DANE
INFORMACIÓN PARA TODOS
70 AÑOS

Use of non-traditional data sources for the SDGs

Colombia's experience

December 2023

Experimental Statistics

Institutional arrangements for utilization

ESTADÍSTICAS EXPERIMENTALES

SDG Indicator 11.7.1 Average proportion of built-up area in cities, corresponding to open spaces for public use by all, disaggregated by age group, sex and people with disabilities

INFORMATION AVAILABLE

In Colombia there was no information on this indicator, which is part of the global list of indicators of the Sustainable Development Goals -SDG. A methodology was developed that uses the DEGURBA method to delimit cities, uses satellite image classification methods to calculate land consumption and additionally uses Open Street Maps as a source of information.

Within the list of indicators of the 2030 Agenda, this indicator is part of:
Goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable.
Goal 11.7:By 2030, provide universal access to safe, inclusive and accessible green spaces and public spaces, in particular for women and children, older people and people with disabilities

Technical information

Presentation	02-Dec-2021	Discharge
SDG indicator methodological sheet 11.7.1	02-Dec-2021	Discharge
ODS Indicator Geoviewer 11.7.1	02-Dec-2021	Go to Geovisor

- An experimental statistic is one that comes from projects that are being undertaken and have innovative aspects, either by taking advantage of new sources of information, the statistical methodology used or a new topic not previously measured.
- They are considered experimental as they offer new ways of quantitatively characterizing phenomena in the three dimensions of Sustainable Development: economic, sociodemographic and environmental of the country.
- ***https://www.dane.gov.co/index.php/estadisticas-por-tema/estadisticas-experimentales***

Integration of statistical and geospatial information
Context

- Key benefits of using Earth observations, in National Statistical Offices, for the calculation of SDG indicators:

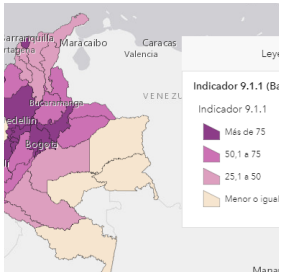
I. The possibility of **deriving indicators from the SDGs**, which would otherwise be technically or financially difficult to calculate.

II. Reduce the frequency of surveys and associated costs by **providing information at a higher level of disaggregation.**

III. Provide breakdown and granularity of indicators, ensuring they are spatially oriented.

Integration of statistical and geospatial information
Applications

- Calculation of indicators of the Sustainable Development Goals (SDGs) from the integration of statistical and geospatial information.

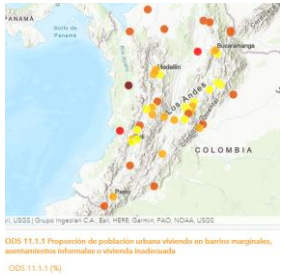


SDG 9.1.1

Proportion of rural population living within 2 km of an available road for a year-round.

Inputs: georeferenced housing from the 2018 census; Road coverage.

Processes: Geographical Information System (GIS) and its geoprocessing models.

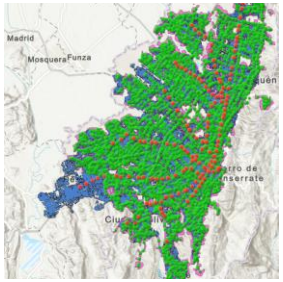


SDG 11.1.1

Proportion of urban population living in slums, informal settlements or inadequate housing.

Inputs: 2018 census microdata; Geographical coverage from external sources.

Processes: operations in spatial databases (PostgreSQL – PostGIS)

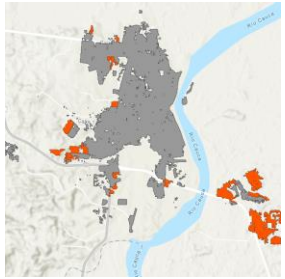


SDG 11.2.1

Proportion of population with easy access to public transport, disaggregated by sex, age and persons with disabilities.

Inputs: Sentinel-2 satellite imagery; WorldPop, georeferenced information on transport systems.

Processes: supervised image classification; network analysis (accessibility); GIS geoprocesses.

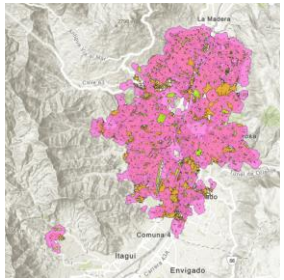


SDG 11.3.1

Relationship between the rate of land consumption and the rate of population growth.

Inputs: Landsat 8 and Sentinel-2 satellite imagery; Population projections.

Processes: supervised image classification; Calculation of fees.



SDG 11.7.1

Average proportion of built-up area in cities, corresponding to open spaces for the public use of all, disaggregated by age group, sex and persons with disabilities.

Inputs: Sentinel-2 satellite imagery; Open Street Map; Georeferenced census population.

Processes: supervised image classification; network analysis (accessibility); GIS geoprocesses.

SDG indicators and data disaggregation

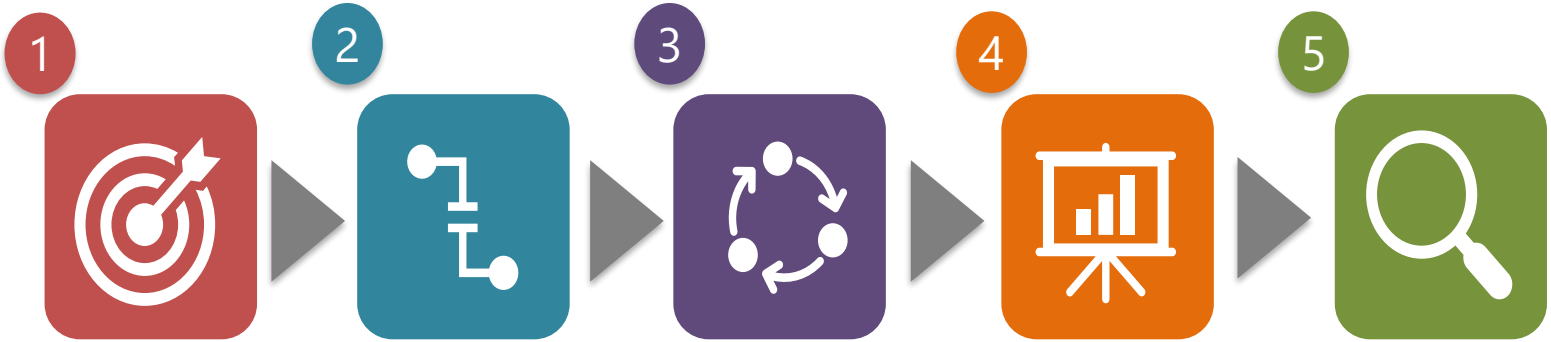
Sustainable Development Goal indicators should be disaggregated, where relevant, by:

- **Income**
- **Sex**
- **Age**
- **Race**
- **Ethnicity**
- **Migratory status**
- **Disability**
- **Geographic location**

or other characteristics, in accordance with the Fundamental Principles of Official Statistics (General Assembly resolution 68/261).

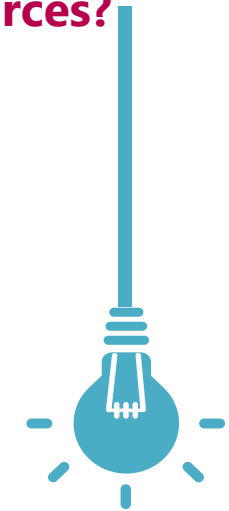
Steps for the implementation of SAE

To optimize public resources and at the same time respond to requests for information, DANE has been working in the application of small area estimation methodologies. It requires the following steps:



1. Identification of variables of interest and their domains
2. Definition of information sources and construction of covariables
3. Definition of the SAE model
4. Consistency, comparation with official figures and validation by experts
5. Process definition and publication of results

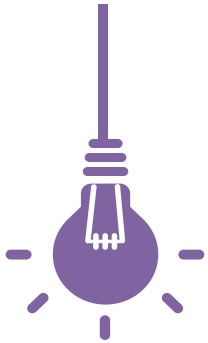
What information do we have available for the integration of data sources?



Census

CENSO NACIONAL DE POBLACIÓN Y VIVIENDA 2018 · COLOMBIA

3er CENSO NACIONAL AGROPECUARIO
Hay campo para todos



Administrative and statistical records

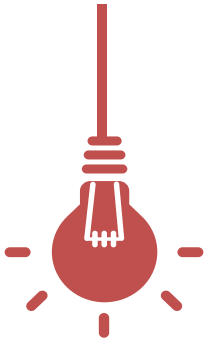
Subsidies
Labor Formality
Access to health care
Access to education

Surveys

GEIH MERCADO LABORAL

ECV

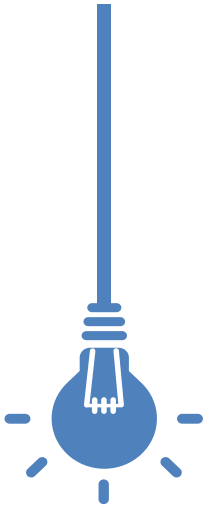
Social, economic and environmental surveys



Geospatial information

Geoportal

Satellite images



Alternative sources

Web scraping
APIS
Texts
Documents
Social media

Poverty mapping

Integration of alternative sources of information in the statistical process

First approach we worked on:

Currently DANE measures:

- MPI index statistics at the department-level using household surveys (annually).
- MPI statistics at the municipality-level using census data (every 10 years).

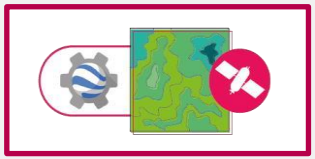
Goal:

- Measure MPI statistics at the municipality-level every year.

Sources:

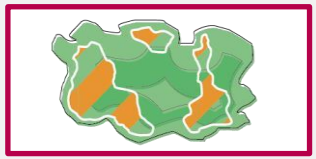
- Household surveys.
- Spatially detailed Census data.
- Geospatial covariate datasets.

Methodology:



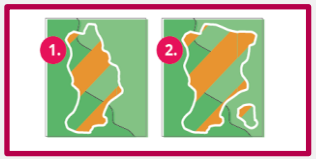
Compile

- Geospatial covariate datasets (eg. nighttime light consumption, vegetation index, accessibility via road to towns and cities).



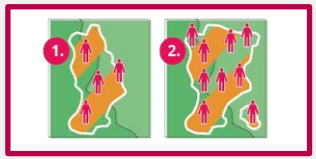
Input

- Survey clusters displaying the cluster-level MPI headcount ratio.



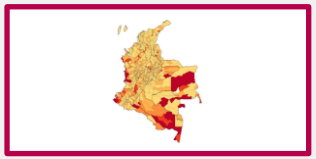
Modelling

- Generalized linear mixed model (model-based geostatistics).
- Bayesian geostatistical model.



Estimate

- The population living in poverty at the cluster level.



Results and validation

- Mapping MPI headcount ratio at the micro- scale (cluster-level) and macro – scale (municipality-level).
- Assess models' predictive performance.

Other applications

Currently, we are working on two small area estimation applications:

1. **Estimation of monetary poverty at the municipal level:** using a household level EBP - Bayesian model.

2. **Municipal level estimation of the Food Insecurity Experience Scale FIES**: This estimation we developed in cooperation with FAO and making use of a Rash model. For the estimation of the model, we are applying an area model (Fay-Herriot) that includes the error component of the sampling design and the Rash model, the main source of information is the National Quality of Life Survey 2022 and statistics at the municipal level produced by DANE or other government entities.

SDG 16 Indicators complementary measurements using social media

Objective

Obtain complementary measurements for SDG 16 Indicators using Facebook.

- **SDG 10.3.1/16.b.1** *Proportion of population reporting having personally felt discriminated against or harassed in the previous 12 months on the basis of a ground of a discrimination prohibited under international human rights law.*
- **SDG 16.7.2** *Proportion of population who believe decision-making is inclusive and responsive, by sex, age, disability and population group.*

Web scraping collection period.

- 2013-2022.

Sample

- Discrimination: 719.902 comments
  8.744 comments with prob.>0.5
  503.553 users
  8.177 users with prob. >0.5
- Inclusiveness: 187.995 comments
  62.000 comments with prob >0.6
  124.302 users
  50.794 com. users with prob >0.6
- Responsiveness : 583.507 comments
  275.360 comments with prob .>0.6
  405.693 users
  219.372 com. users with prob >0.6

Target population-scraped profiles.

- 66 profiles
- Categories: athletes, politicians, economists, public order, artists, public figures.

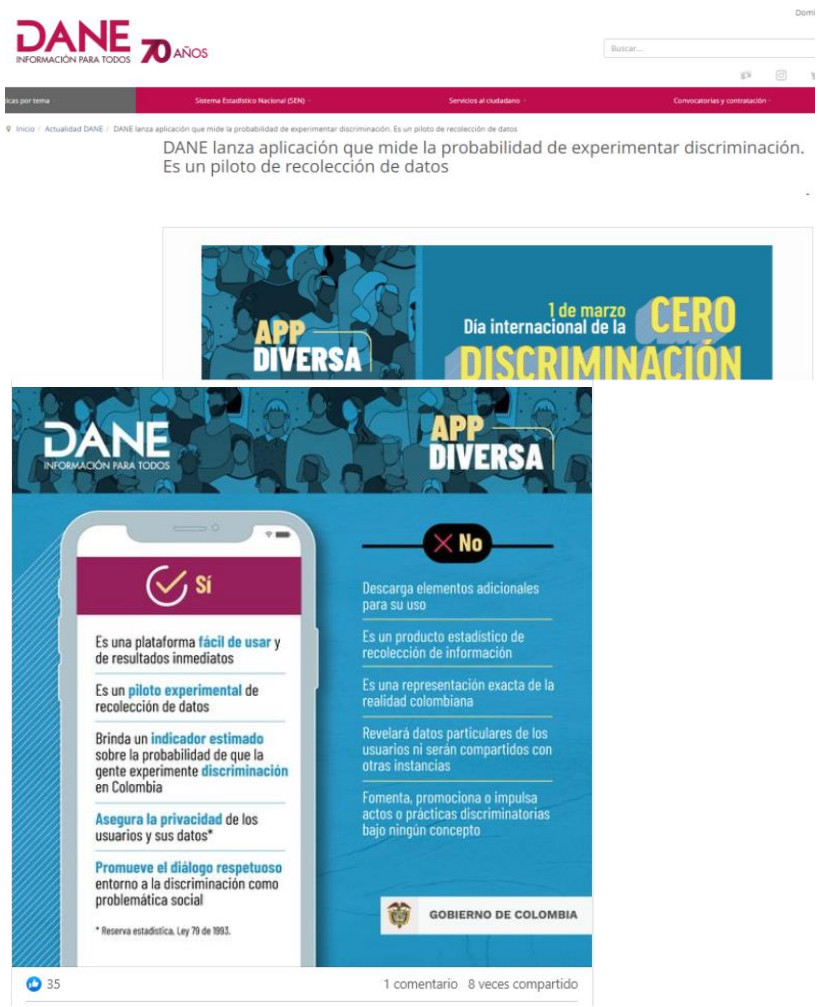
Challenges

- Facebook data presents enormous challenges in data collection, processing and analysis, as well as in terms of data privacy.
- The representativeness of the data remains a challenge.
- In general terms, social media could be used as a complementary or contextual data (e.g., complementary data could be understood in terms of what kind of tendencies and phenomena related to discrimination occur in social media).

AppDiversa

Objectives:

- Develop an alternative strategy for collecting official statistical information, based on the development of web applications and their distribution/promotion on social media.
- Validate it through the collection of data for SDG Indicator 16.b.1 (discrimination), as an application case.
- Measure the probability that individuals will experience discrimination events.
- Promote awareness of discrimination as a social problem.



Challenges in leveraging Big Data

Challenge 1: Insufficient technological infrastructure.

- Using Oracle Cloud for cloud processing.
- Limitations in the generation of local data lakes.

Challenge 2: Lack of technical capacity in Big Data and AI.

- High staff turnover and recruitment difficulties.
- Need for specialized equipment.

Challenge 3: Limited negotiating capacity with the private sector.

- Case studies: Mobile phone data, access to LinkedIn and social networks.
- Technological gaps and the need to improve relationships.

Challenge 4: Ethical use of web data sources.

- Legal problems and clauses of use.
- Ensure safe treatment and privacy of users.

Challenge 5: Change of culture within the NSO

- The strengthening and institutional change derived from this innovation should be understood as a process.
- Duplication of tasks, strengthening and empowerment of areas within the office to lead the process (e.g., systems)

DANE
INFORMACIÓN PARA TODOS
70 AÑOS

Use of non-traditional data sources for the SDGs

Colombia's experience

December 2023