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

Jiří Novák CC BY-NC-ND (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

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.

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