DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

Proyecto de estimación en áreas pequeñas (SAE)

Subdirección

28 de noviembre de 2023

1

¿Por qué usar SAE en DANE?

2

¿Qué aplicaciones con SAE hemos realizado en DANE?

3

¿Qué estamos haciendo?

4

¿Qué Sigue?

COLOMBIA POTENCIA DE LA VIDA

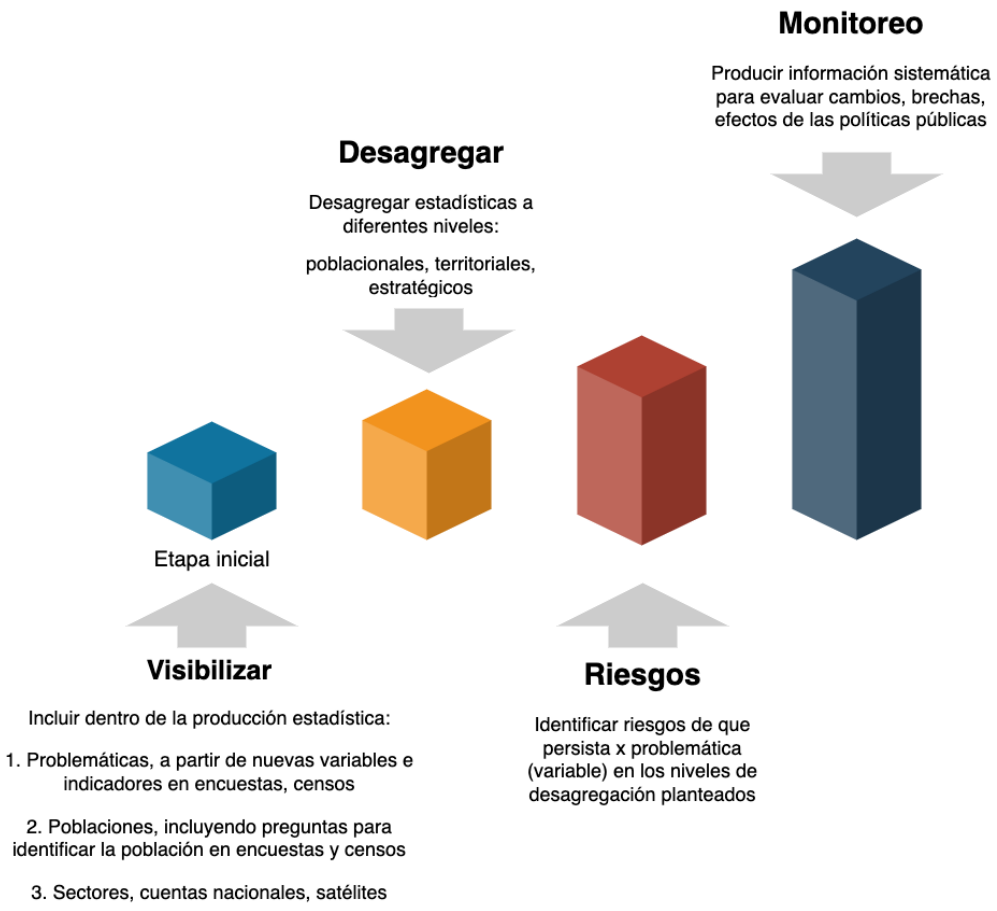
DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

¿Por qué usar SAE en DANE?

Toma de decisiones basadas en información

Dada la alta disponibilidad de datos, se genera una mayor demanda de análisis con el fin de **tomar decisiones basados en información**. En política pública se ha incrementado la demanda de información por parte de otras entidades, la academia, organismos no gubernamentales y la ciudadanía en **publicar información con mayor desagregación y detalle**. Esto con el fin de realizar ejercicios que permitan apoyar el ciclo de política pública, por ejemplo:

Interés estadístico en el Plan Nacional de Desarrollo - PND 2022-2026



El PND 2022-2026 de Colombia busca dar **mayor visibilidad en las estadísticas oficiales de grupos poblacionales de los cuales regularmente no se reporta información** a causa de la baja representatividad de estos grupos en las encuestas, como: población víctima del conflicto armado, población LGBTIQ+, etnicidad, campesinos, entre otros.

Además, de obtener **mayores niveles de desagregación en dominios geográficos y temáticos** con el fin de monitorear riesgos y realizar el seguimiento, monitoreo y evaluación de los cambios y brechas sobre las variables impactadas por las políticas planteadas por el Gobierno.

Niveles de desagregación definidos en el PND 2022-2026

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

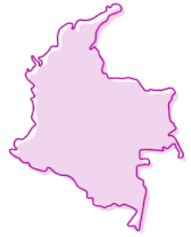
Género
Mujer
LGBTIQ+

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

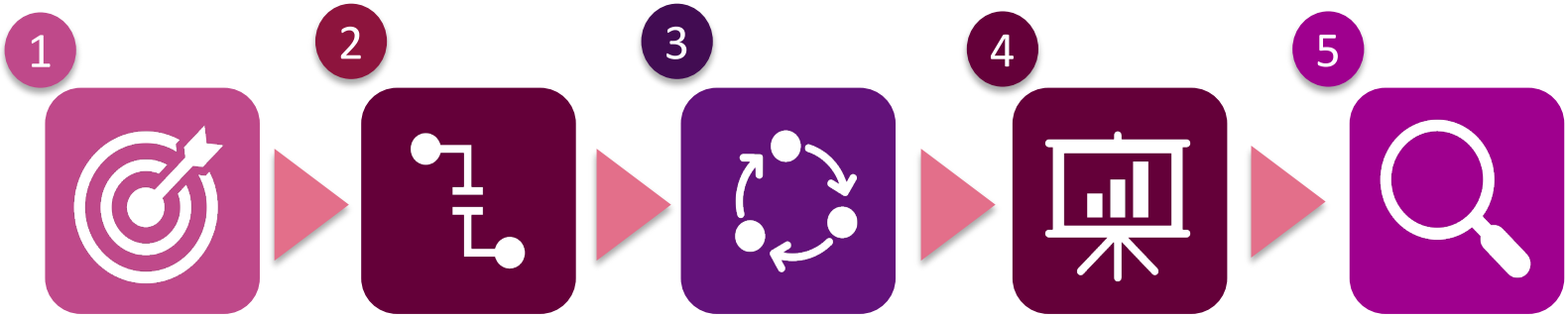
Los indicadores de los Objetivos de Desarrollo Sostenible deberían desglosarse, siempre que fuera pertinente, por:

- **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).

Pasos para la implementación de SAE en la entidad

Con el fin de optimizar los recursos públicos y a la vez responder a las solicitudes de información, se inicia con la aplicación de metodologías de estimación de áreas pequeñas dentro de DANE. Lo cual implica para cada una de las aplicaciones en SAE la consecución de los siguientes pasos:



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 pequeñas
4. Consistencia, comparación con las cifras oficiales y validación de expertos
5. Definición del proceso y publicación de resultados

¿Qué información tenemos disponible para la integración de fuentes de datos?

Censos

CENSO NACIONAL DE POBLACIÓN Y VIVIENDA 2018 · COLOMBIA

3er CENSO NACIONAL AGROPECUARIO
Hay campo para todos

Registros administrativos y estadísticos

Subsidios
Formalidad laboral
Acceso a salud
Acceso a educación

Encuestas

GEIH MERCADO LABORAL

ECV

Encuestas sociales, económicas y ambientales

Información geoespacial

Geoportal

Imágenes satelitales

Fuentes alternas

Web scraping
APIS
Textos
Documentos
Redes sociales

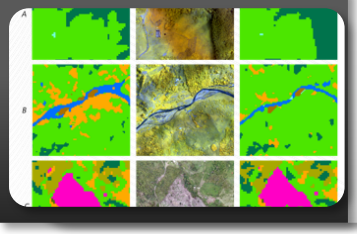
COLOMBIA POTENCIA DE LA VIDA

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

¿Qué aplicaciones hemos hecho en DANE?

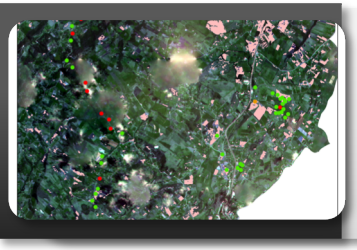
Fortalecimiento de marcos usando imágenes satelitales

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



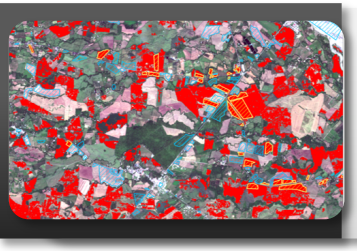
Logros 2020

- 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.

Mapa de pobreza multidimensional a nivel manzana/vereda

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.

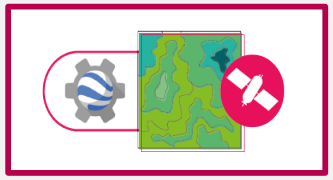
Indicador:

- IPM a nivel municipal anula

Sources:

- Encuesta de hogares
- Censo
- Datos Geoespaciales

Metodología:

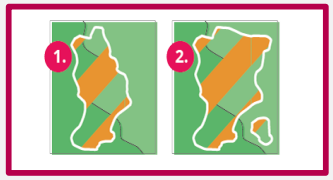


Recolectar

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

Entrada

- Clusters de la encuesta con la medida agregada de IPM

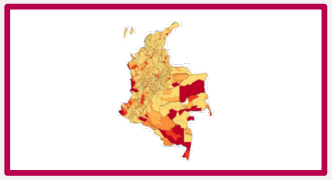


Modelo

- 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

Mapa municipal de pobreza monetaria 2018

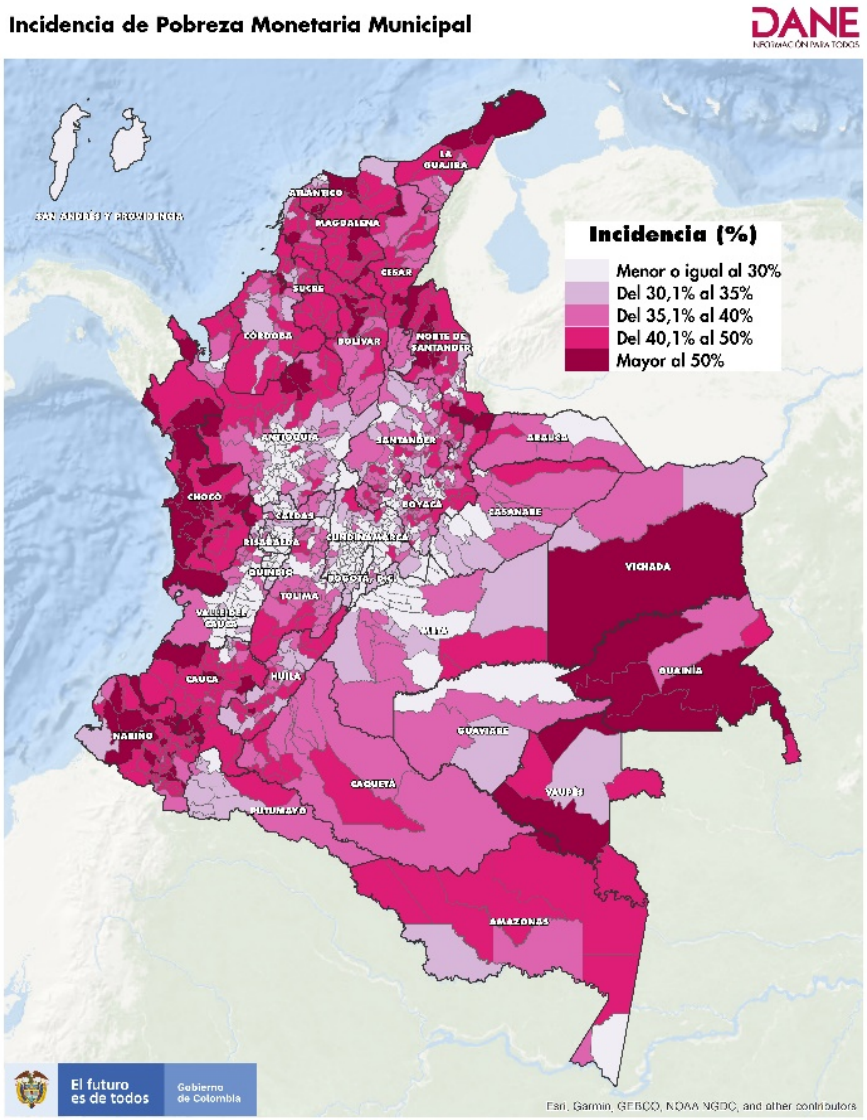
En cooperación con Data4Now se realizó la **estimación de pobreza monetaria para el año 2018 usando un modelo de área EBP (*Empirical Best Predictor*)**, usando el lenguaje de programación R. Para la construcción de covariables para el modelo se realizó la integración del Censo Nacional de Población y Vivienda 2018 junto con registros administrativos, como: SISBEN, BDUA, Familias en Acción, Jóvenes en Acción, SIMAT, RELAB.

Como resultados, se encuentro que:

- 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.

Este mapa no ha sido publicado dado que se encuentra en proceso de validación de los expertos de pobreza.



Estimación municipal de la emigración internacional en Colombia

En Colombia, los indicadores desagregados de alta calidad a nivel municipal solo están disponibles después de cada censo de población, es decir, cada 10 o 15 años. En los años intermedios, es normal que se produzcan cambios en los patrones de inmigración y en la intensidad del fenómeno. Por ello, las encuestas por muestreo proporcionan datos que permiten obtener estimaciones para períodos intermedios (estimaciones intercensales) y para períodos posteriores a los censos, en los que se dispone de los resultados de las encuestas (períodos poscensales).

Para la estimación la proporción de hogares con al menos un miembro habitual viviendo en el extranjero – PHMLA en períodos intercensales a nivel municipal se aplicó un modelo de área, Fay-Herriot, usando como fuente de información principal, la encuesta de demografía y salud del años 2015 (DHS2015).

Journal of Official Statistics, Vol. 37, No. 3, 2021, pp. 771–789, http://dx.doi.org/10.2478/JOS-2021-0034

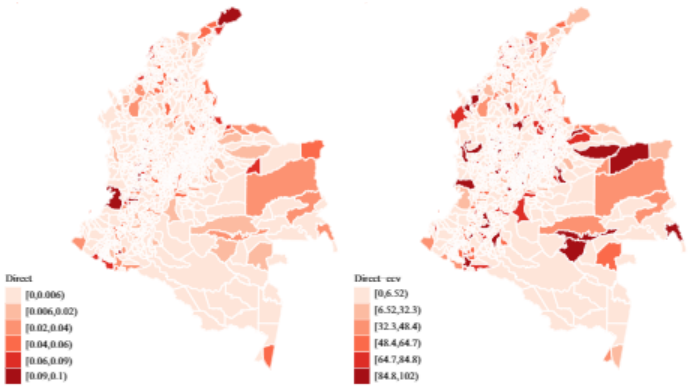
Fay-Herriot Model-Based Prediction Alternatives for Estimating Households with Emigrated Members

Jairo Fúquene-Patiño[1], *César Cristancho*[2], *Mariana Ospina*[2], *and Domingo Morales Gonzalez*[3]

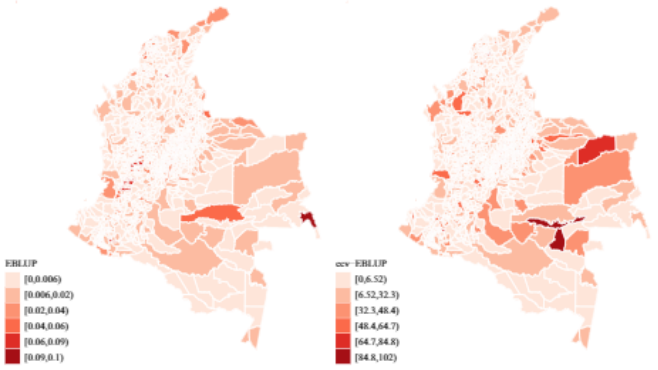
This article proposes a new methodology for estimating the proportions of households that had experience of international migration at the municipal level in Colombia. The Colombian National Statistical Office usually produces estimations of internal migration based on the results of population censuses, but there is a lack of disaggregated information about the main small areas of origin of the population that emigrates from Colombia. The proposed methodology uses frequentist and Bayesian approaches based on a Fay-Herriot model and is illustrated by one example with a dependent variable from the Demographic and Health Survey 2015 and covariables available from the population census 2005. The proposed alternative produces proportion estimates that are consistent with sample sizes and the main internal immigration trends in Colombia. Additionally, the estimated coefficients of variation are lower than 20% for municipalities for both frequentist and Bayesian approaches and large demographically-relevant capital cities and therefore estimates may be considered to be reliable. Finally, we illustrate how the proposed alternative leads to important reductions of the estimated coefficients of variations for the areas with very small sample sizes.

Key words: Small area estimation; international migration; Fay-Herriot model; coefficient of variation; direct estimator; model-based estimator; hierarchical Bayes prediction.

Estimaciones directas y sus correspondientes CV



Estimaciones EBLUP y sus correspondientes CV



Uso de métodos anticonceptivos para población indígena

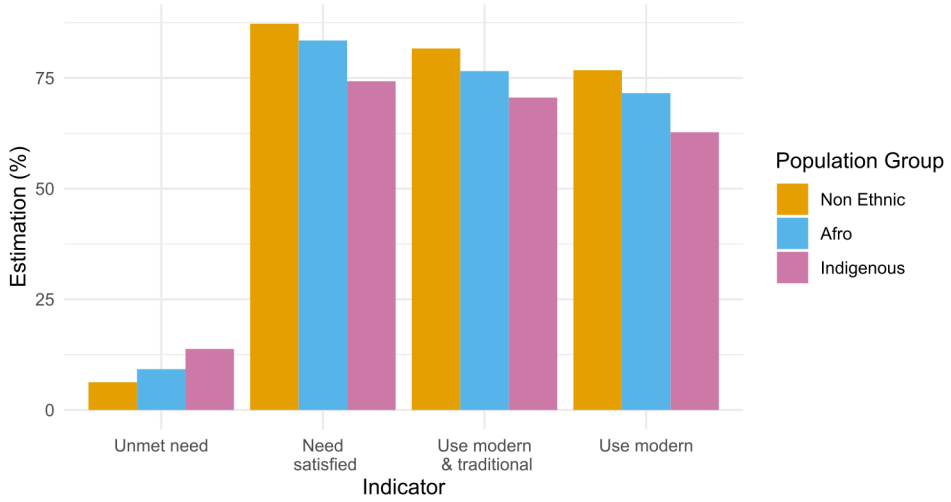
En cooperación con CEPAL y UNFPA se realizó la estimación del uso de métodos anticonceptivos de acuerdo con el grupo etnico desagregado a nivel departamental, para esto se utilizo un modelo de unidad, EBP (*Empirical Best Predictor*) y se usó como insumo el Censo Nacional de Población y Vivienda y la Encuesta de Demografía y Salud. Para lo cuál se aplicaron los siguientes pasos:

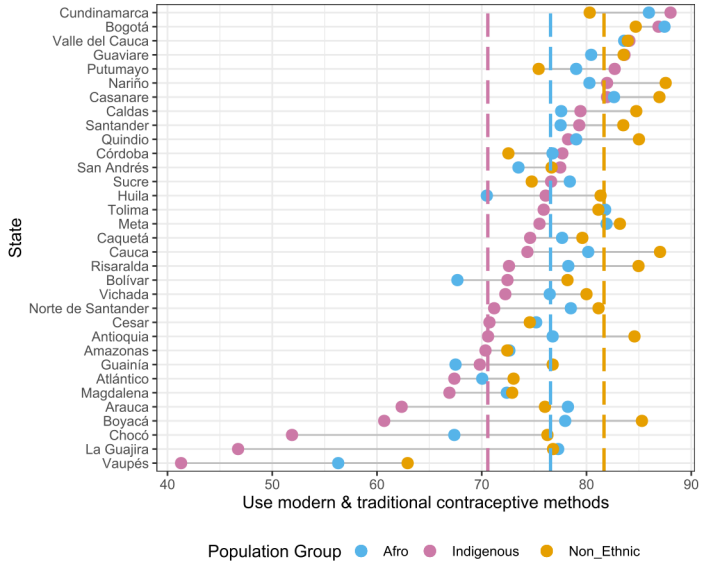
01 Homologación de variables

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

03 Usar los parámetros estimados con la información auxiliar del censo

04 Estimación de los dominios de interés





COLOMBIA POTENCIA DE LA VIDA

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

¿Qué estamos haciendo?

SAE – Master Data Base

Dirección de Geo-estadística

Procesar variables a partir de imágenes satelitales e información de sahepefile

Dirección de Registros Estadísticos *

Cruces con registros administrativos: Transferencias monetarias, Formalidad laboral,Afiliación a salud, funcionarios estado, Matricula, RUV

Otras Fuentes

Dirección Censos y demografía

- Registro base de personas
- Censo Nacional de Población y Vivienda 2018

Dirección de Metodología y Producción Estadística

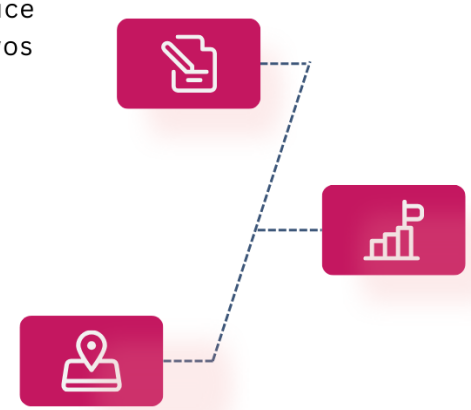
Encuestas
GEIH
ENCV

Definición de tablas

Temáticas: Población, ambientales, educación, salud, económicas del municipio, Sisbén, Transferencias Monetarias, Trabajo, Conformación hogar, Víctimas, Policía .....

1. Base a nivel hogar -Total

Información de censo, registros estadísticos y cruce con registros administrativos



2. Base a nivel hogar - Muestra (GEIH-ENCV...)

Información de encuesta, registros estadísticos y cruce con registros administrativos

3. Base municipal

Información geoespacial, variables de otras entidades, agregaciones de otras unidades de observación. Establecimientos educativos

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

Reglas de negocio

Realizamos el mapeo de diferentes fuentes de información identificando: registros administrativos, cubos, imágenes satelitales y encuestas. Encontrando cerca de 300 variables en diferentes temáticas y diferentes unidad de observación: municipios, viviendas, hogares, personas.

El siguiente diagrama describe el proceso de extracción e integración de variables para pasar a la 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

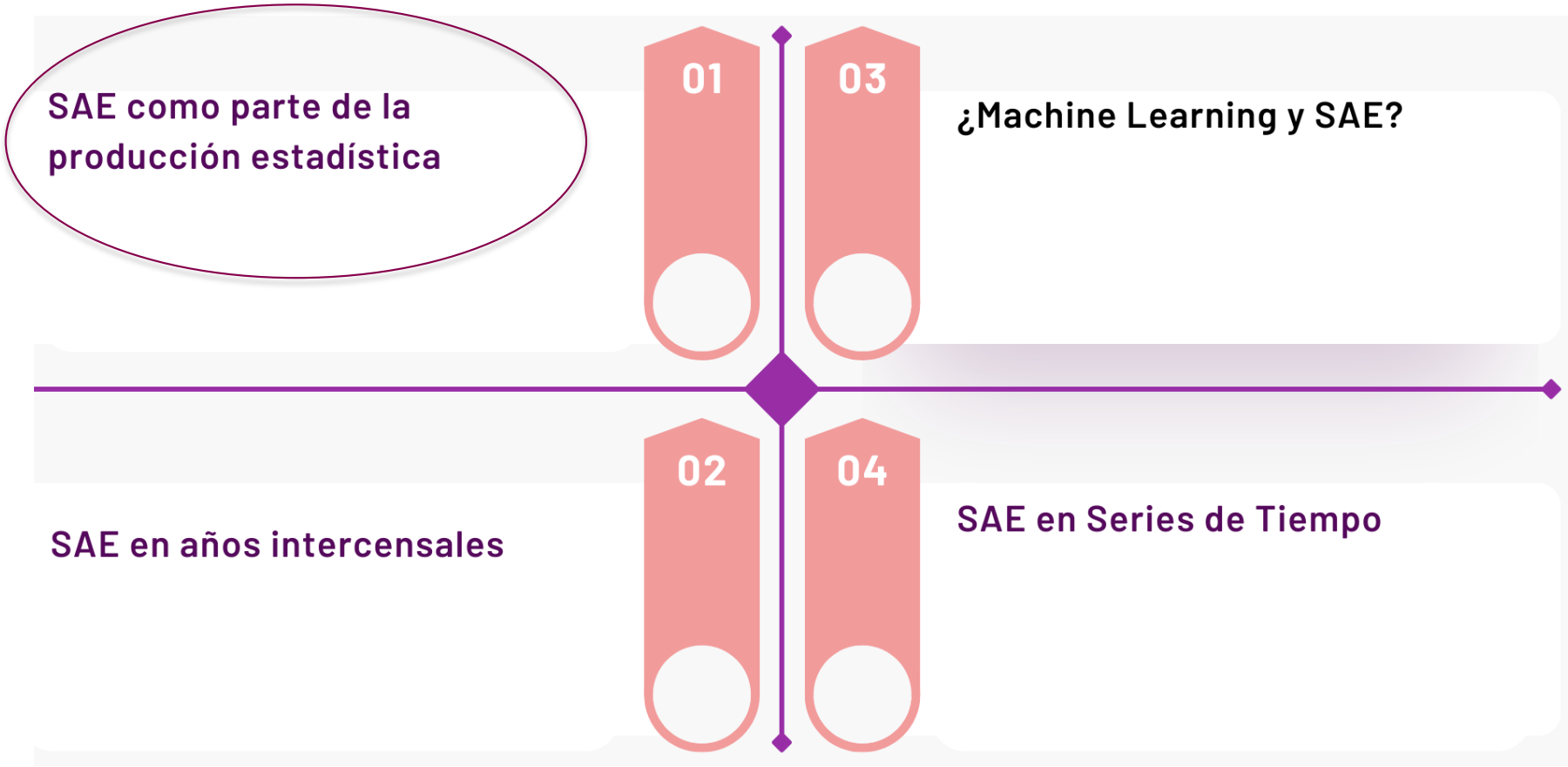
¿En qué estamos trabajando?

COLOMBIA POTENCIA DE LA VIDA

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

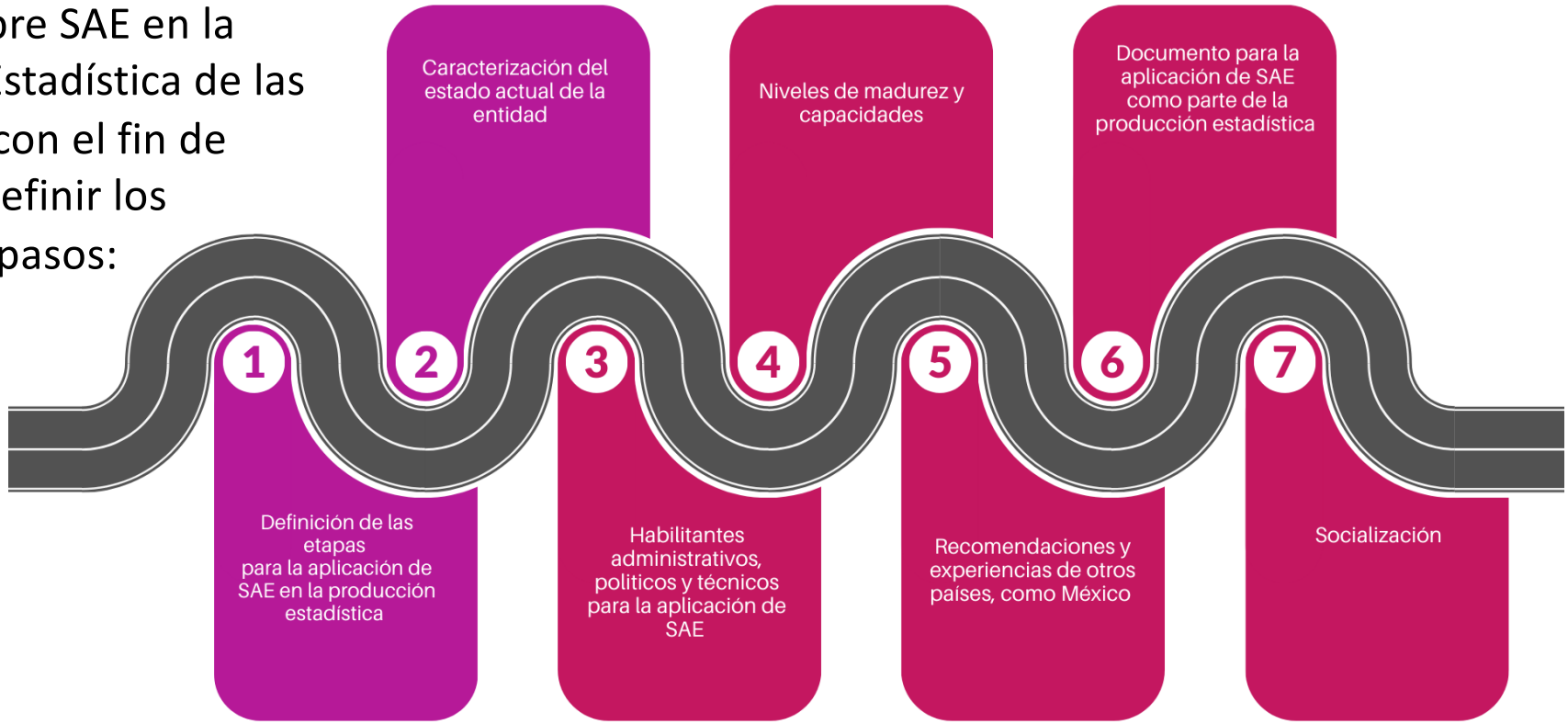
¿Qué sigue?

Seminario regional sobre metodologías de estimación en áreas pequeñas y desagregación de datos



Grupo de trabajo CEA

Colombia y Venezuela coordinarán el grupo de trabajo sobre SAE en la Comisión Estadística de las Américas, con el fin de realizar y definir los siguientes pasos:



COLOMBIA POTENCIA DE LA VIDA

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

Gracias

Subdirección

/DANEColombia
@DANE_Colombia
@DANEColombia
/DANEColombia