Statistical Journal of the IAOS 36 (2020) 83–88
DOI 10.3233/SJI-190527
IOS Press



# Use of new technologies for the 2020 population and housing census round

Edgar Vielma Orozco[a,*], Sabino Navarro Campos[b] and Isaac Salcedo Campos[c]
[a]*General of Sociodemographic Statistics, National Institute of Statistics and Geography, Mexico*
[b]*Development of Informatics Systems, National Institute of Statistics and Geography, Mexico*
[c]*Field Operations, National Institute of Statistics and Geography, Mexico*

**Abstract.** For the first time in Mexico, the 2020 Population and Housing Census data collection will be carried out through a CAPI (Computer-Assisted Personal Interviewing) scheme as the main enumeration method, but it will also include the CATI (Computer-Assisted Telephone Interviewing) and the CAWI (Computer-Assisted Web Interviewing) methods. These innovations, given the census structure size and the rapid changes in technologies, are a significant challenge for INEGI. Progress in census planning and field tests results will be presented, including the main challenges to be faced, the innovations considered for their implementation, as well as the successful experiences on the use of technologies for geo-referencing the information both in the data collection stage and for results dissemination.

Keywords: Population, census, innovations, technologies

## 1. Introduction

The purpose of this paper is to present the main characteristics of the Population and Housing Census 2020 (Census 2020); the IT strategy to be implemented, emphasizing innovations, in order to increase efficiency and control in each of its stages; as well as the main characteristics and the results obtained during the field tests carried out to determine if the defined procedures are sufficiently solid to carry out this census project.

One of the most important technological innovations of this project is the use of mobile computing devices (MD) for data collection, an area in which INEGI already has the experience of previous statistical operations (surveys, economic census). In addition, the implementation of operating procedures supported by a set of computer tools at the service of the staff, achieves the automation of several processes.

*Corresponding author: Edgar Vielma Orozco, General of Sociodemographic Statistics, National Institute of Statistics and Geography, Mexico. Tel.: +52 5552781000, Ext: 1997; E-mail: edgar.vielma@inegi.org.mx.

## 2. Main characteristics of the Population and Housing Census 2020

The 2020 Census in Mexico, like its predecessors, will be a "*de jure*" census, which means that the population is enumerated at its place of usual residence. The units of observation of the census are the usual residents of the country, private dwellings and collective living quarters. The proxy respondent is the head of the dwelling or in his/her absence, a person of 18 or over who is a usual resident of the dwelling and who knows the information of its residents. The data collection is planned to take place in March 2020.

Regarding the data collection forms, two population and housing questionnaires are used: a short form, consisting of around 35 questions; and a long form, designed for a probabilistic sample of around 4 million inhabited private dwellings, with approximately 100 questions, which includes the entire short form.

Additionally, the 2020 Census will provide statistics on the characteristics of the blocks in localities of 5 thousand and more inhabitants of the country (Urban Surroundings Questionnaire), as well as the infrastructure and socioeconomic characteristics of the lo-

calities with less than 5 thousand inhabitants (Locality Questionnaire). The Social Assistance Housing Census (CAAS, for its acronym in Spanish) is also planned, which collects the characteristics of the resident population and other users, the people who work in those institutions, the buildings and the services they provide.

In order to obtain a higher response rate, a multimodal approach will be implemented, with the main method being the personal interview assisted by a mobile device. Self-enumeration via the Internet or the telephone-assisted interview will be available upon request to the enumerator or in case of non-response after several visits of the interviewer. The use of printed questionnaires is considered only for places where the mobile device cannot be used.

## 3. Use of computer tools in the different stages of the Census 2020

In order to support the conceptual design tasks, two web applications were developed, the first one called "Public Consultation to Users of the 2020 Census", which aimed to facilitate the task of identifying users information needs, allowing them to expose systemically and simply the objectives and importance of their requirement. At the managerial level, it allowed to follow up on each of the requests. The second application is the "Conceptual Infrastructure of Census and Counts", which objective is to be a repository of the conceptual framework for population and housing censuses, allowing the recording and updating of conceptual and operational information on the variables that have been collected in the census questionnaires, as well as reviewing changes over time.

During the operational design, the procedures and technical schemes for the data collection activities are established, as well as the administrative, organizational, control and monitoring aspects. Since the 1990 Census, INEGI has been acquiring experience in the development of systems for these tasks and for the year 2020, it is proposed to use three of them:

1. The Delineation of Areas of Responsibility System, which makes it possible to distribute equal workloads to the field staff, automatically generating groups or segments of urban blocks or rural localities grouped by their number of dwellings and communication routes, among other factors, managing to generate areas of operational responsibility with an optimal collection order. Also, it allows staff to make manual adjustments to the delineation, applying their experience and knowledge in the field.
2. The Training System has the purpose of providing the operational personnel with the capacities and tools necessary for the execution of their functions, through automated learning modules that contain the teaching materials. This system allows giving remote advice, sending the learning activities and monitor progress in training.
3. The OPERA System, used in several stages of the information generation process, allows for timely follow-up of the progress in shaping the operational structure, systematizing the recruitment and selection of personnel, from the registration of applicants via the Internet until their hiring, and follow up the logistics of the operation by controlling the materials and resources necessary for field activities and management (vehicles, computers, questionnaires, brochures, among others).

The collection of information covers the set of activities to obtain the data from each unit of observation, based on the program and established work procedures, with an operational structure and controls that promote effectiveness of each of the actions. This implies the implementation of pre-designed schemes for data collection, such as the preparation and distribution of support materials (cartography, manuals, instructions and catalogs), the integration of human resources, communication and consultation, in the framework of a detailed program of activities, the setting of an organizational structure and controls that should be continuously monitored from the planning phase until the closure of the census.

The *Census Administrator* is the main application used on MD for data collection, in which the workload of each operative staff is visualized and managed, allowing reassigning them if necessary. It includes a Capture Module for the timely registration of each building (Buildings List), Urban Surroundings Questionnaire, Locality Questionnaire, short and long questionnaires for the inhabited private dwellings. In this application, local reports of progress and coverage can be obtained; it includes a supervision module in which a sample of areas for the revision of the classification of the dwellings and the inhabited condition is implemented. Through verification mechanisms, it is possible to validate the quality of the information recorded by the interviewer. It also has a Cartographic Module, which is used to record the cartographic updates at the block, town and road levels, allowing to digitize the polygons of new areas, cancel those that are no longer in the field and graphically represent changes

such as fusions and divisions of areas, allowing an immediate update of the cartography. The user authentication process is achieved with a unique user and password assigned to each member of the field personnel. The collected information is sent to the central servers for timely availability of census information. Data is also stored encrypted on MD to ensure confidentiality. Questionnaires modules include basic validations and alerts in case of inconsistencies, permitting in-site rectification of data and therefore improving the quality of information. The integration of cartography module and questionnaires modules in the same application allows the direct linkage of the census data to the geographical area of collection, and assigns a unique code to each dwelling that will be identified by an adhesive label that includes a unique QR (Quick response) code placed by the enumerator on a visible place on the door or facade. Accordingly, the data collection with MD seeks to improve the quality and timeliness of information, but will also be cost-effective by drastically reducing paper questionnaires and other printed forms for operative control (cost of printing, computers, salaries of data entry clerks, data capture offices).

In addition, the Coverage Verification Module will be implemented for the 2020 Census, with the objective of facilitating the activities of the staff and, if necessary, collecting the information of dwellings that could not be enumerated during the census. This module contains the main functionalities of *Census Administrator* but will be pre-loaded with the results of enumeration. One of the main advantages of the use of computer applications for the verification stage is the correct location of the dwellings, which will be achieved by the location of the geographical area, the address but also by reading with the MD the unique QR code on the dwelling's label. This functionality is also designed to support the Post-Enumeration Survey, which unlike the verification module will not be pre-loaded with information, but will conduct again the listing of dwellings and persons. Thus, the *Census Administrator* will incorporate the computer tools required to perform the Post-Enumeration Survey and a module for the comparison between the data obtained by the enumeration and the post-enumeration, for possible in-field rectifications after the data collection of the survey.

This census also aims to promote self-enumeration via the internet or by telephone. Accordingly, a web application to fill out the short form will be available to the population. The data collection by internet or telephone is based on the provision of unique identification codes to the informants though invitations letters, which are previously associated with the dwelling by the enumerator. It should be mentioned that this system will be adapted to collect the information on collective living quarters and on the staff of the Mexican Foreign Service.

The OPERA system, on the other hand, includes the module to monitor the integration of data collected in the field, allowing the generation, analysis and monitoring of indicators of coverage, speed and productivity, important information for management decision making during the census operation.

The data processing strategies include the corresponding design of the coding, validation and results generation systems with their corresponding quality controls. In order to carry out data processing, the Processing Monitoring and Control System will be used to monitor each of the stages. The automatic coding, by means of algorithms that recognize the textual information of the questionnaires, assigns a code of the predefined catalogs and later assigns the workloads for the computer-assisted coding. Once the coding has been completed, the automatic validation of the records is carried out. This system also includes testers of the validation criteria in order to ensure the correct application of the pre-defined algorithms. During each stage, the application provides progress reports in real time. Additionally, reports on the quality of the information are generated, and a log of the changes made to the data is created. In parallel, the assignment of the geographical codes of the National Geostatistical Framework is followed up. Finally, this stage makes available the database for the activity of figures release, which consists of a statistical review of the historical changes of sociodemographic indicators at the state, municipal and by locality size. For this activity, the Libera System will be used, which allows the control of the revision of each indicator but also serves as a repository for the documentation that supports changes found in some statistics.

Data processing for the publication of results is mainly done using predefined algorithms programmed ad-hoc in statistical analysis software, and in the case of products that require editing for presentation, automated processes are generated to facilitate the task, considering the editorial provisions of INEGI. Three web systems allow more experienced users to access information in greater detail: Consultation of Census Information System, which facilitate the interpretation of sociodemographic phenomena through the generation of thematic maps; the Territorial Integra-

tion and Locality Consultation System, which facilitates the query of historical information of each inhabited locality of the country; and the Sociodemographic Panorama of Mexico, which is a dynamic consultation system that integrates, as a synthesis, relevant data about the basic demographic, social and economic characteristics of the population and housing in Mexico. In addition, a series of microdata files of the census sample are available in different formats to facilitate their use.

## 4. Field tests

In the framework of the planning stage of the Census 2020, two field tests were carried out in order to determine, in the operational field, the optimal characteristics of the mobile devices, the required computer tools and to define the operative procedures according to this new paradigm.

In order to test the operation of mobile devices and computer tools in different climates and areas (urban/rural) the test of operational strategy and computer equipment was conducted. During six weeks, Enumeration operations were carried out, verification and self-enumeration, and the flow of information between each of these schemes were tested. For this test, a public bidding was carried out in which six devices of five different models and brands were purchased, in such a way that the computer tools could be tested on equipment with various technical characteristics, but all with Wi-Fi Technology, Micro USB, GPS, as well as voice and data services (SIM). Computer tools, for both Windows and Android, were developed in order to determine the most suitable operating system for its operation.

Concerning the systems for the preparation of operational planning, recruitment and selection of staff, training, logistical and operational monitoring, and integration of information to the central database, among others, the field test returned quite acceptable results, providing elements that allowed optimizing its performance. Regarding the capture of the GPS coordinates of the supervisor's route, a high precision was obtained, because its visual representation is similar to the shape of the covered areas, with a 3.8 meters average error. This method guarantees that the supervisor effectively made the route in the areas of his assignment; this data also allows determining the time spent and the direction of the path.

Among the main areas of opportunity, in computer science, the need was determined to implement laboratories to perform code testing and the use of best practices for the development of applications for MD with Android operating systems; the need for a tool for remote erase or reset of the MD was also detected for the cases of robbery or loss, as well as the remote updating of the applications; in addition to the implementation of a more robust security scheme, mainly in mobile devices and in the transfer of data to the central repository.

Later, with the purpose of emulating in all the senses the activities of the Census, the Pilot Test of the Population and Housing Census 2020 was carried out, with a sample of more than 44 thousand dwellings and a duration of 20 working days. The offices installed corresponded to a municipal coordination, that is to say, one of the around 1,600 sections in which the national territory is divided for data collection. All the planned operations were carried out, including the post-enumeration and the special operation for the enumeration of collective living quarters, as well as the operative verification of the inhabited condition of the buildings, which were not verified in the 2017 test.

For this operation, with the intention of testing the public bidding process to acquire the devices that will be used in the 2020 Census, a similar process was carried out to purchase the required mobile devices. Unlike the 2017 Test, only one brand and model that complies with the minimum characteristics and technical specifications required and defined from the results of 2017 was purchased. One of the bases of this tender was a test of the devices, in which the information transfer was proved through the use of a USB OTG memory, as well as the duration of at least eight hours of the battery, for which ad-hoc tools were used and processes were implemented to guarantee the best device selection. The result was the purchase of 243 devices of the best offer.

Delphi Rio RAD Studio was used for the development of computer tools. It provides a technological platform more oriented to mobile applications, particularly for devices with Android operating system, in such a way that systems were optimized for the training of the personnel, the processes of information exchange between the different operatives, the collection module of the Census Administrator, the synchronization mechanisms for data transfer, the security schemes from the access to the computer tools, the information encryption in the MD and during its transfer, until its integration in the centralized database. The modules for the generation of progress and control reports were also improved, as well as the problems related to

the use of the MD's digital camera for reading the QR codes and the use of the GPS for the registration of the GPS coordinates.

## 5. Technological architecture and information security

In the Census 2020 tests, a strategy for the transfer of information was implemented, in which tools are used for the generation of packed files integrating security passwords, and databases encrypted with the AES256 method algorithm (Advanced Encryption Standard with 256-bit keys). These encrypted data packages are sent through the Information Integration Module of the OPERA System, hosted on a secure institutional website, subsequently loaded in mass storage servers and, finally, integrated to the central database server.

The services of telecommunications, institutional network, databases, and mass storage servers are provided by the computer infrastructure area, who is responsible for keeping these services available. The service availability scheme is configured through an Application Load Balancer, which equally distributes each processing request between the application servers and available databases. In addition, a series of security standards and service availability tests allow planned stability throughout the whole process. Visual Studio NET is used for the development of Web systems, which is a complete set of tools for the generation of this type of applications; the AngularJS development framework is also used, since it provides techniques for the creation of SPA applications (Single Page Application), which benefits the performance by making lighter requests to the application server, by only requiring specific parts of the page. On the other hand, the use of JWT (JSON Web Token) implements a more secure communication and facilitates the administration of the users, assigning roles and permissions with ease. Oracle 12c is used to manage the database, providing greater security and availability of information.

## 6. Main technological innovations

The innovations for the realization of the 2020 Census include the use of mobile devices in the data collection; the possibility for the population to self-enumerate via the Internet or through a telephone-assisted interview; the use of QR codes; as well as the capture of the GPS coordinates during the routes of the work areas by the personnel.

The data collection of the 2020 Census, as already mentioned, will implement a multimodal approach, which contemplates the Personal Computer-Assisted Interview as the main procedure and in which phablet-type mobile devices will be used. This contributes directly to the congruence of data and affects the coverage of the areas, since it makes possible to integrate a primary data-validation during the interview, thanks to its automation based on the answers of the informant.

To carry out a correct identification of buildings and georeferencing of the information, the interviewer will list the dwellings and affix a label on them, the design of which includes an individual unique QR code. The QR code will also be used in the self-enumeration mode, since the interviewer will leave invitation letters in the dwellings with no response or upon request of the informant. As mentioned, these invitations contain a QR code that includes a unique username and password for each dwelling, in addition to the necessary instructions for the registration of information via the Internet and an assistance telephone number. This unique identification of buildings and dwellings, associated with geographical areas and identified by unique codes, will permit integrate the information from enumeration, verification and self-enumeration, allowing implementing controls for coverage and eventual duplication.

The use of GPS serves several purposes, among which can be listed: to indicate the location of the staff in the field, not only related to the work areas but also to the northwest corner where the reconnaissance route and/or the data collection should start; to record the GPS coordinates of the personnel paths in the work areas, mainly those of the supervisor and the interviewer, to facilitate the work of supervision, verification and post-enumeration, since these coordinates can be represented in the geographic module for access to the work areas of the staff responsible for these operations; and georeferencing each of the buildings and dwellings of the country. These GPS coordinates in conjunction with the QR code of the labels will allow updating the National Inventory of Dwellings and will be the start of the statistical registration of the real estate in the country.

## 7. Discussion and conclusion

The devices undoubtedly offer advantages over the use of paper. The adoption of new operating proce-

dures and information processing, which start even before the data collection itself with the digital cartography update, requires special attention and care.

As an institution, INEGI faces a reduction of time for the preparation of the different phases of the program. With respect to the IT applications, it is necessary to accelerate the life cycle of the development in order to minimize the maintenance stage during the operation, and to revise the flow of the information. This implies the realization of a greater number of tests, both in controlled conditions and in the field, which range from fieldwork to the production of information in order to reduce the time for the results dissemination.

Additionally, the country's security conditions require preparing plans for the mitigation of risks in the field, such as paper-assisted interviews, so the traditional processing methods must be available for these cases, even on a much smaller scale. Therefore, in the planning process, it is necessary to consider multimodal schemes for data collection and for the integration of the databases coming from the devices, paper questionnaires and self-enumeration by Internet. All these efforts will allow improving the consistency of the information, to publish results in a timely manner and generate savings, mainly by eliminating the data capture phase.

## References

[1] National Institute of Statistics and Geography (INEGI). Síntesis metodológica y conceptual del Censo de Población y Vivienda 2010. Aguascalientes, México: INEGI; 2011.

[2] National Institute of Statistics and Geography (INEGI). Encuesta Intercensal 2015: síntesis metodológica y conceptual. Aguascalientes, México: INEGI; 2015.

[3] National Institute of Statistics and Geography (INEGI). Aspectos generales del Censo de Población y Vivienda 2020. *Consulta pública 2017.* Aguascalientes, México: INEGI; 2017. (Internal document).