Karlstad Applied Analysis Seminar (2025)

Lotte Pater,
Dienst Uitvoering Onderwijs, Ministry of Education,
Culture and Science, and University of Groningen, NL

25 Mar 2025

Practical Synthetic Data Generation for Statistical Inference

Abstract

In the present information age governments, universities, hospitals and many others collect large amounts of personal data. These data can be very useful for research and decision making, but often is not shared due to legal and ethical privacy implications. Synthetic data is a technique used to combine data usage with data privacy. A synthetic version of a dataset is created, ideally with the same statistical distributions as the real dataset (so that it is still useful) but without any personal information (so that there are no privacy risks). I work with an interdisciplinary team in the Dutch government to implement synthetic data operations. In this talk I will tackle two questions:

1. How can you generate potentially useful synthetic data? We use Classification and Regression Trees (CART) as implemented in the R package synthpop. This technique consistently comes out as generating the most useful synthetic data (i.e. similar to the real data) in the comparative literature. I'll explain what makes this technique tick and contrast it with the more popular General Adversional Networks (GAN) class of models. I will also compare it with the conceptually similar technique by Nicklas Jävergård et al. (2024).

2. How can you make potentially useful synthetic data actually useful? Generating synthetic data might seem like mainly a statistical problem. But in actually applying it, we also encountered legal, ethical, software development, governance, political and communication challenges. In many cases, combining mathematics with other skills was necessary to solve a problem. I will highlight some general experiences and talk about how we set down a privacy approach for synthetic data – a problem that combines mathematics, law, ethics and communication.