United Nations | Department of Economic and Social Affairs

2020 round of population and housing censuses: Census Editing and Validation

Meryem Demirci

UN Statistics Division

Content

- Overview of 2020 World Population and Housing Census Programme and Census Methodological Framework
- Introduction to the Handbook on Census Editing Revision 2
- Editing process stages
- Types of edit
- Editing during data collection
- Imputation and assessment of the quality of imputation

2020 World Population and Housing Census Programme

❑ The Economic and Social Council of the United Nations endorsed the draft Resolution on 2020 World Population and Housing Census Programme on 10 June 2015 ECOSOC

❖ Mandates of UNSD:

- prepare international statistical standards, methods and guidelines for the conduct of PHCs
- coordinate activities among stakeholders in assisting Member States to plan and carry out PHCs
- monitor and regularly report to the Statistical Commission on implementation of the Programme
- provide technical assistance to countries in conducting censuses

UNSD Methodological framework

Principles and Recommendations for Population and Registration, Revision 3 (2015) ENG and FR

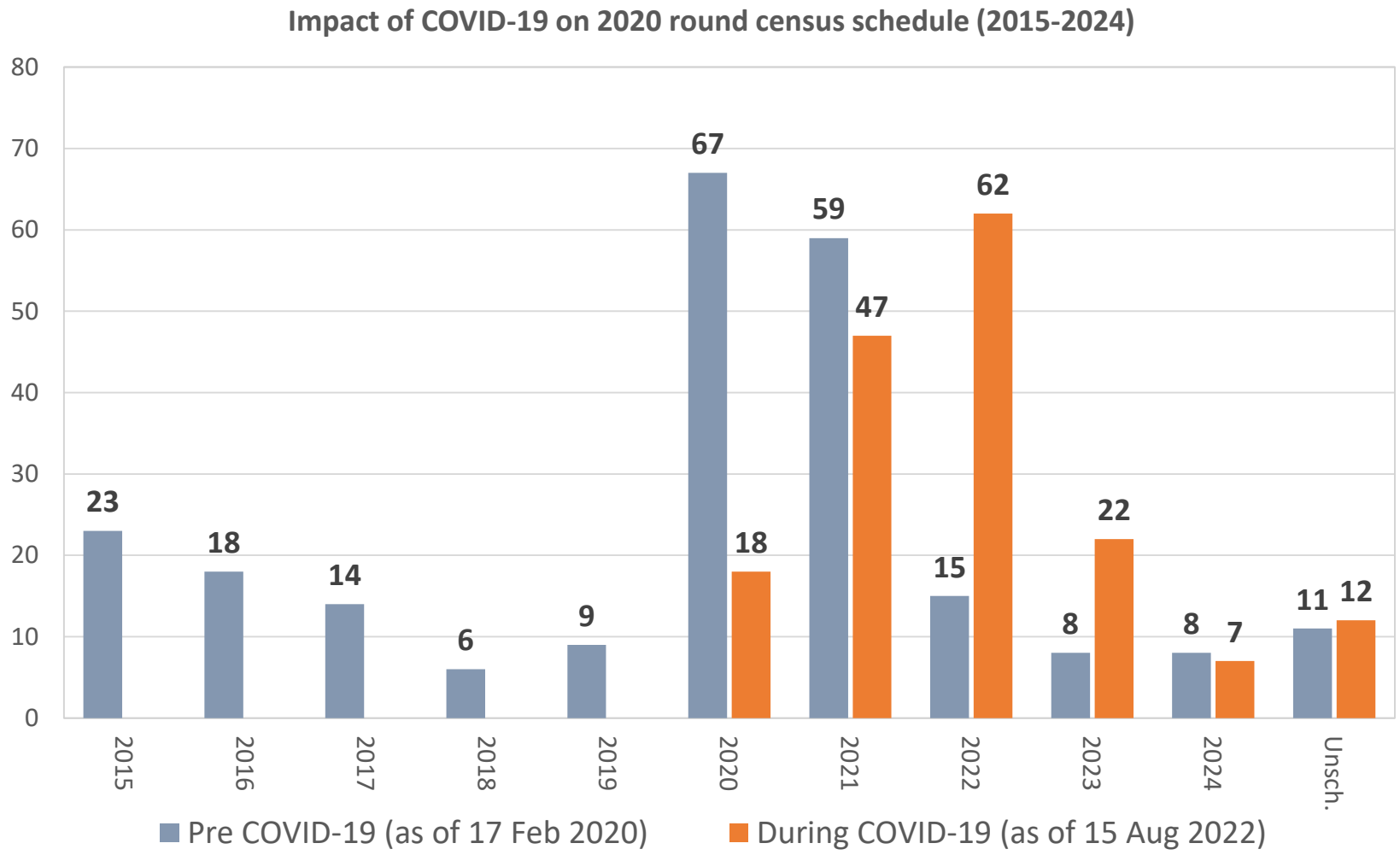
Handbook on Management of Population and Housing Censuses Revision 2 (2016) ENG and FR

- Handbook on Measuring International Migration through censuses (2017) New
- Guidelines on the Use of Electronic Data Collection Technologies in PHC (2019) New
- Handbook on Editing for Population and Housing Censuses Revision 2 (2019)
- Technical report on Measuring SDG Indicators through PHC and CRVS Data (2021) New
- Technical Report on Post Enumeration Surveys (2010)
- Handbook on Registers-Based Population and Housing Censuses (Upcoming) New

https://unstats.un.org/unsd/demographic-social/census/index.cshtml#methods

Impact of the Covid-19 pandemic on the implementation of 2020 round of censuses

- UNSD organized two Expert Group Meetings to document the impact of the pandemic on census-taking

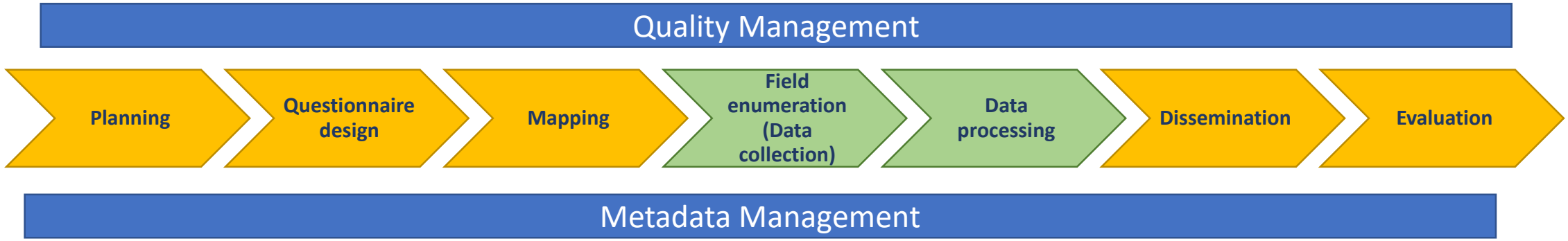


UN Handbook on Census Editing Revision 2

- The second revision of the handbook provides information on the editing process and the use of various approaches to census editing, Main topics covered:
  - Errors in the census processes
  - Editing process steps and types of edit
  - Real-time editing (editing during data collection)
  - Editing applications
  - Structure edits (geography edits, coverage checks, correspondence between housing and population records)
  - Edits for population topics
  - Edits for housing topics

https://unstats.un.org/unsd/publication/SeriesF/seriesf_82rev2e.pdf

Generic Census Processes



- The processes of quality management and metadata management are the overarching processes
- Quality management refers mainly to the quality assurance programme which is developed for monitoring and assessing the quality of operational activities and census data
- The over-arching process of metadata management refers to the creation, use and archiving of operational information and statistical metadata throughout the census process

Errors in the census processes

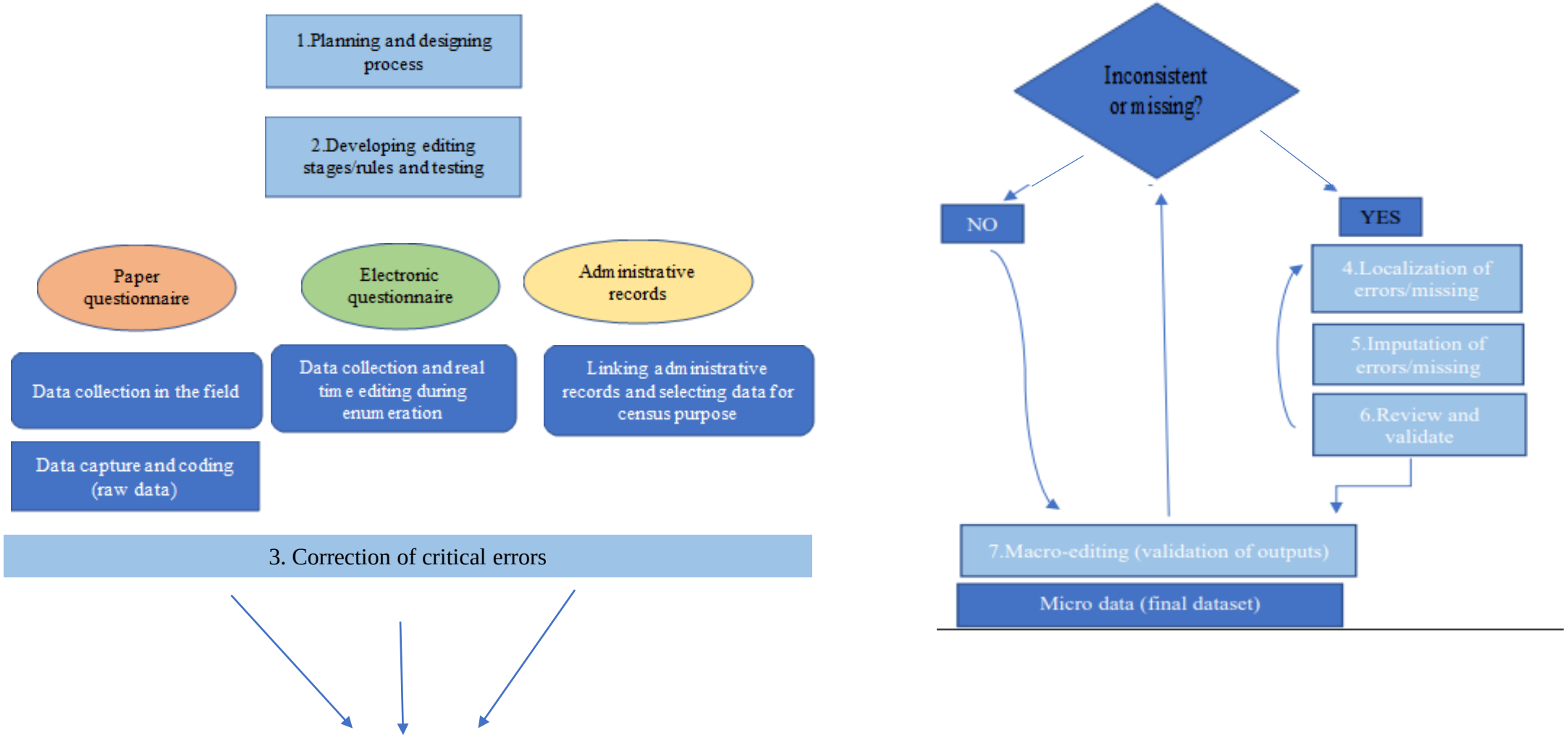
- Census data suffer from many sources of error that may be classified, generally, as coverage errors and content errors.
  - **Coverage errors**
    - arise from omissions or duplications of persons or housing units in the census enumeration
    - sources of coverage error:
      - incomplete or inaccurate maps or lists of enumeration areas or living quarters,
      - failure by enumerators to canvass all the units in their assignment areas, duplicate counting,
      - omission of persons who are not willing to be enumerated,
      - erroneous treatment of certain categories of persons

Errors in the census processes

- **Content errors**
  - Errors in questionnaire design
  - Enumerator errors
  - Respondent errors
  - Coding errors
  - Data entry errors
  - Errors in computer editing process
  - Errors in Master Files

United Nations | Department of Economic and Social Affairs

Editing Process Stages



Editing Process Stages

Step 1. Planning and designing:

- All data processing stages, including data capture (during the enumeration and/or data processing), editing, coding and imputation, should be planned in an integrated way within the overall census program

Step 2. Developing and testing:

- Depending on choices made, sufficient time must be allocated for the development or configuration of software applications, and then for the testing of procedures, workflows and the interactions between different applications or systems within the program

Step 3. Correction of critical errors:

- Certain critical errors which have the potential of blocking further processing should be detected at an early stage in data processing.

  For instance, if multiple data collection modes are used, converting the data in a common format and duplications should be dealt with in this step

Editing Process Stages

Step 4. Identifying inconsistent or missing values:

- Once the critical errors are defined and corrected, it is straightforward to check whether the values are inconsistent in the sense that some of the edits are violated. Basic principle is that the data in each record and between records should be made to satisfy all edits with minimum changes possible in items.

Step 5. Imputation:

- Imputation is the process of resolving problems concerning missing, invalid or inconsistent responses identified during editing. Imputation procedures are of crucial importance to make sure that census results are of high quality and internally consistent.

Step 6. Review and validate:

- This sub-process examines data to identify potential problems, errors and discrepancies in editing and imputation. It can be run iteratively for validating data using quality indicators.

Editing Process Stages

Step 7. Macro editing:

- After micro editing which corrects the data at the record level, macro-editing stage, which analyses aggregate data (totals), starts for the purpose of verifying whether results to be released seem reliable.
- Macro-editing can be accomplished by comparing aggregated data with the results of previous censuses and other relevant data sources such as household surveys and administrative registers.
- If methods used for validation of outputs give unusual or unexpected results, then it is possible to go back to the micro-editing procedures.

Types of micro editing

- **Validity edits** look at one question field or cell at a time and check invalid characters and values, such as;
  - essential fields have been completed (e.g., no field is left blank for key variables, such as sex, age for electronic questionnaires)
  - specified units of measure have been properly used, such as number of children
- **Range edits** are similar to validity edits in that they look at one field at a time. The purpose of this type of edit is to ensure that the values, ratios and calculations fall within the pre-established limits, for example, minimum and maximum age for giving a birth.
- **Consistency edits** compare different answers from the same record or different records to ensure that they are coherent with one another, for example number of household members should be equal to the number of individual records
- **Duplication edits** examine one full record at a time. These types of edits check for duplicated records, making certain that a respondent or a household has only been recorded once.

Editing during data collection

- The use of electronic data collection technologies in censuses allow data editing interactively during data collection. The goal of editing at the time of data collection is to take advantage of the collection application to improve the quality of the data and reduce the costs of the post-collection process

- The response item has a built-in edit rule to inform the user about an error in case the rule is not satisfied. Typically, validation in data collection application is notified in two ways:
  - an error message which means the situation is unacceptable and must be changed in order to continue – **HARD EDITS**
  - a warning which notifies the possibility of incorrectness. These kinds of edits do not prevent the user from moving to next question or submitting the questionnaire. In this case, a failure can be resolved either through changing the response or postponing to later stages by taking no **action-SOFT EDITS**

Imputation

- Imputations generally fall into one of four categories:
  - **Deterministic imputation** - where only one correct value exists, for example, fertility questions are asked only to females.
  - **Model based imputation** - use of averages, medians, regression equations, etc. to impute a value.
  - **Deck imputation** - A donor is used to supply the missing value. The "nearest neighbour" search technique is often used to expedite the search for a donor record. In this search technique, the deck of donors comes from the same dataset and shows similarities.
  - **Mixed imputation** - In most systems there usually is a mixture of imputation methods used in some fixed rank fashion for all items. For example, first a deterministic approach is used. If it is not successful, then a deck approach is tried.

Assessment of the quality of the process

- The editing process is a complex and iterative exercises which may introduce new errors. The quality of the editing and imputation procedures should be systematically monitored and evaluated by a specifically dedicated team for this purpose.
- For monitoring the application of editing rules and imputation procedures, a number of indicators can be used. The following indicators are suggested:
  - **Edit failure rate:** It is the number of detected errors divided by total errors (detected and undetected errors) in the input file- varies between 1 and 0 (low value shows the rule fails and must be improved.
  - **Adjustment rate:** The number of households or people created under absent household imputation- if used- divided by total the number of households or people.
  - **Imputation rate:** It is the percentage of the imputed records in total records. This indicator should be calculated for each variable
  - **Dissimilarity index**: Degree of change of two distributions (observed and total including imputed values) at the variable level-ranges from 0 to 100, which implies maximum dissimilarity.

Index of dissimilarity

- To assess the degree of change induced by imputation on the initial distribution of variables

$$ID = \frac{1}{2}\sum_{k=1}^{K}\left| f_{y_k} - f^{*}_{y_k} \right|$$

Where;
k : categories of the variable
f : percentage distribution of the variable before imputation
f^{*} : percentage distribution of the variable after imputation

Index of dissimilarity

$$ID = \frac{1}{2}\sum_{k=1}^{K}\left| f_{y_k} - f^{*}_{y_k} \right|$$

$0 \leq ID \leq 100$

- It assumes a 0 value when the two distributions before and after imputation are equal
- It is greater than 0 when they are different and reaches its maximum value of 100 when there is maximum dissimilarity between the two distributions
  - *when both are concentrated in one category which is different from each other*

Assessing the performance of imputation

Population and Housing Census of Albania, 2011 Census

	Imputation Rate	Dissimilarity Index
Citizenship (Albania)	10.98	11.0
Had live-born children	9.81	1.3
Family nucleus	9.49	5.8
Highest completed level of education	7.55	2.2
Place of residence, 2001	5.72	2.6
Had a job last week of September	5.32	2.7
Ownership of the dwelling	7.3	6.0
Use of agricultural land	3.1	3.1

Individual dataset

Number of records = 2,800, 138

Number of variables = 66

Household dataset

Number of records = 722,262

Number of variables = 30

Imputation Rate: Number of imputed records/ Total nunber of records*100

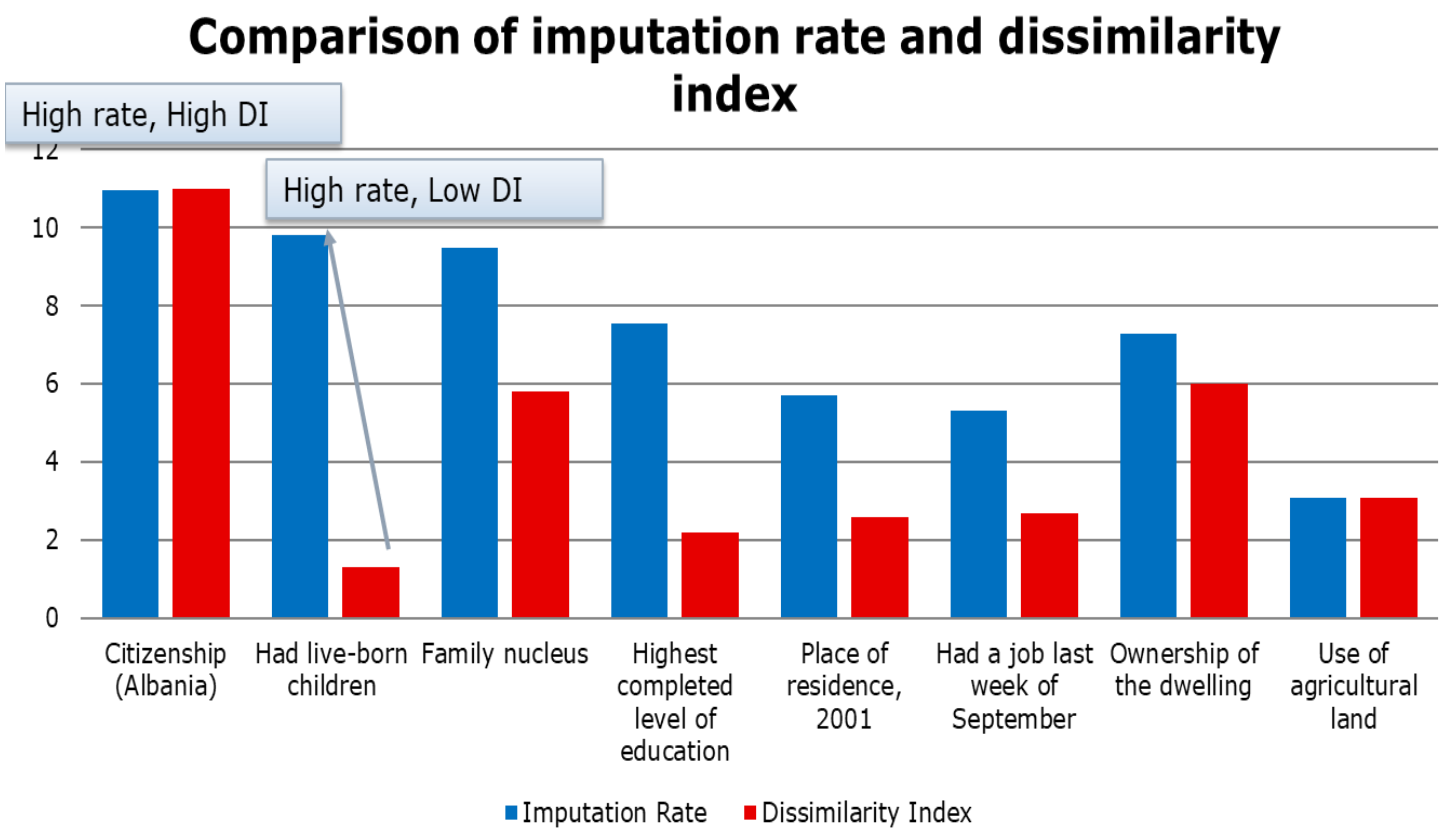
Source: Albania, Quality Dimensions of 2011 Population and Housing Census, May 2014

Assessment of the quality of the process

- This index, calculated as:

$$ID = \frac{1}{2}\sum_{k=1}^{K}\left|f_{y_k} - f^*_{y_k}\right|$$

where: k = categories of the variable:
f= percentage distribution of the variable before imputation;
f*= percentage distribution of the variable after imputation)
This index ranges from 0 to 100, where 100 implies maximum dissimilarity



Source: Albania, Quality Dimensions of 2011 Population and Housing Census, May 2014

Conclusions

- The editing process is an **iterative and interactive process** that includes procedures and processes used for detecting and handling errors in data.
  - These procedures and processes are designed depending on the method of data capture: manual data capture, optical data capture or electronic data capture.
- **Over-editing** has a negative impact on the editing process in several ways, including timeliness, cost, and the distortion of true values.
  - When further editing has a negligible impact on the final results, it is called overediting, and should be avoided.
- **Testing of the editing process** is critical to the process of developing the most effective rules for collecting reliable data.
  - Testing is an iterative process consisting of testing, making changes and corrections and re-testing, until the editing rules and imputation procedures are as much as possible error-free