Data Anonymization for Open Science

useR! 2024

Jiří Novák[1,2] Marko Miletić[3] Oscar Thees[2] Alžběta Beranová[4]

[1]University of Zurich [2]University of Applied Sciences Northwestern Switzerland

[3]Bern University of Applied Sciences [4]Czech Statistical Office

July 8, 2024

About Speakers

Jiří Novák

- Ph.D. in Statistics with topic Simulation of Synthetic Microdata
- Statistical confidentiality of Czech Census 2021

About Speakers

Jiří Novák

- Ph.D. in Statistics with topic Simulation of Synthetic Microdata
- Statistical confidentiality of Czech Census 2021

Marko Miletić

- XXX
- XXX

About Speakers

Jiří Novák

- Ph.D. in Statistics with topic Simulation of Synthetic Microdata
- Statistical confidentiality of Czech Census 2021

Marko Miletić

- XXX
- XXX

Oscar Thees

- XXX
- XXX

About Speakers

Jiří Novák

- Ph.D. in Statistics with topic Simulation of Synthetic Microdata
- Statistical confidentiality of Czech Census 2021

Marko Miletić

- XXX
- XXX

Oscar Thees

- XXX
- XXX

Alžběta Beranová

- XXX

Data Anonymization in which context?

This tutorial is about Data Anonymization in the context of the field of **Statistical Disclosure Control** (SDC).

SDC is also known as Statistical disclosure limitation or Disclosure avoidance.

Statistical Disclosure Control seeks to protect statistical data in such a way that they can be released without giving away confidential information that can be linked to specific individuals or entities.

Importance of Data Anonymization

There are several main reasons:

1. **Principle** It is a fundamental principle of Official Statistics that the statistical records of individual persons, businesses, or events used to produce Official Statistics are strictly confidential and to be used only for statistical purposes.
2. **Legal** Legislation imposes a legal obligation to protect individual business and personal data. Legal frameworks regulate what is allowed and what is not allowed regarding the publication of private information.
3. **Quality** Respondents need confidence in the preservation of the confidentiality of individual information. If they do not trust the confidentiality of the data, they may not provide accurate information.
4. **Ethical** Disclosing information that can be linked to specific individuals or entities is unethical.

Relevance to Open Science

Open Science, **Open Access**, **Open Data** are important trends in the scientific community.

Research data that results from publicly funded research should be **FAIR**: **findable**, **accessible**, **interoperable**, **reusable**

- therefore replicable, transparent, trustworthy
- Principle: **As open as possible, as closed as necessary**
- Enables data sharing and collaboration
- Facilitates reproducible research
- Balances transparency with privacy

Commission Recommendation (EU) 2018/790 on access to and preservation of scientific information

Outputs to protect

Different outputs require different approaches to SDC and different mixtures of tools.

- **Macrodata** (Tabular data)
- **Microdata**
- **Dynamic databases**
- **Statistical analyses**

Disclaimer: Imposing a single solution for all types of data is not possible.
This tutorial will focus on Microdata and Tabular data.

Key Concepts

Key Concepts are:

- **Disclosure**
    - A disclosure occurs when a person or an organisation recognises or learns somethingthat they did not know already about another person or organisation, via released data.

Key Concepts

Key Concepts are:

- **Disclosure**
  - A disclosure occurs when a person or an organisation recognises or learns somethingthat they did not know already about another person or organisation, via released data.
- **Re-identification risk**
  - Re-identification risk is the risk that an intruder can link a record in the released data to a specific individual in the population.

Key Concepts

Key Concepts are:

- **Disclosure**
  - A disclosure occurs when a person or an organisation recognises or learns somethingthat they did not know already about another person or organisation, via released data.
- **Re-identification risk**
  - Re-identification risk is the risk that an intruder can link a record in the released data to a specific individual in the population.
- **Data utility**
  - Data utility is the usefulness of the data for the intended purpose.

Disclosure

A disclosure occurs when a person or an organisation recognises or learns something that they did not know already about another person or organisation, via released data.

Types of disclosure risk:

(1) **Identity disclosure** Revealing the identity of an individual.

(2) **Attribute disclosure** Revealing sensitive attributes of an individual.

(3) **Inferential disclosure** Making inferences about an individual based on the released data.

Disclosure

Types of disclosure risk:

(1) **Identity disclosure**

Residency	Age	Sex	Occupation
Salzburg	50	Male	Professor

(2) **Attribute disclosure**

Group	Males	Females	Total
Football fans	22	0	22
Non Football fan	93	85	178
Total	115	85	200

Risk and utility

SDC seeks to optimise the trade-off between the disclosure risk and the utility of the protected released data.

- **Risk**: the probability of a disclosure event occurring.
- **Utility**: the usefulness of the data for the intended purpose.

The goal is to find a balance between risk and utility.

Risk-utility trade-off

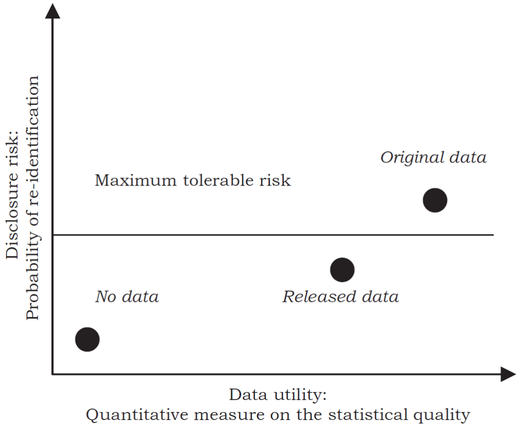

Figure 1: R-U confidentiality map (Duncan et al.,2001)

Disclosure risk

A unit is at risk of disclosure when it cannot be confused with several other units in the data set.

- **k-anonymity** A data set is said to satisfy k-anonymity for $k > 1$ if, for each combination of values of quasi-identifiers (e.g. name, address, age, gender, etc.), at least k records exist in the data set sharing that combination.
- Ensures that each record is indistinguishable from at least k-1 other records with respect to the quasi-identifiers.

More robust aproaches:

- **l-Diversity** Extends k-anonymity by ensuring that the sensitive attribute has at least l well-represented values
- **t-Closeness** Ensures that the distribution of the sensitive attribute in any equivalence class is close to the distribution in the entire dataset

Attacker Scenarios

Data Intruder: An attacker who tries to re-identify individuals using the anonymized dataset and auxiliary information. For example, consider a dataset of anonymized medical records. An intruder might use publicly available information, like voter registration lists, to link unique combinations of quasi-identifiers (such as age, gender, and zip code) to re-identify individuals.

Data Linkage: This scenario involves an attacker who combines multiple datasets to enhance the chances of re-identification. For example, if an attacker has access to an anonymized dataset from a hospital and another dataset from a social media platform, they might link these datasets through common quasi-identifiers to re-identify patients.

Background Knowledge: An attacker might use their own knowledge about certain individuals to identify them in a dataset. For example, if someone knows a particular person's age, job title, and city, they might find a matching record in an anonymized employment dataset, thereby re-identifying that individual.

Variables

1. **Identifiers** - variables that can directly identify an individual
2. **Quasi-identifiers** or **key variables** - these variables don't identify individuals on their own but can do so when combined with other quasi-identifiers
3. **Confidential outcome variables** - variables that contain sensitive information that should be protected
4. **Non-confidential outcome variables** - these are variables that are not sensitive and don't risk the privacy of individuals if disclosed

Disclosure control methods

1. **Masking original data**
    1. **Non-perturbative masking -** Methods that alter data to hide identities without changing its actual values
    2. **Perturbative masking -** Methods that add noise or alter data values to prevent identification
2. **Generating synthetic data**
    1. **Parametric methods -** Techniques that use statistical models based on the data's distribution to generate synthetic data.
    2. **Non-parametric methods** Techniques that do not assume an underlying distribution, using methods like bootstrapping to generate synthetic data.
    3. **Generative Adversarial Networks (GANs)** Advanced machine learning models that generate highly realistic synthetic data by training two neural networks in tandem.

Packages for SDC - Microdata (Unit-level data)

sdcMicro can be used to anonymize data, i.e. to create anonymized files for public and scientific use. It implements a wide range of methods for anonymizing categorical and continuous (key) variables. The package also contains a graphical user interface, which is available by calling the function sdcGUI.

simPop using linear and robust regression methods, random forests (and many more methods) to simulate synthetic data from given complex data. It is also suitable to produce synthetic data when the data have hierarchical and cluster information (such as persons in households) as well as when the data had been collected with a complex sampling design. It makes use of parallel computing internally.

synthpop using regression tree methods to simulate synthetic data from given data. It is suitable to produce synthetic data when the data have no hierarchical and cluster information (such as households) as well as when the data does not collected with a complex sampling design.

Packages for SDC - Tabular data (Aggregated data)

sdcTable can be used to provide confidential (hierarchical) tabular data. It includes the HITAS and the HYPERCUBE technique and uses linear programming packages (Rglpk and lpSolveAPI) for solving (a large amount of) linear programs.

sdcSpatial can be used to smooth or/and suppress raster cells in a map. This is useful when plotting raster-based counts on a map. sdcHierarchies provides methods to generate, modify, import and convert nested hierarchies that are often used when defining inputs for statistical disclosure control methods.

SmallCountRounding can be used to protect frequency tables by rounding necessary inner cells so that cross-classifications to be published are safe.

GaussSuppression can be used to protect tables by suppression using the Gaussian elimination secondary suppression algorithm.

Non-perturbation methods

Non-perturbative masking does not rely on distortion of the original data but on partial suppressions or reductions of detail.

Method	Continuous data	Categorical data
Sampling		X
Global recoding	X	X
Top and bottom coding	X	X
Local suppression		X

Table 1: Non-perturbative methods vs. data types

sdcMicro

- sdcMicro is an R package for statistical disclosure control.
- https://cran.r-project.org/web/packages/sdcMicro/index.html
- https://github.com/sdcTools/sdcMicro

sdcMicro

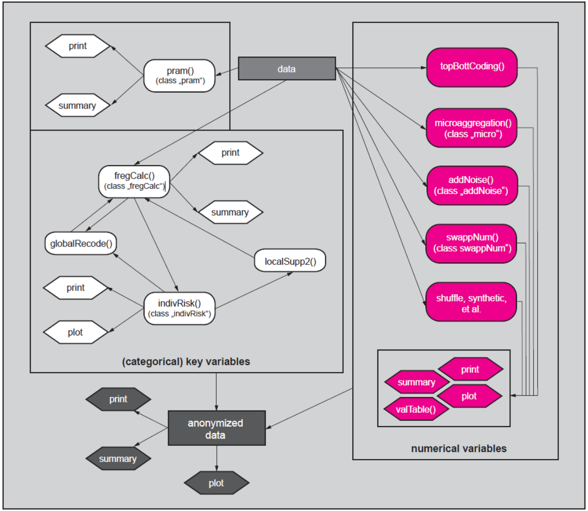

Figure 2: Certain procedures in package *sdcMicro* and their relationship

Perturbation methods

- sdcMicro

Synthetic methods

- Introduction to synthetic data
- for Jiri/Oscar

Synthetic methods: synthpop

- Generating synthetic data with synthpop
- for Jiri/Oscar

Synthetic methods: synthpop

Nowok B, Raab GM, Dibben C (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), 1-26. doi:10.18637/jss.v074.i11. URL **https://www.jstatsoft.org/article/view/v074i11**

Synthetic methods: simPop

- Generating synthetic data with simPop
- for Jiri/Oscar

Synthetic methods: simPop

Meindl B, Templ M, Alfons A, Kowarik A (2016). simPop: Simulation of Synthetic Popula- tions for Survey Data Considering Auxiliary Information. R package version 0.3.0, URL **https://CRAN.R-project.org/package=simPop**

Parametric and non-parametric methods.

Synthetic methods: GANS

- Generating synthetic data with GANs
- for Marco

Thank you for your attention

SwissAnon

Swiss Data Anonymization Competence Center

https://swissanon.ch