Valentina VASILE, INE-RA

Silvia PISICA, NIS

DATA QUALITY- from enumeration to final database. Challenges and limits

International conference

"Population Census, 2020 round and post 2020: from traditions to modernism"

8-9 April 2019, Bucharest

Main challenge

- Methodology - adapted to the new era
- Database – from multiple sources

- Limitations
  - the possibility and particular situation in each country
  - data relevance for users- focuses on improving coverage and data quality

Session IV:

- Role of the post-enumeration survey and methods for correcting / improving the census results by applying the results of post enumeration survey and other census aspects

- Technical aspects regarding the application of statistical methods for estimating certain indicators

Preliminary aspects
- from one size to multidimensional approach and use -

- From "classical" - to "modern" enumeration method – individual record -F2F "Pen and Paper Interviews" ➔ digital one "Computer Assisted Personal Interview" (tablets, web)
- From one source to several - direct and indirect questionnaire, self-enumeration, wider administrative database -all kinds of registers as much as possible)
- no GIS ➔ to with GIS

+ diversification of users

Police-makers

Mass media for information

Experts:

- different specialists i.e. the application of statistical methods for estimation based on census data (Ana Maria Ciuhu & Roxana Glavan)

- researchers i.e. the challenges for censuses in high emigration societies (prof. Dumitru Sandu)

Organization of statistical research includes:

- quality control of the data collection process
- measures of the impact of the data adjustment / imputation activity

Data imputation and adjustment

- Identification of NA / missing data
- Identifying Outliers-Aberrant Values
- Imputation methods
  - N / A
    - Manually correcting data takes place when selecting a more reasonable / good value is done by a person.
    - Automatic data correction takes place as a result of the action of your computer.
  - Aberrant values
- simulation package in R

Automated imputations - the following types:

- **A= Deterministically** - if there is a value that is considered correct, it is attributed to missing values (NA) or aberrant values;
- **B=Model-based** - use of averages, medians, regression equations to impute a value.
- **C=Deck** - A donor is used to impute missing value.
- **D=Hot Deck** - The donor is searched for in the same research as the welcoming record. "Near Neighbor" is often used to speed up the search for a donor. In this search technique, the donor displays similarities with the welcoming record (correlates with the donated data).
- **E=Cold deck** - Similar to the hot deck, except that the data is in a similar research done previously or in administrative sources.
- **F=Mixed** - a combination of methods i.e A, then D, after B and if all fail manual imputation by human intervention occurs (GEIS used by Canadian Statistics)

2011 census in Romania – under registration = 1,183 thousand.

- For these individuals were imputed total individual records respectively, household, dwelling and building (if not already in the database).

- Therefore, methods have been applied to ensure the completeness of census data, using the indirect collection method from administrative sources and statistical methods of imputation of records.

2011 census - post-enumeration technics for improving the database quality

- **the census micro-data database**
  - **Individual enumeration** (F2F) by personal interview (or proxy interview with a person who knew the situation of a specific person)
  - For the under-coverage, by **indirect data** collection from administrative data sources (record imputation).
    - The administrative data source used included individual records about persons for which **the personal interviews were not done during the data collection period**
    - the records found in administrative data sources were imputed in census microdata database only if **it was determinate from several administrative data sources the 12 months continuously presence on the Romania's territory.**
    - a record missing from the initial census database and present in an administrative data source was included in the census micro-data database **only if it was enough evidence that person had the usual residence for at least 12 months inside Romania** around the census reference date.
  - **Record imputation - After processing the individual forms, the under-registration found in 2011 Census's provisional results processing stage was confirmed.**

Post-enumeration additional technics were applied

➔ to ensure **the census data completeness**

2 methods :

- **1. indirect collection from administrative sources –**
  - The administrative data sources were used "**in cascade**", one after one.
  - The information found in administrative data sources were use for item imputation of census's variables of the total imputed records.
- **2. statistical techniques for imputation of data**

the administrative data sources

- **The National Register of Personal Data** (RNEP) – managed by the Directorate for Persons Record and Databases Management;
- Statement on obligations to pay social security contributions and income tax, and the nominal records on **insured persons - D112** – managed by the National Agency for Fiscal Administration;
- **Record of Employees** - IM – managed by the Labor Inspectorate;
- CNPP database – managed by the National House of **Public Pensions** (CNPP);
- CNAS database – managed by the National **Health Insurance** House (CNAS);
- Tax Registration Statement / **Statement for individuals** who carry out economic activities independently or liberal professions **– D70** - managed by the National Agency for Fiscal Administration;
- Record of **beneficiaries of state child allowance, family allowance and help the guaranteed minimum aid** – managed by the National Agency for Payments and Social Inspection (ANPSI);
- Database of **students enrolled** in the 2011-2012 school year – managed by the Ministry of National Education.

The missing information – The principle of **item imputation**

= *the best source for imputation and best criteria to find a good proxy for the missing information*

2 main categories of item imputation :

- A. Some variables were imputed from the administrative data sources where we found it; we consider this **an indirect collection of data**, not an imputation, because these data are information declared by individuals, so it represent valid values, not artificial ones (as it is the case of item imputation).
- B. The second category refers to **statistical imputations**. Depending on type of variables, we applied the following kind of item imputation:
  - for variables referring to individuals = the **hot-deck donor** method;
  - for the qualitative variables related to dwellings = the method of **most frequent value** from the cell a specific dwelling is part of ;
  - for the quantitative variables related to dwellings = the method of **cell's average** for that cell a specific dwelling is part of ;

Duplicates – were deleted

- were found several records with the same individual numerical code (most of the cases we found pairs of records)
  - within the same county or
  - in different counties

Coverage assessment

- 2002-2011 - no data sources on emigration according to usual residence criterion were available.
- Migration pattern in Romania for mobile workers = to increase the period of working abroad – **multiannual basis** ➔ the usual resident population was over-evaluated due to under-evaluation of international migration (i.e. the totality of persons leaving Romania for more than 12 months, establishing usual residence abroad, but maintaining the legal residence – domicile - in Romania).
- For census data's completeness – the method used = was to look for not-enumerated records in administrative data sources, using time, usual residence and twelve months criteria.

Coverage assessment for:

Under- coverage	• Over-coverage
A. Post-enumeration survey (The sample was established in view of the two basic principles of selection: maximizing representativeness and minimizing costs) B. Ad-hoc registration of addresses where the enumerators found non-response C. Mass media messages during data collection period D. Comparison with Romanian population register, administrative sources and Eurostat mirror statistics	was measured after the data entry: - for persons using the Personal Numeric Code (CNP) by finding more than one unique CNP in the P, TP and PPI files; - for dwellings, using the addresses; - for buildings, using the unique building code; The over-coverage was solved deleting the second, third etc. record from the census micro-data database for each identification variable present more than one time in the census micro-data database.

Post-enumeration survey

- Recording the information was performed just as it was at the Census, i.e. by "face-to-face" interviews, based on the statements of persons inside the housing unit
- the sample - was established from that selection base, by means of a two-stage probabilistic selection**: sectors** were chosen in the first stage of the survey **and the dwellings** were chosen in the second stage.
- The survey sample selection was done in two stages:
  - centrally, at the National Institute of Statistics, by choosing the survey sectors, by means of unrepeated random selection; to assess the Census **completeness and coverage degree**
  - in each county, by selecting a third of the dwellings included in the survey, for which essential features relating to all households and individuals that make them up were recorded in order to **verify the quality of Census data**.

Final data validation

- Errors detected & solved
- the whole set of tabulation was prepared.
- Last errors showed by tabulation were solved
- hypercubes were executed on the clean census micro-data database.

2021 census in Romania

- may be collected in electronic format to reduce the period needed for processing the information, to ensure transparency and to publish, even during the census, the people count,
- Use of tablets for data collection and/or web
- widely use of administrative sources that exist at central level
- Usual residence (since 2012)

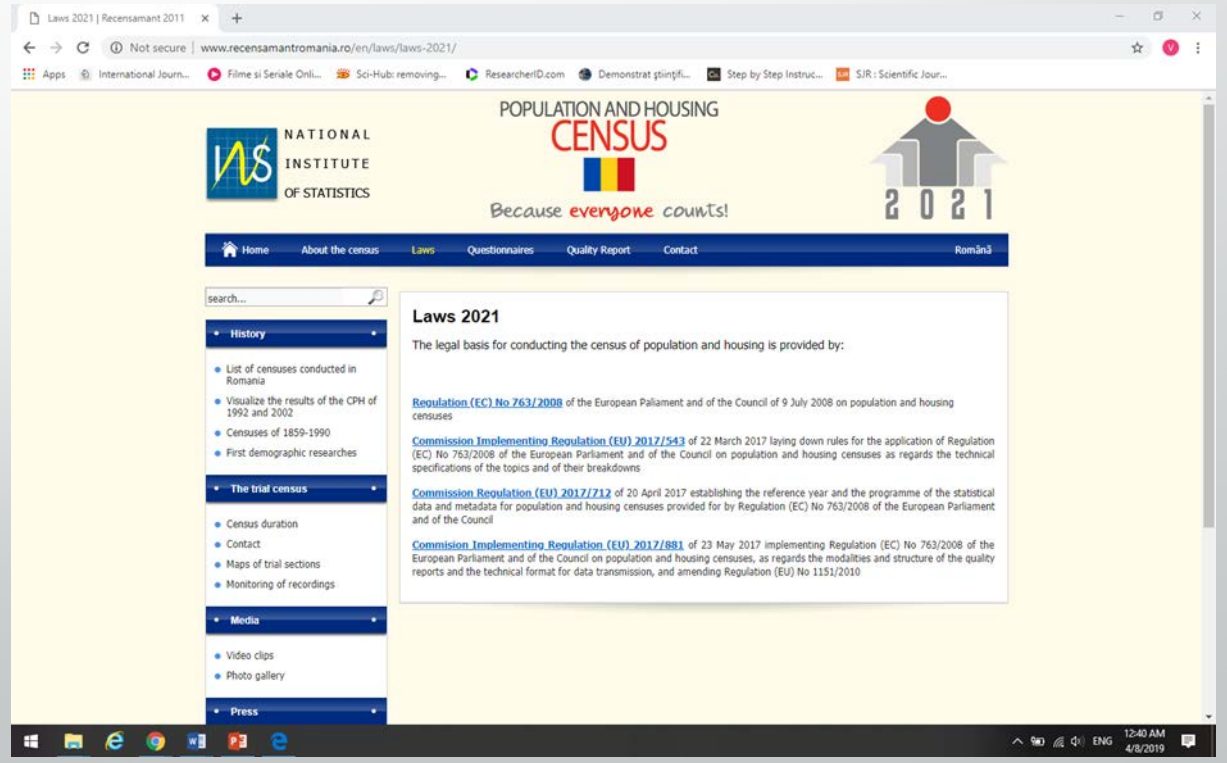
Source: https://www.capital.ro/datele-pentru-recensamantul-populatiei-si-al-locuintelor-ar-putea-fi-colectate-in-format-electronic.html 23 mai 2018, and https://www.romania-insider.com/data-romania-next-census-electronic based on Tudorel Andrei declaration – Agerpres

UNECE table for census typology Romania 2021

- Combined census - (registers + full collection for selected variables)
- Use of internet response

(estimated take up

at least 20-30%)



Lesson learned from 2011 census

- the census should no longer use solely the traditional data collection method (face-to-face interview), but a mixture of methods (self-registration on the web, information taking over from administrative sources and the traditional collection with the enumerators support).

- Source: Meeting of the Management Group on Statistical Cooperation 14-15 March 2013 Luxembourg, - Feedback from the Conference on Population and Housing censuses (the Budva Initiative Group) p.14

In the “digital census” case do we need post-enumeration technics for improving the database quality ?

- **QUALITY CONTROL IN THE CENSUS DATA COLLECTION PHASE IS VITAL TO PRODUCE QUALITY STATISTICS**

- Individual enumeration (F2F) using digitalisation
- Self-enumeration using web
- proxy interview with a person who knew the situation of a specific person
- indirect data collection from administrative data sources

=➔ YES