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
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
Data laboratory BELab

- https://www.bde.es/bde/es/areas/analisis-economia/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.


Unsafe data that DOES allow re-identification of individual units

**Unsafe microdata**

**Unsafe tables**

---

**SDC tools**

- mu-argus or sdcMicro
- tau-argus or sdcTable

---

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

**Safe microdata for research or publication**

---

**Output control**

---

**Safe tables for publication**
Unsafe data that DOES allow re-identification of individual units

<table>
<thead>
<tr>
<th>Company</th>
<th>Activity</th>
<th>Location</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telefonica</td>
<td>Telecom</td>
<td>Madrid</td>
<td>1 mill</td>
</tr>
<tr>
<td>Taller Pérez</td>
<td>Motor</td>
<td>Patones</td>
<td>100</td>
</tr>
</tbody>
</table>

Unsafe microdata

Frequency table

<table>
<thead>
<tr>
<th></th>
<th>Madrid</th>
<th>Patones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telecom</td>
<td>300</td>
<td>2</td>
</tr>
<tr>
<td>Motor</td>
<td>500</td>
<td>1</td>
</tr>
</tbody>
</table>

Average turnover (magnitude table)

<table>
<thead>
<tr>
<th></th>
<th>Madrid</th>
<th>Patones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telecom</td>
<td>1.1 mill</td>
<td>50</td>
</tr>
<tr>
<td>Motor</td>
<td>3 mill</td>
<td>100</td>
</tr>
</tbody>
</table>

Unsafe tables

SDC tools

mu-argus or sdcMicro

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

Safe microdata for research or publication

<table>
<thead>
<tr>
<th>Company</th>
<th>Activity</th>
<th>Location</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>Telecom</td>
<td>Madrid</td>
<td>&gt;500 mil</td>
</tr>
<tr>
<td>?</td>
<td>Motor</td>
<td>Patones</td>
<td>-</td>
</tr>
</tbody>
</table>

Output control

Average turnover (magnitude table)

<table>
<thead>
<tr>
<th></th>
<th>Madrid</th>
<th>Patones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telecom</td>
<td>&gt;1 mill, &lt;2 mill</td>
<td>&lt;100</td>
</tr>
<tr>
<td>Motor</td>
<td>&gt; 2 mill</td>
<td>&lt;500</td>
</tr>
</tbody>
</table>

Unsafe tables

Safe tables for publication
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.
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
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:

<table>
<thead>
<tr>
<th><strong>Deterministic SDC methods</strong></th>
<th><strong>Probabilistic SDC methods</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical key variables</td>
<td>Recoding</td>
</tr>
<tr>
<td>Local suppression</td>
<td></td>
</tr>
<tr>
<td>Continuous key variables</td>
<td>Micro-aggregation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SDC tools for tabular data and microdata protection

- 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 for tabular data and microdata protection

- **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**.

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

5. Select and run **SDC algorithm** (primary and secondary suppression, recoding, etc)
6. Generate an anonymization report summarizing the process and results.

**t-ARGUS Report**

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

**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
- IPObTauHTaS version is 4.2.4.1
- Using SCIP
- SCIP version is 3.1110000
- Using SoFlex 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**

<table>
<thead>
<tr>
<th>Status</th>
<th>Number of cells</th>
<th>Number of respondents</th>
<th>Response value</th>
<th>Cost value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safe</td>
<td>5152</td>
<td>1472648</td>
<td>2988357328</td>
<td>2988357328</td>
</tr>
<tr>
<td>Safe (manual)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Unsafe</td>
<td>17061</td>
<td>152983</td>
<td>1252095455</td>
<td>1252095455</td>
</tr>
<tr>
<td>Unsafe (request)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Key variables (categorical and continuous) are manually identified by the domain expert. High risk samples are identified and SDC methods applied to minimize risk.
Protecting microdata with sdcMicro
Example with a BELab dataset (MCB)

- Disclosure risk assessment
- Anonymization methods to reduce risk
- Information loss and data utility assessment
- Report generation

Risk measures
The output on this page is based on the categorical key variables is the current problem.

What kind of results do you want to show?
- Risk measures
- Risky observations
- Plot of risk

Plot showing distribution of individual re-identification risk levels

Anonymized data

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.

Variable selection

Additional parameters

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

- **Git repository:** [https://github.com/sdcTools](https://github.com/sdcTools)
- **User support:** [https://sdctools.github.io/UserSupport/](https://sdctools.github.io/UserSupport/)
- **Eurostat courses:** [https://ec.europa.eu/eurostat/cros/content/estp-training-offer_en](https://ec.europa.eu/eurostat/cros/content/estp-training-offer_en)

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