DANE
INFORMACIÓN PARA TODOS

International Conference on Big Data

High Level Panel (Session 1)

El futuro es de todos

Gobierno de Colombia

August 2020

INFORMACIÓN PARA TODOS

Integration of statistical and geospatial information

Inputs

Geoespatial

- Fundamental data.
- Supplementary data.
- New data sources.

Statistical

- Censuses. Surveys.
- Administrative registers.
- Big data and other sources.

Principles

- Accessible & usable.
- Statistical and geospatial interoperability.
- Common geographies for dissemination of statistics.
- Geocoded unit record data in a data management environment.
- Used of fundamental geospatial infrastructure and geocoding.

Key elements

- Standards and good practices.
- National laws and policy.
- Technical infrastructure.
- Institutional collaboration.

Outputs

- Integration.
- Harmonized and standardise information.
- Interoperability and comparability.

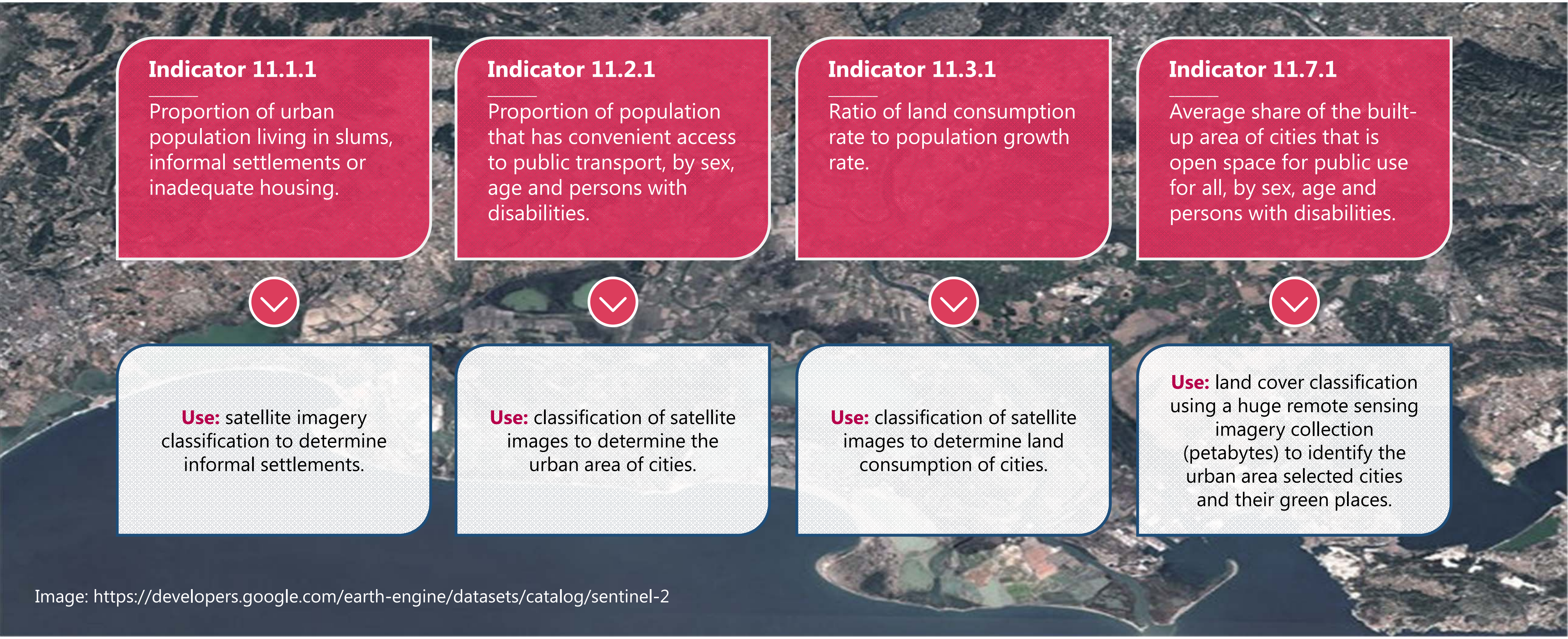
These serve as inputs for:

- Analysis.
- Diffusion.
- Decision making.

Source: Adapted from "THE GLOBAL STATISTICAL GEOSPATIAL FRAMEWORK. WORKING PAPER - FOR EG-ISGI CONSULTATION. UNITED NATIONS EXPERT GROUP ON THE INTEGRATION OF STATISTICAL AND GEOSPATIAL INFORMATION"

Use of satellite images for computing the SDG 11 indicators

Using satellite images, it is possible to obtain historical and updated data on land cover to analyze the expansion of urban agglomerations.



Indicator 11.1.1

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

Use: satellite imagery classification to determine informal settlements.

Indicator 11.2.1

Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities.

Use: classification of satellite images to determine the urban area of cities.

Indicator 11.3.1

Ratio of land consumption rate to population growth rate.

Use: classification of satellite images to determine land consumption of cities.

Indicator 11.7.1

Average share of the built-up area of cities that is open space for public use for all, by sex, age and persons with disabilities.

Use: land cover classification using a huge remote sensing imagery collection (petabytes) to identify the urban area selected cities and their green places.

Image: https://developers.google.com/earth-engine/datasets/catalog/sentinel-2

Poverty mapping

Integration of alternative sources of information in the statistical process

First approach we are currently working on:

1 NO POVERTY

Currently DANE measures:

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

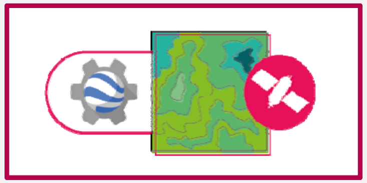
Goal:

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

Sources:

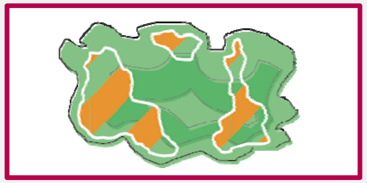
- Household surveys.
- Spatially detailed Census data.
- Geospatial covariate datasets.

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).
- Asses models' predictive performance.

Night-time lights as a proxy of economic activity during COVID-19 outbreak

Integration and comparison between night-time data and socioeconomic measures from firms

We are currently working on:

Currently DANE firm measures (per month):

- Production.
- Total production personnel.
- Power consumption.

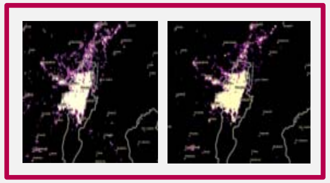
Goal:

- Measure changes in nighttime TOA radiance on a date before the COVID-19 outbreak and on a date during the lockdown. Correlate these changes with shifts in the economic variables by means of an econometric model.

Sources:

- Monthly Manufacturing Survey.
- Spatially detailed Census and firm data.
- Night-time light datasets (VNP46A1 - Radiance).

Methodology:



Compile night-time light data

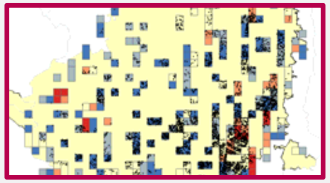
- Night-time light datasets from NASA's VNP46A1 sensor in dates before (07/02/2020) and during (27/04/2020) the outbreak lockdown.

Process and compare images

- Process the images (select suitable dates, apply cloud and vegetation masks, obtain the radiance difference between the two images, focusing on pixels with positive and negative values.

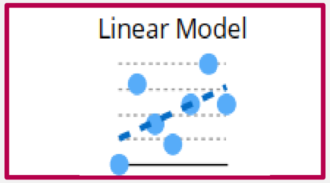
Georeference firms

- Locate the surveyed firms and for each one, calculate power consumption ratio.



Correlate

- Compare firm variables month by month and correlate them with changes in nighttime lights by building an econometric model.



Results and validation

- Assess model results.

Early estimation of manufacturing production

Integration of administrative registers and survey´s

First approach we are currently working on:

Currently DANE measures:

- Estimate anual industry production growth with a lag of 45 days.

Goal:

- Early anual estimation of manufacturing production level with a lag of 15 days.

Sources:

- Manufacturing industry survey.
- Google trends.
- Energy consumption administrative register.

Methodology:

$$\Delta\%Y_t = \beta\Delta\%X_t$$

$$\varepsilon_t = \alpha\varepsilon_{t+1} + \epsilon$$

Compile

- Covariate datasets (i.e, energy consumption index, Google trends).

Input

- Manufacturing enterprises survey.
- Energy consumption administrative register.

Modelling

- Forecast evaluation.
- Integrated moving average model.

Estimate

- Annual production index growth.

Results and validation

- Asses models' predictive performance.

Dissemination of geo-referenced statistical information:

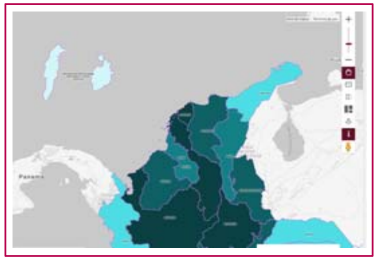
1. Vulnerability and mobility Geovisor

- Comorbidities.
- Adult population.
- Overcrowding.
- Population Density.
- Intergenerational risk.

2. Multidimensional Poverty Index Geovisor

- Educational conditions.
- Conditions of childhood and youth.
- Labor conditions.
- Health.
- Housing conditions and public services.

3. Population Census Geovisor



- Population.
- Dwellings.
- Socio-demographic indicators.
- Housing conditions and public services.

Geovisor

Microdata classification algorithm

Integration of social security register and household survey

First approach we are currently working on:

Currently DANE measures:

- In march and april, DANE measured the unemployment rate, but due to the changes in data collection, the infomality rate is not available for these periods due to a lack of information.

Goal:

- Estimate a dummy informality variable to impute the microdata of the household survey to recover the informality series.

Sources:

- Social security register – PILA.
- Household survey – GEIH.

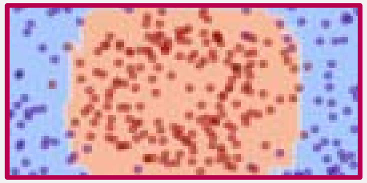
Methodology:

Compile

- Link PILA-GEIH.
- PILA status and GEIH covariates.

Modelling

- Machine learning.
- Classification model – Random Forest.



Estimate

- Random Forest model.
- Impute GEIH microdata

Results and validation

- Symmetric confusión matrix.
- Precision, Recall y F1 score greater than 0,8.
- The dummy indicator of the match is in the top 10 of predictors.

Other Big data initiatives

Current projects focused on the use and exploitation of alternative sources

1 Data imputation with machine learning methods

- Imputation of missing or inconsistent data from statistical operations for the purpose of modernizing the statistical production process.
- Pilot project with economic surveys.

2 Analysis of anomalies for the follow-up of the economic census

- Machine learning for the identification of possible failures in compliance with the collection protocol and inconsistencies.
- Prototypes developed and socialized.

3 Measuring the perception of trust in NSOs through the use of non-traditional sources

- Generate indicators to measure the perception of trust in NSOs and their statistical products.
- Currently developing the methodological document.

4 Now-casting method to predict the behavior of the GDP in the near future

- Define and test a model for estimating the 2019 GDP growth rate.
- Exploratory exercises to learn about the scope of **gnews** and **gtrends.**

DANE
INFORMACIÓN PARA TODOS

www.dane.gov.co

El futuro es de todos

Gobierno de Colombia