UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Automatic Data Editing and Imputation. Experience in the 2020 Mexican Census

Edgar Vielma Orozco

General Director of Sociodemographic Statistics

Mexico´s Statistics and Geography National Institute (INEGI, by its acronym in Spanish)

EDGAR.VIELMA@inegi.org.mx

Abstract

The Population and Housing Censuses provide relevant information for policymaking, research, and planning, so its results should be disseminated with the greatest opportunity. This paper describes the automatic data editing and imputation experience in the Mexican Census in the context of the Covid-19 pandemic. The methodology is described with particular emphasis on the review editing results strategy, carried out at the local offices of the Institute, whose maximum level of disaggregation was the municipality, and the imputation techniques used for the first time in the Census.

2020 Census automatic editing process

General characteristics

The 2020 Population and Housing Census (2020 Census) was a de jure census. The population was counted according to their regular or legal residence. The units of observation were:

- Usual Resident
- Private dwellings
- International migrant

Two types of questionnaires were defined, the Basic Questionnaire, with 38 questions, and the Expanded Questionnaire, with 103 questions, including those of the Basic Questionnaire. The Expanded Questionnaire was applied in a little more than four million households selected using a probabilistic sampling. The greater number of questions in the Expanded Questionnaire made it possible to generate more detailed information on some sociodemographic phenomena so that users had a unique source due to its level of disaggregation.

The information was collected through direct interviews with the support of computer applications on Mobile Computing Devices (MCD) and, when required for operational reasons, through printed questionnaires. As complementary methods, self-enumeration via the Internet and telephone-assisted interviewing were implemented.

It is essential to highlight that the use of MCDs for data collection was the main innovation of the 2020 Census, having as an advantage the access to data with greater timeliness, the implementation of basic validations that contributed to the unity of the information coming from the field, and the use of tools such as the Global Positioning System (GPS), which helped the staff in their correct location in the field.

Content problems and errors

The information collected is always subject to errors of different types, such as omissions and inconsistencies derived from recording errors by the interviewer or inconsistent responses provided by the respondents. To solve this type of problem, conceptually related variables were reviewed during the editing process, seeking to assign the most logical value for the behavior found or, if not, to set the code to "Not specified".

In the Mobile Computing Device for data collection, basic validations were included for recording responses. Still, it was necessary to find a balance between validating everything, which would considerably increase the interview time and would have required equipment with greater processing capacity, or not validating anything, which would mean that the data collection advantages in MCD would not be adequately considered. Thus, it was decided to validate only some fundamental responses and to complement this by sending alert messages to the interviewer so that he/she could edit the record and leave the response blank if this was impossible.

The validations implemented in the MCDs included the mandatory response in the variables of *Sex*, *Age*, and *Relationship*; it also ensured the flow of the questionnaire respecting the question passes and age cut-offs, in addition to including congruence validations, such as providing that the bedrooms were less than or equal to the number of rooms, identifying directly from the list of persons the father, mother or spouse, as appropriate, and adding them to the list of persons in case they had been omitted. It also validated that the dates included in the questionnaire were not after the interview.

As for the information captured in printed questionnaires, it was caught in image format, i.e., no editing was performed during this process. Verifying all the questionnaire packages was done to minimize capture errors. For this purpose, a system was designed that, using an acceptance sampling, selected the sample of questionnaires from each captured package to proceed to their recapture; the system alerted in case of detecting differences between what was caught for the first time and the verification, which made it possible to identify which was the correct data and if there was an error in the original capture.

If the number of capture errors in the sample exceeded the established tolerance (around 0.3% error), the package was rejected, which meant that it had to be captured a second time.

Given the diversity of capture media used, it was necessary to determine the priority of information integration when data from the same observation unit was available in different media to avoid duplicate records. Priority was assigned to face-to-face interviews and then to the other media. Thus, the order was as follows: 1) interview captured with MCD, 2) printed questionnaire, 3) telephone interview and self-enumeration by the Internet.

Strategy for editing Census 2020 data

Methodology for automatic editing

The theoretical vector methodology has been used at the Institute since 1990 to define the criteria and processing of editing census data. This methodology is exhaustive and resolves in a general way the editing of the data by type of error to be corrected. It is used to rescue some elements that do not respond to the variable in question but do respond to other variables (auxiliary variables). The assignment of the values in the variables is deterministic.

The theoretical vector methodology facilitates the detection of inconsistencies since, with it, one has control of the values that a set of operationally or conceptually related variables can take through the generation of all possible response combinations between the variables.

To control the combinations that are generated, what is known as an *addressing function* is built, which allows assigning a single value greater than zero, also known as an image, to each combination that is generated.

In the design of the general editing process, each validation criterion offered a solution derived from the logic of the questions and the flow of the questionnaire, respecting the informants' responses as much as possible. In cases where there was not enough data to give consistency to the information, according to the predefined valid response codes, special codes were assigned from the "Not specified" category. Three hundred eleven criteria were designed, which meant a little more than a thousand treatments on the population's characteristics

and the dwellings for the 2020 Census. The design also contemplated the definition of tools that would allow following up and monitoring the application of these criteria to the information.

The execution of the process of automatic editing and generation of the reports used to review the consistency of the data after the execution was carried out in the central offices of the Institute to have control over the adjustments to the criteria and that these be applied homogeneously to the information of all entities.

Control, monitor, and review edited data

For the tracking of the edited data, different tools were designed and programmed to review the reports of the results of the edition at the different levels of geographical disaggregation contemplated in the process: national, state, and municipal.

The reports for the review made it possible to verify that the editing process worked correctly and that, at the end of it, the information was free of inconsistencies. The reports provided an overview of the information before and after applying the scheduled treatments to assess their impact on the census information.

The analysis reports defined for the review of the editing process were as follows.

- *Control of figures: frequency of data and ranges, initial treatments, and control figures of universes*

The *data frequency and range report* presented the count or frequency, and the percentage of each variable's response options, considering both the Basic and Extended Questionnaires. The count was made for the information: a) that was received from the previous coding process, b) from the database to which the initial treatments were applied and that was used to apply the editing process, and c) the database resulting from the process.

The initial treatments were applied in a process before the edition, and the database was prepared, recovering, as far as possible, the values for the pivot variables of the edition (*Sex*, *Age*, and *Housing class*) and the records without information were removed. The purpose of this analysis report was to keep track of the records to which the initial treatments are applied.

The report of *universe control figures* made it possible to verify that no information was lost during the execution of the edition.

- *Changes by variable and register*

During the execution of the process, this log or trace was generated that accounts for the changes of each variable for each database record and the editing criteria that made the change. It allowed to focus on the changes and identify the adjustments to the criteria during the editing process.

- *Crossing of variables*

To review the consistency and coherence of the information obtained with the editing process, tables with crossings of two or more variables were generated to identify invalid behaviors that contradicted the logical relationship of the variables involved.

- *Omission and unspecified reports*

Two analysis reports were used to analyze the level of assignment of values in the variables: *omission* and *not specified*. The first allowed control over input omissions to the editing process for each of the variables of the two capture instruments. Each variable had an associated tolerance for omissions, which was the reference parameter to assess whether the percentages of this type of response were within limits allowed. The second permitted control of the "Not specified" codes assigned in the process for each variable. In this case, each variable had a tolerance associated with it to ensure that the percentages of values with the "Not Specified" code were within the permitted limits.

- *Input-output matrices and percent changes*

There were two types of information analysis reports to analyze the changes in value: input-output matrices and percentage changes. Both present the information in absolute numbers and percentages.

The input-output matrices punctually showed changes or movements in the distribution of possible responses in a variable before and after the editing process. Matrices were generated for all variables of the survey instruments.

The percentage change report allowed us to evaluate the magnitude of the changes made by the editing process for each variable.

Control and monitoring structure:

While the design and execution of data editing were carried out at the central offices, the review of the analysis reports to monitor the quality of the information resulting from the process was carried out at the Institute's state coordinating offices for the state and municipal levels. Data at the national level were reviewed at Central Offices.

Impact of the SARS-CoV-2 Pandemic

Sanitary measures and change of strategy

The editing of the 2020 Census data was conducted entirely during the time of the SARS-COV-2 pandemic. Training of state officials who would review edited data began on March 23, when health authorities declared the start of the social distancing (with the suspension of non-essential activities) and concluded on April 27.

The training of analysts at the Institute's state offices began on March 30, one day before the health authorities declared the suspension of censuses and surveys and concluded on April 3. As a result, the staff of the permanent and operational structure carried out their activities at home, which made it necessary to modify the planned scheme for the operation and the follow-up. Due to its remote access, the computer system used for automatic editing had to be adapted to guarantee the confidentiality of the information. Once health conditions permitted, the activities were resumed in the state coordinating offices with the minimum essential personnel assistance, concluding the process in December 2020.

Home office and additional resources

To work from home, it was necessary to provide a VPN (Virtual Private Network) connection to access the INEGI network for all personnel involved in the process and ensure that they had communication and collaborative software to interact with the rest of their team. An e-mail account was also required for each analyst, a situation that had not been planned but was necessary to send them official communications and for the remote exchange of information between the analysts and their state validation manager.

Handling of information

To protect the confidentiality of the information handled outside the Institute's facilities, all personnel involved in census activities signed a confidentiality commitment, which established that workers could not transfer or generate copies of work files; if they did so, they could be sanctioned under the provisions of the legislation for public servants.

Among the different processing stages, the one that could imply the most significant risk for the premature disclosure of information until that confidential moment was the analysis of the results of the automatic edition since those in charge of the analysis were working at their homes, where any person outside INEGI could eventually have access to information such as the total population in each of the municipalities of their federal entity, the aggregates of the characteristics of the dwellings and the population with different levels of geographic disaggregation.

To avoid the disclosure of relevant information from the 2020 Census, it was decided to anonymize the national information as follows:

1. From the total number of municipalities in a state, one municipality was extracted, preventing the operative personnel from having complete information on the state.
2. The 32 municipalities extracted from the 32 states formed a new grouping reviewed at the central offices.
3. The remaining municipalities in each state were assigned a new fictitious geographic code so that it would not be possible to know to which municipality the information corresponded.
4. This regrouping made it possible to maintain the confidentiality of the 2020 Census results. Likewise, a table of equivalences was generated before the generation of reports not to delay the process and not affect the programming of the criteria and their analysis tools (which included the geographic keys of municipalities).

Execution of the process

The impact of health measures in the last days of the 2020 Census survey and the suspension of censuses and surveys by the health authorities delayed the planned flow of data.

The editing process was run nine times, with each run including the new information that became available. In the first five runs, revisions were made at the national and state levels, while revisions at the municipal level began with the sixth run. After the ninth run, the Basic Questionnaire information was free of inconsistencies. The database was released in December 2020. One more run allowed validation of the Expanded Questionnaire information, which was released in February 2021.

Imputation

It is universally accepted that, in the execution of all population and housing censuses and counts worldwide, there are missing observation units due to the impossibility of contacting informants or because they refuse to provide their information.[1] In this regard, in the document *Principles and Recommendations for Population and Housing Censuses, the United Nations. Revision 3* recognizes the need to make imputations.[2] In the case of Mexico, although strategies have been designed to minimize these types of situations, they continue to occur, even more so in the context of the pandemic present during the development of the 2020 Population and Housing Census.

According to the United Nations, imputations can be performed using any of the following techniques:[3]

- Deterministic: based on existing information from the observation unit.
- Model-based: the information is imputed due to the application of statistical models (regressions, averages, or other statistics).
- Using information from donor observation units: missing information is assigned from the data of a responding observation unit. This can be done either for specific variables or all missing information.
- Mixed Models: considers applying more than one of the above techniques.

To account for the total population of the country in those dwellings where it was not possible to interview due to refusal of the population to the interviewer or lack of response to the self-enumeration, and for which it was identified that there are residents, the imputation of population and housing information was performed by assigning all data from a dwelling with information, randomly selected, within the same Basic Geostatistical Area (BGEA). That is, all the information from a dwelling for which no information was obtained was taken from another nearby dwelling, the one most likely to have similar characteristics (nearest neighbor technique).[4]

---

[1] *Cfr.*, UN, 2017, *Principles and Recommendations for Population and Housing Censuses. Revision 3.*
[2] *Ibid.*
[3] *Vid.*, UN, 2021, *Handbook on Population and Housing Census Editing. Revision 2*, p. 108.
[4] *Ibid.*, p. 310.

This criterion was applied if there were enough dwellings with information so as not to generate biases in the population structures, socio-demographic characteristics, or characteristics of the dwellings. In the case of not having enough donor dwellings, in the dwellings without information, the total number of persons was assigned according to the municipal average, and in all the characteristics of both population and dwellings, the codes corresponding to "Not specified" were assigned.

The imputation of the information of the unanswered dwellings cannot be done through any deterministic method or model since there is no information on the dwelling, only the fact that it is inhabited. International recommendations do not suggest the imputation of each characteristic of the dwellings or the persons separately since this would generate multiple inconsistencies,[5] so for the imputation, we took the totality of characteristics of the donor dwelling and its occupants.

Likewise, in those areas where it was not possible to collect information due to lack of access, processes were carried out to determine the total number of dwellings that should be imputed in the blocks where it was not possible to carry out the survey, based on historical information and satellite images; in these cases, only the total number of people and their sex were imputed, without considering a nearby donor to avoid biases in the information.

Unlike previous events, the verification period was not carried out immediately after the enumeration due to the COVID-19 pandemic, so the population mobility between the conclusion of the enumeration and the beginning of the verification was more significant than expected in terms of their place of residence. Therefore, an analysis was carried out to maintain as pending dwellings those that were identified as inhabited during the enumeration, but for which information could not be obtained due to the absence of residents or refusals and that were subsequently classified as not inhabited to maintain the reference moment of the census information.

It is essential to point out that the population censuses and counts of several countries, including Mexico, traditionally present a phenomenon of omission in the declaration of minors. For its detection, this project analyzed the administrative records of vital statistics of the last decade (Vital Statistics on Registered Births based on birth certificates issued by the civil registries of each state, Vital Statistics on Registered Deaths based on death certificates issued by the Ministry of Health) and the information captured in the 2020 Census. It was determined that the imputation of persons under seven years of age would be performed.

This imputation was applied only in dwellings where a woman of reproductive age resided, who reported surviving children in the fertility section of the Census questionnaire, and who were not declared usual residents of the dwelling. For this criterion, only one child was imputed per dwelling, even if there was more than one woman with surviving children reported during the interview but not included as residents, or if more than one surviving child was registered and none was present in the dwelling.

This criterion was applied exclusively in municipalities in which, according to the analysis, were omitted children and in no case did the total number of children (reported and imputed) exceed the total number of persons between zero and six years of age according to vital statistics, analyzing each age individually.

Using imputations facilitates the handling of the information by the users, avoiding biases or erroneous interpretations. The total number of imputed cases is indicated in the predefined tabulations, and all imputed records are identified in the databases.

Conclusions

The SARS-Cov-2 pandemic impacted almost all stages of Mexico's Population and Housing Census 2020. The appearance of the first case just five days before the beginning of the enumeration led to an increase in the non-response rate from the third week of the enumeration because the population wanted to avoid the risk of being infected and causing the refusal to open the door to the census interviewer.

It was necessary to adapt, prolong or implement a second phase of the information processing activities, requiring the modification of systems and work schemes established initially to carry out the activities in the state offices for their execution in the modality of work at home.

As already mentioned, it was at this stage that the confidentiality of the results could have been put at risk, so the analysis of the results of the edition was the stage where the most outstanding care was taken to

[5] *Ibid.*, p. 439.

preserve confidentiality. The changes made to the systems and the performance of the verification stage from June to August 2020, as well as the performance of the imputation for the first-time using information from a donor, led to this stage being extended until December 2020.

Concerning the imputation of information in the private inhabited dwellings where it was not possible to capture data, the use of a donor for the first time within the same BGEA contributed to having a better estimate of the resident population in these dwellings since it maintains homogeneity for the total resident population in each of the dwellings.

Likewise, given that the imputed information is consistent with the characteristics of the dwellings in the area, it avoids having high percentages of data with "Not specified" codes. The distribution of the characteristics of the dwellings and persons is not impacted; that is, the proportionality of this information was maintained so that this practice will be replicated for future census events.

Undoubtedly, the health situation faced during the execution of the census activities will allow us to have elements to be prepared for future events, to maximize the use of information technology and process planning, considering the possibility of having to be adjusted in short periods, which will reduce the impact on the presentation of census results.