Statistical Office
in Poznań

Protection of statistical confidentiality of census data in Poland – problems and challenges

Tomasz Klimanek, Andrzej Młodak, Tomasz Józefowski

Introduction

- The National Population and Housing Census in Poland was conducted from 1st April to 30th September 2021. providing a very large collection of data.
- An efficient and safe dissemination of census outputs requires broad and specialised actions aimed at protecting statistical confidentiality and ensuring maximum data utility for end users.
- We present the main tools and methods to be used in statistical disclosure control (SDC) of the census data and indicate the key problems and challenges that can be encountered in this field.

Presentation outline

- 2021 Census
  - basic information about the census
  - data to be released
  - types of possible users
- Data protection
  - basic principles
  - microdata protection
  - tabular data protection
- Problems
  - methodological challenges
  - technical and IT questions
- Summary

2021 Census

Basic information about the census

- online self-enumeration (CAWI), supplemented by phone interviews (CATI) and direct interviews (CAPI), which was seriously restricted owing to COVID-19
- data collection was supported by administrative data sources
- the census database (about 38 million records) contains information about demographic characteristics, economic activity, education, disability, migrations, households, families and dwellings

2021 Census

Data to be released

- **microdata**
  - data for **scientific use** – selected according to user specification, to be used for scientific studies and analyses, available in a protected environment
  - data for **public use** – anonymized data without certain variables, available for any interested person (online, on a CD or other storage media)
- **tabular data**
  - tables and analyses contained in official publications and databases (Local Data Bank, Geostatistical Portal, etc.) – after statistical disclosure control
  - additional, non-standard tables and analyses prepared on request

2021 Census

Prospective users

- government and local government agencies need data to conduct their statutory activities, in special cases they also need microdata about entities of the public finance sector; individual sampling frame data can be transmitted to another authority engaged in a given survey
- scientific and research organizations and institutions with an official status of scientific units
- non-governmental organizations
- economic entities
- pupils, students
- other users

Data protection

Basic SDC principles:

- rules established in the Polish law on public statistics:
  - 3-anonymity
  - 1,75 dominance
- anonymization by removing direct identifiers (e.g. personal ID number, name, given name, exact address) or replacing them with artificial codes to protect confidentiality of personal information
- removal of technical variables used only to verify the completeness of census data

Data protection

Microdata protection

- work is under way in line with international recommendations
- methods to be implemented
  - for **categorical variables**
    - ❑ targeted record swapping
    - ❑ microaggregation based on the Gower distance
    - ❑ post-randomization (PRAM)
  - for **continuous variables**
    - ❑ correlated noise addition
    - ❑ microaggregation based on the Gower distance
  - synthetic data

Data protection

Tabular data protection

- current work focuses on primary and secondary suppression to ensure compliance with the 3-anonymity and (1,75) – dominance rule; a SAS algorithm created for this purpose
- an experiment to check the 3-anonymity rule: the following approaches were analysed:
  - ❑ Edora, F. (2016), I Have a Secret: How Can I Hide Small Numbers from Public View? MWSUG 2016 - Paper RF-10, https://www.mwsug.org/proceedings/2016/RF/MWSUG-2016-RF10.pdf
  - ❑ Batkhan, L. (2018), Implementing Privacy Protection-Compliant SAS® Aggregate Reports, Paper SAS2022-2018, https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2022-2018.pdf

Data protection

Tabular data protection: an experiment for 3-anonymity

- finally, Batkhan's option was chosen
- the three-dimensional table included the following variables: district (LAU 1 region), education and biological disability (yes/no)
- a special SAS algorithm was created, containing macros for primary suppression and iterative secondary suppression according to the 3-anonimity rule

Data protection

- Tabular data protection
  - an experiment for 3-anonymity
    - constructed macros:

```
%macro suppress(in=,supclass1=,supclass2=);
   proc sort data=&in(keep=&tclass1--cnt_orig);
      by &supclass1 &supclass2 count;
   run;
   %let sup_flag = 0;
   data &in ;
      set &in;
      by &supclass1 &supclass2 count;
      if first.&supclass1 or first.&supclass2 then
      do;
/* initialize number and sum of suppressed cells */
         _supn_ = 0;
         _supsum_ = 0;
      end;
      /* enhanced suppression flag */
      if (_supn_ ne 0) then avg_flg = (_supsum_/_supn_
<= &sup_avg);
      /* apply suppression criteria */
      if (count>0) and (count<=&sup_max or _supn_=1 or
avg_flg) then
      do; /* suppress cell */
         count = .S;
         call symputx('sup_flag',1);
      end;
      if (count eq .S) then
      do; /* increment number and sum of suppressed
cells */
         _supn_ + 1;
         _supsum_ + cnt_orig;
      end;
   run;
%mend suppress;
/* suppression iteration macro definition */
%macro iterate_suppression (class1=, class2=, class3=);
%global order;
%let order = &class1, &class2, &class3;
   %let nextclass1 = &class1;
   %let nextclass2 = &class2;
%do %while (&sup_flag);
/* iterate suppression across dimension */
%suppress(in=&procmeanscounts, supclass1=&nextclass1,
supclass2=&nextclass2);
/* swap dimension */
%if (&nextclass1 eq &class1) and (&nextclass2 eq
&class2) %then %do;
 %let nextclass1 = &class2;
 %let nextclass2 = &class3;
%end;
%else %if (&nextclass1 eq &class2) and (&nextclass2 eq
&class3) %then %do;
 %let nextclass1 = &class3;
 %let nextclass2 = &class1;
 %end; %else %do;
   %let nextclass1 = &class1;
   %let nextclass2 = &class2;
   %end;
%end;
%mend iterate_suppression;
```

Data protection

Tabular data protection

- the recommended **cell-key method** is also considered as a possibility
- other possible solutions that could be used:
  - ❑ rounding
  - ❑ hypercube
  - ❑ Controlled Tabular Adjustment (CTA)

Problems

Methodological challenges

- selection of key variables (with the most sensitive information and contributing most to the risk of disclosure)
- treatment of variables regarding a given subpopulation
- dealing with missing data
- complex assessment of disclosure risk
  - ❑ different methods for categorical and continuous variables; whereas the database to be made available for users can contain both types; so a comprehensive approach to the problem is expected
  - ❑ It is hard to predict and assess the possibility of unit identification by combining the disclosed file and external data sources

Problems

Technical and IT questions

- the efficient implementation of SDC requires high capacity software that can process very big data files
- It must be easy to customize method parameters
- the R environment offers the broadest possibilities of creating efficient SDC algorithms
- problem: computer memory available to making computations in R is limited and it is not always possible to increase it, especially when the file is large.

Problems

Technical and IT questions

- in R, efficient computation of disclosure risk (and some measures of information loss) for microdata can only be performed on objects of **sdcMicroObj** class but some operations (e.g. targeted record swapping) are made in other packages, which don't use this format or use it in a restricted way; so, the final disclosure risk cannot be calculated accurately (original data are compared with SDC changes last stored in the sdcMicroObj object)
- The **sdcMicro** package now contains some measures of information loss, which don't have this inconvenience

Summary

- Efficient performance of statistical disclosure control for census data should take the following aspects into account:
  - valid regulations concerning statistical confidentiality and data disclosure
  - specificity and large size of original census database
  - a variety and big number of tables to be published or disclosed on request
  - trade-off between minimization of disclosure risk and minimization of information loss due to suppression or perturbation of data
  - utility and capacity of used software

Summary

SDC scenario for microdata:

- select variables to be protected
- divide these variables into key, PRAM and continuous ones,
- conduct preliminary assessment of disclosure risk
- apply TRS to hierarchical key variables
- choose the optimal method for remaining key variables
- use PRAM for another categorical variables
- use relevant perturbation of continuous variables
- assess disclosure risk and information loss

SDC scenario for tabular data: basic rules, linear programming or the cell-key method.

Statistical Office
in Poznań

Thank you very much for your attention!