DANE
INFORMACIÓN PARA TODOS

DATA INNOVATION
DATA4NOW - COLOMBIA

KAREN CHAVEZ

GOBIERNO DE COLOMBIA

13th Meeting of the IAEG-SDGs
Bangkok, Thailand
November 2022

Experimental Statistics

DANE's approach for innovations

Are the ones derived from projects that have at least one of the following innovative aspects:

- new sources of information
- the statistical methodology used
- a new topic not previously measured.

Experimental statistics are **official statistics in Colombia**

1. **Improve the availability of relevant and timely statistics** with the required levels of disaggregation

2. **Integrate traditional and non-traditional sources of information**

3. **Generate capacities in the entity's staff** to avoid continued dependence on external agents to advance projects on new sources. Take advantage of scale economies

Published Experimental Statistics

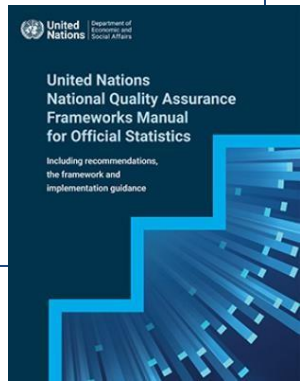
- SDG Indicator 1.2.2
- SDG Indicator 9.1.1
- SDG Indicator 11.1.1
- SDG Indicator 11.2.1
- SDG Indicator 11.3.1
- SDG Indicator 11.7.1
- Early Estimator for Manufacturing Industry in Colombia
- Estimation of population volume of the ethnic groups of Colombia

https://www.dane.gov.co/index.php/estadisticas-por-tema/estadisticas-experimentales

Quality Attributes

1. **Relevance**
2. **Opportunity**
3. **Accessibility**
4. **Interpretability**
5. **Coherence**
6. **Transparency**
7. Accuracy
8. Comparability
9. Continuity
10. Credibility

Attributes required for experimental statistics



DATA FOR NOW

United Nations | Department of Economic and Social Affairs Statistics

Global Partnership for Sustainable Development Data

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data

SUSTAINABLE DEVELOPMENT SOLUTIONS NETWORK

Working areas

Estimation of Poverty measures at subnational levels

- Multidimensional Poverty Index - MPI
- Monetary Poverty.

Interinstitutional work to establish key measures for public policy

- Measuring home to school accesibility and analysis of results
- Create a System of information for Statistics in Education

Work with civil society

Use of social media to extract data and process it using NLP to calculate indicators associated with perception of discrimination and political representativeness.

LGBTI+ Registry

- Support with resources for an expert consultant on LGBTI+ issues.
- Support with resources to support the platform's processes.

SIPSA (System of Prices and Supply in the Agricultural Sector)

Design a digital data collection mechanism for the statistical operation, SIPSA.

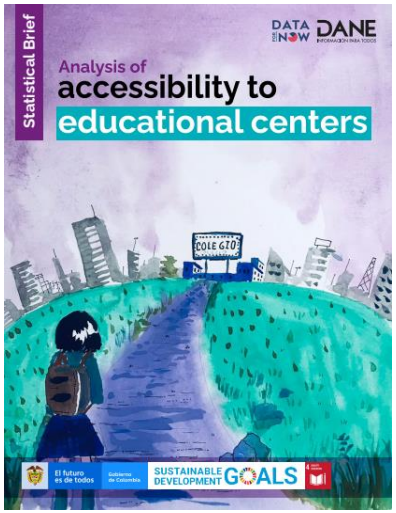
Cross-cutting areas

- IT architecture and data management
- Geospatial Information

Use of administrative data and distance techniques

Measuring home to school accessibility and analysis of results

- Work with the Ministry of education to identify needs in measuring home-to-school accessibility
- Explore descriptive statistics and a document published in English and Spanish
- Characterize the impact of accessibility on school dropout and attendance rates.



https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/serie-notas-estadisticas

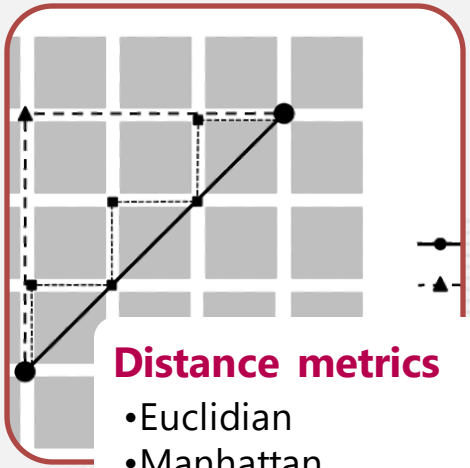
D4N Support:

- Consultant expert in measuring distances with satellite imagery.
- Consultant expert in Data Linking to merge data bases
- Consultant expert in IT infrastructure to interoperate the diverse data sources

DATA FOR NOW

4 QUALITY EDUCATION

Process and results



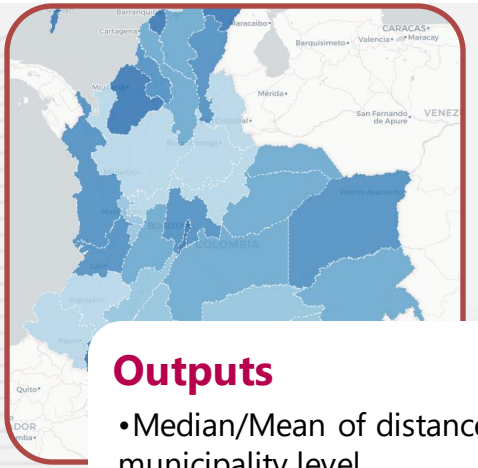
Distance metrics

- Euclidian
- Manhattan
- Routing

L	M	N	O	P	Q
ati_mgn	long_mgn	coord_x	coord_y	distancia_euclidiana	distancia_manhattar
6,3052182	-75,544525	-75,5453329	6,306702945	187,6962201	253,5722198
6,311398	-75,548363	-75,5453329	6,306702945	620,2464311	854,4469604
6,6781378	-75,22287	-75,5453329	6,306702945	54544,14668	76743,54688
6,3110008	-75,548042	-75,5453329	6,306702945	563,9476826	775,0056152
6,3102498	-75,547516	-75,5453329	6,306702945	462,3432074	633,7593384
6,3056369	-75,54377	-75,5453329	6,306702945	209,4977885	290,8425293
6,3079638	-75,54599	-75,5453329	6,306702945	157,8935525	212,1266785
6,31142	-75,548103	-75,5453329	6,306702945	607,3261326	828,1052856
6,3086729	-75,545059	-75,5453329	6,306702945	221,1313194	248,1464844
6,3076591	-75,545303	-75,5453329	6,306702945	106,3711211	109,0467072
6,3102698	-75,547729	-75,5453329	6,306702945	476,9009458	659,5436401
6,3089681	-75,543869	-75,5453329	6,306702945	299,3624856	412,4571533
6,3061938	-75,550423	-75,5453329	6,306702945	565,4094516	619,5536499
6,3063402	-75,551407	-75,5453329	6,306702945	672,5323155	712,2363281
6,309031	-75,543793	-75,5453329	6,306702945	309,8042936	427,8226624
6,4621491	-75,554474	-75,7028633	6,033594312	50396,97594	63807,39844
6,3078432	-75,544624	-75,5453329	6,306702945	149,0457787	204,5270844
6,3080339	-75,544868	-75,5453329	6,306702945	156,6617541	198,6127625
6,3108702	-75,547(				
6,3065319	-75,54484				
6,3077092	-75,54686				
6,3079491	-75,54597				
6,3093171	-75,57417				
6,309762	-75,54949				

Coverage challenges

- Euclidean and Manhattan: close to 70%.
- Routing: close 30% in urban areas



Outputs

- Median/Mean of distances at municipality level.
- Choropleth maps with the mean/medians.
- Dropout rates by distance deciles

Main results

We found evidence of positive correlation between distance deciles and dropout rates. When disaggregating results in public and private schools, we found that differences in dropout rates between distance deciles were only relevant in public schools with no significant differences for private schools.

Guide for the inclusion of the differential and intersectional in the statistical production of the national statistical system

Rationale

For a Statistical Operation specific for LGBTI+ Population

1. Adequate identification and characterization of the LGBTI+ population are necessary to inform the design and monitoring of plans, programs, and relevant public policies to achieve the realization of their rights.

2. Statistical visibility is important to promote recognition and respect.

3. As explained in the OECD report "Society at a Glance, 2019", discrimination against this population implies economic and social costs for countries

INFORMACIÓN PARA TODOS

Statistical Operation on Sexual and Gender Diversity

Phase 1: Voluntary registry for the visibility of sexual and gender diversity in Colombia.

Objective: Compilation of a database for aggregated statistics.

- This registry requests basic contact information, sexual orientation and gender identity, and geographic location.
- It allows us to identify the number of LGBTIQ+ people in Colombia and the number of people in each group.

D4N Support:

Consultant expert in LGBTIQ+ issues, to mediate dialogues with different actors on the questionnaire for the identification of the LGBTIQ+ population, given that it is the first specialized statistic operation for this population in Colombia.

Phase 2: Vitual Survey

As a result of the Phase 1 dialogues, a survey was developed and will be conducted through a digital tool for the data collection .

D4N Support:

A consultant to develop a digital Tool to collect the data

1. Thanks to Internet coverage, there will be a higher response rate in areas that are difficult to access.
2. Generation of confidence in respondents when disclosing information.
3. Elimination of the surveyors' bias, who sometimes assume answers based on previous answers.

INFORMACIÓN PARA TODOS

Calculation of proxy SDG indicators 16.b.1 and 16.7.2

For the 16.b.1: Percentage of Facebook users, by gender, with comments that include discriminatory language.

For the 16.7.2: Facebook users with comments that include language related to inclusive and decision making.

D4N Support:

Consultant expert in NLP methods
Consultant expert in social media data
Consultant to manage and process the data

- The usage of data from socia media is quite feasible. Still standardization protocols for collect, process, analyse and guarantee the confidentiality of gathered data are needed.
- NLP models are useful at least to gather complementary information for official statistics. Furthermore, NLP models could be used for other projects where images or text analysis are required.
- Strengthen the technical capacities in the staff of the NSO is needed.
- Methodology is already published in our experimental statistics website.
- Currently we are working with the Presidential Counselor to publish the results in a dashboard, and summarized results to authomate the process .

INFORMACIÓN PARA TODOS

Building a digital data collection tool for SIPSA

SIPSA (acronym in Spanish for Information System of Prices and Supply in the Agricultural Sector)

Information on prices and supply are directly related with food insecurity

D4N Support:

Consultant expert in NLP methods
Consultant expert in processing images and audios

Some notes

- This is an ongoing process we have just started the work with one of the consultants
- SIPSA collection processes occur in the morning (2am-6am) and there is no too many time to collect all data and currently we collect data manually in supply centers
- We aim to improve the quality of the data collected by processing audios and images
- We aim to authomate the collection process for producing diary reports that we publish

DANE
INFORMACIÓN PARA TODOS

GOBIERNO DE COLOMBIA

MANY THANKS!

direccion@dane.gov.co