UNIUNEA EUROPEANĂ

Programul Operațional Capacitate Administrativă
Competența face diferența

Instrumente Structurale 2014-2020

ROMANIA

Reimbursable Advisory Services Agreement on
Romania Capacity Building for Statistics (P167217)

OUTPUT No. 10b

Report on advisory services provided to Recipient on the technical assistance and best practice recommendations for SDC, data confidentiality, ways to secure micro-data and aggregate data

October 2022

Revised version November 2022

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Project co-financed from the European Social Fund through the Operational Programme for Administrative Capacity 2014-2020

Disclaimer

This report is a product of the staff of the World Bank. The findings, interpretation, and conclusions expressed in this paper do not necessarily reflect the views of the Executive Directors of the World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work and does not assume responsibility for any errors, omissions, or discrepancies in the information, or liability with respect to the use of or failure to use the information, methods, processes, or conclusions set forth. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.
This report does not necessarily represent the position of the European Union or the Romanian Government.

This report was delivered in October 2022 and revised version in November 2022 under the Reimbursable Advisory Services Agreement on Romania Capacity Building for Statistics (P167217) signed between the Romanian National Institute of Statistics and the International Bank for Reconstruction and Development on September 17, 2019. It contributes to Output 10 under the above-mentioned agreement.

Table of Contents

List of Acronyms

CKM	Cell Key Method
EU	European Union
INS	National Institute of Statistics
NSO	National Statistics Offices
PHC	Population Housing Census
R	R programing language
RAS	Reimbursable Advisory Services
SDC	Statistical Disclosure Control
WB	World Bank

Introduction

The purpose of this report is to present the **technical assistance and best practice recommendations for SDC, data confidentiality, and ways to secure micro-data and aggregate data.** This is part of the deliverables under the Reimbursable Advisory Services (RAS) Agreement on Romania Capacity Building for Statistics (P167217). The project is implemented by the National Institute of Statistics (INS) with support from the World Bank.

Under the output described by this report, the Bank team provided technical assistance on how to perform Statistical Disclosure Control (SDC) through on-the-job training in order to learn methods to ensure data confidentiality and eliminate the risk of disclosure of confidential data, methods applied at the level of microdata (individual data), but also at the level of aggregated data. The activities were carried out during June 28 to July 1, and November 21-22, 2022.

The hands-on technical assistance provided under this output had the goal to equip participants with knowledge and tools that would enable them to apply SDC methods to official statistics datasets, and census data in particular, and to enable them to decide which methods are most appropriate in specific use cases.

The report presents:

- the prerequisites that are important for the application of SDC; they were analyzed and agreed with the INS and the participants in the activities.
- all the methods considered for the anonymization of the PHC 2021 data are listed; the list is extended to a few methods in which the participants were interested.
- list of the provided R scripts for the use of the INS staff.
- conclusions and recommendations for the anonymization of the population and housing census database
- the list of available resources provided to INS as Annex.

The technical assistance provided complements the information and practices carried out during the training on SDC performed in December 2021 and reflected in the corresponding report about the detailed discussion and explanations of the anonymization methods, delivered in January 2022 as part of Output 12 under the RAS.

1. The goal and objectives of on-the-job training on SDC

The overarching theme of the on-the-job training was the application of methods and tools that can be used to protect census datasets, with a focus on the training and methods needs of the participants. The goal was to equip participants with knowledge and tools that would enable them to apply SDC methods to official statistics datasets, and census data in particular, and to enable them to decide which methods are most appropriate in specific use cases.

The intent of the assistance provided was to deepen the understanding of the methods through applications and on-the-job training. In particular, the methods were intended to be (and were) applied to PHC databases and recommendations were made (see last chapter).

The objectives of the technical assistance envisaged that by the end of the on-the-job training the participants:

- must have a strong command of the principal SDC methods (pre and post tabular methods), both from theoretical and practical perspective.
- must be able to apply from start to finish the main steps of a data anonymization and in addition, participants must be able to explain why they opted for a particular SDC method, along with describing the advantages and disadvantages of that method; and
- should be proficient users of at least one software tool for both pre and post tabular methods.

2. Preliminary actions

The following prerequisites were clarified before the technical assistance and during the on-the-job training:

a) Generally, for a proper guidance on best practice to INS to apply SDC it is important to be aware of all statistics produced and hierarchies in the data, when statistics calculated for finer geographic areas, for example, add up to figures for larger geographic areas. It has been clarified that detailed statistics at 1 sqkm grid village (200 inhabitants) must be provided to Eurostat referring to Commission Implementing Regulation (EU) 2018/1799 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km2 grid. All in all, these are 13 tables (on total population, sex, age groups, employed persons, place of birth, place of residence) which should be disseminated at grid level (presented in the Annex II of the regulation no.2018/1799). In addition, hypercubes (tabular data) must be transmitted to Eurostat that refer to Commission Regulation (EU) 2017/712 establishing the reference year and the programme of the statistical data and metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council.

b) In order to apply SDC in a way that anonymizes the results of potential new analyses and data sharing options, it is important to be aware of what potential results might be produced in addition to current needs. Thus, another requirement was to clarify whether the delivery obligations (from/for Eurostat) are to be published exclusively from the data sets material or whether there are also additional evaluations/tables that should be published (nationally) and from which new outputs could be produced in the future.

The output of the discussion was that the data should be prepared to be ready to be published at village level, ethnical and cultural (ethnicity, mother tongue and religion) NUTS3 level (municipalities, cities, communes). Only time will tell whether further data deliveries will be required.

c) For micro data but also tabular data protection specifics: some variables include a lot of categories, e.g., occupational code. How and where those categorizations can be generalized to end up with fewer ones? This question has not been clarified by INS. As SDC is first time being applied by INS, they need to establish how to set this occupational code – 6 digits but published at 4 digits (discussion needed with experts). For other variables, such as age groups, the choice of age groups was clarified.

d) There were no tendencies (after the training carried out in December 2021) regarding the choice of methods for tabular data and/or microdata. In addition, it was mentioned how other countries apply SDC methods for population and housing census data. Austria, for example, applies target swapping.

e) The need for the installation of R and RStudio with the associated packages for data anonymization (sdcTable, sdcMicro, simPop) was clarified, by that R, RStudio and the SDC packages are available at INS.

f) The participants in the on-the-job training are the same persons in the training held in 2021, to ensure a consolidation of their information and skills. Additionally, were included 3 newcomers. They all have at least basic knowledge in R.

g) Was clarified that support in theory and application is given, and the issues are discussed. Also, it has been clarified that the whole work to anonymize the population and housing data is the responsibility of INS and will remain to INS.

h) It has been decided to discuss SDC methods in the light of the PHC 2011. In particular SDC should be applied on sets (samples) of data from 2011 (on the job training) then INS repeat the process for same samples and for others sets to learn by doing the path and process. It has been considered to apply SDC also on actual collected data (e.g., from localities where the census is completed through self-enumeration). INS will replicate the process and learn how to do it and once the data is ready collected at the end of census and validated, they can apply the method to actual PHC 2021 data collected. It should be noted that the PHC 2011 subset was delivered to WB experts on the first day of the on-the-job training.

i) The applied SDC methods on the data sets chosen by INS, the results and specific technical aspects observed during testing will receive and answers and explanations during two online sessions to be caried out in November 21-22, 2022.

3. Main aspects covered during the training

In depth, the on-the-job training covered the following topics with applications in R on the PHC databases:

- (Target) Record swapping
- Cell Key Method (CKM)
- Microdata anonymization including risk assessment, anonymization methods and evaluation of the data utility
- SDC spatial

These methods were explained in detail in the report submitted after the training held in December 2021 (see Output 12). During the on-the-job training, these methods were practically applied.

Due to the large interest of the participants, the following topics were also covered:

- Synthetic data generation
- Anonymization of spatial data

Both topics were not part of the training in 2021. Slides and code on these topics were provided, see the Annex.

Moreover, theoretical and practical discussions took place for:

- Short repetition of methods
- Introduction to the provided report of the training carried out in December 2021. We note that the report prepared in December 2021 was not received and viewed by the participants until the on-the-job-training time therefore, was also discussed the content of the main parts of this report.
- Advantages and disadvantages of methods for certain data sets and method selection for certain data sets

4. The R code materials provided

The materials that were prepared, presented and provided to INS in the form of R code by the Bank team are presented further in this section.

A. Data preparation (file prep-data.R)

Script that imports the subsets on the PHC 2011 and PHC 2021 to R including some data preparation.

B. Application of the cell key method (file practical_cellkey_a.R)

On the PHC 2011 sample was showed functionality of cellKey-package. Note that the training on cellkey was done in the first day of the training-on-the-job, where we still did not had access to the new PHC 2021 data set. The remaining training was then applied on the PHC 2021 data.

The training on the CK method involved a step-by-step approach of how to protect a table

- read sample dataset + adding record keys
- recode dataset in order to be able to define a table "gender x agegroups x occupation"
- create required hierarchies
- define perturbation parameters
- perturb table
- extract and save perturbed results

C. Application of the cell key method II (file practical_cellkey_b.R)

One of the required tables that need to be submitted for the census have been perturbed using the cell key method, particularly point 12 at Commission Regulation (EU) 2017/712, that is population by sex and marital status, by age groups - categories of localities. For didactical

reasons this script was developed during the course with active involvement of the audience and guided by the WB experts.

D. Data preparation needed before applying record swapping (file recodes.R)

It applies some recordings of the variables.

E. Target record swapping (file recordSwapping_a.R)

Record swapping was applied on a subset of the PHC 2021 database.

F. Spatial SDC (file sdcSpatial.R)

The script shows how to find out which locations are considered sensitive / unsafe for publishing and application of protection methods that reduce sensitivity and enhance spatial patterns by smoothing, coarsing and removement of sensitive locations.

G. Application of traditional anonymization methods and synthetic data simulation on survey data and PHC 2021 data (file training_2022_class_sdc.R)

This longer script contains the application of SDC methods, but also shows the simulation of synthetic data. It has four parts:

- traditional anonymization (SDC) on one survey data
- traditional SDC on the CENSUS data
- synthetic data on one survey data
- synthetic data for the CENSUS data

H. Utility function to deal with households (file utils.R)

This utility function used to extract households from the PHC database.

5. Conclusions and recommendations

The technical assistance and on-the-job training were necessary to demonstrate how SDC methods are applied in practice to population and housing census data and, in general, to other data sets.

It was observed by the Bank team that this practical course was even more useful for the participants than the training held in December 2021 and in fact represented the way of consolidating the knowledge and applying the specific SDC tools. Due to the large topic and complex methods, the more theoretical training initially provided was nevertheless necessary to get to know the basics and to allow the participants to be able to fully concentrate on the practical examples in this course. This could be a good practice for all similar specific production tools for learning and applying further in the activity of INS.

The feedback received directly from participants was overall very positive and confirms the approach of experts that the material and exercises that prepared for and discussed throughout the training will also be helpful in the real-world work of the participants. The participants should be able to translate the R code developed by the Bank team to their practical problems and to anonymize the PHC database.

Discussions that had taken place and parts of the scripts that were provided, were developed and maintained in a guided manner with the participants. This was done for didactic reasons, to actively involve the participants in a practical way. This allowed "instant-feedback" and supported discussions.

The participants were also interested how to protect other data sets. Recommendations were given as well during the course for certain other data sets of INS. This also supported discussions and gave the participants a broader view on this topic.

Finally, some recommendations are made. Notably, these recommendations are also included in the report submitted in December 2021 (see Output 12), but they were tightened and are now based on the new information received from the INS.

The main recommendation could be to use record swapping for protecting the PHC database. Several reasons substantiate the recommendation, but the final decision is the INS attribute, which could consider the following:

a) It is obviously known that in addition to the delivery obligations additional data (additional variables, other breakdowns, hierarchies, etc.) will be requested; this also means that in reality it is unclear what will actually be published from the census micro data. All tables and other delivery obligations (hypercube + grid level) for Eurostat will be only a part of the whole publication.

b) This means that the only solution to anonymize the data consistently is a (single) micro data perturbation (such as swapping). This perturbed dataset is then from which all other publications, tables, hypercubes, and others, are produced.

c) This approach also solves the problem of "villages" being parts of localities and other hierarchy problems; otherwise, it is impossible to produce consistent (and safe) anonymization.

d) It should be clearly stated that it makes little sense to work with primary and secondary suppression in the hypercubes or grids, to suppress a lot of information and then not to be able to guarantee that: a) linked cells always have the same status (open/suppressed); and b) therefore single suppressed cells can be uncovered. Such an approach (cell suppression) is extremely time-consuming and often leads to questions afterwards, which are difficult to explain (e.g., inconsistencies).

e) There is a reason why other NSOs prefer to use record swapping for the census data. The regulations regarding the required data hypercubes lead to some of the data cubes, that are essentially micro data (large percentage of cells with single units/persons). Furthermore, many cells across the data cubes overlap.

f) A huge benefit of the variant is also that one could offer later for researchers e.g., statistics on smaller grid level, without having to change anything.

g) Special care should be taken to communicate swapping to managers, stakeholders, lawyers, and public.

h) One disadvantage of record swapping is that it is the most difficult method to apply since the software is sophisticated to use.

i) Clarifications still need to be made for the INS on the details of the categories of some variables. This especially includes the possible recoding of 6-digit code levels on occupation to, e.g., 2 or 4-digit codes. The WB experts strongly recommend that INS to also evaluate data utility and disclosure risk after recoding.

The primary actions to be taken include the adaption of experts' provided code for the full PHC database, and final decision on the choice of method.

Annexes

1. The on-the-job training participants were provided with the R codes mentioned in Section 4 (A-H) and all slides in a Google Drive folder shared by INS: https://drive.google.com/drive/folders/1XOS09XB5DOEjnk2VWlhS6ZORfbzw3kJ5?usp=sharing_eip_m&ts=62bc30bc (access only by invitation from the INS). It also includes detailed slides on new topics such as synthetic data generation and spatial SDC, as these two methods were not part of the 2021 training.

2. Code on sdcMicro, sdcTable and simPop, maintained and written by the Bank team are available at https://cran.r-project.org .

3. The code for the cell key method is available on https://github.com/sdcTools/cellKey .

4. The report of the training delivered in December 2021 detailing the methods was made available and is also included in the drive folder given above.

Competence makes a difference!

Project selected under the Administrative Capacity Operational Program, co-financed by European Union from the European Social Fund