# Big Data for the National Agricultural Census,Colombia2014

Sandra, Rodriguez; Sandra, Moreno; Yineth, Acosta; Cesar, Maldonado; William, Martinez; Anibal, Montero; Alvaro, Murillo.
syrodriguezf@dane.gov.co, slmorenom@dane.gov.co,
yacostab@dane.gov.co,camaldonadom@dane.gov.co, wialmadi@gmail.com,
amontero@dane.gov.co, amurillor@dane.gov.co
National Administrative Department of Statistics –DANE
Carrera 59 No. 26-70 Interior I - CAN
Bogotá, Colombia
dane@dane.gov.co

DOI: 10.1481/icasVII.2016.f29c

## ABSTRACT

National Statistics Offices (NSO) are including new information sources, as Big Data, specifically Earth Observation (EO) data into their production processes, in order to improve official statistics for different subjects. Under this approach, DANE, the NSO of Colombia, carried out a pilot project to complement the information on land cover for the agricultural statistical framework of Colombia in three municipalities of the Atlántico department and their multi-temporal analysis. Processing was carried out with available Landsat satellite images for the years 2005, 2010 and 2014 usingErdas, InterIMAGE and ArcGIS software. As a result, the variable land cover was imputed for the 20% of the agricultural statistical framework units without information. It is concluded that, in order to get more disaggregated coverages, further research must be done including higher resolution images and multi-temporal analysis in periods shorter than one year.

**Keywords:**Satellite images, land cover, object-basedimage analysis, data imputation.

## 1. Introduction

43 years after the realization of the last agricultural census in Colombia, and with the aim to know the present situation of agriculture and livestock in the country, during 2013 and 2014 DANE carried out the third National Agricultural Census (CNA for its acronym in Spanish).During the country-wide field collection process more than 113 million hectares were covered, in the 1101 municipalities distributed across the 32 departments that integrate the country.It is important to highlight that besides the administrative division of the territory, 181Afro-descendant communities' lands, 773 indigenous territories and 56 national natural parks were included in the census operational design.

Nevertheless, census coverage was under 70% in some municipalities because of:

- Inability tocontact producers orsuitable informants.
- Uninhabited properties.
- Inaccessibilitydue to geographical and public order conditions
- Rejection to census enumerators.

Following these shortcomings and for the construction of the National Agricultural Framework, DANE carried out a pilot project to complement the information about land cover using Big Data specifically remote sensing data.

Big Data is defined as "*any large and complex collection of data sets that becomes difficult to process using traditional data processing applications", and therefore includes data from Earth Observations satellites. In many cases Big Data is considered to be unstructured data or data that is used for another aim than the original creator of the data intended. In the case of Earth Observation it is actually purposefully created to be highly structured and to measure specific aspects of the earth. The complexity of EO data is to extract a meaningful form of information about a real object, value, state or condition from an electromagnetic signal measured in space"*(Task Team Satellite Imagery, Remote Sensing and Geo-spatial data, 2016).

Through the extraction of land cover data from satellite images, the goal is to impute the variable "land cover" for the analysis units of the National Agricultural Framework. This will provide detailed and high quality agricultural statistical information for policies formulation and implementation, as those currently is being implemented in Colombia, for example the Land Law and Rural Development.

## 2. Background

### National Agricultural Census - CNA

CNA was conducted during 2013 and 2014, to collect information about agriculture, forest, aquaculture and fisheries;as well as socio-demographic information fromagricultural producers and rural inhabitants. This was a declarative census, in which information was provided by suitable respondent, who knowsabout the activities undertaken at each observation unit. The census analyticalunit is the Agricultural Production Unit (UPA for its acronym in Spanish)which is defined as the area in which a single producer[1] makes decisions and takes responsibility over the productive activity. The UPA has no size limits, and it can be formed by part of a rural property, a complete rural property, or a group of continuous rural properties.

The National Agricultural Framework is a spatial representation of the UPAs and non-agricultural units (UPNA) located in rural area; and therefore, it is the basis for sampling design in agricultural surveys and any statistical operation with rural domain.The Framework includes information about thelocation (X, Y coordinates), identification (property code and politic-administrative information), and agriculturespecificdata from each unit.Given the importance of the framework for DANE regular statistical production, the information of the 3rd National Agricultural Census is fundamental for updating this tool.

Based on the information from the 3rd CNA, ten classes for land use and land cover were stablished, which are grouped into six categories (Table 1). For each UPA was assigned a predominant land coverage defined by the coverage whose participation percentage is equal or greater than70% of the UPA total area. If the condition was not fulfilled, 'mixed' predominant coveragewas assigned.

[1]The concept of agricultural producer includes individuals (natural person), as well as companies or formal partnerships (legal person)

| GENERAL CATEGORY | CLASS | DESCRIPTION |
|---|---|---|
| Temporary | Temporary crops | "Temporary crops are those which are both sown and harvested during the same agricultural year, sometimes more than once." (FAOSTAT, 2015) |
| | Fallow land | Area that had crops during the last 12 months, but had no crops the day of the interview. |
| | Roosting area | Area that had crops in the last 3 years, but had no crops in the 12 months prior the interview. |
| Permanent | Permanent crops | "Permanent crops are sown or planted once, and then occupy the land for some years and need not be replanted after each annual harvest." (FAOSTAT, 2015) |
| Grassland | Grassland | Land on which the vegetation is dominated by grasses, glasslike plants and herbaceous. |
| Non-agricultural | Non-agricultural | Land where the properties are dedicated to develop non-agricultural activities, such as industry, commerce and services. |
| Forest land | Forest plantation | "Forest stands established by planting or/and seeding in the process of afforestation or reforestation" (FAO, 2016). |
| Other land cover and uses | Other land cover | Soils covered by natural waters, *páramo* vegetation, bare soil, rocky outcrops, opencast mines, etc. |
| | Stubble | Land area where more than three years have passed since last crop and is covered by shrubby vegetation. Census day had no crops. |
| | Natural forest | Land with different classes and associations of trees, shrub, herbaceous, and other plants which are not classified as a forest plantation. |

**Table 1.** Land use and land cover categories for the National Agricultural Framework.

## Related works

Remote sensing in agriculture has been applied since the seventies, covering a wide variety of fields (Carfagna, 1999), (Hanuschak & Delincé, 2004) and (Gallego, 2006); among them, vegetation monitoring and land cover mapping is highlighted.

Other remote sensing applications are related to the survey design (sampling frame and stratification), crop identification, planted areas estimation, crop status in large areas and production estimation. Remote sensing complemented with GIS exhibits great potential for creating information layers on crops that may include location at a centimeter scale for precision agriculture (Craig & Atkinson, 2013).

Furthermore, United Nations Statistics Division has recognized the need for further investigating the benefits and challenges of using Big Data for official statistics. To address this matter, the Working Group on Satellite Images and Geospatial Data was established in 2014; in this context,several pilot projects were developed, including the application of remote sensing for agricultural statistics production, presented by the Australian Bureau of Statistics (ABS).

## 3. Methodology

In order to impute the information on land use and land cover to the UPA that lacked of this information,a remote sensing methodology was implemented. In this case, the steps were as follows:

Step 1. Defining the areas of study. The main considerations were related to land cover defined by the CNA.

- Municipalities with slope less than 20° (to avoid topographic correction on the images)
- Municipalities that show natural and/or anthropic precedents of changes on land cover that could be of interest for a multi-temporal analysis.

Three municipalities metthese criteria:*Campo de la Cruz*, *Santa Lucía and Suan*, located on the Caribbean coast, northern Colombia (Figure 1).


During 2010 - 2011 an intensive rainy season generated floods and landslides in different zones of Colombia. One of the biggest effects was generated by the rupture of the *Canal del Dique*dike, which caused flooding in several towns, including the ones chosen for the study.(CEPAL, 2012).

Figure 1. Study areais 852 km$^2$ (828 km$^2$ rural area).

Step 2. Satellite data selection. The imagery acquisition was done under these criteria:

- Total coverage of the study area, to avoid the need to generate a satellite mosaic[2].
- Similar sensors for the years 2005, 2010, and 2014.
- Images taken over the same area and the epoch of census field work(2014, first trimester) for 2005, 2010, and 2014.

The Landsat images met those criteria and they were downloaded free of charge from the Earth Explorer website[3] (United States Geological Survey, USGS)see Table 2.However, Landsat 7 images captured since May/2003 have gaps generated by a failure in the SLC sensor. Therefore, a filling procedure was done by using two images from the same sensor.

[2]A mosaic is one image created by merging several individual images of adjacent areas.

[3]http://earthexplorer.usgs.gov/

| Program/Sensor | Path-Row | Date |
|---|---|---|
| Landsat7ETM+ | 9-53 | 23January2005 |
| Landsat7ETM+ | 9-53 | 24February 2005 |
| Landsat5TM | 9-53 | 29January2010 |
| Landsat8OLI | 9-53 | 9February2014 |

**Table2**.Landsat imagery chosen for the project

Step 3. Satellite images pre-processing.This procedure included the following activities:

- Filling the gaps for the 2005 images.The Phase 2 Gap-Fillalgorithm proposed by the USGS and NASA was used. This algorithm fills the gaps of one primary image using another image with valid data in the gap zone.A pixel with no data from the primary image is replaced by the adjusted value of same pixel of the secundary image.The pixel is adjusted using the value of the standard deviation and mean of neighboring pixels (valid pixels in both images). Every spectral band was processed separately.

- Layerstacking. This procedureconsistsof the combination of multiple separate bands in a single image. As the number of bands of the Landsat sensors (ETM+, TM and OLI) is different, only the bands common to all images were used. In this project, this process was made with the *"Layer Stack"*Erdas's tool.

- Defining LANDSAT image subset.A portion of a larger image was generated covering the study area.

- Radiometric resolution scaling:The Landsat image for 2014 has a radiometric resolution of 10 bits, and the 2005 and 2010 images have an 8 bits value. Since it is desirable to have all the images in the same rank of digital levels, a scaling radiometric resolution procedure was applied to the 2014 image to ensure the same radiometric resolution, i. e. 8 bits.

Step 4. Satellite images classification
Digital image classification involves grouping pixels to represent land cover features. There are several image classification methods and for the purposes of this project, object-based imageanalysis was chosen.

This methodis carried outthroughtwo general procedures:

- Image segmentation: This process creates objects by grouping similar pixels. It starts with each pixel forming one image object, which is sequentially merged to formbigger ones.For this process, the TA_Baatzsegmenter algorithm, which creates objects based on colour and shape (Baatz& Schäpe, 2000), was used, along with the InterIMAGE free software.

- Classification: In this process, each image object is allocated to a particular class, according to the criteria and rules defined.

Step 5. Global accuracy assessment.

To evaluate how well the classification represents the real world, the following taskswere executed:

- Estimation of the sample size (n) by using the binomial distribution formula.(Rossiter, 2014).

$$n = \frac{Np(1-p)}{(N-1)\frac{d^2}{Z^2_{(1-\alpha/2)}} + p(1-p}$$

Where:

n s mple size.
N universe (total image objects)
D sampling error (5%)
A confidence level (95%), and
*p* *apriori* precision (90%)

- Distribution of the random sample with the "create random points" ArcGIS tool.
- Visual classification of the sample, using Google Earth and orthophotos, taken as close to the date of the satellite images as possible.
- Comparison between visualand automatic classifications (InterIMAGE results) for sample points.
- Estimation of the coincidence percentage between the visual and automatic classification.

Step 6. Data Imputation.

To impute land cover information to the UPAs with no data, a model was implemented using ArcGIS tools. The model takes the land cover map obtained from the image automatic classification, and split the map into UPAs. As a result, every UPA without land cover information was allocated a predominant land cover.

Step 7.Change detection

In order to quantify the changes in land cover during the period of study, a pixel by pixelcomparison was conducted between the 2005 -2010, and 2010 -2014land cover maps. As a result, a change detection matrixwas created; which is a table that allows to measure changes among different land-cover/land-use classes over a time period (Jensen, 2005).

## 4. Resultsand discussion

The classeswere defined considering the CNAgeneral categories (Table 1). Nevertheless,the categories "permanent crops" and "forest" were excluded due to their small area size and quantity. The non-agricultural land category cannot be identified by satellite images, since it is mainly related to industry and services uses.

Table 3 shows the decision rules defined to classify each objects into the land cover categories. The objects which did not meet any decision rule were assigned to the Temporary category.

| **General category** | | **Decision rule** | | |
|---|---|---|---|---|
| | | **2014** | **2010** | **2005** |
| Oher land cover and | Water | ratio(band5)<0.21 | mean(band4)≤ 50.4 | mean(band4)≤ 53.5 |
| | Bare | bandMeanDiv(band5,band2)<2.3 | mean(band1)> 83.5 | mean(band1) ≥89 |

| | | | | |
|---|---|---|---|---|
| uses | soil | ratio(band5)≥ 0.21 | | ratio(band5)<0.228 |
| | Stubble/ Natural Forest | ratio(band5) ≥0.32 | mean(band2) ≤ 37<br>mean(band4)> 50.4 | mean(band2)≤ 80<br>ratio(band2)<0.2 |
| Grassland | | ratio(band3)>0.11<br>ratio(band5)≥ 0.21 | ratio(band5) ≥0.285 | ratio(band5) ≥ 0.228 |

**Table 3.** Decision Rules used in each year.

The global accuracy of the classification for each yearis presented in Table 4. The highest accuracy was obtained for the year 2014.

| Year | Samples | Coincidence percentage |
|---|---|---|
| 2005 | 135 | 57% |
| 2010 | 135 | 70% |
| 2014 | 136 | 84% |

**Table 4.** Global accuracy of the classification.

Figures 2, 3 and 4show the land cover distribution for the years 2005, 2010 and 2014.

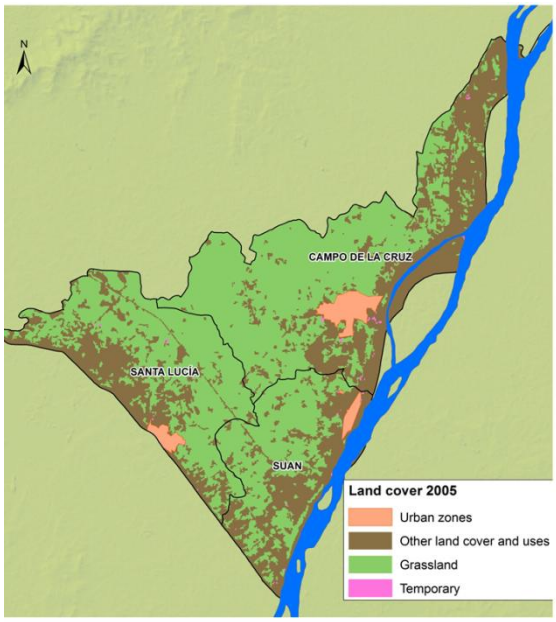


**Figure 2**. Land cover year 2005

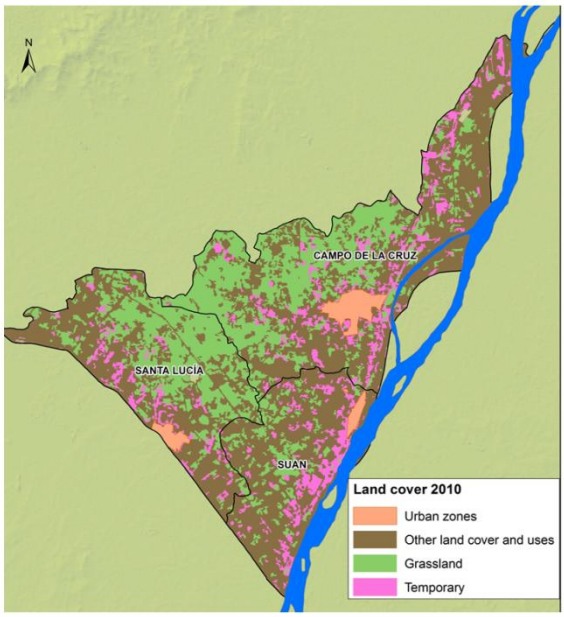


**Figure 3**. Land cover year 2010

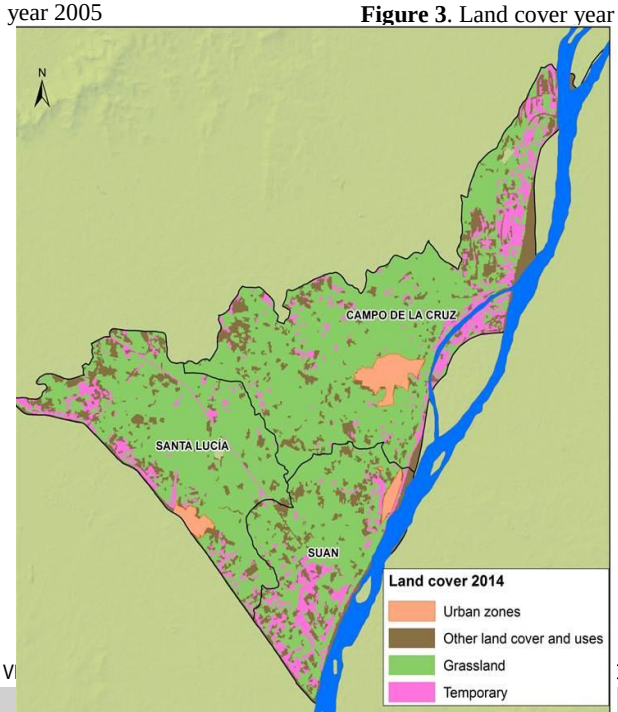

**Figure 4.**Land Cover 2014

With the 2014land cover map obtained from classification,it was possible to impute the predominant land cover to598 UPAs without census information (out of 2.975). Figure 5 shows the land cover map obtained from census results and imputation. Grassland areas are prevailing in the three municipalities;temporary and permanent cropsareas are located mainly in Suan and the riverbank.

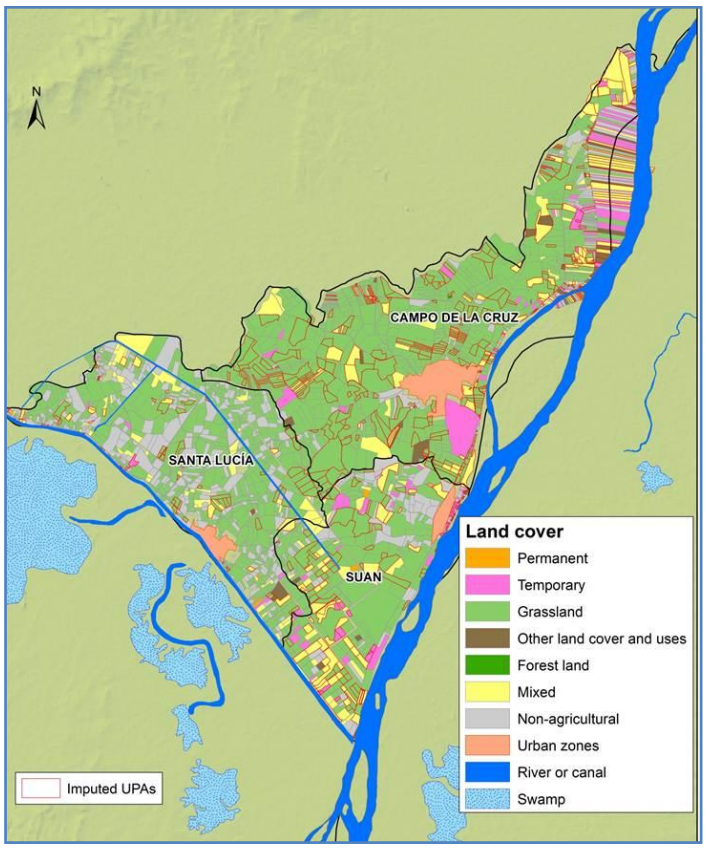


**Figure 5.**Predominantcoverage by UPAs. 2014

Once the post-classification and comparison methods were applied, changes were detected asreportedin Table5. From 2005 to 2010, there was an increase in land classified as 'Other land cover and uses', which includes Stubble and Natural forest. However, the most significant land cover changes are identified during the period 2010-2014, asit shows an important increase in grassland, and decrease in Oher land cover and uses.

| 2005-2010 | Temporary | Grassland | Oher land cover |
|---|---|---|---|

| | | | and uses |
|---|---|---|---|
| **Temporary** | 22% | 17% | 61% |
| **Grassland** | 9% | 46% | 45% |
| **Other land cover and uses** | 13% | 11% | 76% |

| 2010-2014 | Temporary | Grassland | Oher land cover and uses |
|---|---|---|---|
| **Temporary** | 8% | 79% | 12% |
| **Grassland** | 6% | 86% | 8% |
| **Other land cover and uses** | 16% | 58% | 25% |

**Table 5.**Change detection matrix.

A possible explanation to the 2005-2010 phenomenon is that during this time, the more common crops were scattered mango and guava trees. Due to the characteristics of these crops, they were classified as Stubble and Natural forest. It concludes that patterns of fruit trees and Other land cover and uses are similar. The increase in this category can be related to the rise of fruit trees.On the other hand, for the 2010-2014 changes, they might have been caused by the 2010 -2011 floods, as this climate phenomenon and the subsequent increase in soil moisture, may have induced a faster degradation of agricultural and productive soils.

## 5. Conclusions

Satellite images are a valuable source of information on land cover and they provide suitable complementary information for areas where field data are not available.More detailed results require satellite images with higher resolution (spatial and spectral) preferably collected at different months during the same year. This could be useful to make decisions based on productive cycles (temporary or permanent), crop areas, sowing and harvest periods, and to improve the results of the CNA.

The study results prove the change detection matrix to be an useful tool to measure changes in land cover; however its precision depends on the classification accuracy. In addition, it is important to highlight that an adequate use of image classification methods requires knowing the study area, in order to correctly define land cover classes and decision rules. Also, although the classification software InterIMAGE is of free access and easy to use, documentation on it use is limited.

In addition, it is important to explore remote sensing automatic techniques that allow replicate the process in other areas of the country. This could lead to face new challenges as dealing with images and algorithms for cloudy areas, development of topographic correction methods and assessment of computational efficiency for Big Data.

Finally, it was prove that Big Data from satellite imagery is considered a powerful tool to overcome lack information, and with greatpotential for the usein National Statistics Offices.

## REFERENCES

Baatz, M., &Schäpe, A. (2000). Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation. *Angewandte Geographische Informationsverarbeitung XII. Beiträge zum AGIT-Symposium Salzburg*, 12-23.

Carfagna, E. (1999). Using remote sensing for agricultural statistics. (P. t. 52nd Session, Ed.) *Bulletin de I'Institut International de Statistique*.

CEPAL. (2012). *Valoración de daños y pérdidas Ola invernal en Colombia 2010-2011*.Retrieved 04 14, 2016, from http://www.cepal.org/publicaciones/xml/0/47330/olainvernalcolombia2010-2011.pdf

Craig, M., & Atkinson, D. (2013). *A literature review of crop area estimation.* Retrieved 04 15, 2016, from UN-FAO: http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/GS_SAC_2013/Improving_methods_for_crops_estimates/Crop_Area_Estimation_Lit_review.pdf

Gallego, F. (2006). *Review of the Main Remote Sensing Methods for Crop Area Estimates Agriculture unit.* Retrieved 04 23, 2016, from Compilation of ISPRS WG VIII/10 Workshop 2006, Remote Sensing Support to Crop Yield Forecast and Area Estimates, Stresa, Italy, Agriculture Unit, IPSC, JRC: http://www.isprs.org/proceedings/XXXVI/8-W48/

Hanuschak, G., & Delincé, J. (2004). *Utilization of Remotely Sensed Data and Geographic Information Systems (GIS) for Agricultural Statistics in the United States and the European Union. Third International Conference on Agricultural Statistics, ICAS III – MEXSAI – Measuring Sustainable Agri.* Retrieved 04 22, 2016, from http://www.nass.usda.gov/mexsai/papersabstracts.htm.

Jensen, J. (2005). *Introductory Digital Image Processing: A Remote Sensing Perspective.* Pearson - Prentice-Hall.

Rossiter. (2014). Technical Note: Statistical Methods for Accuracy Assesment of Classified Thematic Maps. *Department of Earth Systems Analysis University of Twente, Faculty of Geo Information Science & Earth Observation (ITC), 25*, 107.

SDSN. (2016, 01 20). *Indicators report*. Retrieved from http://indicators.report/indicators/i-68/

Task Team Satellite Imagery, Remote Sensing and Geo-spatial data. (2016). Draft of Report on earth observations to the UN Statistics Commission. Chapter 2 - "Sources". 43 p.