Register-based census Statistical processing
Case of Slovenia

Danilo Dolenc

GCC-STAT Regional Workshop

Muscat, Oman, 22-24 September 2019

Statistical process – general scheme

- 1. Defining content, structure and sources of basic and auxiliary productional census tables
- 2. Preparation of input datasets
- 3. Integration of datasets
- 4. Editing
- 5. Data warehousing

Defining content, structure, sources (1)

- Three basic ORACLE tables are created
  - POPulation, DWelling, BUILding
- Several auxiliary ORACLE tables for specific domains or derived variables are created
  - HOUsehold, EDUcation, ACTivity_status, MIGration, FERtility….

Defining content, structure, sources (2)

- Six types of variables determined:
  - Input identifiers
  - Output (final) identifiers
  - Census topics
    - Could be more than one variable for a topic
  - Working (processing) variables
  - Final derived variables – warehouse only
  - Territorial metadata

Defining content, structure, sources (3)

- Obligatory metadata for every variable
  - Data source (name of input dataset)
    - The same name of the variable
  - Editable or not
  - Format and lenght
  - Classification used
    - Stored in our classification server
      - Publicly available

Defining content, structure, sources (4)

- Timeliness of sources
  - Most refer to census reference day (1 January)
  - Some refer to another reference day
  - Few refer to previous calendar year
  - Also some historical / longitudinal / previous census data are used
- Availability of sources
  - Four stage data processing

Data sources (1)

Data source content	**Ref. day / period**	**Availability in months**
Central Population Register	1.1.Y	T+3
Household Register	1.1.Y	T
Real Estate Register	1.1.Y	T / T+12
Statistical Register of Employment	1.1.Y	T+2
Business Register	1.1.Y	T+1
Registered unemployed persons	1.1.Y	T+1
Primary and secondary enrolment	30.9.Y-1	T+5
Tertiary enrolment	1.10.Y-1	T+5
Scholarship recipients	1.1.Y	T+5
Pension recipients	1.1.Y	T+3

T = 1 January (reference day)

Data sources (2)

Data source content	Ref. day / period	Availability in months
Persons in health insurance	1.1.Y	T+1
Social transfer recipients	Y-1	T+6
Income tax payers	Y-1	T+10
Tertiary education graduates	1989+	T+6
Matura graduates	2002+	T+6
Previous census	1.1.Y-3	T
Statistical survey on migration	2002+	T+7
Statistical survey on birth	2002+	T+6
Population one year before	1.1.Y-1	T

T = 1 January (reference day)

Preparation of input datasets (1)

- More than 30 datasets
  - Most common - .txt or .csv
  - Direct connection to another ORACLE database
- Datasets prepared by:
  - Subject matter methodologists
    - Statistical (final already edited data)
  - Persons responsible for administrative sources
    - Administrative raw data (not changed)
    - Administrative edited data

Preparation of input datasets (2)

- Basic rules for input datasets
  - No duplicates of key identifier
  - Only valid identifiers used
  - Re-coding (if necessary) to prescribed classification
  - Full coverage

Preparation of input datasets (3)

- Statistical data (example)
  - Usual population quarterly derived from CPR three months after reference day
    - Input – 2.6 million records
    - Output – 2,084,301 (1 April 2019)
  - Usual population = census population

Preparation of input datasets (4)

- Administrative raw data(example)
  - Household Register data
    - No sense to edit data before data integration
    - Usual residence already derived before using household data
      - Full match of address and PIN

Preparation of input datasets (5)

- Administrative edited data(example)
  - Income Tax Payers
    - Labour force status derived on the basis of type of income

Data integration

- Integration of datasets using unique identifiers in ORACLE
  - Fundamental principle for each variable – number of sources foreseen
    - One source – direct load to basic tables
    - Two or more sources – load to auxiliary tables first
      - After data integration and editing – load to basic tables
- Data integration is the most important step in register-based census processing

Editing

- Four main procedures employed
  - Automated correction / derivation of missing identifiers by using tailor-made programmes
    - To link tables
  - Manual corrections using interface
    - Only household / family data
  - Automated corrections using generalised own developed metadriven application in SAS
  - Imputations using generalised own developed metadriven application in SAS

Data warehouse

- Setting up four final databases for dissemination
  - PERSONS
  - HOUSEHOLDS
  - FAMILIES
  - DWELLINGS
- Variables from basic ORACLE tables
  - PERSONS - 79
- Derived variables
  - PERSONS - 47

Basic principles of data processing

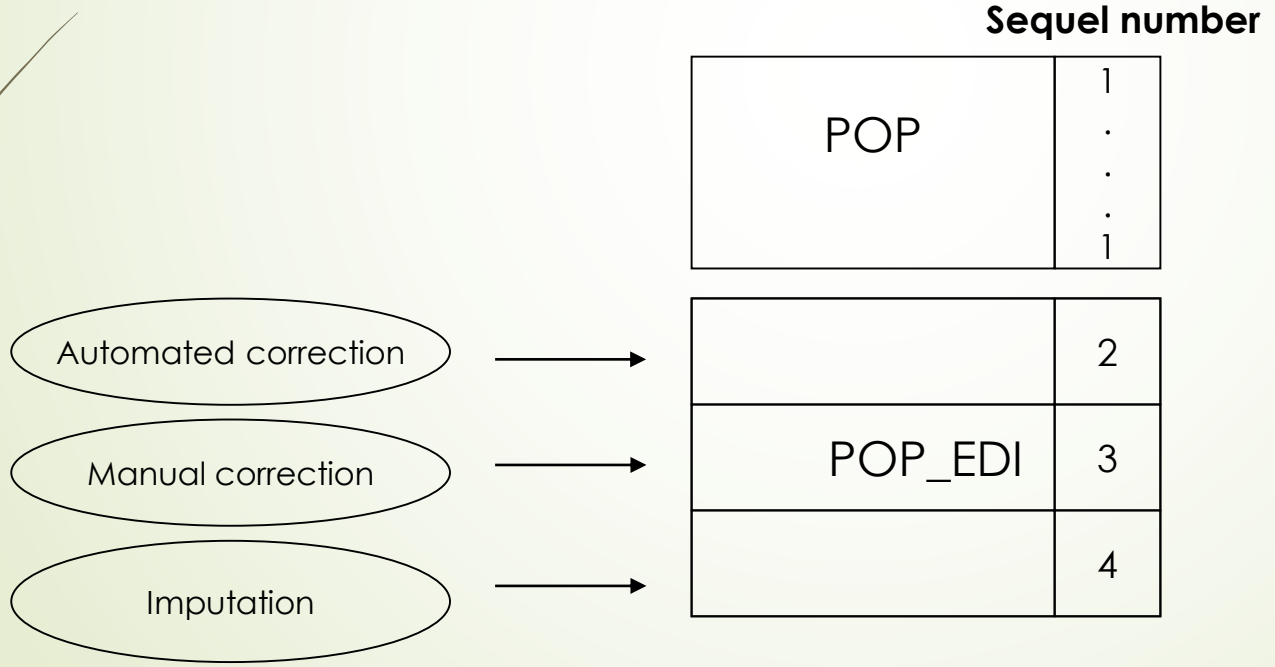
- Gradual (step by step) data processing
  - All data sources are not available at the same time - integration of the input data had to be adapted to the timeliness of the sources
- Traceability
  - Corrections did not replace the old values - new version of the record is created
- Repeatability
  - Each step in the process is repeated as many times as necessary with the same outputs

Gradual data processing - steps

- Basic population data – T + 4 months
- Household / family data – T + 9 months
- Migration, fertility and socio-economic characteristics – T + 11 months
- Housing data (occupied and non-occupied dwellings) – T + 18 months
- **No change of data after every step**

Traceability (1)

- Basic table (POP) created – initial input data
- Auxiliary table (POP_EDI) – new version of the record created in case of editing
- Any change of record is labelled by sequel number



Traceability (2)

- Another ‘‘mirror‘‘ table (POP_STATUS) is created to trace the status of the change of each variable

POP

ID	V1	V2	…	Vn

POP_STATUS

ID	V1_s	V2_s	…	Vn_s

Traceability (3)

- The status gives information on the type of the change in the process

Initial data

ID	V1	V2
111	300	01
112	2	Null

ID	V1_s	V2_s
111	21.11	31.11
112	21.12	Null

Automated correction

ID	V1	V2
111	3	01
112	2	Null

ID	V1_s	V2_s
111	22.13	31.11
112	21.12	Null

Imputation

ID	V1	V2
111	3	01
112	2	04

ID	V1_s	V2_s
111	22.13	31.11
112	21.12	41.14

Traceability (4)

- Relation between tables and views
- The key: statistical identifier of person