GESIS Workshop Series  
  
“Synthetic Data: Generation and Evaluation"

|  |  |
| --- | --- |
| Lecturers: | Thom Volker, MSc. |
| Affiliation: | Utrecht University, Department of Methodology and Statistics |
| Email: | t.b.volker@uu.nl |

Date: 13.11.2023-15.11.2023

Time: 09.00-13.00

Location: online

About the lecturers

Thom Volker is a PhD candidate in Methodology and Statistics at Utrecht University, and holds Master’s degrees in statistics and sociology. His PhD research focuses on privacy-aware data synthesis and integrates techniques from statistics and computer science to enhance the generation and evaluation of realistic, safe, and sharable synthetic data. Besides this, he is interested in a broad range of topics related to statistics and computational social science and worked on projects related to Bayesian methods for research synthesis, multiple imputation of missing data, text analysis and game theory.

Course description

In the current age of open science, sharing research code and data is often required when publishing a scientific paper. Moreover, the open dissemination of research data is a potential gold mine for answering many research questions. However, privacy and confidentiality constraints often impede the open dissemination of research data. Synthetic data can be an excellent solution to this problem: the real data is kept secret, but a "fake" version of the data is made available. This synthetic dataset can serve many purposes. For example, it allows those in the process of obtaining access to the real data set to get familiar with the structure of the data, and it allows reviewers (or other researchers) to rerun scripts and assess whether the original analysis code is reproducible and runs as intended. Additionally, the synthetic data itself can be used to run completely different analyses, unrelated to the original research problem. In this course, you will learn what synthetic data is, how to generate synthetic data, how to evaluate its quality in terms of utility and remaining privacy risks, and how to obtain statistically valid results from analyses on this data.

In three half days, we will cover the origins of synthetic data (including its relation to multiple imputation of missing data), practice with generating our own synthetic version of a realistic scientific dataset, and evaluate its quality and disclosure risks. We will discuss how to make inferences from synthetic data and work on increasing the synthetic data quality through advanced modelling or solving practical problems that arise when working with complex data structures (for example, how to deal with deterministic systems/composite variables or logical constraints). On the final day there will be room for individual consultation.

The course will have a hands-on format, with more time scheduled for practicals (+ discussion) than lectures (approximately a 60/40 division). In principle, (social) scientific datasets are provided for all practicals, but participants can also bring their own data (this might not be ideal if (1) this data is so privacy-sensitive that instructors cannot look at it; (2) the dataset is so large that running code takes too long). All practical exercises are in R, but only little programming experience is required (a recent ‘introduction to R’-course or some working experience with R or another scripting language suffices). A good understanding of basic statistics will definitely be beneficial (i.e., working experience with regression analysis).

Keywords

Synthetic data, data generation, disclosure risk, statistical disclosure limitation

Categories (please tick all that apply)

|  |  |  |
| --- | --- | --- |
| **Topics** | **Software** | **Level** |
| Computational Social Science | MPlus | Beginner |
| Data Analysis | Python | Advanced |
| Research Data Management | R |  |
| Survey Methodology | Stata |  |

Course prerequisites

* Experience with R or another scripting/programming language (e.g., some basic understanding of data structures in R (e.g., numeric, factor and character variable types, basic data wrangling and running regression analyses).
* Understanding of basic statistics (working experience with regression modelling).

Target group

Participants will find the course useful if:

* they want understand the idea of synthetic data
* want to be able to generate high-quality synthetic data
* want to evaluate utility and disclosure risks of generated data
* want to share a secure version of their privacy-sensitive data with collaborators and in replication archives
* want to adhere to open science principles (including open data) but are restricted by privacy-issues

Course and learning objectives

By the end of the course participants will:

* have a good understanding of the concept of synthetic data
* know the advantages and disadvantages of synthetic data
* be able to independently generate high-quality synthetic data
* be able to independently evaluate the quality of synthetic data and the remaining disclosure risks

Organizational structure of the course

Each day will consist of two blocks of two hours, containing a live lecture of approximately 45 minutes and a hands-on practical (that can be completed individually or in small groups), and a discussion of approximately 60 minutes. The lecturer will be available for questions during the practicals. On the last day, there will be some time for individual consultation (but project-related questions can also be asked during breaks or before/after class).

Software requirements

*Make sure to have a recent R (and RStudio) installation. Required packages will be announced in due time.*

Recommended literature to look at in advance

Not applicable. If you feel like your knowledge on regression analysis is rusty, I advise to look at the book “Regression and Other Stories” (2020) by Andrew Gelman, Jennifer Hill, and Aki Vehari (Ch. 7 specifically, but Ch. 6, 9 and 10 might also be interesting). If you want get informed on synthetic data before the start of the course, I advise the book “Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation (2011) by Jörg Drechsler.

Day-to-day schedule and literature

**Day 1:** Introduction to synthetic data

*Session 1*

* Lecture 1 (45 min): Introduction, getting to know each other background and general idea of synthetic data
* Practical 1 (60 min): Getting familiar with the concept of synthetic data; introduction to data simulation in R

15 minute break

*Session 2*

* Lecture 2 (45 min): Generating synthetic data, modelling strategies for synthetic data, advantages and disadvantages of synthetic data compared to other disclosure limitation techniques.
* Practical 2 (60 min): Generating synthetic data in R: Different modelling strategies (e.g., parametric and non-parametric approaches) for different problems (e.g., different variable types, different levels of privacy and utility).

**Day 2:** Analyzing and evaluating synthetic data

*Session 3*

* Lecture 3 (45 min): Inferences from synthetic data.
* Practical 3 (60 min): Analyzing synthetic data and making correct inferences (i.e., getting the variances right).

15 minute break

*Session 4*

* Lecture 4 (45 min): Assessing utility and privacy risks in synthetic data.
* Practical 4 (60 min): Applying different approaches for evaluating the quality of synthetic data (different utility measures, general and specific data utility).

**Day 3:** Advanced topics in data synthesis

*Session 5*

* Lecture 5 (45 min): Improving synthetic data quality (advanced modelling approaches, dealing with complex data structures).
* Practical 5 (60 min): Applying advanced data synthesis techniques in R.

15 minute break

*Session 6*

* Lecture 6 (45 min): Advanced topics in data synthesis (for example, generative adversarial models, differential privacy, multilevel data; exact content to be determined after consulting workshop participants).
* Practical 6 / Q&A (60 min): Time to ask questions related to personal projects, but there is also an optional practical on advanced data synthesis techniques, or room for a discussion on such approaches.

Additional recommended literature

Will be listed at the end of each lecture, an exhaustive list will be provided at the end of the course.