GOBIERNO DE COLOMBIA

DANE
INFORMACIÓN PARA TODOS
70 AÑOS

Construcción de estadísticas experimentales con SAE

05 de junio de 2023

Ecosistema de datos de las estadísticas oficiales - SEN

- **Sistema Estadístico Nacional – SEN**
- **Subsistema Departamental**
- **Subsistema Municipal**
- **Usuarios de información estadística**

*** Subsistema de otras entidades territoriales**

- Provincias
- Áreas metropolitanas
- Territorios indígenas
- Territorio colectivo de comunidades negras
- Asociaciones de municipios
- Territorios PDET

[1] **Artículo 155 de la Ley 1955 de 2019.** Desde la oferta, además de las entidades del orden nacional y subnacional que producen información estadística; están las personas jurídicas, públicas o privadas, que prestan servicios públicos y las personas jurídicas que posean, produzcan o administren registros administrativos en el desarrollo de su objeto social, que sean insumos necesarios para la producción de estadísticas oficiales. Dentro de las organizaciones de la sociedad civil (OSC) se incluyen: las organizaciones no gubernamentales (ONG), entidades sin ánimo de lucro (ESAL) las cuales a su vez, incluyen las cámaras de comercio; gremios; organizaciones sociales, comunitarias, sindicales, de profesionales, étnicas, académicas; entre otras.

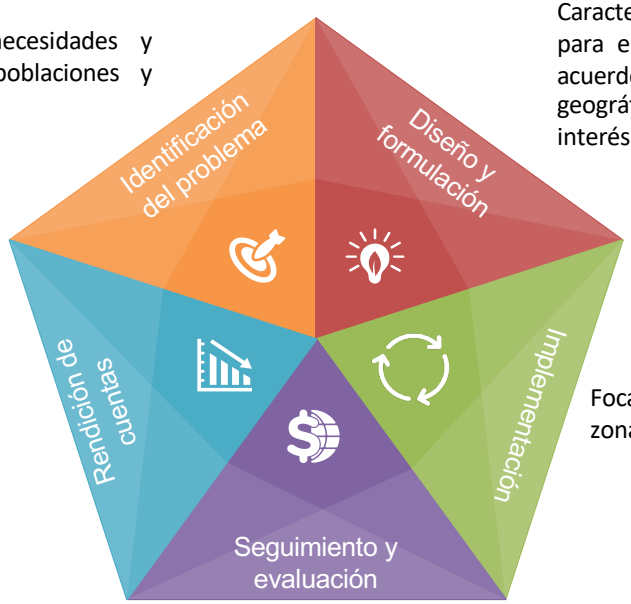
INFORMACIÓN PARA TODOS

Toma de decisiones basadas en información

Dada la alta disponibilidad de datos se genera de una mayor demanda de análisis con el fin de **tomar decisiones basados en datos**. En la política pública se está dando el mismo enfoque lo que aumenta la demanda de información por parte de otras entidades, la academia, organismos gubernamentales y la ciudadanía en **publicar información con mayor desagregación y detalle**.

Identificación de necesidades y brechas entre subpoblaciones y dominios de interés

Caracterización poblacional y geográfica para el diseño diferencial de políticas, de acuerdo con los grupos poblacionales y geográficos de interés para la variable de interés

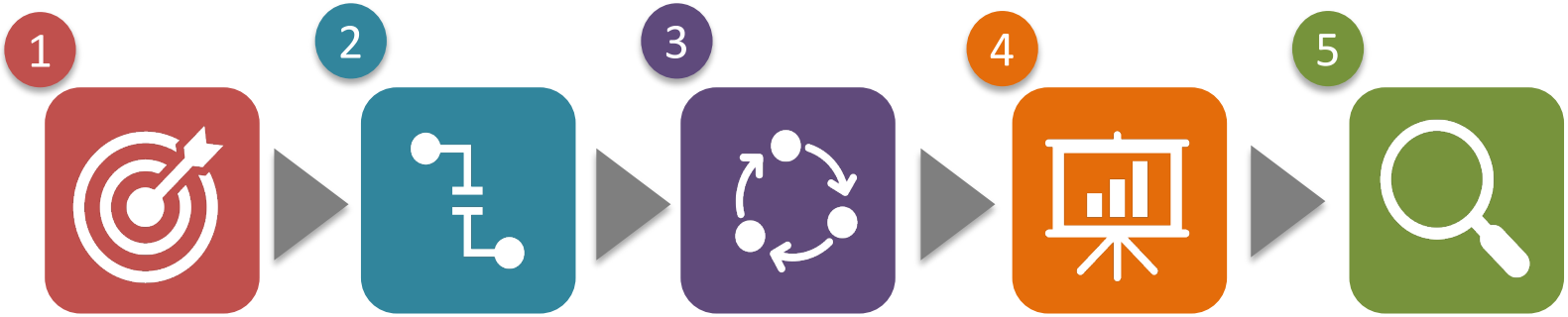


Publicación de estadísticas oficiales

Focalización geográfica: Definición de zonas prioritarias

Uso de microdatos para la medición de resultados y efectos de la política pública.

Pasos para la implementación



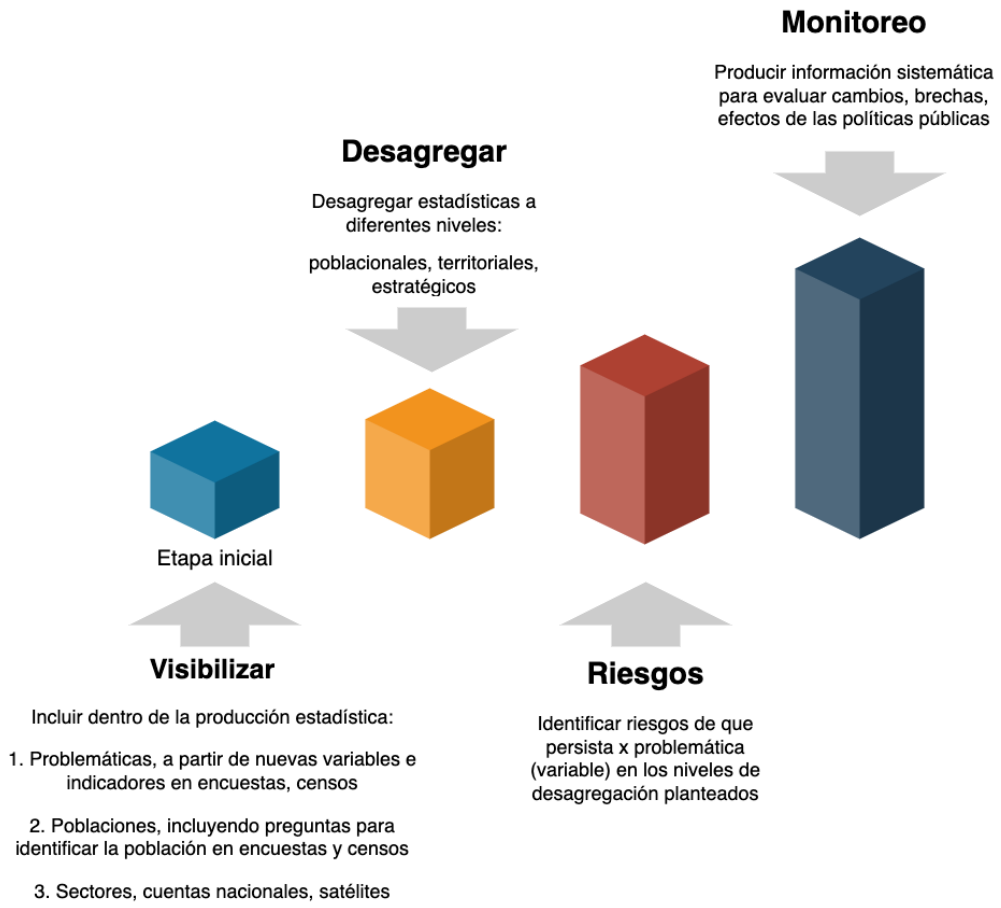
1. Identificación de la variables de interés y sus dominios
2. Definición de las fuentes de información y construcción de covariables
3. Definición del modelo de estimación de áreas peuqeñas
4. Consistencia, comparación con las cifras oficiales y validación de expertos
5. Definición del proceso y publicación de resultados

INFORMACIÓN PARA TODOS

1

Identificación de los planes de interés y sus dominios

Interés estadístico del nuevo Plan Nacional de Desarrollo - PND 2022-2026



Las principales necesidades de información y análisis de datos son:

Censo Económico – 2024.

Conteo Intercensal de Población – 2025.

Mayor granularidad de las estimaciones. Estadísticas oficiales para grupos poblacionales más específicos como víctimas, género, etnicidad, pero también estimaciones a niveles más desagregados de nivel geográfica como municipios, localidades, barrios para todas las OOEE no sólo del DANE sino del SEN para el seguimiento del cumplimiento de los indicadores.

Niveles de desagregación definidos en el PND 2022-2026

De acuerdo con las bases del Plan Nacional de Desarrollo 2022-2026 **“Colombia potencia mundial de la vida”,** se cuenta con los siguientes actores diferenciales para el cambio:

Genero
Mujer
LGBTIQ+

Ciclo de vida
Niños y niñas
Jóvenes

Discapacidad

Víctimas

Grupos étnicos

Territorio
Municipios
Ruralidad
PDET

Campesinos

Economía Popular

Niveles de desagreación requeridos a través de los ODS

Los indicadores de los Objetivos de Desarrollo Sostenible deberían desglosarse, siempre que fuera pertinente, por:

- **Ingreso**
- **Sexo**
- **Edad**
- **Raza**
- **Etnicidad**
- **Estado migratorio**
- **Discapacidad**
- **Ubicación geográfica**

u otras características, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales (resolución 68/261 de la Asamblea General).

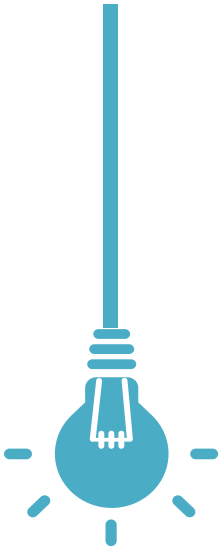
INFORMACIÓN PARA TODOS

2

Definición de las fuentes de información y construcción de covariables

INFORMACIÓN PARA TODOS

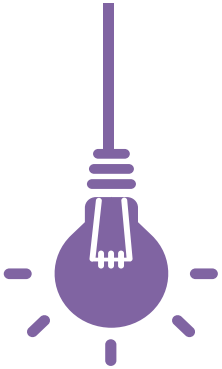
Fuentes de información disponibles para la integración de datos



Censos

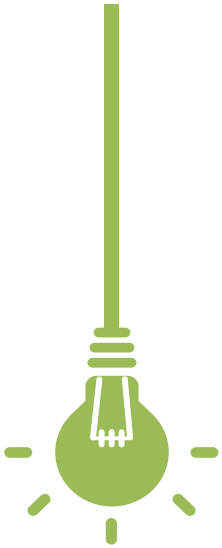
CENSO NACIONAL DE POBLACIÓN Y VIVIENDA 2018 · COLOMBIA

3er CENSO NACIONAL AGROPECUARIO
Hay campo para todos



Registros administrativos

RELAB
Variables municipales

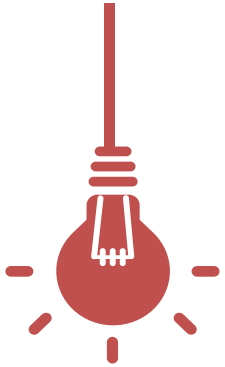


Encuestas

GEIH MERCADO LABORAL

ECV

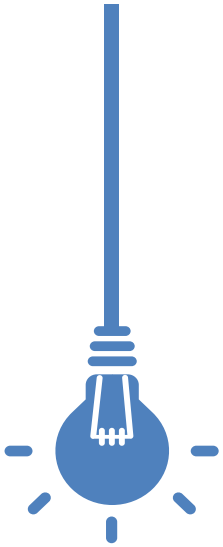
ENA



Información geoespacial

Geoportal

Imágenes satelitales



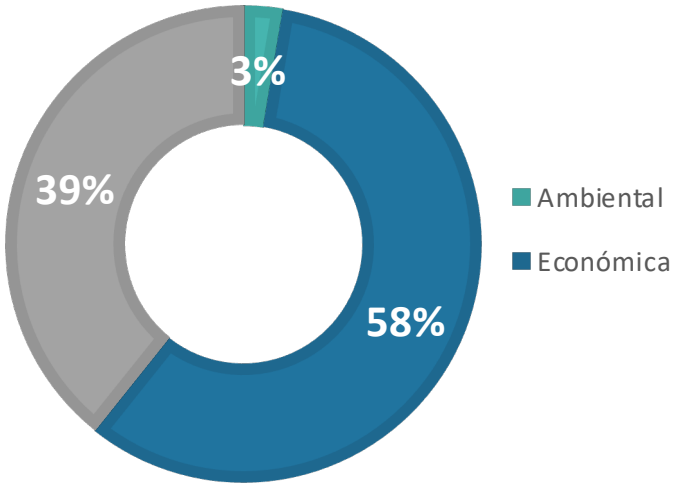
Fuentes alternas

Web scraping
APPS
IA
Big Data

Fuentes secundarias disponibles y su uso estadístico

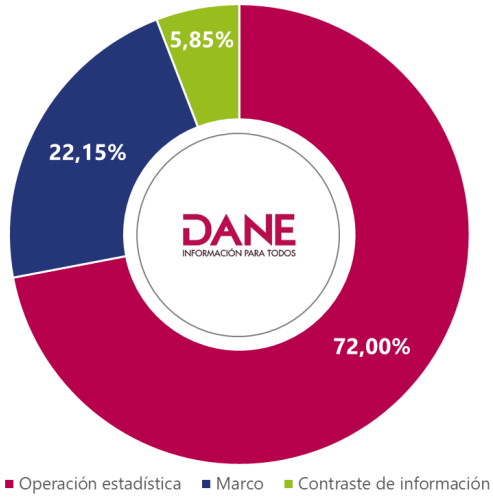
Los datos e información estadística del SEN son activos disponibles para la investigación aplicada y la toma de decisiones.

Participación de los RR.AA. por temática



El país tiene **527 Registros administrativos** producidos por **95 entidades.**

Participación de los principales usos de RR.AA. por el DANE



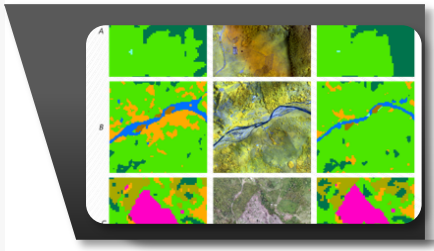
El DANE aprovecha estadísticamente **233 Registros administrativos.**

* OO.EE y RR.AA en estado activo

Fuente: Inventario de usos de registros administrativos 2022, revisado al 3 de octubre de 2022

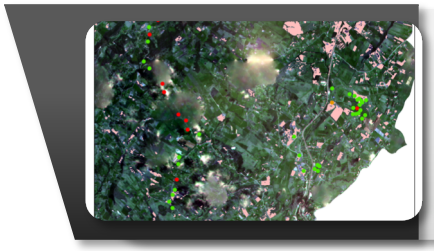
2. Fortalecimiento de marcos

Aplicación de técnicas alternativas sobre **imágenes satelitales y de dron**, para la **actualización y fortalecimiento del Marco Maestro Rural Agropecuario (MMRA)**



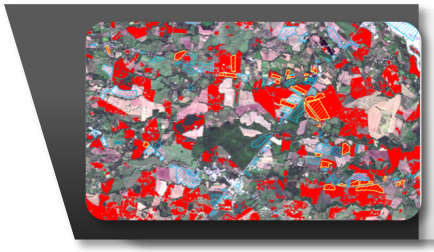
Logros 2020

- Con base en imágenes Sentinel-2, imágenes de drone, y muestras de entrenamiento recolectadas en campo, **se aplicó el algoritmo de *Random Forest* para la clasificación supervisada basada en píxeles de las coberturas del suelo**.
- **Mejoras en la precisión de clasificación** (hasta el 3%) y en la delimitación de coberturas.
- Método reemplaza fotointerpretación masiva previa por una **clasificación supervisada semiautomatizada a través del procesamiento en la nube y ejecución de scripts**.



Logros 2021

- Clasificación de imágenes satelitales (Sentinel-2, Sentinel-1, PlanetScope) aplicando *Random Forest.*
- Se estimaron 521 ha de **papa** (Villapinzón – Cundinamarca) y 483,6 ha de **fríjol** (San Gil –Santander) con precisiones generales del 82% y 76%, respectivamente.
- Aplicación de los scripts procesamiento para la **clasificación e identificación de grandes coberturas** de la tierra de forma masiva. **Mejoras en la precisión de estimación de coberturas gruesas** como bosques y pastos.



Logros 2022

- Clasificación con modelos ML supervisados *Random Forest* y *XGBoost*.
- Se estimaron más de 5 mil ha de **maíz** y 408 ha de **piña** en la zona de estudio (San Martín y Granada – Meta) con **precisiones generales superiores al 90%,** para la actualización de cultivos de dominio.
- Modelo para la identificación de cultivos de interés disponible para puesta en producción.

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.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

	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

INFORMACIÓN PARA TODOS

3

Definición del modelo de estimación de áreas pequeñas

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.

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. 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.

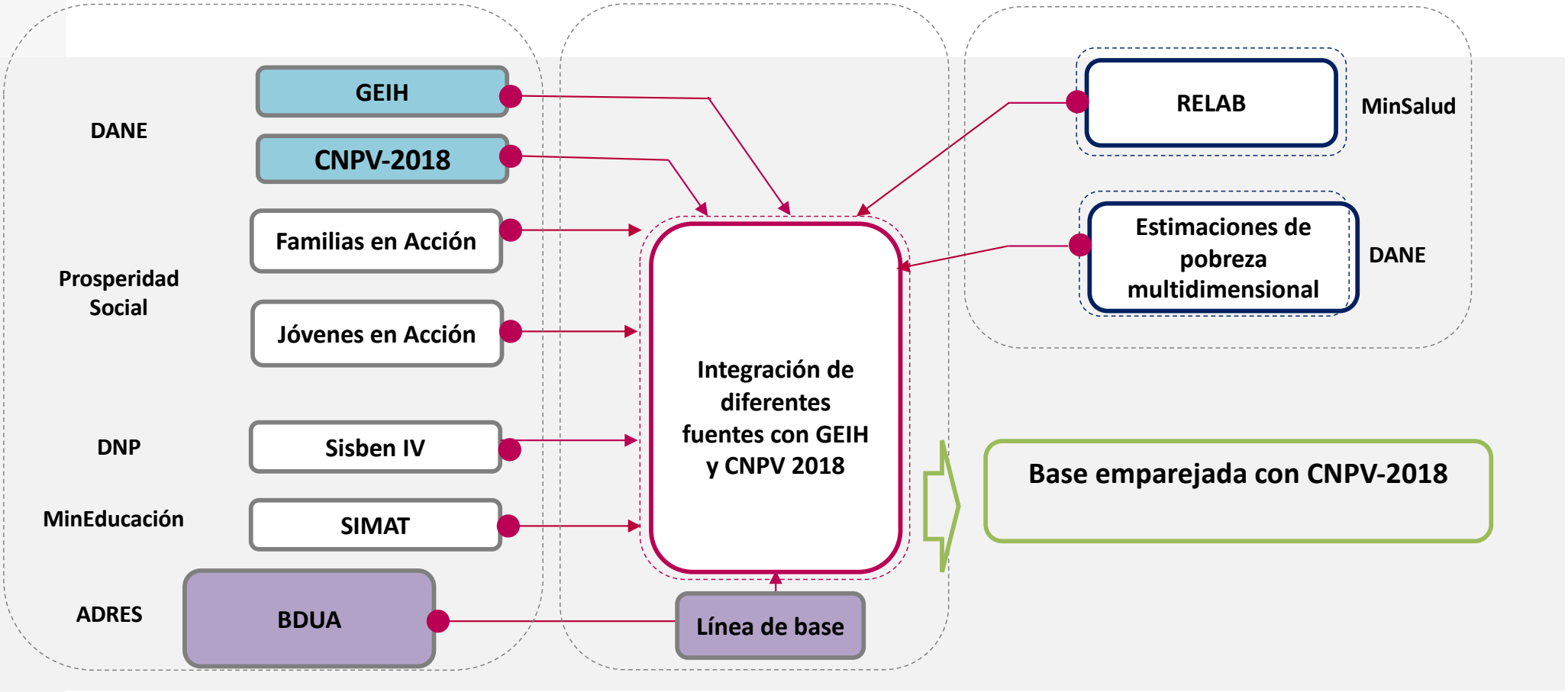
Dado que es una técnica no paramétrica es necesario estimar el ECM utilizando técnicas como Bootstrap

INFORMACIÓN PARA TODOS

3

Algunas aplicaciones en DANE

Fuentes de información



Resultados:

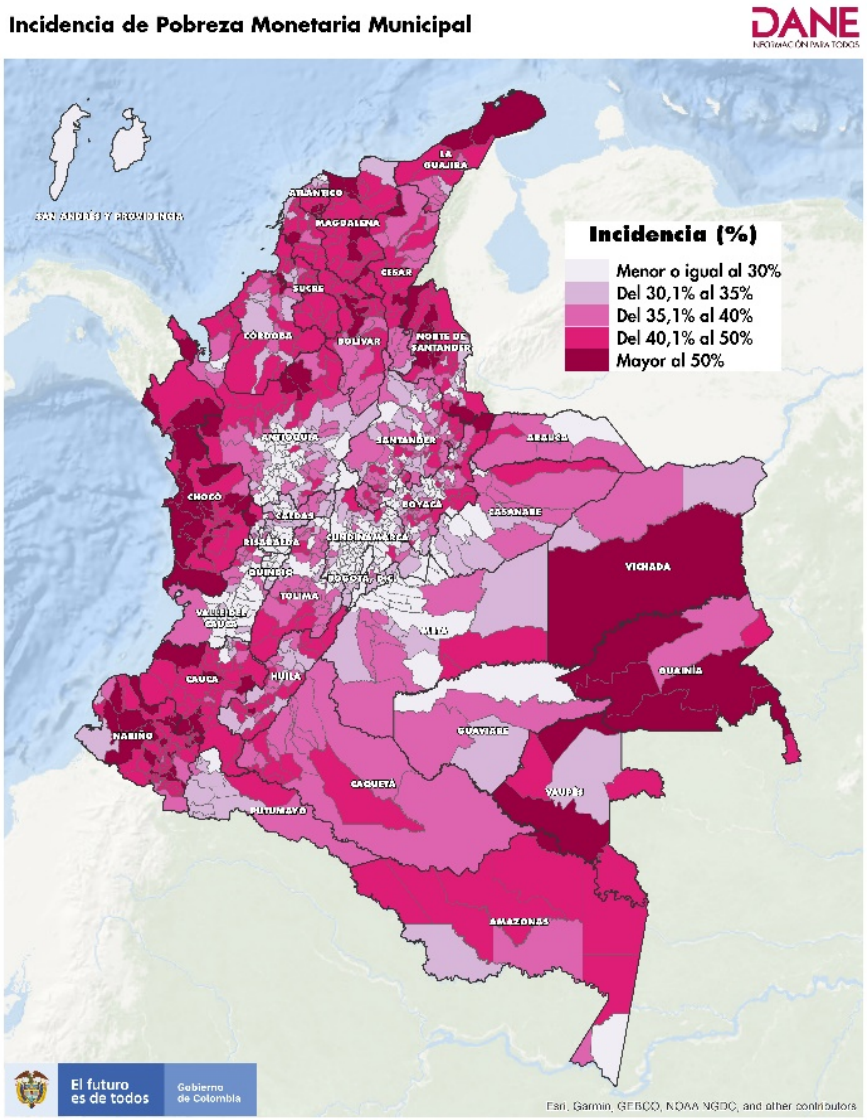
Mapa de la pobreza monetaria estimada con el modelo

- El mapa de la pobreza monetaria de los hogares pronosticada con el modelo de efectos aleatorios para los 1.122 municipios de Colombia

- Más allá de las usuales ciudades capitales de los departamentos, **el mapa muestra patrones de diferenciación geográfica entre los municipios en términos de la pobreza monetaria**.

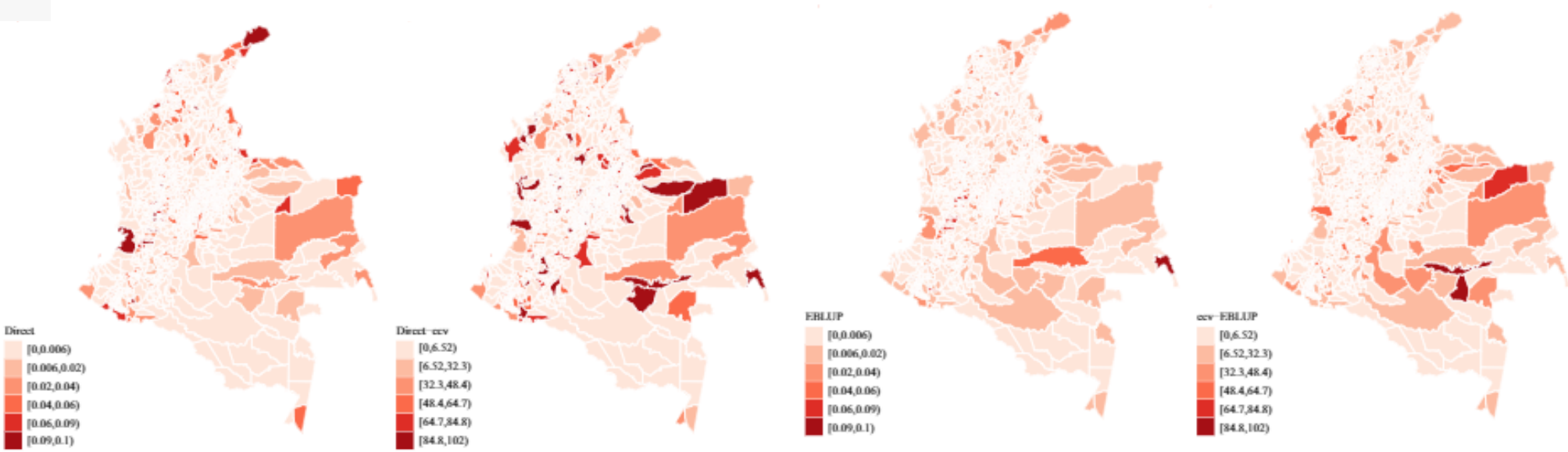
- Estos patrones permiten distinguir aglomeraciones de municipios de baja o alta pobreza monetaria o, por el contrario, municipios que se encuentran en un extremo de la distribución rodeados por municipios en el extremo opuesto.

Nota: Resultados proximamente a ser publicados.



Resultados: estimación de la emigración internacional en áreas pequeñas (municipios) en Colombia

Se aplicó el modelo Fay-Herriot, se ajusta a la DHS 2015 según la información del censo disponible. Se calcula el EBLUP al indicador seleccionado de emigración internacional a nivel municipal, se aplica el método GVF para estimar las varianzas de los estimadores directos de PHMLA.



Fuente: DANE.

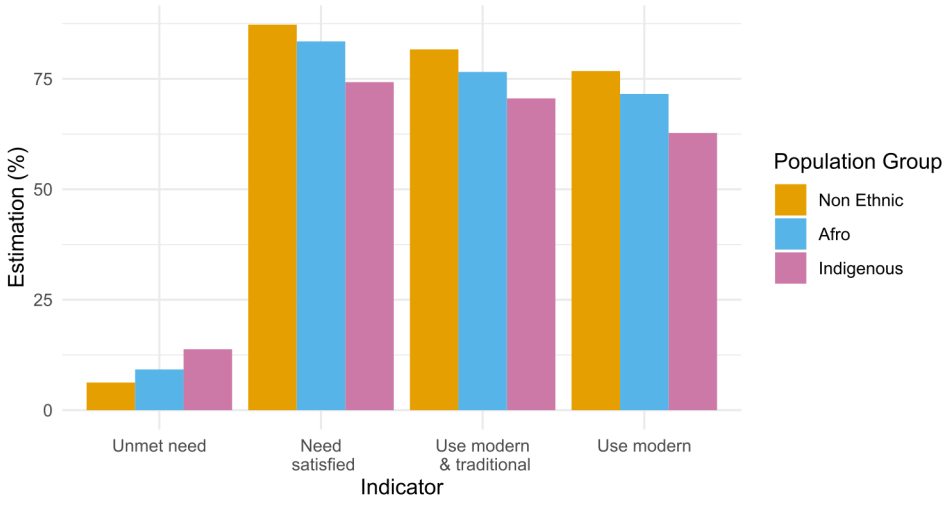
Uso de métodos anticonceptivos para población indígena

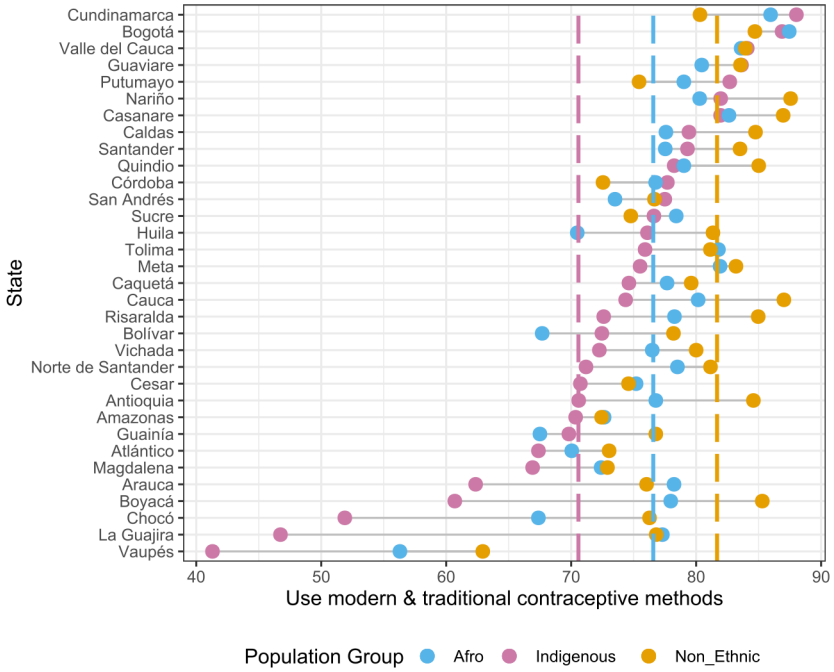
01 Homologación de variables

02 Usar los datos muestrales de la DHS para estimar el modelo

03 Usar los parámetros estimados con la infromación auxiliar del censo

04 Estimación de los dominios de interés





Rao and Molina (2015). Small Area Estimation, New York: John Wiley & Sons.

SAE4SDG Toolkit https://unstats.un.org/wiki/display/SAE4SDG/SAE4SDG

Mapa de pobreza multidimensional

Integración de fuentes alternativas de información en el proceso estadístico

Actualmente el DANE mide:

- **IPM** a nivel departamento usando la encuesta de hogares anualmente
- IPM a nivel municipal usando el censo cada 10 años.

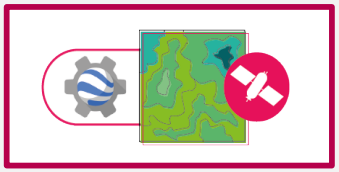
Indicador:

- IPM a nivel municipal anula

Sources:

- Encuesta de hogares
- Censo
- Datos Geoespaciales

Metodología:

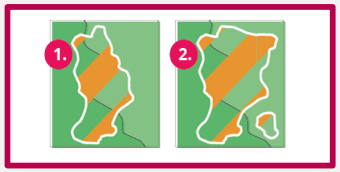


Recolectar

- Información geoespacial como : luces nocturnas, índice de vegetación, vías.

Entrada

- Clusters de la encuesta con la medida agregada de IPM

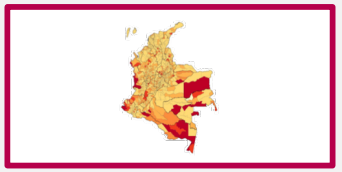


Modelo

- Modelo mixto generalizado (Geoestadístico)
- Modelo geoestadístico bayesiano

Estimación

- Población viviendo en pobreza a nivel de cluster



Resultados y validación

- Mapeo del IPM a nivel de cluster (manzanas/veredas)
- Rendimiento del modelo

INFORMACIÓN PARA TODOS

INFORMACIÓN PARA TODOS

4

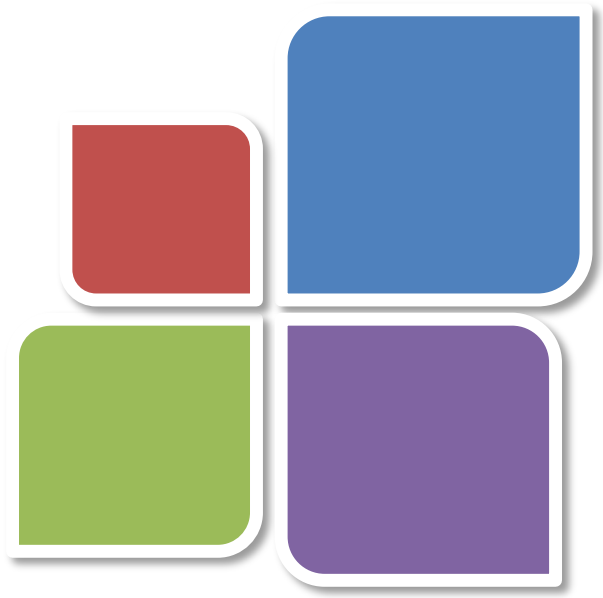
Consistencia, comparación con las cifras oficiales y validación de expertos

¿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

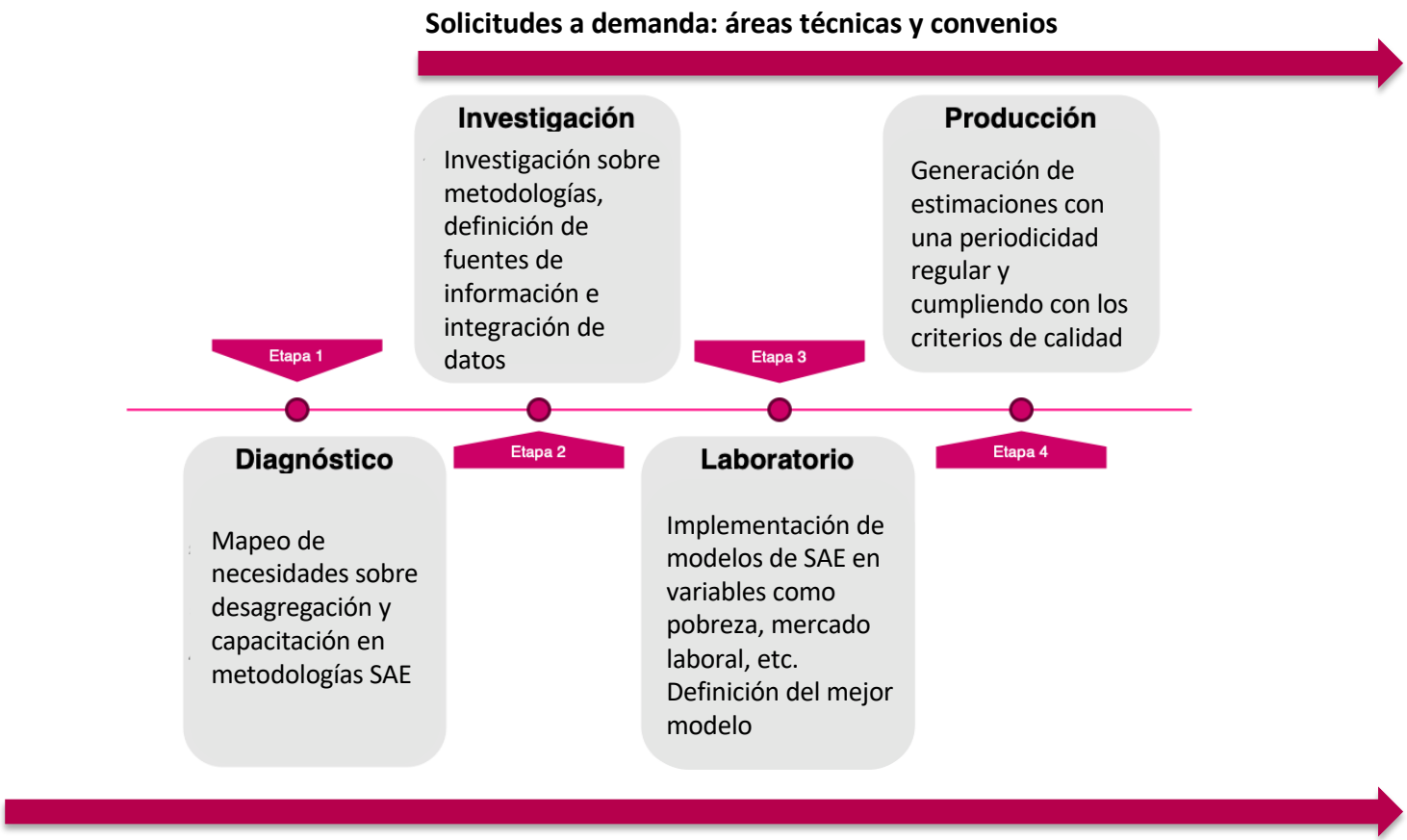
Replicabilidad
Es replicable este modelo a través del tiempo

INFORMACIÓN PARA TODOS

5

Definición del proceso y publicación de resultados

Plan de trabajo SAE



Ciclo de vida del dato

Fuente: basado en OCDE

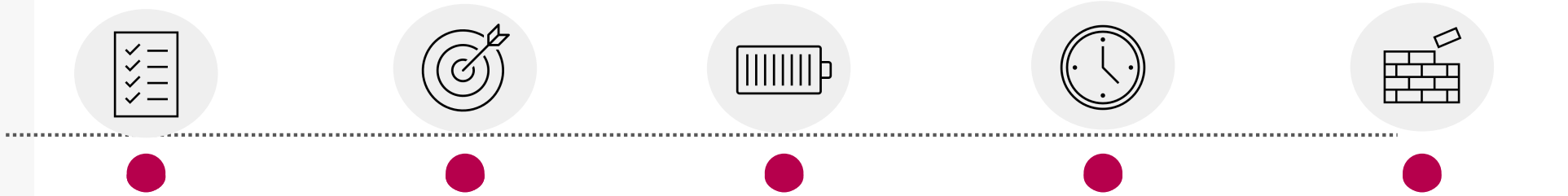
Preparación de datos

Replicable para diferentes variables

1. Inventario de fuentes de información y mapeo de variables para la construcción de indicadores
2. Descarga de información y almacenamiento en servidores DANE
3. Extracción, transformación y carga de variables del modelo de áreas pequeñas
4. Análisis exploratorio de datos y análisis de calidad de datos

Proceso de revisión en Registros Administrativos

5 dimensiones en la revisión de las fuentes para asegurar la calidad del producto final: REBE



Chequeos técnicos

- Legibilidad
- Declaración de cumplimiento (metadatos/diccionario)
- Convertibilidad

Precisión

- Legitimidad de los objetos
- Objetos inconsistentes
- Objetos dudosos
- Errores medición
- Valores inconsistentes
- Combinaciones no posibles

Completitud

- Cobertura temática
- Objetos redundantes o duplicados
- Valores vacíos
- Valores imputados por el proveedor del RR.AA u otra fuente

Puntualidad

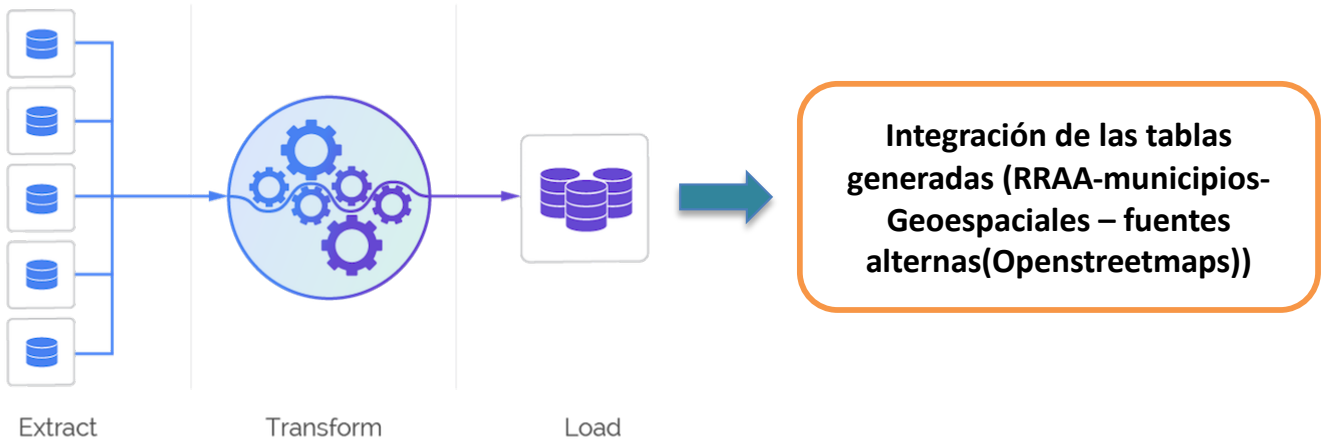
- Puntualidad en la entrega de los datos acordada
- Estabilidad de los atributos en el tiempo

Integrabilidad

- Llaves de integración y comparabilidad de atributos con otras fuentes con mismos atributos

*Tomado de estándares consolidados por la DIRPEN en el marco del BLUE Enterprise and Trade Statistics (2011)

Integración de datos



Retos:

1. Efectividad de los cruce de información (Cobertura). Potenciales sesgos (*Linking Bias*)
2. Actualización de información para definir periodicidad del proceso y pertinencia en la entrega de las fuentes (Gobierno de datos a través del CAD/SEN)
3. Calidad de datos

Construcción y transformación de variables

01

Homologación de variables

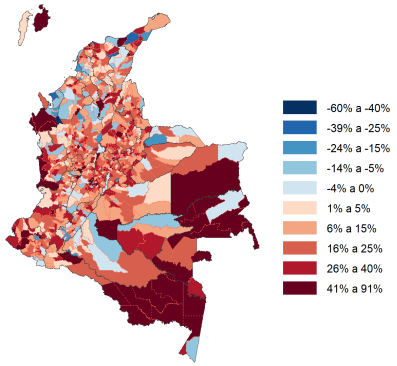
- Manejar el mismo formato
- Estandarización de categorías
- Identificación de variables similares con fraseos o métodos de recolección diferentes

02

Jerarquía de la fuente de información

- Definición de la confiabilidad de la fuente por variable

(2018 Census Adjusted-REBP) / 2018 Census Adjusted



03

Actualización de información

- Complementar fuentes de información:
- Ejemplo: Nivel y grado educativo.

Grado = Línea de base (CNPV2018) + SIMAT (2019) + …

04

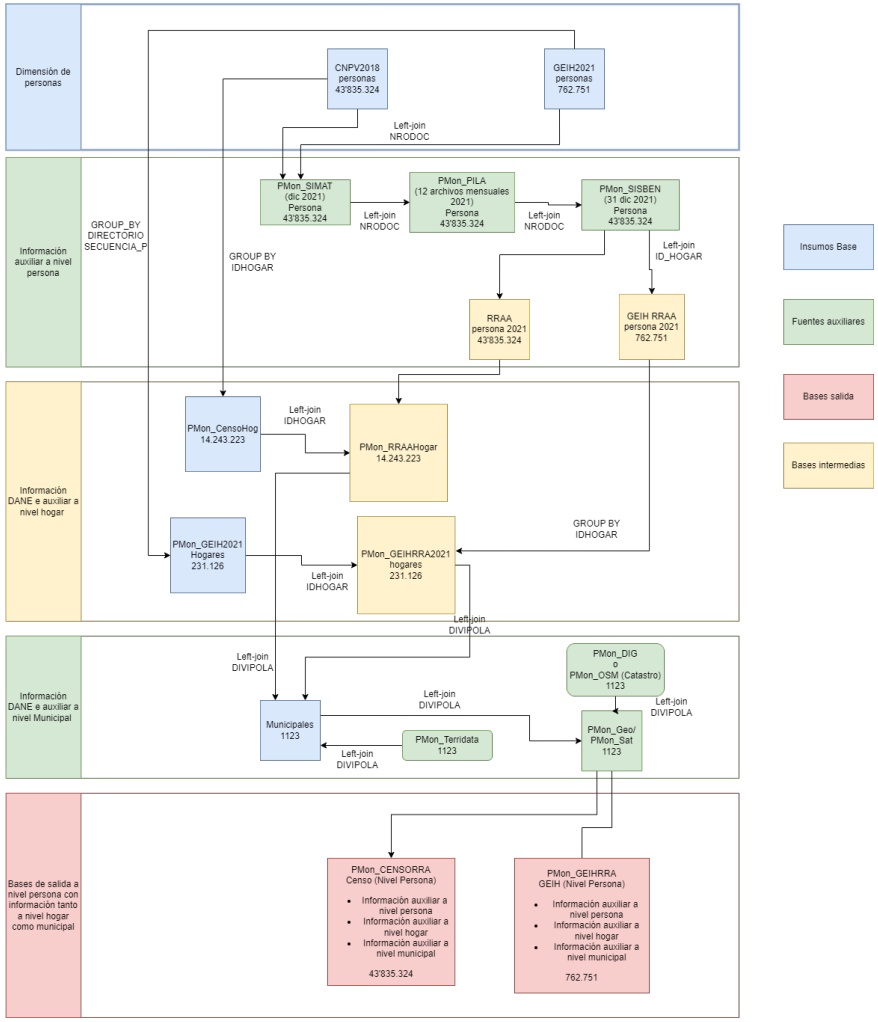
Construcción de variables

- Definición del indicador usado como covariable del modelo de estimación

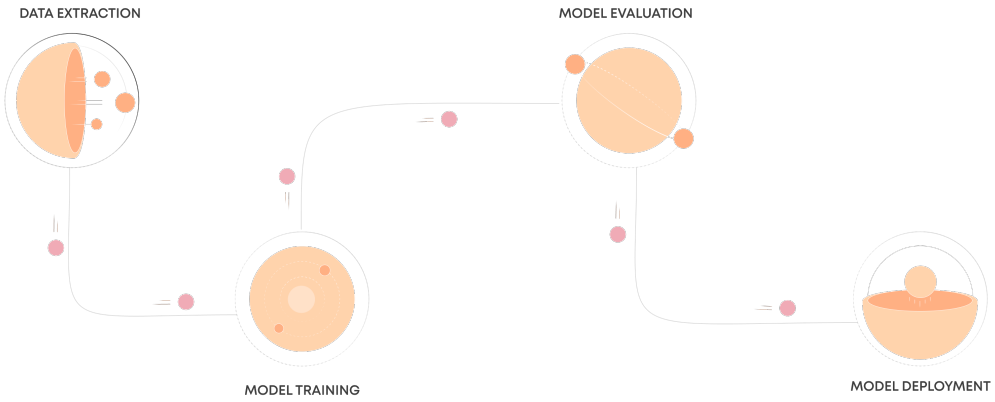
Preparación de datos
Ejemplo – Pobreza monetaria

Procedimiento de tablas a usar en la estimación de pobreza monetaria municipal usando SAE.

Pensando en la replicabilidad del proceso en el siguiente año de estimación



Garantizar actualización y escalamiento – Tránsito a producción estadística



Construcción de flujo de datos para el procesamiento de modelos

Buscando un enfoque de flujo de datos (*pipeline*) 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.