

SOFTWARE TOOLS FOR STATISTICAL DISCLOSURE CONTROL

Herramientas software para el control de la
confidencialidad y del output

Eugenia Koblents
División de Central de Balances
Banco de España

**SEMINARIO SOBRE APLICACIONES Y DESARROLLO DE
BIG DATA Y DATA SCIENCE EN LA BANCA CENTRAL**

3 de junio de 2021

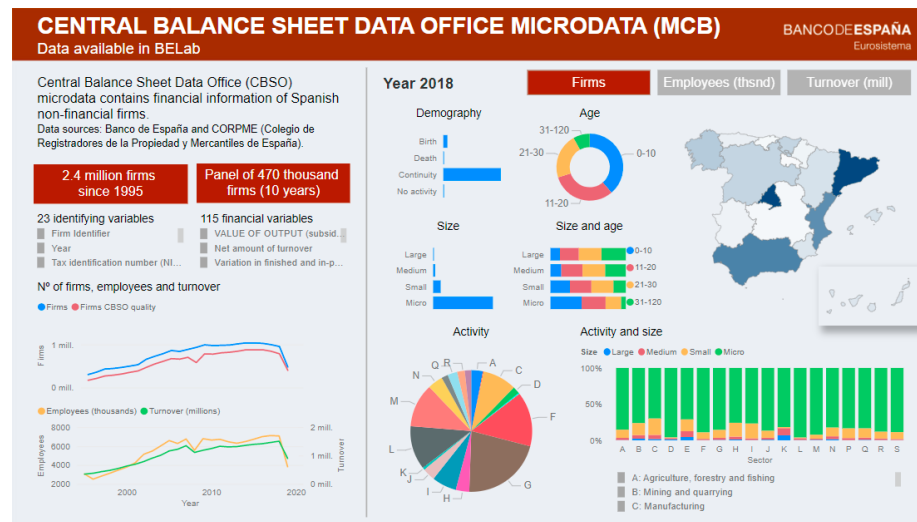
ESTADÍSTICA



ÍNDICE

1. Data laboratory BELab
2. Introduction to SDC
3. SDC workflow for microdata protection
4. SDC tools for data protection:
 - Protecting tabular data with tau-argus
 - Protecting microdata with sdcMicro
5. Resources
6. Summary and conclusions

- ❑ <https://www.bde.es/bde/es/areas/analisis-economi/otros/que-es-belab/>
- ❑ Banco de España launched BELab in **July 2019** to provide access to the research community to high quality microdata, as part of its **strategic plan** (December 2019)
- ❑ **On-site and remote access**
- ❑ **Available datasets:**
 - Non-financial enterprises and corporate groups
 - Debt securities issuers
 - Households surveys
 - The German Federal Employment Agency
- ❑ Interactive **dashboards** for the exploration of available datasets
- ❑ Exploring **anonymization and output control tools** for future use



- ❑ Due to **national laws on privacy**, micro-data cannot be distributed to the public or to researchers whenever re-identification of persons or establishments is possible.
- ❑ The goal of **anonymizing** micro-data and tabular data is to prevent confidential information from being assigned to a specific respondent.
- ❑ **Disclosure**, also known as “**re-identification**”, occurs when an intruder uses some released data to reveal previously unknown information about an individual.
- ❑ **Types of disclosure**: identity disclosure, attribute disclosure, inferential disclosure.
- ❑ Confidentiality can be achieved by applying **statistical disclosure control (SDC)** methods to the data in order to **decrease the disclosure risk** [1].
- ❑ **Software** packages are fundamental for the anonymization of data sets.

[1] Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. Journal of Statistical Software, 67(1), 1-36.

[2] Templ, M., Meindl, B., & Kowarik, A. (2013). Introduction to statistical disclosure control (SDC).

Unsafe data that DOES allow re-identification of individual units

SDC tools

Safe data that DOES NOT allow re-identification of individual units



Unsafe microdata

mu-argus or
sdcMicro



Safe microdata for research or publication

Unsafe tables

tau-argus or
sdcTable

Output control

Safe tables for publication

Unsafe data that DOES allow re-identification of individual units

Company	Activity	Location	Turnover
Telefonica	Telecom	Madrid	1 mill
Taller Pérez	Motor	Patones	100

Unsafe microdata

↓
Frequency table

	Madrid	Patones
Telecom	300	2
Motor	500	1

Average turnover (magnitude table)

	Madrid	Patones
Telecom	1.1 mill	50
Motor	3 mill	100

Unsafe tables

SDC tools

mu-argus or
sdcMicro

Safe data that DOES NOT allow re-identification of individual units

Company	Activity	Location	Turnover
?	Telecom	Madrid	>500 mil
?	Motor	Patones	-

Safe microdata for
research or publication

↓
Output control

Average turnover (magnitude table)

	Madrid	Patones
Telecom	>1 mill, < 2mill	<100
Motor	> 2 mill	<500

Safe tables for publication

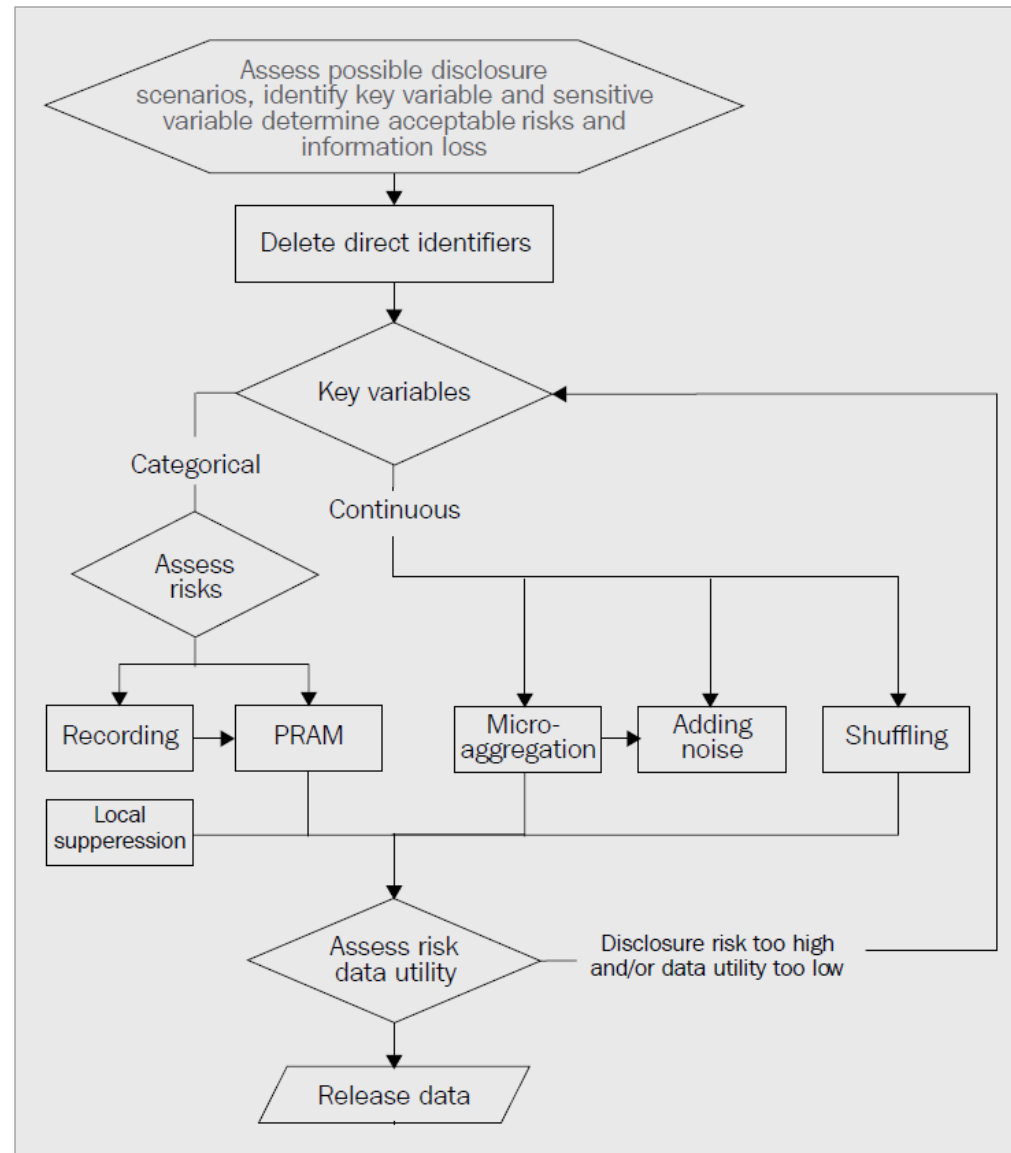
tau-argus or
sdcTable

- ❑ **Identifying variables**, those whose values might lead to re-identification, must be determined:
 - **Direct identifiers** precisely identify statistical units (company name, CIF, address, etc).
 - **Key variables** (categorical or continuous), when considered together, can be used to identify individual units (region, activity, net turnover, total assets, total employment, etc).
 - **Sensitive variables** must not be discovered for any individual unit (insolvency status, etc).
- ❑ Determining **key variables** is a challenge and involves discussions with domain experts and interpretation of national laws.

Select variables ⓘ

Variable name	Key variables
Identificador.de.Empresa	<input type="radio"/> Cat. <input type="radio"/> Cont.
Ejercicio	<input type="radio"/> Cat. <input type="radio"/> Cont.
NIF..periodico.	<input type="radio"/> Cat. <input type="radio"/> Cont.
Nombre.Empresa	<input type="radio"/> Cat. <input type="radio"/> Cont.
Ano.de.Constitucion	<input type="radio"/> Cat. <input type="radio"/> Cont.
Sectorizada	<input type="radio"/> Cat. <input type="radio"/> Cont.
CNAE09..periodico.	<input type="radio"/> Cat. <input type="radio"/> Cont.
Divisiones.CNAE2009..periodico.	<input type="radio"/> Cat. <input type="radio"/> Cont.
Secciones.CNAE2009..periodico.	<input type="radio"/> Cat. <input type="radio"/> Cont.
Tipo.de.Cuestionario	<input type="radio"/> Cat. <input type="radio"/> Cont.

1. **Deletion of direct identifiers**, to guarantee primary confidentiality
 2. **Key and sensitive variables identification**, to address secondary confidentiality
 3. **Individual disclosure risks measurement** based on sample frequency counts (k-anonymity, l-diversity, etc).
 4. **Application of SDC-methods** to modify high-risk observations.
 5. **Disclosure risk and information loss** are recomputed comparing original and modified data.
- ❑ The goal is to **release a safe data set** with low (individual) risks and high data utility.



- ❑ **Anonymization tool** developed by the IT Dept for BELab to guarantee **primary confidentiality**:
 - Replaces **direct identifiers** by anonymous unique identifiers (sha256, sha512 hashing algorithm)
 - Repeatable but irreversible process

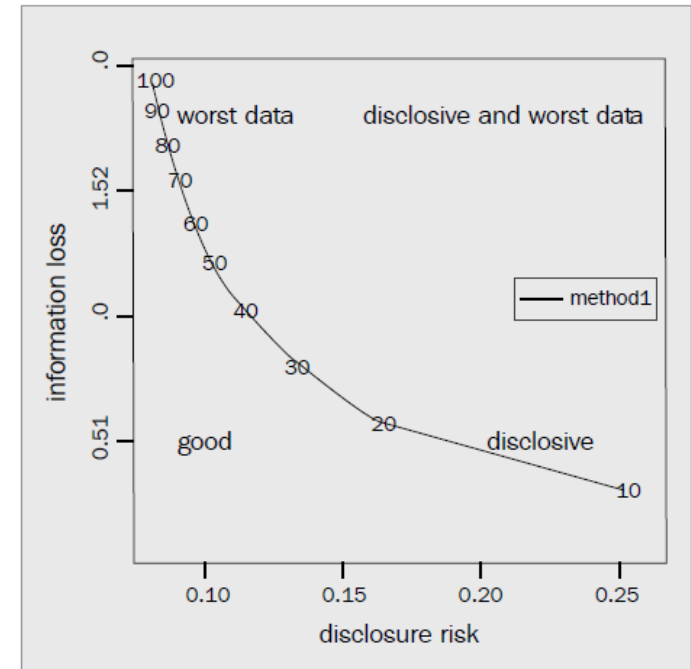
❑ **Steps:**

- Select input file
- Select configuration parameters (seed)
- Select identifying variables
- Run anonymization and save data

- ❑ **Future functionalities:** allow for string substitutions, etc

The screenshot displays the 'LAB Anonimization tools' interface. At the top, there is a blue header with the title 'LAB Anonimization tools'. Below this, the 'Data' section contains three input fields: 'Input File' (C:\TRABAJO\DATOS\MCB\MCB_202011_1995_2019\MCB_2017_Nov2020_esp.excel.csv), 'Config File' (empty), and 'OutputFile' (C:\TRABAJO\DATOS\MCB\MCB_202011_1995_2019\MCB_2017_Nov2020_esp.excel.Anonimized.csv). Below these fields are control buttons: 'Run' (green play icon), 'Stop' (black square icon), 'Clear' (red X icon), and 'Field by Field' (green play icon). The 'Anonimization' section has three tabs: 'Data Encryption', 'Nulling Out', and 'Substitution'. The 'Data Encryption' tab is active, showing 'Encryption Algorithm' set to 'sha256'. A 'Salt' field with a 'Generate Salt' link is also present. Two lists of fields are shown: 'All Fields' (Ejercicio, Año de Constitución, Sectorizada, CNAE09 (periódico), Divisiones CNAE2009 (periódico), Secciones CNAE2009 (periódico), Tipo de Cuestionario, Tamaño Recomendación Europea, Tamaño Estadístico, Indicador de propiedad, Cotización bolsa) and 'Selected Fields' (Identificador de Empresa, NIF (periódico), Nombre Empresa).

- ❑ A **trade-off** between information loss and disclosure risk must be achieved, based on the use case requirements.
- ❑ Very **sensitive data** requires more **aggressive anonymization** to guarantee low disclosure risk.
- ❑ The **access mode** (on-site vs remote access) also determines the degree of anonymization.
- ❑ The complexity of **output control** depends on the anonymization used and the affordable risk.
- ❑ Multiple **SDC methods** for microdata and tabular data protection are available:



	Deterministic SDC methods	Probabilistic SDC methods
Categorical key variables	Recoding Local suppression	Swapping PRAM
Continuous key variables	Micro-aggregation	Adding correlated noise Shuffling

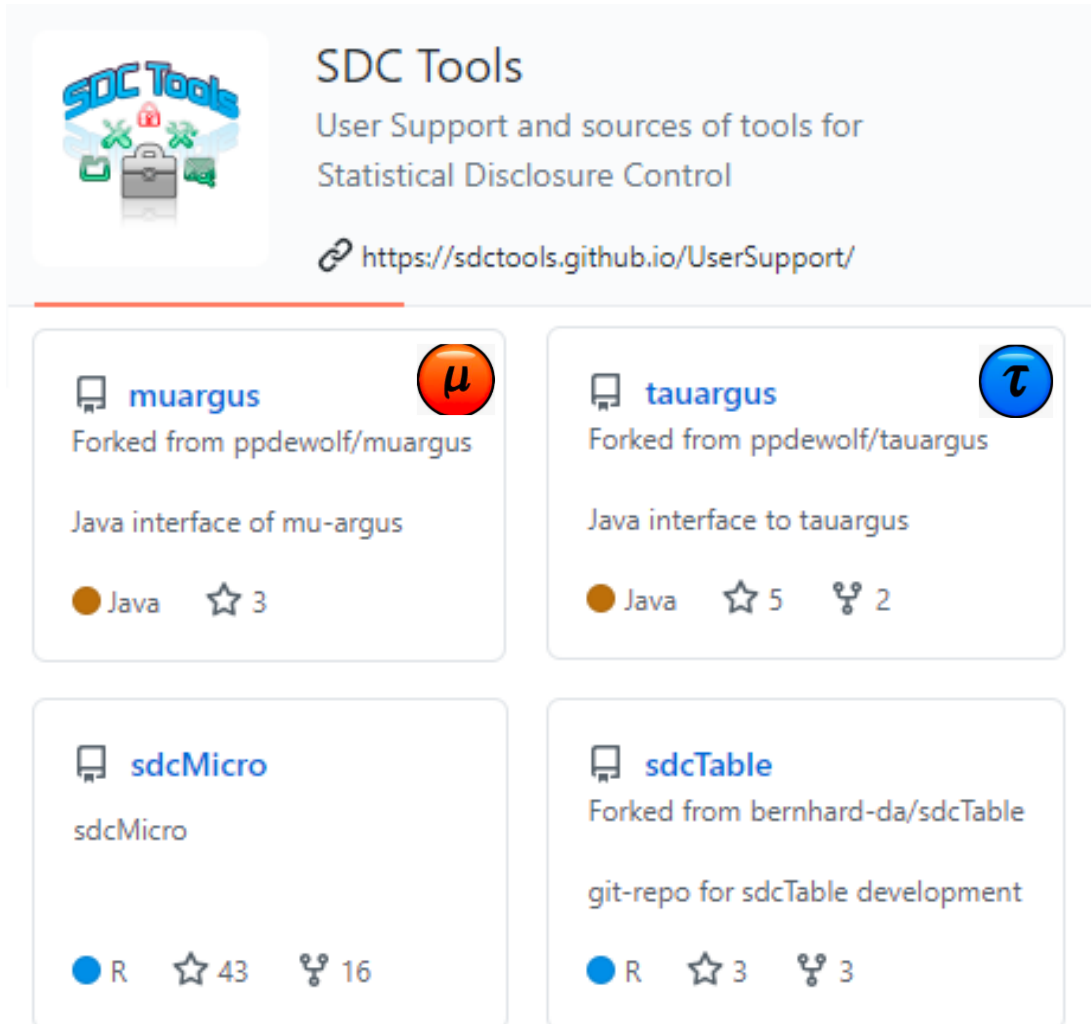
- SDC software is used by **National Statistical Institutes, Eurostat, National Banks,** and other **public bodies**.

- Eurostat** launched a Specific Grant for the user support and maintenance of SDC tools.

- Git repository:**

<https://github.com/sdcTools>

- Tools for **microdata protection**: mu-argus, sdcMicro
- Tools for **tabular data protection**: tau-argus, sdcTable



SDC Tools
User Support and sources of tools for Statistical Disclosure Control
<https://sdctools.github.io/UserSupport/>

muargus (Java) ★ 3
Forked from ppdewolf/muargus
Java interface of mu-argus

tauargus (Java) ★ 5 🍴 2
Forked from ppdewolf/tauargus
Java interface to tauargus

sdcMicro (R) ★ 43 🍴 16
sdcMicro

sdcTable (R) ★ 3 🍴 3
Forked from bernhard-da/sdcTable
git-repo for sdcTable development

□ Tabular data protection: tau-argus vs sdcTable:

- Tau-argus has a **GUI**, sdcTable is command line and requires programming
- We will use **tau-argus** in BELab

□ Microdata protection: mu-argus vs sdcMicro:

- Both libraries have a **GUI**, no programming required
- Mu-argus is similar to tau-argus, learning can be easier
- sdcMicro incorporates more algorithms
- sdcMicro claims to be better optimized for **large datasets**.
- We will probably use **sdcMicro** in BELab

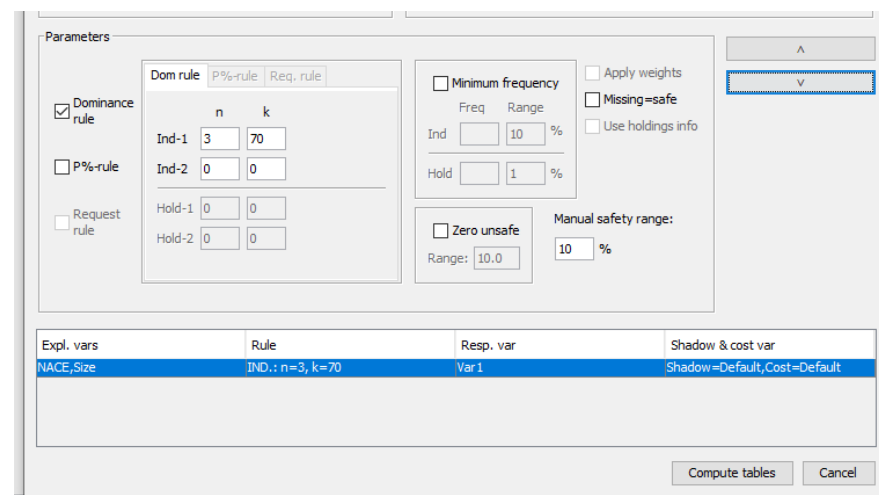
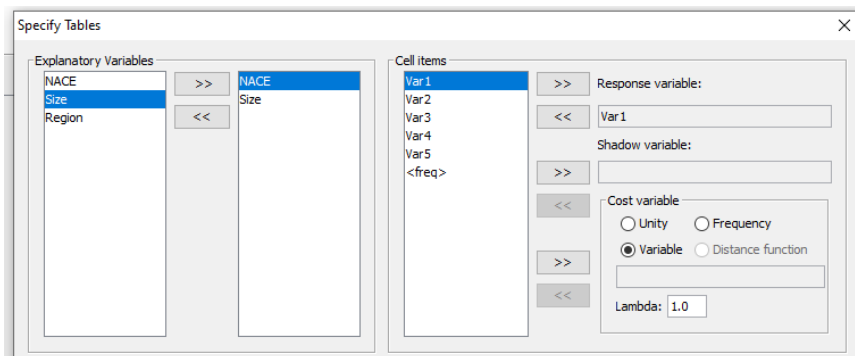
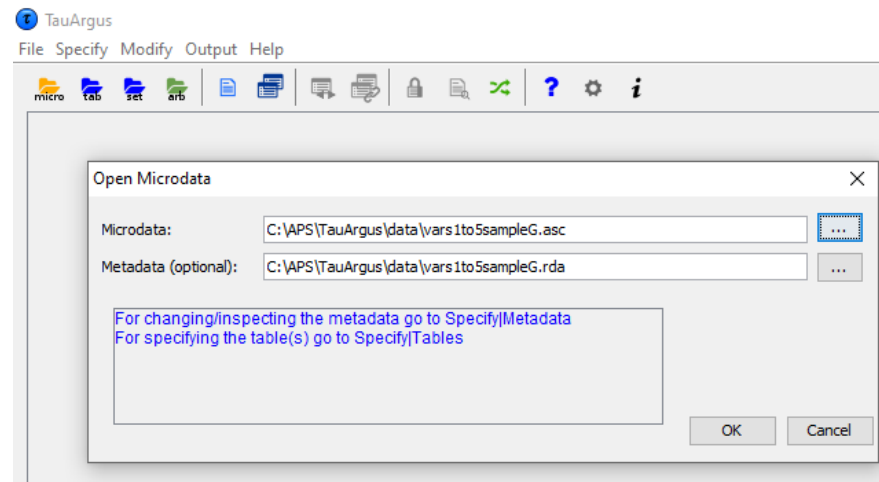
□ Tau-argus and mu-argus are implemented in **Java**, sdcTable and sdcMicro in **R**.

Method	Software	μ -Argus	sdcMicroGUI
		4.2	> 1.1.0
Frequency counts		✓	✓
Individual risk		✓	✓
Individual risk on households		✓	✓
<i>l</i> -diversity			✓
SUDA2			
Global risk		✓	✓
Global risk with log-lin mod.			
Recoding		✓	✓
Local suppression		(✓)	✓
Swapping		(✓)	
PRAM		✓	✓
Adding correlated noise			✓
Micro-aggregation		✓	✓
Shuffling			✓
Utility measures		(✓)	✓
GUI		(✓)	✓
CLI			
Missing values		✓	✓
Cluster designs		✓	✓
Large data			✓
Reporting		✓	✓
Platform independent			✓
Free and open-source			✓

Protecting tabular data with tau-argus

Example with a sample dataset

1. Open microdata (or table)
2. Select variables and **specify tables**
3. Set **anonymization parameters** (dominance rule, P% rule, weights, etc)



4. Tau-argus identifies table cells with high risk of re-identification.

The screenshot shows the TauArgus application interface. The main window displays a table with the following data:

Region										
Var 1: NACE x Size Total										
	- Total	01	02	03	04	05	06	07	08	09
- Total	174134755	1362797	6079566	15366820	53879161	1642562	14137802	8625576	1839742	8916845
+ G45	50048310	262163	1479126	5929448	19552089	382905	3499301	1696402	780665	789110
+ G46	99564359	870771	3476167	8405011	25832846	1071976	8218884	6619388	545517	7786206
+ G47	24522086	229863	1124273	1032361	8494226	187681	2419617	309786	513560	341529

On the right, the 'Cell Information' panel shows the following details for a selected cell:

- Value: 187681
- Status: Unsafe
- Shadow: 187681
- Cost: 187681
- #contributions: 66
- Top n of shadow: 131182, 13531, 5773

5. Select and run SDC algorithm (primary and secondary suppression, recoding, etc)

The screenshot shows the TauArgus application interface with the 'Recode' dialog box open. The dialog box has a 'Recode' title bar and a 'Suppress' section with the following options:

- Hypercube
- Modular
- Optimal
- Network
- CTA
- Rounding

Buttons for 'Suppress', 'Undo', and 'Audit' are visible. The main window displays the same data table as in the previous screenshot, but with some cells highlighted in blue:

Region										
Var 1: NACE x Size Total										
	- Total	01	02	03	04	05	06	07	08	09
- Total	174134755	1362797	6079566	15366820	53879161	1642562	14137802	8625576	1839742	8916845
+ G45	50048310	262163	1479126	5929448	19552089	382905	3499301	1696402	780665	789110
+ G46	99564359	870771	3476167	8405011	25832846	1071976	8218884	6619388	545517	7786206
+ G47	24522086	229863	1124273	1032361	8494226	187681	2419617	309786	513560	341529

6. Generate an anonymization **report** summarizing the process and results.

T-ARGUS Report

Fri May 28 13:46:24 CEST 2021

Original file:	C:\APS\TauArgus\data\vars1to5sampleG.asc
Meta file:	C:\APS\TauArgus\data\vars1to5sampleG.rda
Table file:	C:\APS\TauArgus\data\cosa.txt

Table generated from microdata

Table structure

Type	Var	# codes
Response var:	Var1	
Explanatory var1:	NACE	248
Explanatory var2:	Size	17
Explanatory var3:	Region	12

Sensitivity Rule:

Dominance rule (Individual level) with n = 3 and k = 70%
Manual safety margin: 10%
Missing codes have been considered unsafe

Modular (HITAS) Salazar solution

Solver used: SCIP

ItbTauHITaS version is 4.2.4.1

Using SCIP
SCIP version is 3.110000
using SoPlex 2.0.1

Max time per subtable: 1 minutes

Additional Singleton/Singleton option has been used
Additional Singleton/Multiple option has been used
Additional Min. Frequency option has been used

Time used to protect the table: 10 min 48 sec

Summary of the table

	Status	Number of cells	Number of respondents	Response value	Cost value
1	Safe	5152	1472848	2988357328	2988357328
2	Safe (manual)	0	0	0	0
3	Unsafe	17061	152983	1252095455	1252095455
4	Unsafe (request)	0	0	0	0

- Key variables (categorical and continuous) are manually identified by the domain expert. High risk samples are identified and SDC methods applied to minimize risk.

sdcmicro GUI About/Help **Microdata** Anonymize Risk/Utility Export Data Reproducibility Undo

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
Ano.de.Constitucion	109 (109)	93.615 (93.615)	1 (1)
CNAE09..periodico.	541 (541)	18.861 (18.861)	1 (1)
Cotiza.en.bolsa	7 (7)	1457.714 (1457.714)	3 (3)
Codigo.Postal	2432 (2432)	4.196 (4.196)	1 (1)

Risk measures for categorical key variables

We expect 10079.00 (98.77%) re-identifications in the population, as compared to 10079.00 (98.77%) re-identifications in the original data.
0 observations have a higher risk than the risk in the main part of the data, as compared to 0 observations in the original data. ⓘ

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	9970 (97.707%)	9970 (97.707%)
3-anonymity	10162 (99.588%)	10162 (99.588%)

Disclosure risk assessment



Anonymization methods to reduce risk

sdcmicro GUI

About/Help Microdata Anonymize Risk/Utility Export Data Reproducibility Undo

View/Analyze existing sdcProblem

Show summary

Explore variables

Add linked variables

Create new IDs

Anonymize categorical variables

Recoding

k-Anonymity

PRAM (simple)

PRAM (expert)

Suppress values with high risks

Anonymize numerical variables

Top/bottom coding

Reset SDC problem

Recode categorical key variables

To reduce risk, it is often useful to combine the levels of categorical key variables into a new, combined category. You need to select a categorical key variable and then choose two or more levels, which you want to combine. Once this has been done, a new label for the new category can be assigned.

Note: If you only select only one level, you can rename the selected value.

Choose factor variable
Ano de Constitución

Select levels to recode/combine

Variable selection

Variable name	Type	Additional suppressions by local suppression algorithm
Ano.de.Constitucion	cat. key variable	0
CNAE09...periodico.	cat. key variable	0
Cotiza.en.bo1sa	cat. key variable	0
Codigo.Postal	cat. key variable	0

Additional parameters

Parameter	Value
number of records	10204
alpha	1
random seed	0

k-anonymity

k-anonymity	Modified data	Original data

Information loss and data utility assessment

sdcmicro GUI

About/Help Microdata Anonymize Risk/Utility Export Data Reproducibility Undo

Display information loss based on recodings of categorical key variables

For each categorical key variable, the following key figures are computed:

- The number of categories in original and modified variables.
- The mean size of groups in original and modified variables.
- The size of the smallest category/group in original and modified variables.

Show 10 entries

keyVar	nrCategories.orig	nrCategories.mod	mean.size.orig	mean.size.mod	min.size.orig
Ano.de.Constitucion	109	109	93.615	93.615	
CNAE09...periodico.	541	541	18.861	18.861	
Cotiza.en.bo1sa	7	7	1457.714	1457.714	
Codigo.Postal	2432	2432	4.196	4.196	

Report generation

sdcmicro GUI

About/Help Microdata Anonymize Risk/Utility Export Data Reproducibility Undo

What do you want to export?

Anonymized Data

Anonymization Report

Create anonymization report

A report for internal use (more detailed) or a report for external use (less detailed) is saved to the export directory.

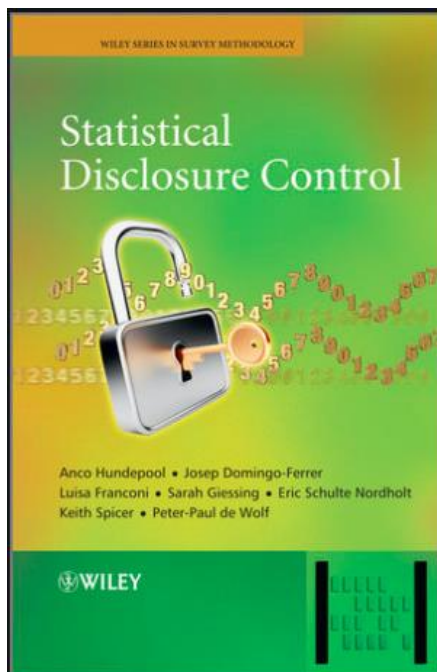
Select type of report

Internal (detailed) external (short overview)

Save report

The report was saved as C:/Users/q32058/Documents/sdcReport_interna1_20210202_1122.html

- ❑ **Git repository:** <https://github.com/sdcTools>
- ❑ **User support:** <https://sdctools.github.io/UserSupport/>
- ❑ **SDC book:** Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2012). *Statistical disclosure control*. John Wiley & Sons.
- ❑ **Eurostat courses:** https://ec.europa.eu/eurostat/cros/content/estp-training-offer_en



DATE	COURSE TITLE	VENUE	COURSE ORGANISER	APPLICATION DEADLINE
23-26 March 2021 4 days	Statistical Disclosure Control	ONLINE	EUROSTAT	25.01.2021
21-22 October 2021 2 days	Output checking in research data centres	Eurostat, Luxembourg	EUROSTAT	23.08.2021
05-10-2021 12-10-2021 19-10-2021 26-10-2021 4 sessions	Big Data tools for data scientists	ONLINE	ICON- INSTITUT Public Sector GmbH	09.08.2021

- ❑ The goal of **Statistical Disclosure Control** is to minimize disclosure risk while maximizing information utility when releasing microdata or tabular data.
- ❑ Powerful and reliable **software tools** for SDC are available, including mu-argus, tau-argus, sdcMicro and sdcTable.
- ❑ Multiple **public institutions** use them (Central Banks, Data Centers, Statistical Institutes, etc).
- ❑ The identification of **key variables** is a challenge and requires expert knowledge of the data.
- ❑ Eurostat **courses** and other learning **resources** are available.
- ❑ **Output control** is still a highly manual process. **Eurostat** is about to release a Stata tool to support output control.
- ❑ **BELab** staff has recently explored existing SDC tools and plans to use them in the near future when sensitive datasets are incorporated to the laboratory.

Thank you for your attention!

