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

Context of Anonymization

Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL) 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.

Why is confidentiality protection important?

There are three main reasons:

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

Motivation

- Legal frameworks regulate what is allowed and what is not allowed with regard to publication of private information.
- Before sensitive statistical databases can be made available to universities for research, confidentiality must be guaranteed.
- Users of statistical outputs should be aware of the reasoning and methodology behind statistical disclosure control.

Outputs

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

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

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

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

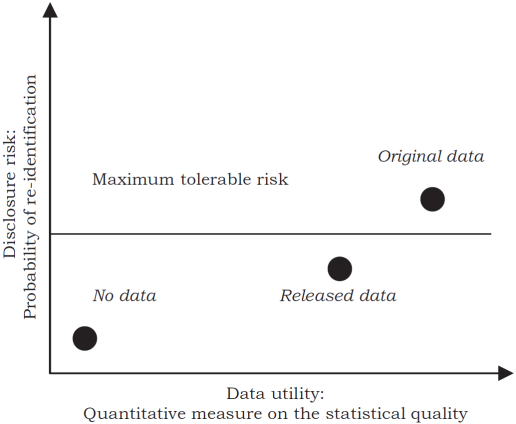

Figure 1: R-U confidentiality map (Duncan et al.,2001)

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.

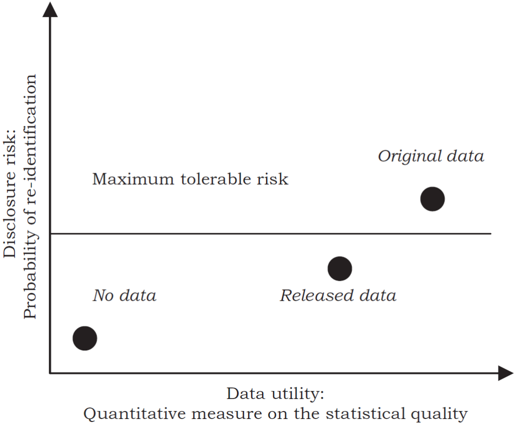

Figure 2: R-U confidentiality map (Duncan et al. 2001)

Disclosure risk

A unit is at risk of disclosure when it cannot be confused with several other units

Re-identification risk

- attacker scenarios and risk measures in more detail using examples

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.

Variables

1. Identifiers
2. Quasi-identifiers or key variables
3. Confidential outcome variables
4. Non-confidential outcome variables

Disclosure control methods

1. Masking original data
   i. Non-perturbative masking.
   ii. Perturbative masking.
2. Generating synthetic data

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.

Table 3: Non-perturbative methods vs. data types.

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

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

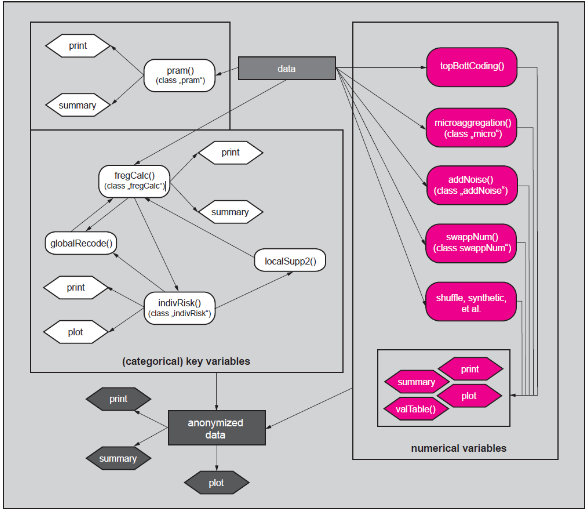

Figure 3: Certain procedures in package *sdcMicro* and their relationship

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: GANS

- Generating synthetic data with GANs
- for Marco

Thank you for your attention

SwissAnon

Swiss Data Anonymization Competence Center

https://swissanon.ch