Center of Excellence on Statistical Disclosure Control

Maël Buron[*]

[*] French National Institute for Statistics and Economic Studies (Insee), mael-luc.buron@insee.fr

Abstract: In October 2014, a Framework Partnership Agreement (FPA) was started to establish a Centre of Excellence for Statistical Disclosure Control (SDC). This partnership consists of eight partners and is coordinated by Statistics Netherlands (CBS). The work this Centre of Excellence will perform is organised by use of Specific Grant Agreements (SGAs) for specific projects. Beginning 2016, an SGA on “User support for and maintenance of SDC tools” was signed between partners of the FPA and Eurostat, partly financed by Eurostat. One activity of this SGA is to ensure support of SDC software. To do so, a SDC Tools User Group and a Help Desk Portal were set up. Another goal is to maintain and develop tools, adding functionalities, new methods, improvements and fixing bugs. To identify the needs, two questionnaires concerning micro data and tabular data protection were sent to members of the Working Group on Methodology. The paper will briefly present the conclusions drawn from the questionnaires. It will then describe the environment set in place for the support and development and the state of affairs of the Centre of Excellence for Statistical Disclosure Control.

Introduction

A Framework of Partnership Agreement (FPA) was created in October 2014 to institute a Centre of Excellence (CoE) on Statistical Disclosure Control (SDC). This partnership is coordinated by Statistics Netherlands (CBS). It consists of eight partners: Statistics Netherlands (CBS), Statistics Austria, Destatis (Germany), Statistics Finland, Insee (France), Genes (France), HCSO (Hungary) and SURS (Slovenia). The FPA gathers together SDC experts from different countries. It lasts four years and has five main objectives:

1. Develop methodological solutions to SDC issues common in the whole ESS and for different statistical domains;
2. Improve and develop SDC tools;
3. Develop guidelines and requirements that are compatible with the current legal frameworks and practices;
4. Support and/or guide National Statistical Institutes (NSIs) to set up and adapt their infrastructure in order to comply with minimum requirements for protection of confidential data;
5. Provide training and courses on particular SDC issues.

Specific Grant Agreements (SGAs) regulate the work of the CoE. These SGAs tackle specific projects. Three SGAs have been launched so far :

1. Public use files for European Statistical System (ESS) micro data
2. User support for and maintenance of SDC tools
3. Harmonised protection of Census data

The first part of this paper describes the FPA and gives an overview of its work in these three different SGAs. The second part of the paper presents on the work in the second SGA: User support for and maintenance of SDC tools. It describes the results of two questionnaires sent to the members of the Working Group on Methodology then focuses on the environment set in place for the support of SDC software in the User group and the Help Desk Portal and its development.

1 State of affairs of the Centre of Excellence for Statistical Disclosure Control

The work of the CoE is organised with SGAs. When one is offered by Eurostat the CoE debates and tries to find solutions. If some partners of the CoE believe it is possible to produce meaningful work in order to provide answers for the SGA, then a formal proposal is made. In this proposal the partners describe the means to tackle the issue raised by the SGA. This proposal also specifies the amount of funding the work requires.

The Working Group of Methodology produced a ranked list of possible SDC projects. Eurostat proposes SGAs from it. These SGAs are specific issues for which the work would benefit the ESS. The funding provided is up to at most 70% of the eligible costs. Therefore, the partners of the FPA that make a proposal for the SGA are also interested in the results of the project. The CoE made proposals for three SGAs.

This first part of the paper gives an overview of the work in these three SGAs. The organisation of the CoE was first presented at the NTTS conference (Smukavec and de Wolf, 2017). A website was set up on the CROS portal to monitor the progress on the different projects : https://ec.europa.eu/eurostat/cros/content/centre-excellence-statistical-disclosure-control-0_en.

1.1 Public use files for European Statistical System (ESS) micro data

Six partners are involved in this first SGA: Statistics Netherlands, Statistics Austria, Statistics Finland, Insee, Destatis, HCSO and SURS. This first SGA started in January 2015 and lasted one year. See de Wolf (2015) for the first results.

1.1.1 A new kind of micro data set useful before researchers' accreditation and during students' training

Eurostat has different micro data sets available to researchers. There are Secure Use Files (ScUFs) and Scientific Use Files (SUFs). The ScUFs are accessible in Eurostat buildings only whereas the SUFs can be used outside. Both of these files are confidential and not fully anonymised. Therefore, these two types of files are only intended for accredited researchers and the files are provided with legal obligations and penalties if the restrictions are breached.

To get access to these micro data sets, the organization has to be a genuine research entity. Researchers of genuine research entity can submit research proposals to Eurostat for specific projects. The access to the micro data is granted if Eurostat accepts the proposal. This accreditation process sometimes takes up to 10 weeks.

Because of the delay in this accreditation process, researchers could profit from a precise description of the dataset available before applying. It would even be better if Eurostat had example of micro data file available. Such a file would need to be completely protected because it would be sent without any obligations. To produce this new kind of micro data called Public Use Files (PUFs) Eurostat launched a project in January 2015.

1.1.2 Different approaches for EU Labour Force Survey (EU-LFS) and EU Statistics on Income and Living Conditions (EU-SILC)

The PUFs are created from SUFs already available at Eurostat. The PUFs are made to provide a first view to the researchers before they apply to the accreditation process. They should also be provided for students during their statistical training. In the SGA proposal the SUFs for the EU-LFS (EU Labour Force Survey) and the EU-SILC (EU Statistics on Income and Living Conditions) to be used for creating those PUFs were singled out by Eurostat base on popularity.

Two different methodologies were used for the EU-SILC and the EU-LFS files. A synthetic data approach was preferred for the EU-SILC while the EU-LFS PUF was created with traditional SDC methods only.

Because the project lasted only one year and because there were no readily available method to set up a synthetic longitudinal dataset, the partners decided to limit the work to the cross-sectional part of the EU-SILC.

For EU-LFS, the partners first created PUFs for the quarterly dataset then combined these to construct a yearly PUF. Only tradition methods were used: global recoding, local suppression and PRAM (Post RAndomization Method, Gouweleeuw et al., 1998).

The PUFs were constructed for EU-SILC 2012 and 2013 and for EU-LFS 2013. They can be downloaded on the CROS portal website for the partners involved in the SGA: https://ec.europa.eu/eurostat/cros/content/puf-public-use-files_en. Eurostat is now working to construct the PUFs and get the consent to publish these for the other Member States.

1.2 User support for and maintenance of SDC tools – brief overview

Five partners are involved in this second SGA: Statistics Netherlands, Statistics Austria, Insee, Destatis and SURS. It started in April 2016 and lasts two years.

A brief overview of the work conducted in this SGA will be given in the following paragraphs and it will be detailed more precisely in the second part of the paper.

1.2.1 Support of SDC software used in the ESS: ARGUS and R-packages

SDC software is critical to protect the privacy of the respondents and the confidentiality of their response when statistical output is disseminated. The main aim of this second SGA in the FPA is to ensure support, testing, maintenance and development of SDC software of interest to the NSIs in the ESS.

Software tools can be seen as an user interface (UI) with underlying methods offered through this UI (which can be for example modular, optimal, CTA, hypercube, local suppression, PRAM, swapping, see Hundepool et al. (2010) or Hundepool et al. (2012) for details on these methods). The following interfaces are used by the NSIs and Eurostat :

1. ARGUS (μ-ARGUS, τ-ARGUS)
2. R-packages (sdcMicro sdcMicroGUI, sdcTable, SimPop)

In these interfaces the underlying methods are sometimes similar but not exactly identical. Another goal of the SGA is to identify such cases of parallelism and to select those who shall be corrected.

The project involves a preparatory phase and an operational phase. On one hand, the preparatory phase an environment was set in place and an initial inventory of the SDC methods offered through both interfaces was constructed. On the other hand, the operational phase involves actual user support, maintenance and development.

1.2.2 Preparatory Phase : environment set up and inventory of functionalities

Several functionalities were set up in this preparatory phase :

1. Code repositories of the underlying methods
2. Code repositories of the UI
3. A formal site to supply the official releases of the software
4. A formal site for the Help Desk Portal for user support

These repositories allow for a structured access control and make it easier to coordinate development of the software. The Help Desk Portal reduces the number of requests directly sent to individual developers and makes it possible for all users to see the problems encountered by others and their solutions or workarounds.

1.2.3 Operational Phase : user support, maintenance and development

In the operation phase, user support, maintenance and development take place relying on the environment set in place before.

User support involves collecting questions and suggestions about the software, and answering to these. It may involve different user interfaces or underlying methods. It will lead to a group of frequently asked questions (FAQ). User support may also lead to new requirements of the software (UI and/or underlying methods).

A short inventory of the SDC software and their underlying methods was first drafted by the partners and it was updated with a questionnaire sent to all Member States where they were invited to tick the functionalities they use often, rarely, or they would like to be implemented.

Maintenance is adjusting UI by correcting small bug fixes or doing minor adjustments, and extracting possible improvements or new functionalities from user support. While development will be adding functionality and methods, or correcting major bug fixes.

1.3 Harmonised protection of Census data

Six partners of the FPA are involved in this third SGA: Statistics Netherlands, Statistics Finland, Insee, Destatis, HCSO and SURS. It was launched in September 2016 and lasts one year.

1.3.1 Census 2021 adds grid data to the already complex SDC case of multi-dimensional hypercubes

Population and Housing Census data is an essential source of statistical information. For the 2011 Census European legislation defined a very detailed large set of multi-dimensional hypercubes. One of the biggest challenges was disclosure control: identification and protection of the confidential cells. Different methods were applied by the Member States and it sometimes led to difficulties for comparisons.

Moreover, the 2021 Census will provide data by grid (1 km²) squares, which alongside the regional breakdowns will lead to two parallel non-nested geographical classifications. Even though the level of detail in the hypercubes will not be as high as the 2011 Census and the number of counts in each grid square will be very limited, it requires a thorough evaluation of the SDC methods applicable to prevent disclosure.

1.3.2 Inventory of the country specific data protection regulations and methods via a questionnaire

Around two third of the Member States consider some variables sensitive, and some countries consider all census variables sensitive while a handful consider none of the census variables as sensitive. The variables that are most often considered sensitive are country of birth and country of citizenship.

The vast majority of the Member States applied SDC methods to the Census 2011 hypercubes. Cell suppression was the most popular method used. About half of the countries evaluated the disclosure risk and information loss.

Many Member States are interested in testing a tool developed by others. Cell suppression with Tau-Argus is the most cited method for 2021 Census. Due to lack of missing a tool to test methods most other methods will only be used by a few countries or none at all. More than half the countries may change their planned SDC methods for Census 2021 if alternative recommendations are given.

1.3.3 Identification of best practices

The project team considered the most important aspects to be that the harmonised SDC method should:

1. keep the structure of hypercubes
2. be able to take care of attribute disclosure risk in a sensible way.
3. keep the information loss minimal, also for detailed hypercube data.
4. be applicable by many Member States, perhaps with slightly different parameters.

Perturbative methods thus seemed superior. Global recoding was excluded due to the fixed hypercubes design. Implementing cell suppression in a consistent way across countries is very difficult due to the major differences between the disclosure risk concepts and the rules used. Another problem arises with the management of differencing risk between hypercubes and grid level data.

A harmonised method should be flexible to meet specific needs and adaptable simply by changing parameters. The project team decided to use a combination of pre-tabular perturbation and post-tabular perturbation. The pre-tabular method selected for the test is targeted record swapping and the post-tabular is random noise. Both include parameters that are not fixed and each countries will be able to decide on them. As these do not lead to suppressed data, data can be combined into European-level data.

1.3.4 Testing and recommendations

The English NSI (ONS) provided the targeted record swapping SAS codes they used in their 2011 Census (see Spicer and Tudor (2009) or Frend et al. (2011) for details on the methodology). The random noise method is called cell-key method and was first adapted by ONS for their 2011 Census based on work from the Australian Bureau of Statistics (ABS) (see Fraser and Wooton (2005)). The partners adapted the codes and tested both pre-tabular and post-tabular method and conducted a utility/risk analysis with specific information loss measures. Then the codes were made available to other

Member States and they are invited to test the methods on their countries' specific data as well.

Recommendations will be developed according to the results of the utility/risk analysis, both for census hypercubes and for grid data taking into account other parallel geographical classifications.

2 User support for and maintenance of SDC tools

As briefly discussed in the paragraph 1.2, a two-year project started in April 2016 aiming at supporting, testing, maintaining and developing SDC software of interest to the NSIs in the ESS.

Fig 1 SDC Tools logo

2.1 Questionnaires and inventories of supported features

As stated in 1.2.1 there is a distinction to be made between the UI and the underlying methods. On the one hand, some underlying methods yields different results in each UI and this should be corrected as much as possible. On the other hand, the UIs provide different services for users. This enables users of SDC software to interfaces the underlying methods with different environments. The R-packages offer to integrate disclosure control in R and this is a strong feature for users that are familiar with R. The ARGUS software is already in use by several NSIs and has been integrated in the statistical processes. These different types of users justify the existence of different UIs.

The partners of the SGA made a first inventory of functionalities. Then two questionnaires were sent to identify the functionalities to be supported in the SGA. One focused on the software for micro data, the other for tabulated data. They were sent to members of the Working Group on Methodology. The results were collected and analysed in the last quarter of 2016. Important core functionalities were defined with the questionnaires and sorted in three groups :

1. A-functionality: each UI should provide it and the results should be identical with each UI
2. B-functionality: can be supported by one UI (rare)
3. C-functionality: less relevant, not to be supported but may remain in the UIs

There were some cases of functionalities difficult to put in one of these three groups based on the results of the questionnaire. The Expert Group on SDC in December 2016 sorted these and acted that the A-functionalities will be added into SDC tools and supported in the SGA. The development work started in 2017 and is planned for 2017 and 2018.

2.2 User group

The user group is informal: there are no official representatives of the different Member States. Anybody using any of the SDC tools supported in the project can join by registering on the JoinUp webpage of the sdcTools project: https://joinup.ec.europa.eu/software/sdctools/home where members can also join a mailing list. This mailing list will be used to provide information on workshops, meetings, new releases, etc.

The main goals of the user group are to:

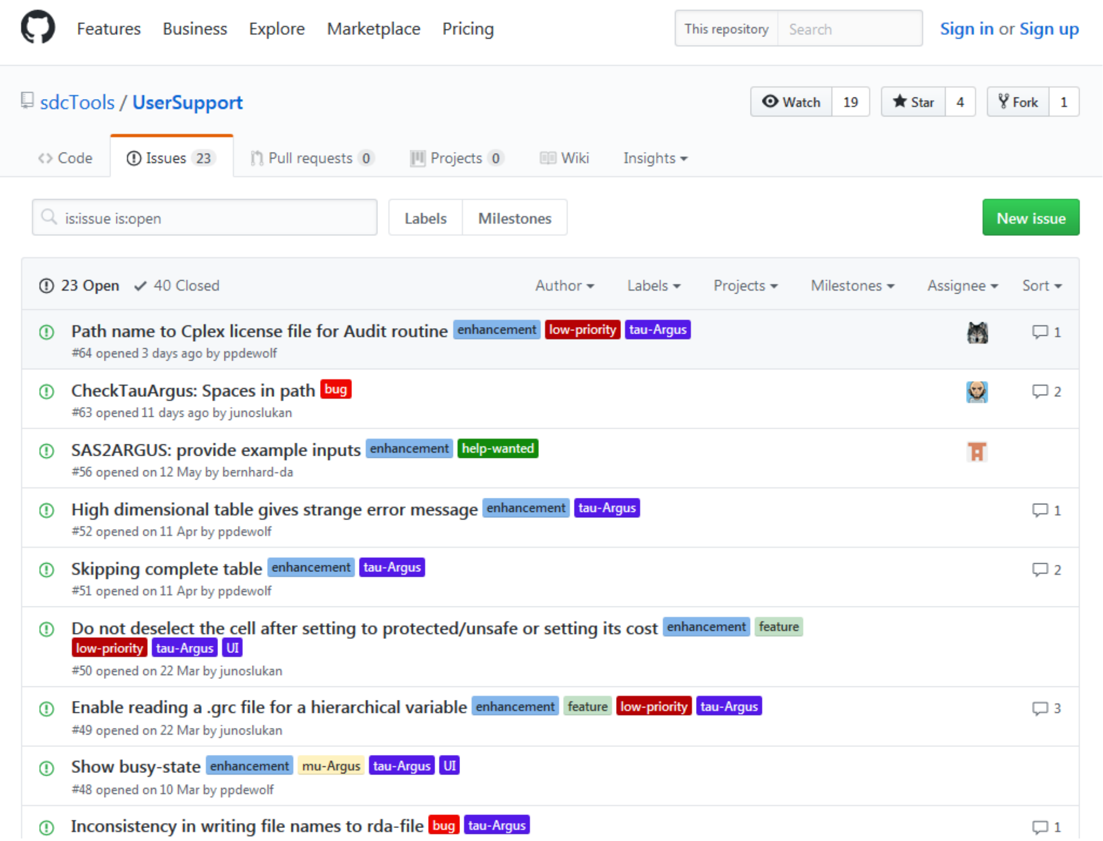
1. test the software. This is done indirectly by using the SDC tools in daily practice.
2. report bugs. Users may encounter strange or unexpected behaviour and are encouraged to report those at the Help desk portal so that developers can assess the issue.
3. exchange information and good practices between users. Making use of the Help desk portal, users can interact and provide answers to issues raised when they already found a work around.
4. suggest improvements and desired new functionalities. A real production environment may yield different ideas of improvements than the developers environment. Users can suggest improvements and new features.

A first meeting of the user group was held in Luxembourg on 2 December 2016. Approximately 20 participants attended. The developers presented the inventory of supported features.

2.3 Help desk portal

The Help Desk Portal is a key aspect of the environment of the user group. It has been set up on Github at : https://github.com/sdcTools/UserSupport/issues . Through this website, users can report bugs or anything strange they encounter while using software. Users can also suggest improvements and new features. Anybody can read this website: it is only necessary to be registered on Github to post an issue.

Even though the manuals for Argus (Hundepool et al. (2014), de Wolf et al. (2014)) and R-packages (Templ et al. (2017), Meindl (2017)) are very detailed, it is useful for users to have this way of communicating and exchanging solutions for problems they encounter.



Features Business Explore Marketplace Pricing
This repository Search
Sign in or Sign up
Watch 19 Star 4 Fork 1
Code Issues 23 Pull requests 0 Projects 0 Wiki Insights
Labels Milestones
New issue
Author Labels Projects Milestones Assignee Sort

sdcTools / UserSupport

is:issue is:open

23 Open ✓ 40 Closed

Path name to Cplex license file for Audit routine enhancement low-priority tau-Argus
#64 opened 3 days ago by ppdewolf — 1

CheckTauArgus: Spaces in path bug
#63 opened 11 days ago by junoslukan — 2

SAS2ARGUS: provide example inputs enhancement help-wanted
#56 opened on 12 May by bernhard-da

High dimensional table gives strange error message enhancement tau-Argus
#52 opened on 11 Apr by ppdewolf — 1

Skipping complete table enhancement tau-Argus
#51 opened on 11 Apr by ppdewolf — 2

Do not deselect the cell after setting to protected/unsafe or setting its cost enhancement feature low-priority tau-Argus UI
#50 opened on 22 Mar by junoslukan

Enable reading a .grc file for a hierarchical variable enhancement feature low-priority tau-Argus
#49 opened on 22 Mar by junoslukan — 3

Show busy-state enhancement mu-Argus tau-Argus UI
#48 opened on 10 Mar by ppdewolf

Inconsistency in writing file names to rda-file bug tau-Argus — 1

Fig 2 User support website

When reporting a bug, users can use coloured labels, for example “blocks-production” if a bug is found that prevents the NSI to go on in its production process. This label will alert the developers that a reply is needed quickly. They will assess the severity of the bug and when a solution will be available.

2.4 Development

Official releases of the software are made available on the Eurostat's JoinUp website : https://joinup.ec.europa.eu/software/sdctools/home. The code repositories for the different UIs and the underlying methods were put on Github at the following link : https://github.com/sdcTools. The major part of the actual development is done by CBS (Argus) and Statistics Austria (R-packages).

Developers set up unit testing with a collection of automated tests. These tests were defined to check the software at each release and guarantee stable versions. It leads to fewer bugs, better and robust code structure, easier restarts and increased confidence for the developers when making changes in the software.

Official releases are intended to be as backward compatible as possible. If this is not the case it will be recognisable in the change of the higher level of the version number and will be clearly stated in the release notes. Official releases are considered to be adequate for production environments.

Conclusion

The FPA is a flexible organisation structure. The CoE on SDC is the first CoE created with this structure. It makes it possible to start without any fixed work agenda and to make use of SGAs to define the work when necessary. The FPA is organized with a stable group of partners. It changes only slightly between each SGAs. The FPA only tackles SGAs if the partners believe a solution can be developed. A drawback of this organization structure is the difficulty of adding new partner on a specific project.

Three SGAs were launched in the FPA.

The SGA on the public use files for European Statistical System (ESS) micro data leads to the publication of PUFs for EU-LFS and EU-SILC and Eurostat is working to extend the work to provide files for all member states.

The SGA on user support for and maintenance of SDC tools is in its operational phase: the environment for user support and development is set up on Github and the community of users can interact easily, developers have started developing the functionalities identified and it lead to new releases of the software. This SGA lasts until April 2018 while the other were one-year project.

The SGA on harmonised protection of Census data is being finalized : after testing promising pre-tabular and post-tabular methods and conducting a risk/utility analysis of their effects on census data, recommendations to help census experts tackle the issue of disclosure control in the Census 2021 dissemination are being written both for hypercubes and grid data.

References

de Wolf, P.P. (2015) *Public Use Files of EU-SILC and EU-LFS data*, Joint UNECE/Eurostat Work session on statistical data confidentiality.

de Wolf, P.P., Hundepool, A., Giessing, S., Salazar, J., Castro, J. (2014) *User's Manual for τ-ARGUS, version 4.1*, http://neon.vb.cbs.nl/casc/Software/TauManualV4.1.pdf

Fraser, B., Wooton, J. (2005) *A proposed method for confidentialising tabular output to protect against differencing* Joint UNECE/Eurostat Work session on statistical data confidentiality.

Frend, J., Abrahams, C., Forbes, A., Groom, P., Spicer, K., Tudor, C. Youens, P.(2012). *Statistical Disclosure Control in the 2011 UK Census: Swapping Certainty for Safety*, ESSnet workshop on SDC.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P.P. (1998) *Post Randomisation for Statistical Disclosure Control: Theory and Implementation,* Journal of Official Statistics

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G., de Wolf, P.P. (2012). *Handbook on Statistical disclosure control*, ESSnet.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., de Wolf, P.P. (2012). *Statistical disclosure control*, Wiley Series in Survey Methodology.

Hundepool, A., de Wolf, P.P., van de Wetering, A., Bakker, J., Reedijk, A., Franconi, L., Polettini, S., Capobianchi, A., Domingo-Ferrer, J. (2014) *User's Manual for μ-ARGUS, version 5.1*, http://neon.vb.cbs.nl/casc/Software/MUmanual5.1.pdf

Meindl, B. , *User's Manual for sdcTable*. May 18, 2017. https://cran.r-project.org/web/packages/sdcTable/sdcTable.pdf

Smukavec, A., de Wolf, P.P. (2017). *Centre of Excellence on Statistical Disclosure Control*. NTTS conference.

Spicer, K., Tudor, C. (2009). *Balancing Risk and Utility – Statistical Disclosure Control for the 2011 UK Census*, Joint UNECE/Eurostat Work session on statistical data confidentiality.

Templ, M., Kowarik, A., Meindl, B., *User's Manual for sdcMicro*. May 25, 2017. https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf