo pleno y productivo y el trabajo decente para todas las mujeres y los hombres, incluidos los jóvenes y las personas con discapacidad, así como la igualdad de remuneración por trabajo de igual valor.
  - Ingreso medio por hora de mujeres y hombres empleados, desglosado por ocupación, edad y personas con discapacidad.

Algunas metas del ODS8 (Empleo decente)

- De aquí a 2030, lograr el empleo pleno y productivo y el trabajo decente para todas las mujeres y los hombres, incluidos los jóvenes y las personas con discapacidad, así como la igualdad de remuneración por trabajo de igual valor.
  - Tasa de desempleo, desglosada por sexo, edad y personas con discapacidad.

Principio fundamental de la desagregación de datos

Los indicadores de los Objetivos de Desarrollo Sostenible deberán desglosarse, siempre que sea pertinente, por ingreso, sexo, edad, raza, etnicidad, estado migratorio, discapacidad y ubicación geográfica, u otras características, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales.

Resolución de la Asamblea General - 68/261

No dejar a nadie atrás

Figura2: Desagregación de indicadores en los 17 ODS

Principios fundamentales de las estadísticas oficiales

La confianza esencial del público en la integridad de los sistemas estadísticos oficiales y la credibilidad que este otorga a las estadísticas dependen en gran medida del respeto de los valores y principios fundamentales que son la base de toda sociedad que procura entenderse a sí misma y respetar los derechos de sus miembros y que, en este contexto, son cruciales la independencia profesional y la rendición de cuentas de los organismos de estadística.

Resolución de la Asamblea General - 68/261

No dejar a nadie atrás

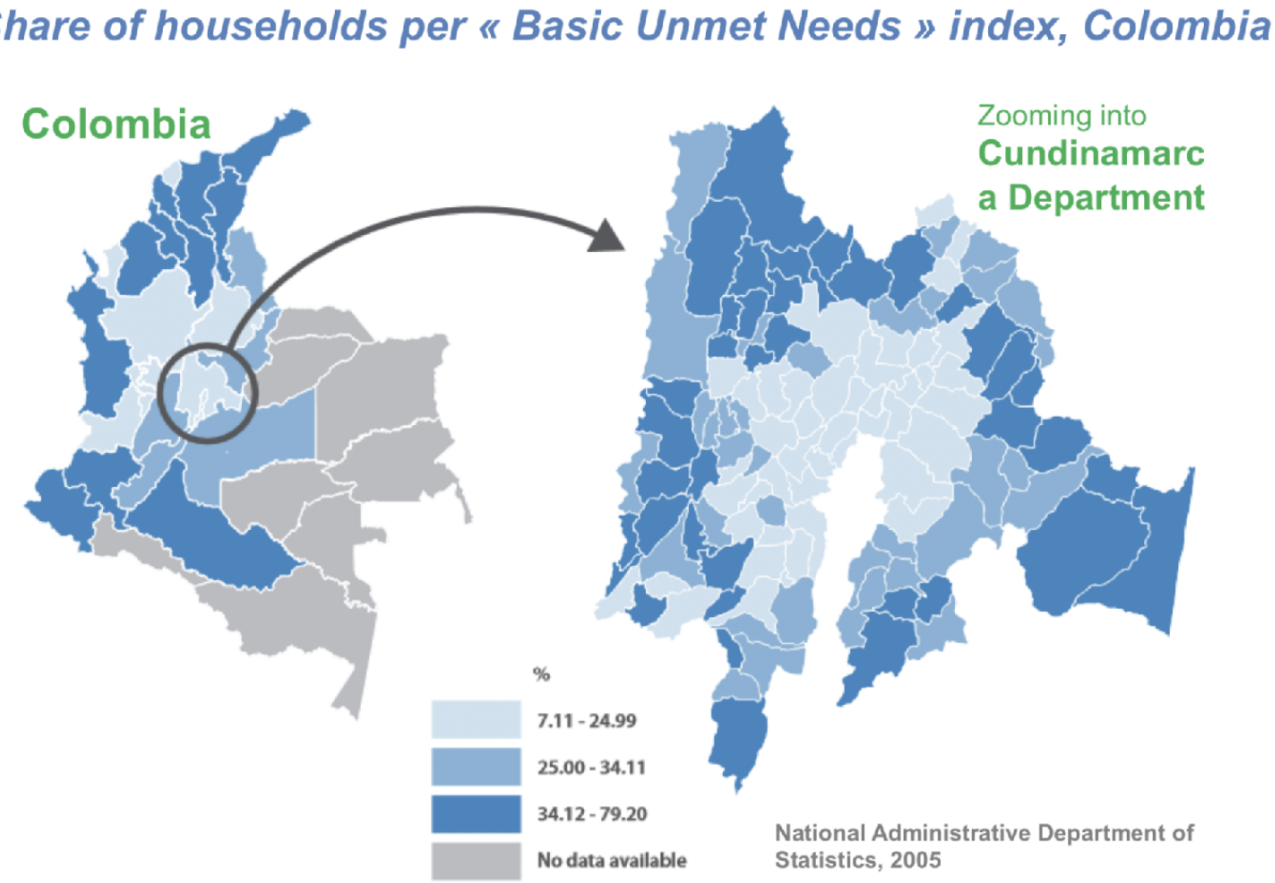

Figura3: Desagregación de un indicador en áreas pequeñas en Colombia. Fuente: UNSD

Algunas metas del ODS17 (Alianzas para lograr los objetivos)

- De aquí a 2020, mejorar el apoyo a la creación de capacidad prestado a los países en desarrollo, incluidos los países menos adelantados y los pequeños Estados insulares en desarrollo, para aumentar significativamente la disponibilidad de datos oportunos, fiables y de gran calidad desglosados por ingresos, sexo, edad, raza, origen étnico, estatus migratorio, discapacidad, ubicación geográfica y otras características pertinentes en los contextos nacionales.

Limitaciones de las encuestas

¿Qué es el coeficiente de variación?

El coeficiente de variación es una medida de error relativo a un estimador, se define como:

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

Muchas veces se expresa como un porcentaje, aunque no está acotado a la derecha, y por eso es conveniente a la hora de hablar de la precisión de una estadística que viene de una encuesta.

Uso del coeficiente de variación

Sarndal et. al.(2003) afirma que un estadístico puede expresar su opinión acerca de que *un coeficiente de variación del 2 % es bueno, considerando las restricciones de la encuesta, mientras que un valor del coeficiente de variación de 9 % puede ser considerado inaceptable.*

De esta forma, muchos institutos nacionales de estadística alrededor del mundo han considerado que las precisiones de las estadísticas resultantes de una encuesta estén supeditadas al comportamiento de su coeficiente de variación.

Alertas sobre el coeficiente de variación

Interpretación	Semaforización	Viviendas / Hogares DGES / DGEGSPyJ
Buena		[0%, 15%)
Aceptable		[15%, 25%)
Con reserva		>= 25%

Figura4: Fuente: INEGI

Alertas sobre el coeficiente de variación

Coeficiente de variación (%)	Número de Observaciones	
	Bajo	Alto
[20 , 100]	Estimador no confiable	Estimador no confiable
[15 , 20)	Estimador no confiable	Descriptivo
[5 , 15)	Descriptivo	Estimador confiable
(0 , 5)	Estimador confiable	Estimador confiable

Figura5: Fuente: INE - Chile

Estándares de alerta en algunos países (encuestas de hogares)

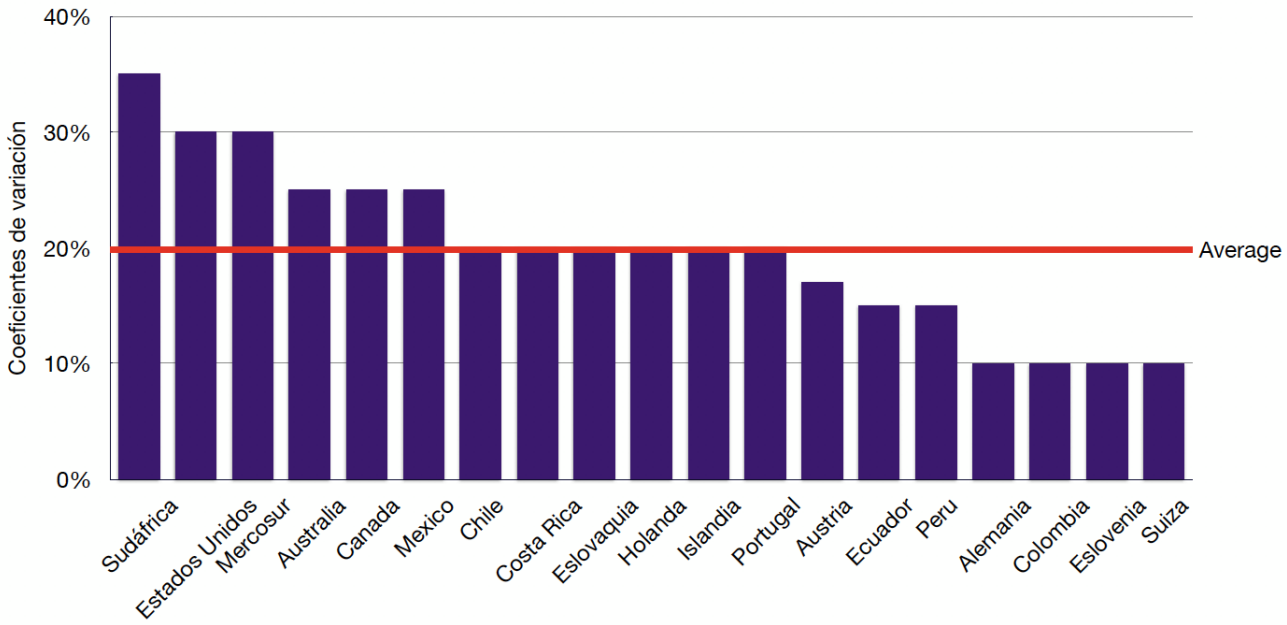

Figura6: Alertas sobre los coeficientes de variación

Algunas alertas definidas en la publicación

Cuando se sobrepasa el umbral del coeficiente de variación aparecen algunas de las siguientes alertas:

- No se publica
- Usar con precaución.
- Las estimaciones requieren revisiones, no son precisas y se deben usar con precaución.
- Poco confiable, menos preciso.
- No cumple con los estándares de publicación.
- Con reserva, referencial, cuestionable.
- Valores muy aleatorios, estimación pobre.

Dominios de estudio y subpoblaciones de interés

Una encuesta se planea con el fin de generar información precisa y confiable en los dominios de estudio que se han predefinido. Sin embargo, existen subgrupos poblacionales que la encuesta no abordó en su diseño, y sobre los cuales se quisiera una mayor precisión.

- Incidencia de la pobreza desagregado por departamento o provincia (tamaño de muestra conocido y planificado).
- Tasa de desocupación desagregada por sexo (tamaño de muestra aleatorio, pero planificado).
- Tasa de asistencia neta estudiantil en primaria desagregada por quintiles de ingreso (tamaño de muestra aleatorio).

Precisión de los estimadores

Debido a que una encuesta es una investigación parcial sobre una población finita, es necesario saber que:

- A partir de una encuesta, no se calculan indicadores, sino que se estiman con ayuda de los datos de la encuesta.
- Es necesario calcular el grado de error que se comete al no poder realizar una investigación exhaustiva. Este error es conocido como el error de muestreo.
- La precisión de un estimador está supeditada al intervalo de confianza.

Entre más angosto sea el intervalo, más precisión se genera y por ende se tiene un menor error de muestreo.

El intervalo de confianza en subpoblaciones

Si el parámetro de interés sobre el cual se busca realizar la inferencia es θ_d, y se ha definido una subpoblación de interés U_d, entonces un intervalo del 95 % de confianza sobre esa subpoblación está dado por la siguiente expresión:

$$\left(\hat{\theta} - t_{0975,gl} \times se(\hat{\theta}) \quad , \quad \hat{\theta} + t_{0975,gl} \times se(\hat{\theta})\right)$$

El uso del coeficiente de variación como indicador de la confiabilidad de las estadísticas provenientes de encuestas de hogares debería ser complementado con algunas otras medidas que permitan crear reglas de confiabilidad y precisión.

El intervalo de confianza

Nótese que la longitud de los intervalos de confianza induce la seguridad de que un estimador es preciso:

- La incidencia de la pobreza en el departamento del país se estimó en 5.2 %, con un intervalo de confianza de (5.15 %, 5.25 %).
- La tasa de desocupación en el país para los hombres se ubicó en en 7.5 %, con un intervalo de confianza de (7.1 %, 7.9 %); mientras que para las mujeres se ubicó en 9.2 %, con intervalo de confianza de (8.8 %, 9.6 %).
- La tasa de asistencia neta estudiantil en primaria para el último quintil de ingreso se estimó en 85 %, con un intervalo de confianza de (48.2 %, 100.0 %).

El tamaño de muestra

- El tamaño de muestra afecta de manera indirecta la amplitud del intervalo de confianza, a través del error estándar que generalmente decrece a medida que el tamaño de muestra se hace más grande.
- Un tamaño de muestra adecuado garantiza la convergencia en distribución de los estimadores a la distribución teórica de donde se calculan los percentiles.
- Por ejemplo, es posible plantear que todas las estimaciones basadas en un tamaño de muestra menor a un umbral predefinido deberían ser suprimidas o marcadas como no confiables.

El tamaño de muestra efectivo

- En las encuestas de hogares, con diseños de muestreo complejos, no existe una sucesión de variables que sean independientes e identicamente distribuidas.
- La muestra $y_1, \ldots, y_n$ no es un vector en el espacio n-dimensional, donde se asume que cada componente del vector puede variar por sí mismo.
- La dimensión final del vector $(y_1, \ldots, y_n)$ es mucho menor que n, puesto que existe una forma jerárquica en la selección de los hogares y a la interrelación de la variable de interés con las UPMs

El tamaño de muestra efectivo

El tamaño de muestra efectivo se define como sigue:

$$n_{efectivo} = \frac{n}{Deff}$$

En donde *Deff* es el efecto de diseño que depende de: 1. El número de encuestas promedio que se realizaron en cada UPM. 2. La correlación existente entre la variable de interés y las mismas UPMs.

Es posible considerar que, si el tamaño de muestra efectivo no es mayor a un umbral, entonces la cifra no debería ser considerada para publicación.

Grados de libertad

Son una medida de cuántas unidades independientes de información se tienen en la inferencia. Nótese que:

- En el caso extremo de realizar un censo en cada UPMs, sin importar el número de individuos que componen el conglomerado, el número de unidades independientes será únicamente el número de UPMs seleccionadas en la primera etapa de muestreo.
- En las encuestas de hogares, la variabilidad de la estimación es la contribución del conglomerado a la gran media más una contribución (considerada insignificante) de la segunda etapa de muestreo.

Grados de libertad

En las subpoblaciones los grados de libertad no se consideran fijos sino variables.

$$gl = \sum_{h=1}^{H} \nu_h \times (n_{lh} - 1)$$

Note que ν_h es una variable indicadora que toma el valor uno si el estrato h contiene uno o mas casos de la subpoblación de interés, n_{lh} es el número de UPMs en el estrato. En el caso más general, los grados de libertad se reducen a la siguiente expresión:

$$gl = \#UPMs - \#Estratos$$

Grados de libertad

Por ejemplo, considere por ejemplo el percentil 0.975 para el cual los valores críticos de la distribución t varían con respecto a sus grados de libertad

- $t - student_{gl=1} = 127$
- $t - student_{gl=2} = 430$
- $t - student_{gl=5} = 257$
- $t - student_{gl=40} = 202$
- $t - student_{gl=\infty} = 196$

Es posible considerar que si los grados de libertad inducidos por la subpoblación son menores a un umbral predefinido, la cifra debería ser suprimida.

Ejemplo

Quintil Urbano	sexo	n	Deff	n.eff	gl	Desocupación%	Li %	Ls %	cv %	Alerta
Quinto	Mujer	2055	1.2	1757	309	1.0	0.4	1.6	30.6	*
Quinto	Hombre	1969	1.1	1738	335	1.1	0.5	1.7	26.3	*
Cuarto	Hombre	2245	1.2	1807	347	2.2	1.4	3.0	19.3	
Cuarto	Mujer	2301	1.6	1466	357	4.1	2.7	5.5	17.5	
Tercero	Mujer	2421	1.5	1646	336	6.1	4.3	7.9	15.1	
Segundo	Hombre	2280	1.4	1654	295	5.9	4.3	7.5	13.8	
Tercero	Hombre	2351	1.2	2025	331	4.6	3.4	5.8	13.3	
Segundo	Mujer	2541	1.6	1547	310	10.8	8.0	13.6	13.1	
Primero	Mujer	2862	2.0	1466	266	20.0	15.4	24.6	11.8	
Primero	Hombre	2562	1.6	1610	263	11.9	9.4	14.5	10.9	

Figura7: Desocupación urbana por quintiles de ingreso y sexo

Ejemplo

Quintil Rural	sexo	n	Deff	n.eff	gl	Desocupación%	Li %	Ls %	cv %	Alerta
Primer	Mujer	1788	0.6	2754	140	0.8	0.1	1.5	44.5	*
Cuarto	Hombre	2112	1.7	1223	178	1.8	0.8	2.7	26.7	*
Segundo	Hombre	2281	1.8	1236	156	2.6	1.3	3.9	25.7	*
Primer	Hombre	1704	1.3	1324	137	2.8	1.4	4.2	25.5	*
Quinto	Mujer	1780	1.3	1391	166	2.6	1.4	3.9	23.7	*
Segundo	Mujer	2195	1.4	1579	158	5.3	2.9	7.6	22.7	*
Quinto	Hombre	2127	1.4	1553	171	1.8	1.0	2.6	22.1	*
Tercero	Hombre	2180	2.0	1068	169	3.7	2.1	5.2	21.4	*
Tercero	Mujer	2023	1.4	1411	164	6.5	3.8	9.2	20.8	*
Cuarto	Mujer	1942	1.6	1225	174	7.3	4.6	10.0	19.1	

Figura8: Desocupación rural por quintiles de ingreso y sexo

Uso de métodos SAE

Justificación

- Los estimadores directos, basados solo en unidades de muestreo observadas para cada área pequeña, no son suficientemente confiables.
- Tamaño de muestra pequeño o incluso ninguna unidad observada (falta de información).
- El coeficiente de variación (CV) es demasiado alto para el indicador objetivo a nivel de área.

Incremento del coeficiente de variación

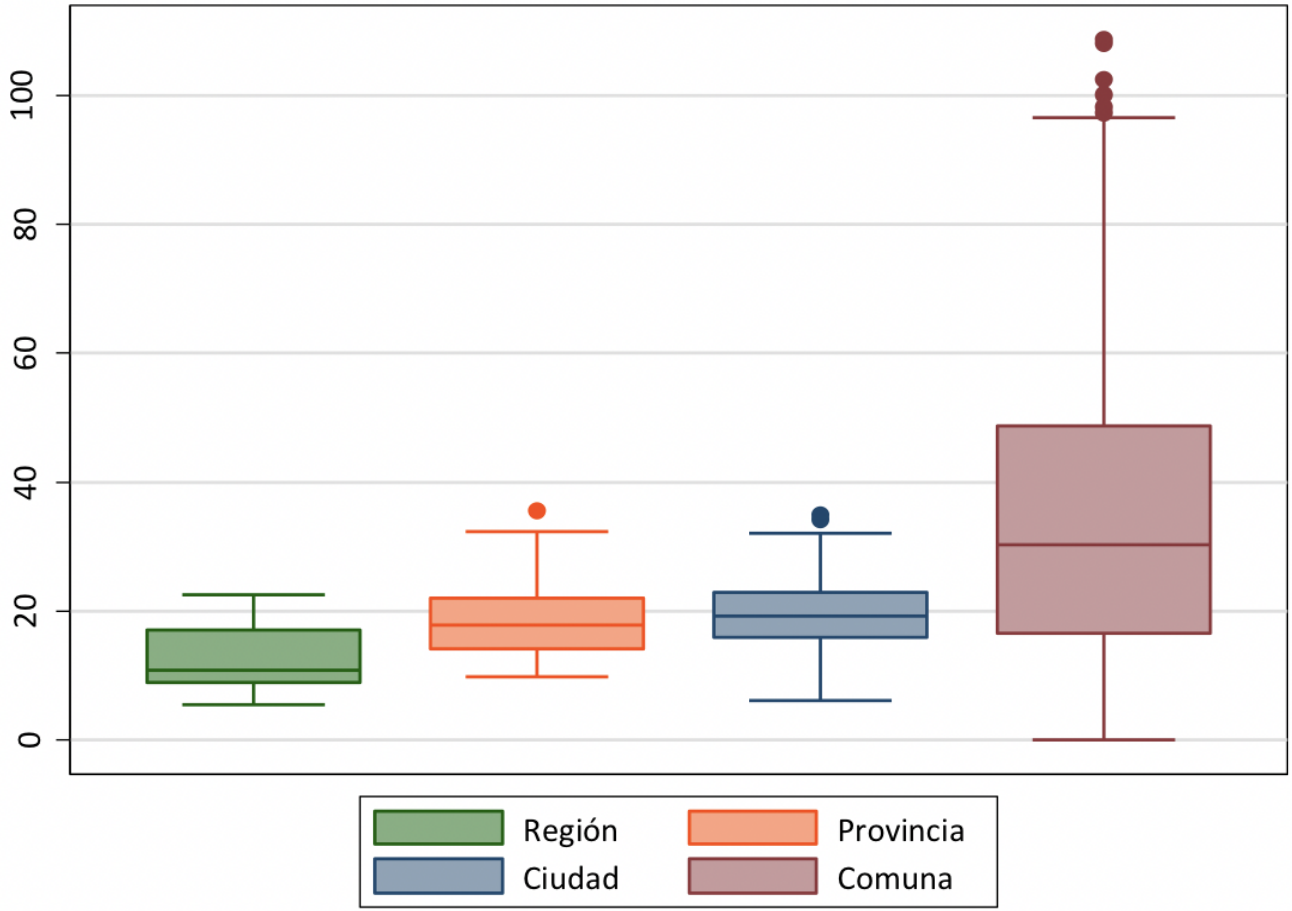

Figura9: Distribución de los coeficientes de variación en Chile

Justificación

Área	Obs	Mean	Std. Dev.	Min	Max
region	15	12,0	4,9	5,5	22,5
provincia	26	19,5	7,1	9,8	35,5
ciudad	33	20,0	6,5	6,0	34,9
comuna	237	34,8	26,0	0,0	108,7

Figura10: Coeficientes de variación en Chile

Justificación

Cuando los estimadores directos no son confiables para algunos dominios de interés, existen dos opciones:

1. Sobremuestreo: aumentar el tamaño de la muestra en los dominios de interés (aumento de los costos).
2. Aplicar técnicas estadísticas que permitan estimaciones confiables en esos dominios, métodos SAE.

Justificación

- Durante la última década ha habido una demanda creciente de (objetivo y subjetivo) indicadores de progreso y bienestar.
- Estas medidas desempeñan un papel central para los responsables de las políticas, para planificar y verificar la efectividad de las mismas.

Ejemplos

- Indicadores de pobreza: en riesgo de pobreza, ingreso de los hogares.
- Indicadores del mercado de trabajo: Tasa de desempleo, Satisfacción con el trabajo, etc.
- Indicadores de salud: esperanza de vida media, porcentaje de población con conductas peligrosas (obesidad, fumadores, etc.)

Justificación

- Para ser informativos y efectivos, estos indicadores deben elegirse a el nivel apropiado de desagregación.
- Los indicadores pueden ser desagregados a lo largo de varias dimensiones, incluyendo áreas geográficas, grupos demográficos, grupos de ingresos / consumo y grupos sociales.

¿Qué es un área pequeña?

- La mayoría de las encuestas nacionales están planificadas para entregar estimaciones confiables a nivel nacional y regional pero a niveles más bajos se reduce la precisión.
- Un área pequeña es un dominio para el cual el tamaño de muestra específico no es suficientemente grande para obtener estimaciones confiables.
- Habitualmente son dominios no planificados y su tamaño de muestra esperado es aleatorio y es más grande a medida que aumenta el tamaño de la población del área.

¿Qué es un área pequeña?

La subpoblación de interés puede ser una zona geográfica o subgrupos socioeconómicos.

- Geográfico: provincias, áreas del mercado de trabajo, municipios, sectores censales para medir por ejemplo la tasa de desempleo a nivel comunal.
- Dominio de subgrupos específicos: edad $\times$ sexo $\times$ raza dentro del ámbito geográfico de una zona, para medir por ejemplo la tasa de desempleo por sexo o edad específica en las zonas urbanas.

¿Qué es un área pequeña?

- La solución es **tomar prestada fuerza** de otras áreas y/o en diferentes ocasiones mediante modelos explícitos o implícitos que explotan la relación entre variables aumentando el tamaño efectivo de la muestra.
- El modelo proporciona un enlace a áreas relacionadas y/o períodos de tiempo a través de información complementaria tales como recuentos de censos (recientes o actuales) o registros administrativos relacionados con la variable objetivo.

Algunos métodos

- Estimador sintético: En el contexto de subpoblaciones, los estimadores se llaman sintéticos cuando éstos se basan en un estimador directo y se estiman a partir de información auxiliar a través de un modelo.
- Estimador compuesto: es una combinación lineal entre un estimador directo y un estimador sintético. Representa un buen compromiso entre las características de los dos componentes.

Algunos métodos

- El estimador compuesto está dado por una combinación lineal de estimador sintético y estimador directo equilibrando el sesgo potencial del estimador sintético contra la inestabilidad del estimador directo (compensación entre precisión y sesgo).
- Las estimaciones más grandes de áreas pequeñas están más cerca de las estimaciones directas mientras que las más pequeñas están más cerca de las estimaciones sintéticas.

Algunos métodos

- Los estimadores SAE se dividen en dos tipos principales dependiendo de cómo se aplican los modelos a los datos dentro de las áreas pequeñas: nivel de área y nivel de unidad.
- Los estimadores de área pequeña se basan en cálculos de nivel de área si los modelos vinculan la variable de interés y con variables auxiliares x específicas del área.

Algunos métodos

- Se llaman modelos a nivel de unidad si se vinculan valores individuales para las variables auxiliares específicas de la unidad.
- Los estimadores basados en áreas pequeñas se calculan a nivel de área si los datos de la unidad no están disponibles.
- También pueden ser calculados si los datos de nivel de unidad están disponibles resumiéndolos en el nivel de área apropiado.

Proceso de estimación

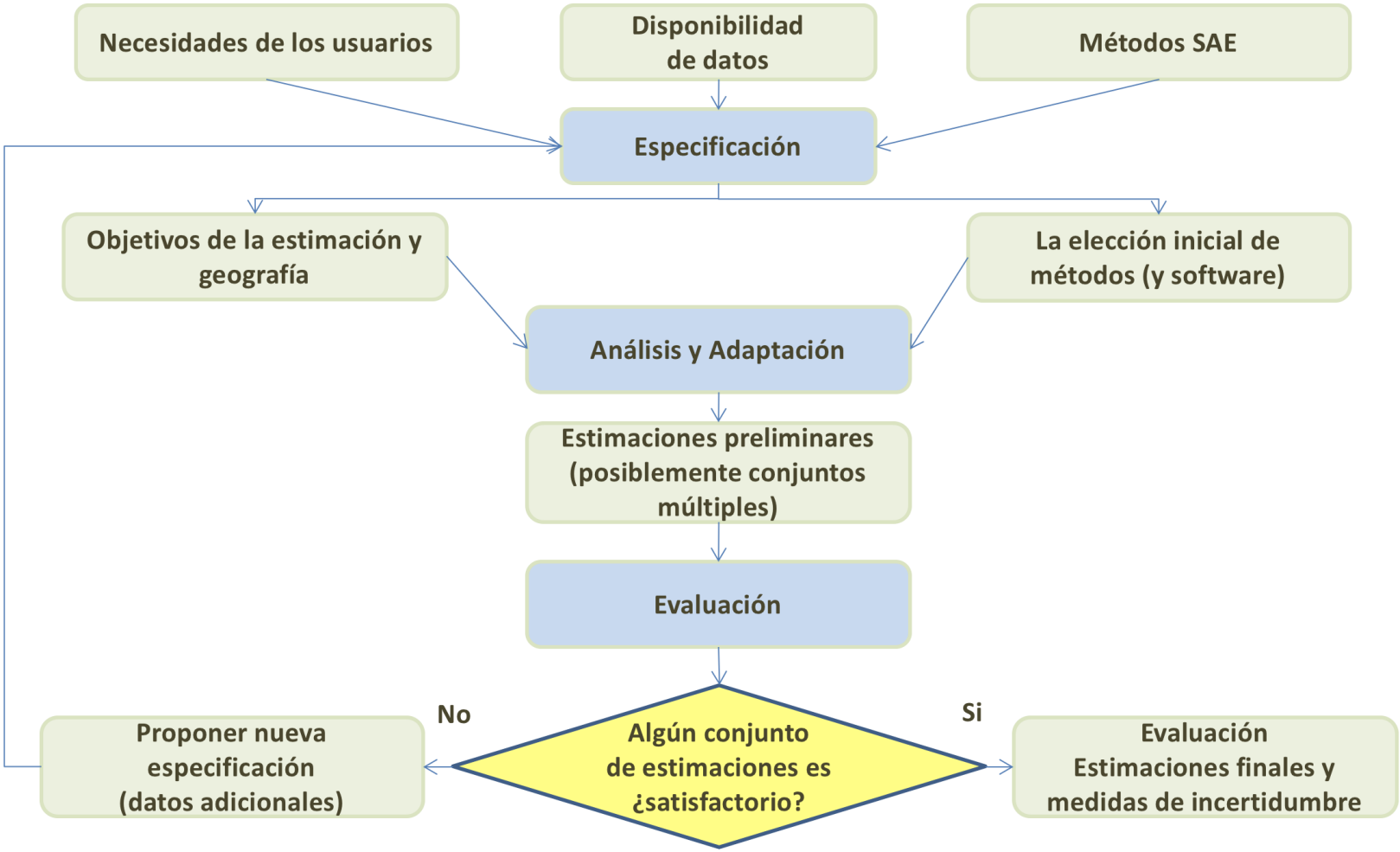

Figura11: Producción de estadísticas con SAE

Algunos riesgos

La producción de estimaciones en área pequeña involucra riesgos que se deben tener en consideración:

- El tamaño de las áreas pequeñas en los términos del número de unidades que les pertenecen es también una consideración importante. Áreas que son demasiado pequeñas pueden presentar problemas de confidencialidad.
- Las estimaciones de área pequeña pueden diferir demasiado de las estadísticas basadas en el conocimiento local.
- Las fuentes de información y el diseño utilizado pueden ser no sustentables en el tiempo.

Algunos riesgos

- El compromiso y la voluntad de la agencia para apoyar estas metodologías a través de sistemas y personal capacitado en la materia.
- Disponibilidad de datos auxiliares correlacionados con la variable de interés.
- Tamaño de muestra de debe ser suficientemente grande para permitir estimaciones confiables mediante el uso de los datos de la encuesta y los datos auxiliares existentes.

Consideraciones

- Todos los métodos SAE requieren datos auxiliares a nivel del área pequeña desde el cual **toman prestada la fuerza**.
- La efectividad de los métodos SAE depende del grado de asociación entre la variable de interés y los datos auxiliares.
- La búsqueda de buenas variables auxiliares es crítica, incluida la construcción imaginativa de tales variables.
- Los datos auxiliares deben medirse de manera consistente a través de las áreas pequeñas, pero pueden incluir estimaciones de muestras grandes con error de muestreo conocido.

Desafíos

- Aumento de las tasas de no respuesta.
- Aumento de costos, menos financiación.
- Aumento de la demanda de estimaciones para dominios pequeños como por raza, etnia o pobreza.
- Aumento de la demanda de estimaciones de áreas pequeñas.
- Aumento de la complejidad en los contenidos de los cuestionarios y por lo tanto la carga de respuesta.
- Aumento de la demanda de análisis secundarios, uso público y archivos de datos de uso restringido.

Algunas aplicaciones actuales

Mapas de pobreza

Estimados por el Banco Mundial en diversos países: Perú, Brasil, Guatemala, Nicaragua, Panamá, Ecuador. Combinan datos provenientes de encuestas con datos censales.

Small Area Income and Poverty Estimates (SAIPE)

El Small Area Income and Poverty Estimates (SAIPE) es un programa del Census Bureau de los Estados Unidos. El programa SAIPE produce estimaciones de pobreza para el total de la población y la media de los ingresos por hogares estimada anualmente para todos los condados y estados.

Estimaciones mensuales de empleo local y estatal

A través del programa de Estadísticas de desempleo del área local, la BLS produce estimaciones mensuales del empleo y desempleo total para aproximadamente 7300 áreas, incluido regiones censales, divisiones, estados, condados y ciudades.

Las estimaciones son basadas en datos de varias fuentes, incluyendo el CPS, el programa de estadísticas de empleo actual (CES), los sistemas de seguro de desempleo de los estados (UI) y el censo decenal.

La experiencia canadiense

Utilizaron una serie de fuentes primarias de datos: Censo Canadiense de población y vivienda, La encuesta de fuerza laboral (LFS) Canadiense y el sistema de seguro gubernamental federal de desempleo (UI). Otros datos de áreas pequeñas de mercado laboral fueron utilizados como información auxiliar.

Dado que el objetivo principal era minimizar el error de estimación del modelo, probaron tres técnicas de estimación: Estimación sintética, SPREE y regresión.

En el Reino Unido

En conjunto con la universidad de Southampton, se realizó una estimación del desempleo basado en la definición de la OIT. Se exploraron distintos enfoques utilizando la alta correlación entre las estimaciones de desempleo de la LFS y el número de demandantes de de beneficios de subsidios para solicitantes de empleo.

El enfoque principal fue desarrollar modelos de regresión que vinculen las estimaciones de desempleo con información de los solicitantes otorgada por las autoridades locales de los distritos (ONS, 2001b).

En Italia

La experiencia de Italia fue un intento de mejorar el rendimiento de la estimación de la tasa de desocupación a nivel sub-regional a través de la LFS del ITSTAT (Italy statistics) utilizando como referencia los dominios que abarcan los estratos que se pueden agregar a los municipios.

La tasa de desempleo se estimó utilizando un modelo lineal mixto con efectos de área espacialmente correlacionados y covariables como el sexo, edad y el desempleo a nivel área del censo anterior.

¡Gracias!

¡Gracias!