United Nations ECE/CES/GE.41/2018/5

Economic and Social Council

Distr.: General
6 July 2018

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twentieth Meeting
Geneva, 26–28 September 2018
Item 7 of the provisional agenda
Geo-spatial information

Use of geocoding technologies and administrative records for the georeferencing of the units of the 2018 National Population and Housing Census in Colombia

Note by the National Administrative Department of Statistics, Colombia*

Summary

In the development of the 2018 National Population and Housing Census – CNPV,[1] the National Administrative Department of Statistics (Departamento Administrativo Nacional de Estadística – DANE) integrated information from administrative records and geocoding services in order to ensure the georeferencing of the dwellings. For this process, an address register geo-referenced database was formed based on the integration of the National Geostatistical Framework (MGN),[2] the national cadastral database with more than 15 million records, the information relating to geographic names (toponyms) corresponding to 3.3 million records and the geo-referenced database pertaining to the national villages with 32,308 records. From the address register database, a normalization and standardization application was developed that allowed linking each dwelling registered in the electronic census[3] to its respective geographic coordinates. For those records that were

* Prepared by Miguel Ángel Cardenas, Pedro Franco, Josué López and Sandra Moreno.

[1] Acronym of National Population and Housing Census in Spanish.

[2] Ibidem. The MGN is a system designed by DANE, to allow referencing the statistical information of the censuses, the surveys by sampling, derived statistics and the information obtained from administrative records, to their geographic location.

[3] DANE is conducting the National Population and Housing Census – CNPV throughout the course of 2018; the collection of the information thereof was designed to be in two phases: the first one, by means of an electronic form (known as electronic census or eCensus) that the households completed

not located in the address register, a georeferencing process was carried out by means of the road network, Geographic Information Systems software and assisted methods on digital cartography. 97% of the information collected by means of the electronic census was georeferenced through the use of these methods. Additionally, these data were arranged in a geographical viewer, which facilitated the analysis and monitoring of the electronic census coverage in the national territory.

I. Introduction

1. The Population Censuses require the generation of georeferenced statistical information at the microdata level so as to meet the demands of geographical disaggregation of the different users, to respond to the commitments derived from the 2030 Agenda and to potentiate the usefulness of statistics at the local level.

2. In order to comply with this demand for information, the National Statistical Offices must take advantage of the Geographic Information System technologies for the collection and georeferencing of the data, strengthen inter-institutional work, especially with the Cartographic Agencies, and link administrative records that allow consolidating the databases of census units.

3. The above is correlated with the first principle proposed by the United Nations (2014) for the integration of statistical and geospatial information, which raises the importance of the use of fundamental geospatial infrastructures and geocoding. The purpose of this principle is to obtain high-quality standardized addresses, an identifier of the property or building or other description of the location and / or a small geographical area or a standardized grid reference to each statistical unit (i.e., at the microdata level).

4. In Colombia, the geospatial component has played a relevant role in the development of statistical operations. Recently, in the execution of the 2018 National Population and Housing Census, DANE optimized the georeferencing process of dwellings by means of the integration of administrative records and the development of geocoding services. As a result, 97% of the households registered electronically were georeferenced with these processes, which reduced processing times and made it possible to track the coverage of the electronic census in a timely and efficient manner.

5. This article presents the experience of Colombia with respect to the use of geocoding technologies and administrative records for the georeferencing of the units of the 2018 National Population and Housing Census. Specifically, it describes the methodology of the process used, the results obtained and the main lessons learned as well as the future challenges that were identified.

II. Conceptual framework and background

A. Conceptual framework

6. Geocoding is the process of assigning coordinates to objects in space from elements that allow their association to a specific place (addresses, points of interest, etc.). In turn, Google Code (2009) defines it as the translation of information available in a natural language into geographic coordinates.

directly, during the period between January and April 2018; the second one through the traditional door-to-door method, with the visit of a DANE interviewer; this phase of the operation is currently being conducted as of April 2018 and is scheduled to be completed by July 2018.

7. Geocoding can be performed through different methods, including automatic (based on road network), assisted geocoding (manual) and by using administrative records. In the first method, georeferencing is performed by means of geocoders available in a specialized GIS-type software, which allow the geographic coordinates (latitude, longitude) equivalent to a dwelling address or nomenclature to be calculated from the existing road network. In assisted geocoding, in the event of not having a previously geo-referenced input, possible mechanisms for manually assigning coordinates are identified from the search of the address or toponyms through services such as Google Street View, Google Maps, Google Earth, OpenStreetMap or local maps servers.

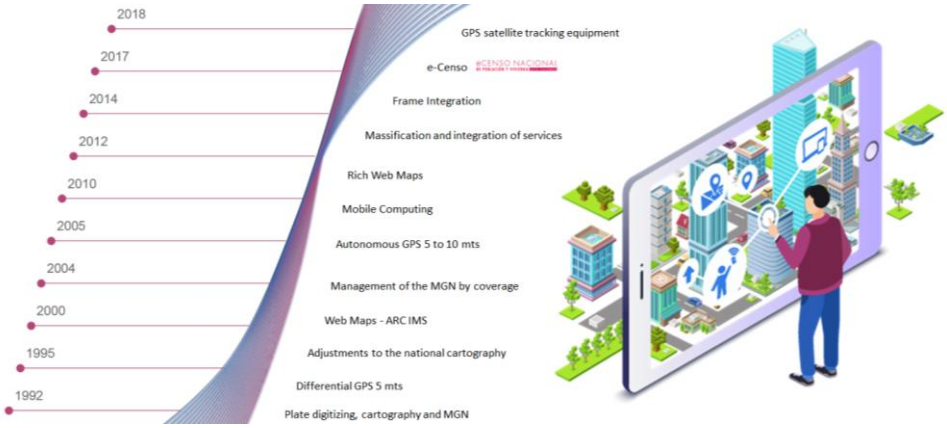
8. The third method related to the use of geo-localized administrative records is based on the comparison between the address of the source (normalized and standardized) and the address registered in the georeferenced database of the cadastral register and place names, once the correspondence of the address is identified, the corresponding coordinates and census code are transferred to the source.

9. These methods allow georeferencing units to point referencing, which according to the United Nations is highly preferable when compared to only associating statistical units with a geographic region (i.e. a polygon). The use of point-type references allows considerable adaptability to change in geographical regions over time or to new geographic delimitations that may arise (UN-GGIM, 2014).

B. Background

10. DANE, in the last 25 years, has carried out innovations that have allowed technological leaps in favour of the transformation and continuous growth of the statistical process (see Figure I). Taking into account that every event which occurs at the geographical or societal level, occurs at a specific geographical point, and that event in turn is associated with statistical information, one of the fields with the greatest advances has been the assigning of statistical information to the geographical area.

Figure I
Evolution of the use of geospatial technologies for georeferencing in DANE



11. In 1992, the digitalization of analog cartography for the design of the National Geostatistical Framework (MGN) was started, in order to reference the statistical information obtained from the 1993 National Population and Housing Census (CNPV), and other sources, such as sample surveys. In 2005, for the conducting of the General Census, devices with Global Positioning System (GPS) capabilities with accuracies between 5 and 10 meters were used, thus obtaining the georeferencing of the coverage units (dwellings

and special accommodation sites), which allowed making information available to users at different levels of geographic aggregation or disaggregation, as applicable.

12. For the conducting of the 2014 Third National Agriculture and Livestock Census, the process pertaining to the integration of frameworks, (i.e. the integration of the MGN with the cadastral rural property framework), started to be developed in order to ensure the definition and delimitation of agriculture and livestock production units. This process currently continues with respect to the definition of the urban and rural coverage units of the 2018 National Population and Housing Census, which is being conducted at this time.

13. Undoubtedly, one of the major innovations for the 2018 CNPV was the implementation of the electronic census, which allowed this process to be carried out from the households themselves through the Internet, achieving an estimated coverage of more than 10%. It is important to mention other methods and inputs that have allowed obtaining geo-referenced information in order to achieve the objectives and goals set out in the institutional mission. The following stand out among them: SmartData strategies; the definition of multi-thematic frameworks; the counts of buildings; the use of Handheld Data Capture Devices (DMC),[4] participatory social mapping workshops for the location of ethnic communities; and the complementation of the MGN by means of administrative records.

III. Methodology

A. Data and inputs

14. The following inputs were used for the georeferencing process of the units of the 2018 National Population and Housing Census.

(a) Cadastre register at the level of each property in the country, with normalized and standardized addresses and at the rural level the names of the georeferenced properties;

(b) National Geostatistical Framework – MGN for 2017: which includes the 1,101 municipalities of the country, 20 areas not organized into municipalities and the San Andrés, Providencia and Santa Catalina Archipelago;

(c) A database of toponyms: Geographical names obtained from digital cartography;

(d) Geo-referenced establishments from the statistical business register (approximately three million records);

(e) Geographic layer at the level of the country's villages: Rural subdivision defined by the municipality based on the organization of the population by self-recognition (with more than 30,000 records);

(f) Basic reference cartography at different scales of the Agustín Codazzi Geographic Institute – IGAC.[5]

B. Methodology

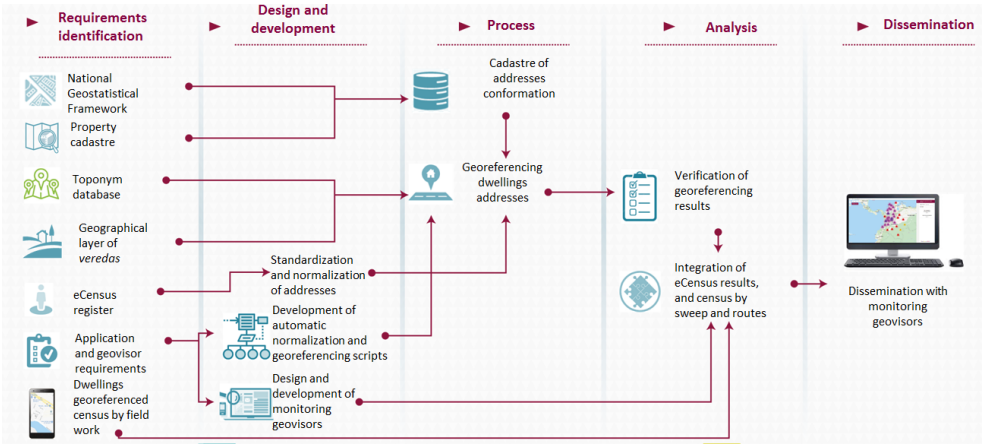
15. The general diagram in Figure II below illustrates the general phases for the development of the georeferencing with respect to census units using geospatial technologies and administrative records.

Figure II

[4] Acronym of Handheld Data Capture Devices in Spanish.

[5] Acronym of Agustín Codazzi Geographic Institute in Spanish

General Process Flow Chart (Source: DANE)



16. In a first phase, the basic requirements with respect to information and applications for the georeferencing of the 2018 CNPV census units were identified. For this purpose, administrative records of the Real Estate Cadastre were collected for the country, the National Database of Place Names, the geographic level of villages, the IGAC's basic reference cartography to different scales, the National Geostatistical Framework and the consolidated database of records from the eCensus with the corresponding standard addresses of the dwellings. This information was the basis for the georeferencing process of the census units.

17. In terms of the applications requirements, the need was identified to develop tools, geoviewers and acquire services in order to monitor the census operation in its electronic and face-to-face census phases (sweep[6] and route[7] methods).

18. In order to ensure that the activities pertaining to control and monitoring of the location with respect to operational personnel by routes were executed as planned, and for the timely making of decisions pertaining to the advance of the field operation, DANE acquired the satellite tracking and location service using SPOT GEN 3 satellite tracking devices.

19. In the design and development phase, algorithms (routines) were prepared for the standardization and normalization of the addresses, i.e., from the structure of the nomenclature established in each municipality; the address record was standardized to facilitate its search in the administrative records. In turn, georeferencing services were optimized based on geographic information systems and road networks. Simultaneously, the requirements for the design and development of the geoviewers were analyzed, identifying the information to be presented, the geographical disaggregation (department, municipality, municipal township, population centers, operational areas, etc.) and the geographic search and visualization options.

20. The activities of the processing phase focused on the conformation of the address register database from the integration of the information of the National Geostatistical Framework and the real estate cadaster, which allowed generating the georeferenced register with respect to the country's dwellings that includes the corresponding DANE census code.

[6] Visit to domiciles door-to-door.

[7] Areas with a difficult access.

21. An essential element that allowed linking each dwelling registered in the eCensus with its respective geographical coordinates was to request the user to enter the address in the web application in a standardized way, as well as to have a normalization and standardization process of addresses. With this information, procedures stored in the Oracle 11g database engine were executed that associate the electronic census address and its correspondence in the address register database (ratio 1: 1), assigning the geographic coordinates and the DANE census code. In the cases in which a direct correspondence was not found, a search process of similar records was carried out in the address register database that allowed determining its approximate location.

22. In instances where it was not possible for a given dwelling to be geo-referenced with the above-mentioned methods, the toponyms and villages database was used to associate it with the additional location variables of the electronic form (neighborhood, ward or village) through methods of analysis by semantics and phonetics available in Oracle, which allowed obtaining an approximate location.

23. For the records that were not georeferenced with the previous method, a second procedure was performed, which was the analysis by automatic geocoding, and it was carried out by means of tools of geographic information systems and road networks – ArcGIS – StreetMap Premium.

24. As a third option, manual georeferencing activities were carried out by means of alternative resources, such as digital cartography searches from other sources (Google Maps and OpenStreet Maps, among others).

25. In the analysis phase, the preliminary results of the execution of the georeferencing routines were iteratively reviewed and subsequently the geo-referenced results of the eCensus were integrated with the records obtained in the field from the face-to-face census.

26. Two geoviewers were developed for monitoring the advance of the census operation with respect to the coverage achieved by:

(a) The sweep method and visualizing the geo-referenced eCensus units;

(b) The routes method.

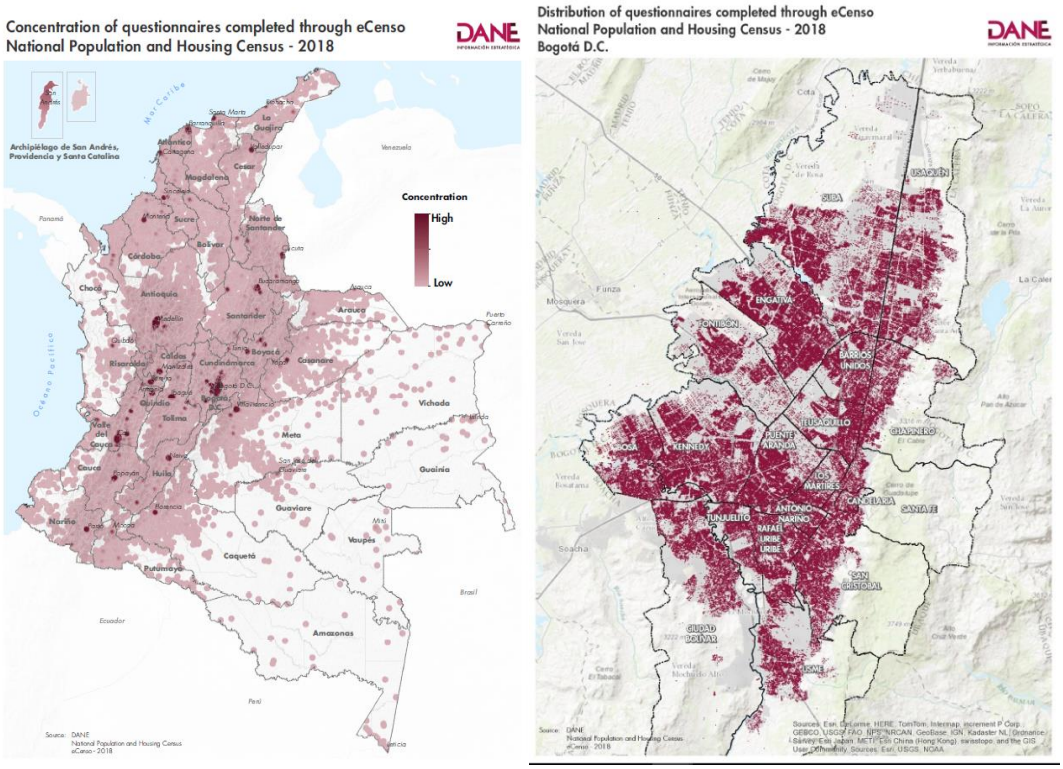
IV. Results

27. The conformation of the address register database allowed obtaining a consolidated amount of more than 15 million georeferenced addresses, with their corresponding DANE code. This input and the use of georeferencing techniques with Geographic Information Systems and assisted techniques were the basis for the georeferencing the 97% of units registered in the eCensus.

28. Figure III below shows the results of the georeferencing with respect to the eCensus; records were obtained from all the municipalities in the country, and the areas with the highest concentration coincide with the areas with the highest population density (left Figure) and on the right map is shows the distribution of registers in Bogotá. This information provided an approximation with respect to the geographic coverage of the eCensus, which allowed optimizing the development of the face-to-face phase.

Figure III

Map referring to the concentration of questionnaires completed through the electronic form of the 2018 National Population and Housing Census in Colombia and Bogotá respectively (Source: DANE)

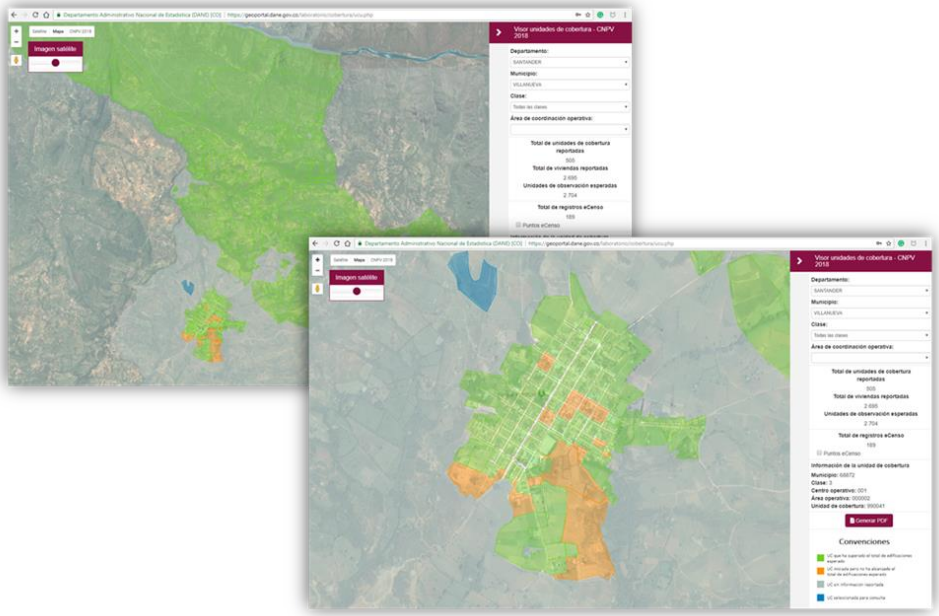


29. As a result of the development of applications for the CNPV, the following geoviewers were implemented to support the monitoring and control of the field operation.

(a) One geoviewer for tracking the progress with respect to the coverage achieved by sweep method and visualizing the geo-referenced eCensus units. With this tool, the follow-up and control of the field operation are achieved by means of monitoring the geographical coverage (progress in the visit to each coverage unit)[8] and the thematic coverage (progress in the identification of dwelling units or buildings). Figure IV shows an image of this geoviewer; the coverage units where the number of households surveyed is less than the number of dwellings estimated according to the framework can be observed in the areas in orange. This information allows identifying areas for control and verification of the operation.

[8] The coverage unit corresponds to the minimum operational geographical unit that corresponds to census block for the urban area and rural lot for the rural area.

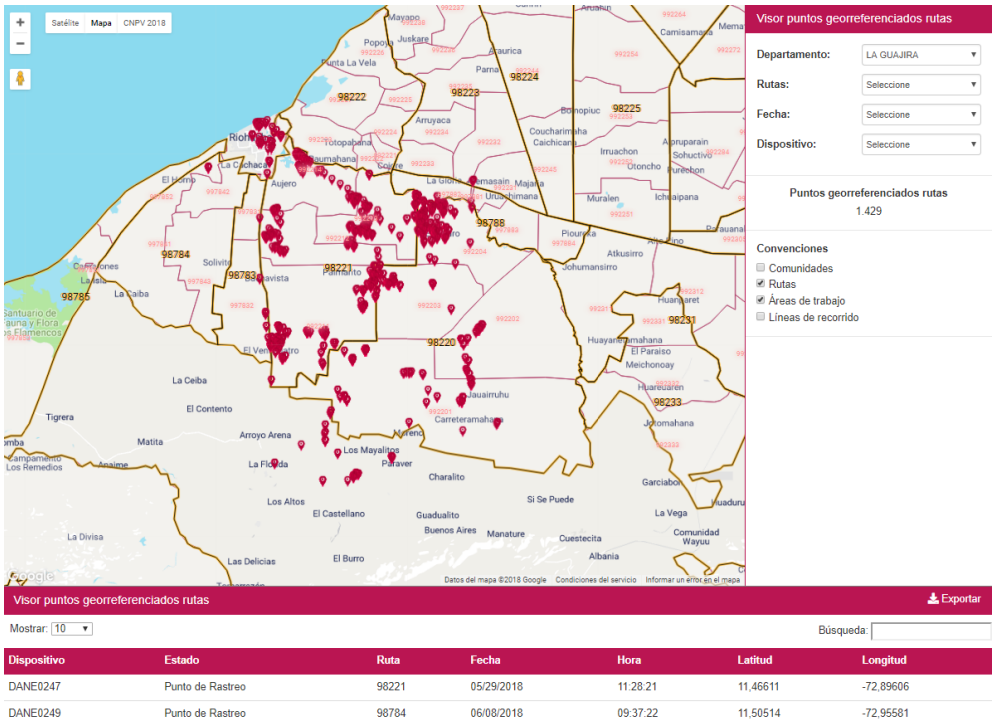
Figure IV
Geoviewer for monitoring the advance with respect to the coverage achieved by the sweep method and visualizing the geo-referenced eCensus units. (Source: DANE)



(b) One geoviewer for tracking the advance with respect to the coverage achieved by the routes method. With this tool, the location sent by the satellite tracking equipment[9] can be visualized and filters can be made by message type, date sent and the code of the tracking device (see Figure V).

[9] The satellite tracking system makes it possible for the interviewers to communicate with DANE by means of messages that are relayed via satellite while they are in remote areas where there is no mobile telephony service coverage.

Figure V
Geoviewer for monitoring the advance with respect to the coverage achieved by the routes method (Source DANE)



V. Conclusions

30. From the integration of administrative records and the National Geostatistical Framework, it was possible to consolidate a cadastral database of addresses that includes the geographical coordinates and their respective DANE code, and that through stored procedures and similarity algorithms, allowed the automatic georeferencing of more than one million households registered through the eCensus.

31. A determining factor in the results obtained with the automatic geocoding by means of the address cadastre was to control the capture of the location data of the electronic questionnaire in a coded and standardized manner, to use the coding of the Political Administrative Division – DIVIPOLA and to include additional variables for the identification of more detailed geographic levels of aggregation such as the village, commune, district, communes, townships and communities.

32. The satellite tracking devices – GPS – have allowed the CNPV operational team to be located in the most isolated areas of the country at all times, specifically where there are limitations on the coverage of mobile phone operators and that make it difficult to monitor as well as the timely making of decisions pertaining to the advance of the field operation.

33. Monitoring geoviewers are effective tools to support the field operation of the Population and Housing Census, whose main input is the georeferencing of the census units by means of various geospatial technologies, allowing the continuous verification of the census coverage and making informed decisions in the right time.

Bibliography

GoogleCode. (2009). Google Maps JavaScript API V2 Services.

UN-GGIM. (2014). United Nations initiative on Global Geospatial Information Management

Recuperado el 26 de 03 de 2018, de http://ggim.un.org/knowledgebase/KnowledgebaseCategory22.aspx