



UNECE Statistical Data Editing meeting

# **Automatic data editing and imputation. Experience in the 2020 Mexican Census**



**MEc. Edgar Vielma Orozco**  
General Director of Socio-demographic Statistics

---

October 2022

---



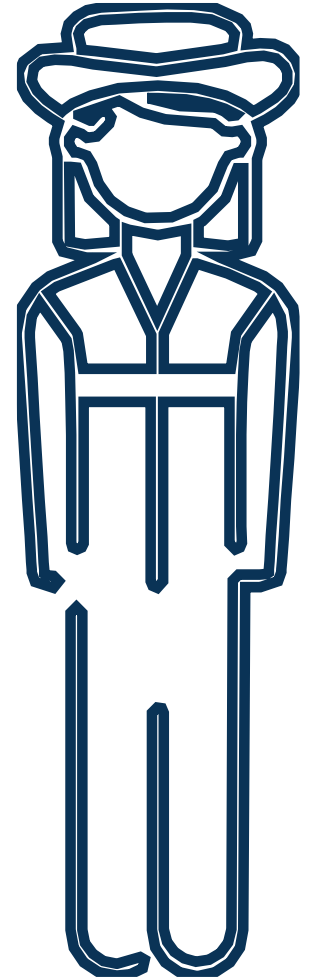
# 2020 Census automatic editing process

● ○ ●

# General characteristics



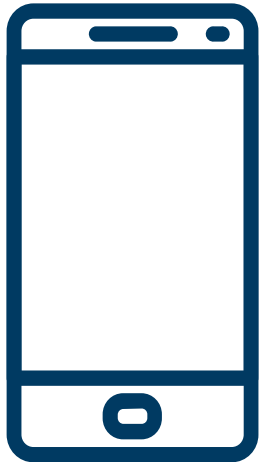
- **Basic Questionnaire**, with 38 questions. **Expanded Questionnaire**, with 103 questions, applied in probabilistic sampling (approx. four million households) to generate more detailed information.
- The information was mainly collected through direct interviews with **Mobile Computing Devices (MCD)** support and, when required for operational reasons, through printed questionnaires.
- Use of MCD 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 consistency of the information, and the use of tools such as the Global Positioning System (GPS).



# Content problems and errors (1/2)



- Information collected is always subject to **errors of different types: omissions and inconsistencies** by the interviewer, or inconsistent responses 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".

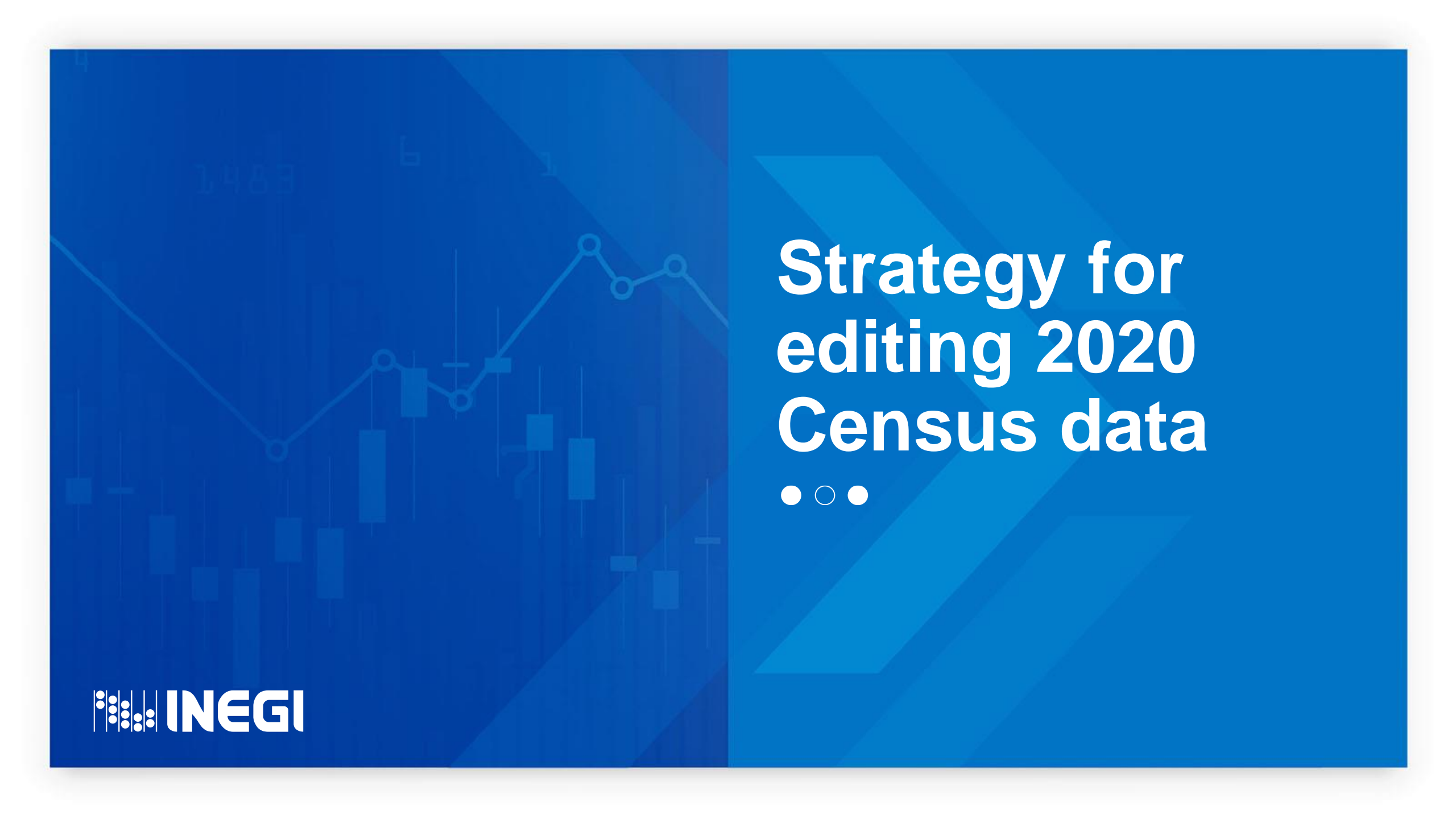


- In the **Mobile Computing Device for data collection, basic validations were included for recording some fundamental responses** and sending alert messages to the interviewer for record editing or leaving the response blank if this was impossible.
- The validations implemented in the MCDs included the mandatory response in the variables of **Sex, Age, and Relationship, assuring the flow of the questionnaire** respecting the question passes and age cutoffs. It also validated that the dates included in the questionnaire were not after the interview.

# Content problems and errors (2/2)



- During the capture of the paper questionnaires, no editing was performed. Verifying all the questionnaire packages was done to minimize capture errors.
- 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 the first time and the verification.
- If the number of capture errors in the sample exceeded the tolerance (around 0.3% error), **the package was rejected**, which meant that it had to be captured a second time.
- **Priority of information integration** when data from the same observation unit was available in different media to avoid duplicate records:
  - 1) interview captured with MCD,
  - 2) printed questionnaire,
  - 3) telephone interview and self-enumeration by the Internet.



# Strategy for editing 2020 Census data

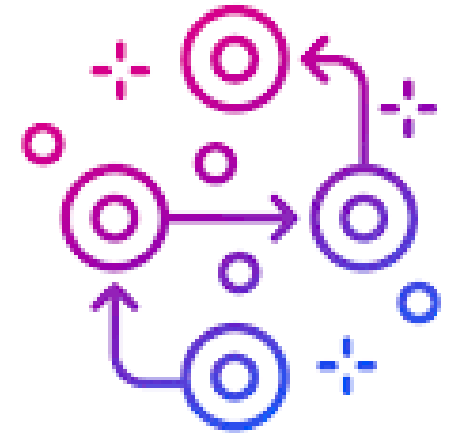
● ○ ●

# Methodology for automatic editing



## The theoretical vector methodology:

- **Facilitates the detection of inconsistencies** since 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.
- In cases where there was not enough data to give consistency to the information, the **"Not specified" code was assigned**.
- **Three hundred eleven criteria were designed**, which meant a little more than a thousand treatments for the characteristics of the population and dwellings for the 2020 Census.
- The execution of the process of automatic editing and generation of the reports **was carried out in the central offices** to have control over the adjustments to the criteria and that these be applied homogeneously.





# Control, monitor, and review edited data



For the **tracking of the edited data, different tools were designed to review the reports of the results** at the different levels of geographical disaggregation: national, state, and local.

Reports **provided an overview of the information before and after applying the automatic 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.
- Changes by variable and register.
- Omission and not specified reports.
- Input-output matrices and percent of changes.

**Control and monitoring structure:** The review of the analysis reports to monitor the quality of the information was carried out at the Institute's state coordinating offices for the state and local 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 reviewed edited data began on **March 23, when health authorities declared the** suspension of non-essential activities and concluded on April 27.
  - analysts started on **March 30, one day before the suspension of censuses and surveys**, and concluded on April 3.
- As a result, the **staff 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**, **all personnel** involved in census activities **signed a confidentiality commitment**.

The **riskiest stage** for **premature disclosure of information** was the **analysis of results** of the automatic edition since analysts were working at their homes, where someone outside INEGI could eventually access information.

To **avoid** the **disclosure** of relevant **information**, it **was** decided to **anonymize** the information as follows:

1. From the total number of municipalities in a state, **one municipality was extracted**, preventing someone 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 **results**.

# Execution of the process



- The **impact of health measures** in the last days of the 2020 Census survey (Expanded Questionnaire) **and the suspension of censuses and surveys** by the health authorities **delayed the planned flow** of data.
- The **editing process was run nine times, including the information available at the execution moment.** 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 (1/4)



- 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 (**AGEB**). 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).
- **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**, the **total number of persons was assigned according to the municipal average**, and the codes corresponding to "Not specified" were assigned to the characteristics of both population and dwellings.
- **The imputation of the information of the dwellings without response cannot be done through any deterministic method or model since there is no information on the dwelling, only the fact that it is inhabited.**



# Imputation (2/4)



- 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, **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.
- 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, those **dwellings** that, during the census, were **identified as inhabited**, but **the information could not be obtained**, and that during the **verification were classified as uninhabited**, to maintain the reference moment of the census, **remained as pending dwellings**.





# Imputation (3/4)



- To detect the **omission in the declaration of minors**, this project analyzed the vital statistics administrative records <sup>1</sup> of the last decade 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**, 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.

<sup>1</sup> 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.

# Imputation (4/4)



- **This criterion was applied exclusively in the municipalities in which, according to the analysis, children were omitted,** and in no case did the total number of children (declared and imputed) exceed the total number of people between zero and six years old, according to vital statistics.
- 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.

The background of the slide is a solid blue color. On the left side, there is a faint, light blue graphic of a financial chart, including a line graph with circular markers and a candlestick chart below it. On the right side, there are several overlapping, semi-transparent geometric shapes, primarily triangles and parallelograms, in various shades of blue, creating a modern, abstract design.

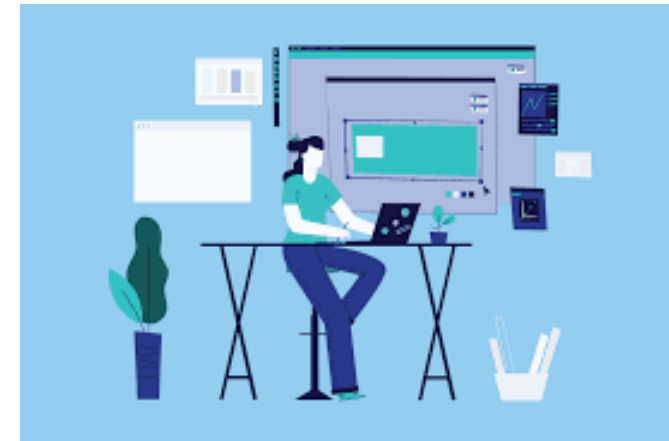
# Conclusions

● ○ ●

# Conclusions



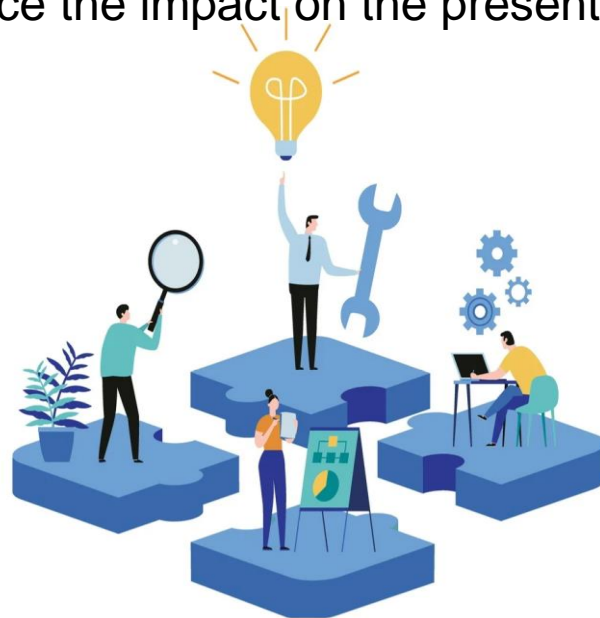
- The **SARS-Cov-2 pandemic** impacted almost all stages of the **2020 Census in México**. The appearance of the first case five days before the beginning of the enumeration led to **an increase in the non-response rate**.
- The data processing stage required more time because there were **modifications to systems and work schemes** (home office). The verification stage was postponed, for example.
- In the **private inhabited dwellings** where it was impossible to capture data, using a donor for the first time within the same AGEB contributed to having a **better estimation of the resident population** in these dwellings.



# Conclusions



- Since the **imputed information is consistent with the characteristics of the dwellings in the area, high percentages of data with "Not specified" codes are avoided. The distribution of the characteristics of dwellings and people was maintained**, so this practice will be replicated in future census events.
- The health situation faced during the execution of the census activities evidenced the need to **take advantage of the available technology and the importance of planning processes** with the ability to adjust to short times to reduce the impact on the presentation of census results.



## Conociendo México

800 111 46 34

[www.inegi.org.mx](http://www.inegi.org.mx)

[atencion.usuarios@inegi.org.mx](mailto:atencion.usuarios@inegi.org.mx)



**INEGI** Informa

# THANK YOU

