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
- 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.

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 and
(2) attribute disclosure.

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

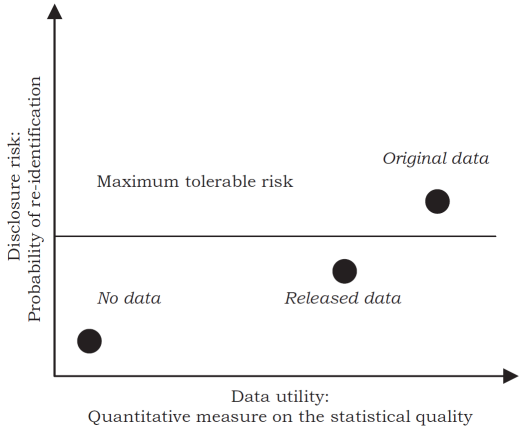

Figure 1: R-U confidentiality map (Duncan et al.,2001)

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.

Variables

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