GOBIERNO DE COLOMBIA

DANE INFORMACIÓN PARA TODOS 70 AÑOS

Planes para la desagregación de variables DANE

Subdirección

21 abril de 2023

Desagregación de variables en encuestas

La **desagregación de variables** en una encuesta por muestreo probabilístico a un dominio de interés, como: municipios, grupos étnicos, genero, entre otros; se puede realizar a través de un estimador directo, siempre y cuando, **la estimación tenga un coeficiente de variación por debajo del 15%.** Con el fin de garantizar que la estimación es representativa para el dominio de interés

Ejemplo:

La estimación de pobreza monetaria de Sogamoso – Boyacá es XX% con un CV del KK%

Variable de interés = pobreza monetaria
Dominio de interés = Sogamoso – Boyacá
Estimación directa = XX%
Coeficiente de variación = CV del KK%

Dominios

Los dominios están definidos como **subgrupos de la población**, con las siguientes características: i) cada elemento pertenece a un único subgrupo, 2) todos los elementos de la población pertenecen a un subgrupo y 3) la suma de subgrupos componente al total de la población.

En el caso de la muestra de una encuesta **podemos tener dominios con muestras pequeñas o sin información**, por lo que es imposible obtener una estimación directa usando los factores de expansión presentes en los microdatos.

DANE INFORMACIÓN PARA TODOS — El futuro es de todos — Gobierno de Colombia

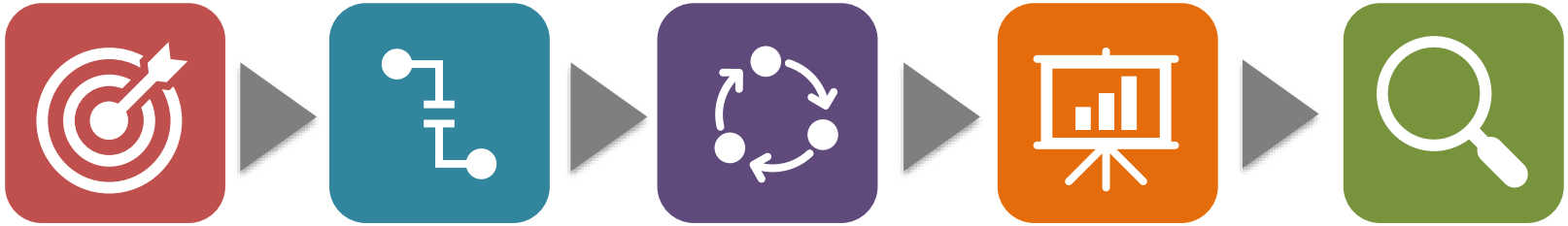
Intervalos de Confianza para la Incidencia de la Pobreza Monetaria

I.C. para la Incidencia Pobreza Monetaria
Principales Dominios y 23 Ciudades y Áreas Metropolitanas (A.M.)
2021

Dominio	Estimación	Error Estándar	Límite Inferior	Límite Superior	Coeficiente de Variación (%)
Armenia	40,5	0,86	38,82	42,19	2,13
Barranquilla A.M.	35,7	0,68	34,33	37,01	1,91
Bogotá	35,8	0,66	34,51	37,09	1,83
Bucaramanga A.M.	35,5	0,95	33,61	37,33	2,68
Cali A.M.	29,3	0,76	27,77	30,74	2,60
Cartagena	40,4	0,77	38,85	41,87	1,91
Cúcuta A.M.	49,0	0,89	47,29	50,77	1,81
Florencia	48,2	0,87	46,49	49,89	1,80
Ibagué	34,3	0,84	32,64	35,93	2,45
Manizales A.M.	30,2	0,80	28,59	31,72	2,65
Medellín A.M.	27,6	0,62	26,34	28,76	2,24
Montería	43,7	0,86	42,01	45,38	1,97
Neiva	42,2	0,95	40,36	44,08	2,25
Pasto	40,1	1,06	38,02	42,19	2,65
Pereira A.M.	35,4	0,80	33,80	36,93	2,26
Popayán	46,3	0,99	44,38	48,27	2,14
Quibdó	64,8	0,88	63,07	66,53	1,36
Riohacha	56,6	0,80	55,02	58,18	1,42
Santa Marta	51,6	0,73	50,13	53,01	1,42
Sincelejo	43,4	0,78	41,89	44,93	1,79
Tunja	37,6	0,95	35,71	39,44	2,53
Valledupar	51,0	0,74	49,54	52,46	1,46
Villavicencio	33,4	0,81	31,84	35,03	2,44

En el caso del dominio de municipios, la GEIH está diseñada para la estimación de la variable pobreza anual solo para las 23 ciudades principales.

Pasos para la implementación



Identificación de la variables de interés y sus dominios

Definición de las fuentes de infromación y construcción de covariables

Definición del modelo de estimación de áreas peuqeñas

Análisis de coeficientes de variación

Consistencia, comparación con las cifras oficiales y validación de expertos

Censo

Para el caso de las estadísticas sociales, el Censo Nacional de Población y Vivienda CNPV se convierte en la fuente de información de referencia para la construcción de un modelo de estimación de áreas pequeñas.

Se actualiza cada de 10 años

Los microdatos contienen información de 44 millones de personas

Contiene información sobre: Salud, Educación, Vivienda Trabajo

Incluye la coordenada del hogar y la pertenencia a grupos etnicos

Periodo de referencia 2018

Registros administrativos

SEN
Sistema Estadístico Nacional-Colombia

Uno de los objetivos del SEN es, propiciar el fortalecimiento y **aprovechamiento de los registros administrativos**, así como el intercambio de información entre los miembros del SEN, como fuente para la **producción de estadísticas oficiales, el mejoramiento de la calidad y la coherencia en las cifras**.

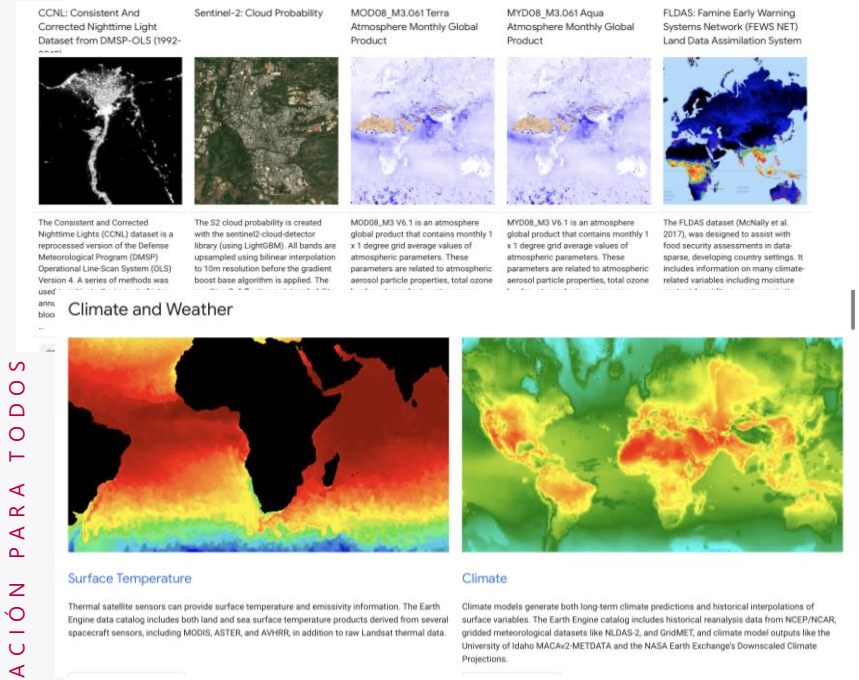
El aprovechamiento estadístico de los RRAA había aumentado antes de la pandemia por COVID-19, la pandemia fue una oportunidad para la innovación.

Uno de los casos es el **RELAB (Registro Estadístico de Relaciones Laborales)**

Estos registros presentan retos en cuanto a la cobertura del total de la población objetivo y la calidad de la información

Fuentes alternas

Google Earth - Imágenes satelitales



Google maps - Sitios de interés

Web scrapping

Estas fuentes también tienen retos de calidad y cobertura

INFORMACIÓN PARA TODOS

Fuentes de información: Modelo a nivel municipal

1. Discapacidad
2. RIPS
3. Equipamientos salud
4. Cobertura

1. Policia Nacional (violencia y seguridad)
2. RNI: víctimas
3. ICFES: puntajes

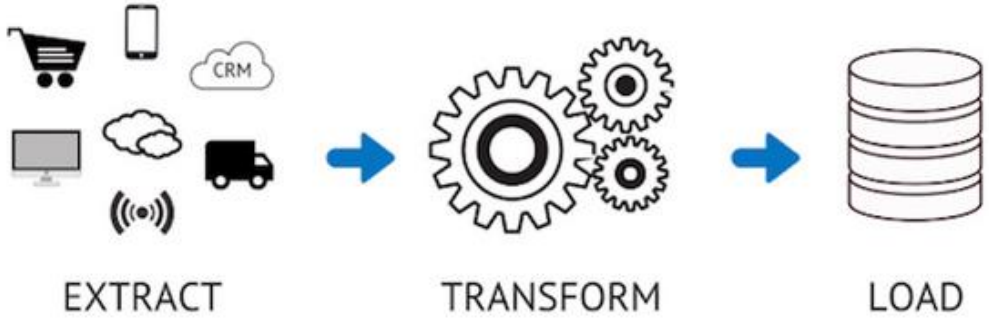
1. Finanzas: Recursos municipales
2. Cobertura neta/bruta en educación
3. Tránsito inmediato a educación superior
4. Caracterización municipal (categoría, ruralidad)
5. Servicios públicos, etc

1. Proyecciones poblacionales
2. Cuentas nacionales: Valor agregado
3. Estadísticas vitales
4. Caracterización geográfica
5. EDUC

Integración de datos

1. Se deberán definir las fuentes a usar para la construcción de co-variables
2. Análisis de calidad de las fuentes y construcción de indicadores
3. Construcción de tabla(s) de entrada al modelo

Un servicio de nube reduciría el tiempo para el procesamiento de la ETL



Fuente: https://mirelman.com/es/integracion-de-datos/

Retos:

1. Efectividad de los cruce de información
2. Actualización de información para definir periodicidad del proceso y pertinencia en la entrega de las fuentes
3. Calidad de datos

Modelos de estimación de áreas pequeñas

La estimación de áreas pequeñas es una técnica estadística que permite desagregar a subpoblaciones de las no se cuenta con una muestra probabilística grande o no hay información para la estimación de dominios a través de inferencia basada en el diseño. Está técnica se divide en dos enfoques para la estimación de parámetros:

Modelos de área

Modelos de unidad

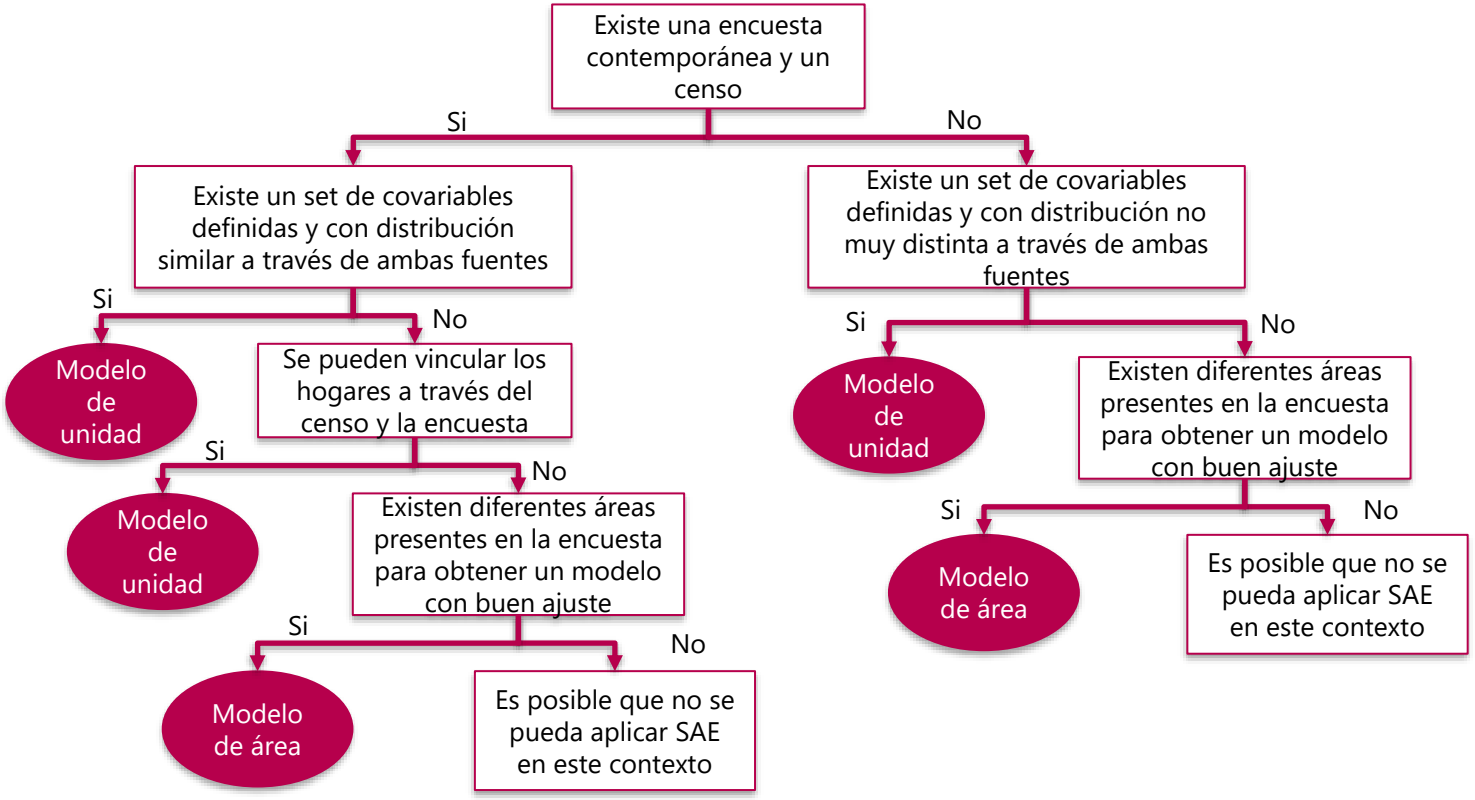
Modelos de área

Las covariables usadas en estos modelos se encuentran agregadas a la subpoblación de interés. Se usan al no tener información a nivel de unidad.

Modelos de unidad

Relacionan los valores unitarios de una variable de estudio con co-variables unitarias

Como elegir el modelo adecuado



Fuente: https://openknowledge.worldbank.org/server/api/core/bitstreams/1d1fcadc-43e3-541b-8949-fea45dd2a528/content

Modelo básico de estimación en áreas pequeñas

Modelo de Fay-Herriot es un modelo de área basado en los modelos lineales mixtos. Donde

Estimación directa

$$y_d = \mu_d + e_d$$

Tradicionalmente se tienen las estimaciones directas que usualmente no pueden estimar todos los dominios de interés

Estimación por modelos

$$\mu_d = x_d\beta + u_d$$

Sin embargo, es posible incluir información auxiliar que permita estimar dar estimaciones adecuados para estos dominios

Modelo de áreas pequeñas

$$y_d = X_d\beta + u_d + e_d$$

El uso de las ventajas del estimador directo y el estimador por modelos conllevan a un modelo mixto, denominado el modelo de áreas pequeñas

¿Cuál modelo seleccionar?

Dado el auge y la importancia que están teniendo los métodos de estimación en áreas pequeñas, es necesario realizar investigación en nuevos métodos u otras propuestas para mejorar los indicadores estimados.

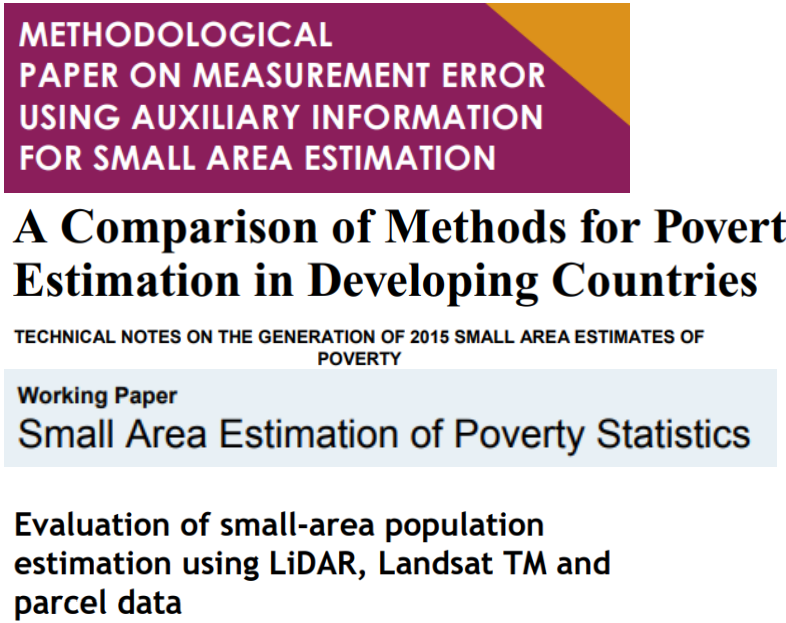
Fuentes de datos para los modelos de áreas pequeñas

Uno de los pasos más importantes de la estimación en áreas pequeñas es la definición de las fuentes de información a utilizar, así se utilizan fuentes censales y de registros administrativos.

	Ventajas	Desventajas
Fuentes censales	Se posee información tanto a nivel de unidad como de área para toda la población Es validado a través de metodologías estadísticas para evitar sesgos	Se actualiza como mínimo cada 10 años Su realización es muy costosa
Registros administrativos	Se actualiza recurrentemente Existen múltiples mediciones a nivel de registros administrativos	Su objetivo es tener un registro de la población por lo cual no fue diseñado para fines estadísticos Puede tener sesgos no medibles

Estimación en áreas pequeñas a partir de registros administrativos

En la bibliografía autores como Villa Juan-Albacea, Zita (2009), Das, S., & Haslett, S. (2019), , las naciones unidas, la CEPAL, entre otros. Recomiendan el uso tanto de información censal como de registros administrativos, dado que, la inclusión de registros administrativos permite una mejor estimación de los modelos en años no censales.



Estimación en áreas pequeñas con modelos de machine learning

Con el fin de aprovechar las ventajas del Machine Learning algunos autores como Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kassteele, J. (2022), Singleton, A., Alexiou, A., & Savani, R. (2020), Krennmair, P., Wurz, N., & Schmid, T. (2022). Han iniciado la investigación de como usar modelos de ML en la estimación en areas pequeñas, como por ejemplo tree-based models.

A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands

Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation

Tree-Based Machine Learning in Small Area Estimation

Estimación en áreas pequeñas con modelos de machine learning

Uno de los principales enfoques del ML es incluir muestras de entrenamiento y de prueba. Dado que buscan predecir correctamente la información sobre datos no observados en las muestras utilizando metodologías como validación cruzada. Uno de los modelos utilizados es el XGBoost.

XGBoost

Ventajas	Desventajas
Es uno de los métodos más usados cuando se trabaja con problemas de gran dimensión tanto en individuos como en variables. Reduce tiempos para la predicción de la variable de interés.	Se denomina de caja negra al no determinar la importancia de variables. Sin embargo, se puede solventar usando métodos locales como LIME.

Dado que es una técnica no paramétrica es necesario estimar el ECM utilizando técnicas como Bootstrap

¿Cómo escoger el mejor modelo?

Tanto para modelos de área como para modelos de unidad existen diferentes propuestas alrededor del mundo y se aplican diferentes estrategias para la estimación del indicador de interés. Sin embargo, surge la necesidad de como escoger el mejor modelo:

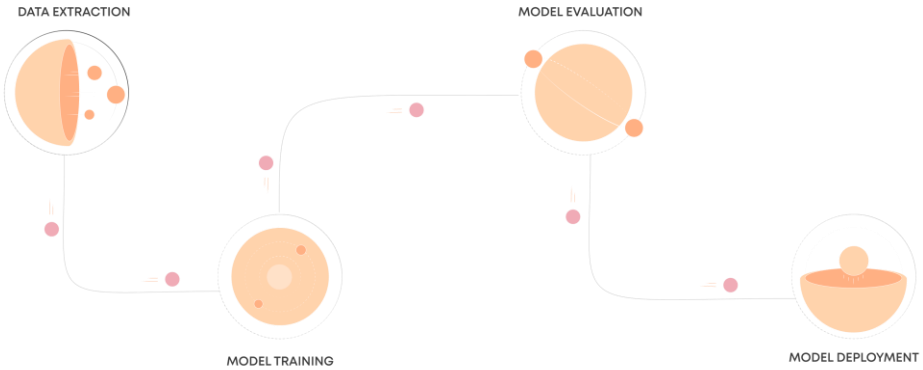
Precisión
El modelo predice bien el comportamiento de mi variable

Disponibilidad de la información
¿Los registros administrativos estarán disponibles a futuro?

Error de las estimaciones
Tengo un error de estimación aceptable para la publicación de resultados

Replicabilidad
Es replicable este modelo a través del tiempo

Garantizar actualización y escalamiento



Construcción de flujo de datos para el procesamiento de modelos

Buscando un enfoque semi – automatizado para la actualización de información

Fuente: https://valohai.com/machine-learning-pipeline/

Bibliografía

1. Corral P., Molina I., Cojocaru A., Segovia S. (2022). Guidelines to Small Area Estimation for Poverty Mapping. World Bank.
2. Das, S., & Haslett, S. (2019). A comparison of methods for poverty estimation in developing countries. *International Statistical Review*, *87*(2), 368-392.
3. Dong, P., Ramesh, S., & Nepali, A. (2010). Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data. *International Journal of Remote Sensing*, *31*(21), 5571-5586.
4. Ghosh, M., & Rao, J. N. (1994). Small area estimation: an appraisal. *Statistical science*, *9*(1), 55-76.
5. Marchetti, Tzavidis, Permanyer, Spain, et al. (2021). METHODOLOGICAL PAPER ON MEASUREMENT ERROR USING AUXILIARY INFORMATION FOR SMALL AREA ESTIMATION, United Nations.
6. Villa Juan-Albacea, Zita (2009) : Small Area Estimation of Poverty Statistics, PIDS Discussion Paper Series, No. 2009-16, Philippine Institute for Development Studies (PIDS), Makati City
7. Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. International Journal of Health Geographics, 21(1), 4.
8. Singleton, A., Alexiou, A., & Savani, R. (2020). Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems*, *82*, 101486.
9. Krennmair, P., Wurz, N., & Schmid, T. (2022). Tree-Based Machine Learning in Small Area Estimation.