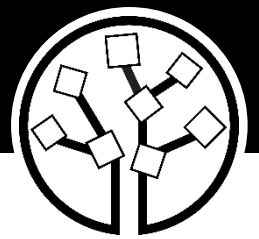


synthpop

# Synthetic data for disclosure control of confidential data


Gillian M Raab

Gillian.raab@ed.ac.uk

University of Edinburgh

Administrative Data Research Centre - Scotland


SLS-DSU

Scottish Longitudinal Study Development & Support Unit

Research Data Scotland

ADR SCOTLAND

*Data-driven change*

Business School December 2024

# Synthetic data for disclosure control?

- Data based on the original data that looks like the original data
- That will give very similar results when a researcher analyses them instead of the real data
- But where, by design, no synthetic record corresponds to an individual record in the original
- Created by modelling the whole statistical distribution  of the original data and taking a random sample from this distribution
- Reproduce the signal, but obfuscate the noise

# What got me involved?

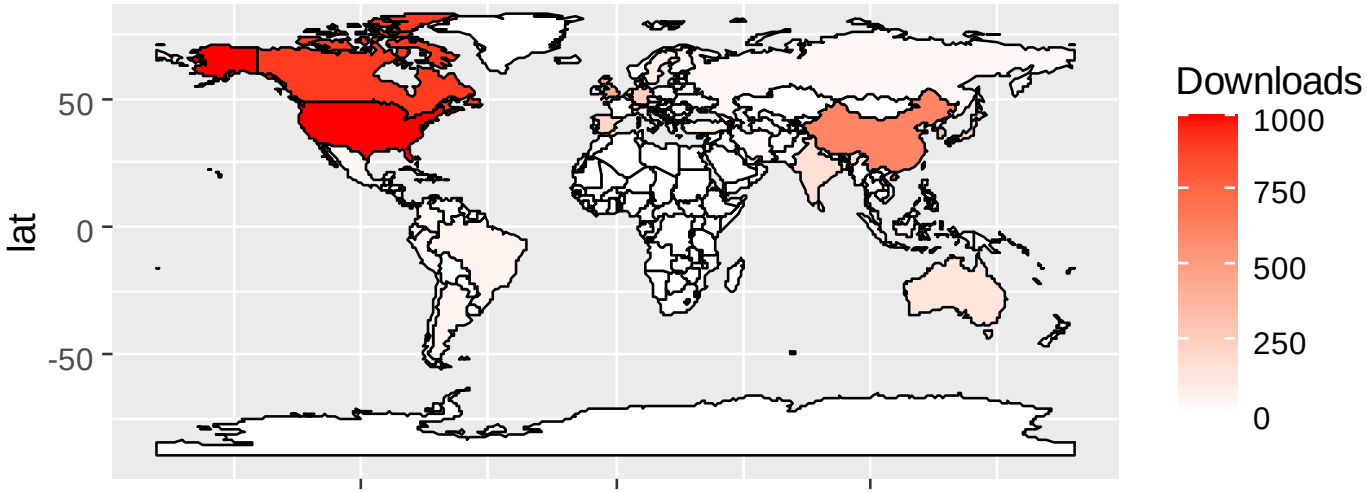
- Scottish Longitudinal  Study
- Census data linked  (5% sample)
  - 1991 2001 2011 soon 2022
- Also linked to many other sources Births, deaths, marriages, widowhoods, Migration, internal and external, Medical data from NHS, School achievement data etc. etc.
- Every user gets an individual extract
- But they need to visit the SLS unit to analyse them
- The SYLLS project (ESRC funded) developed methods to create synthetic versions of their extracts that could be used on their own computers
- The SLS board agreed and a synthetic extract is an option for SLS users (with accreditation and a formal user agreement.
- We also used synthetic data created from linked SLS data in training courses

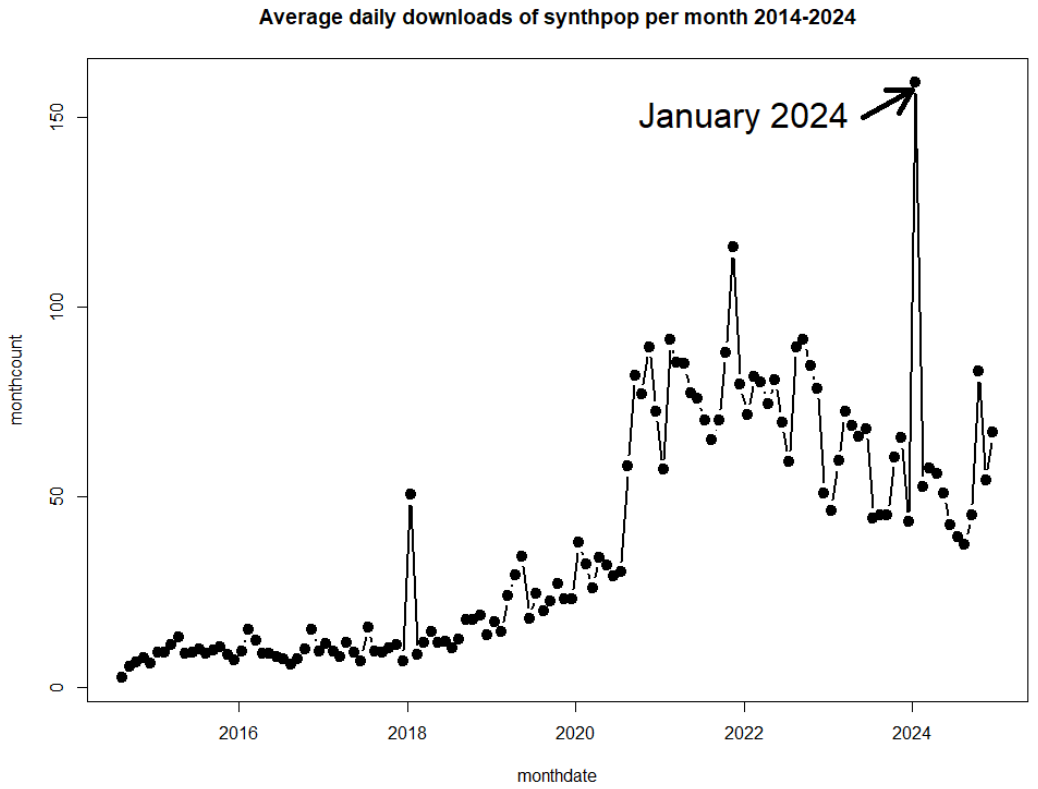
# *Synthpop package for R open source*

Beata Nowok me Chris Dibben

Version 1.1-0 on CRAN website in 2014
Current CRAN version is 1-8-0 from 2022
GitHub most recent version 1.9-0 in 2024

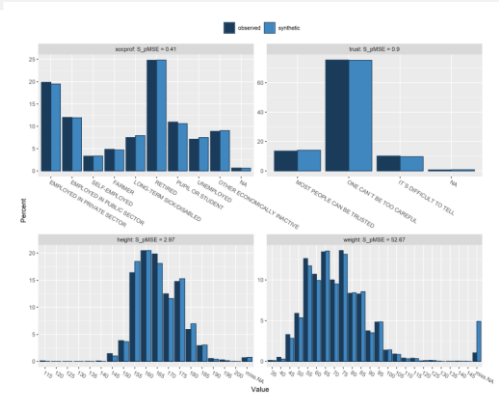
Total of 143K + downloads from main CRAN mirror 2014-2024

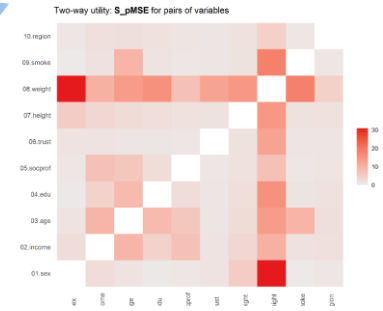


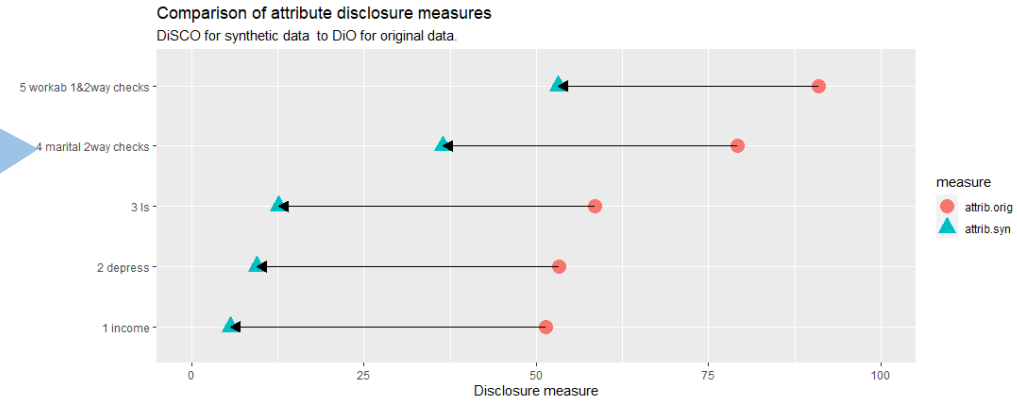
# What does the synthpop package do now?

- Allows the user to create synthetic versions of an original data set
- Using a variety of methods and allowing for customising the results
- Evaluates how well the synthetic data will reproduce results from the original (utility)






- Compares the disclosure risk of the synthetic data to that of the original



# Could it be used for credit-scoring data ?

- Yes
- To get synthetic data with good utility would require work
- The structure of the data would need to be understood and modelled
- Probably need to be approached in two stages
- Not unlike data on hospital admissions
- Could be an interesting project

# Different approaches to creating synthetic data

**STATISTICAL (from 1993)**

Fitting a model to the joint distribution

Simulate from the model

- Retaining the signal
- But not the noise

Large choice of models (usually from conditional distributions) that can be adapted to the properties of the data set

**MACHINE LEARNING (from 2014)**

Specify the algorithm and tune the parameters

Methods include

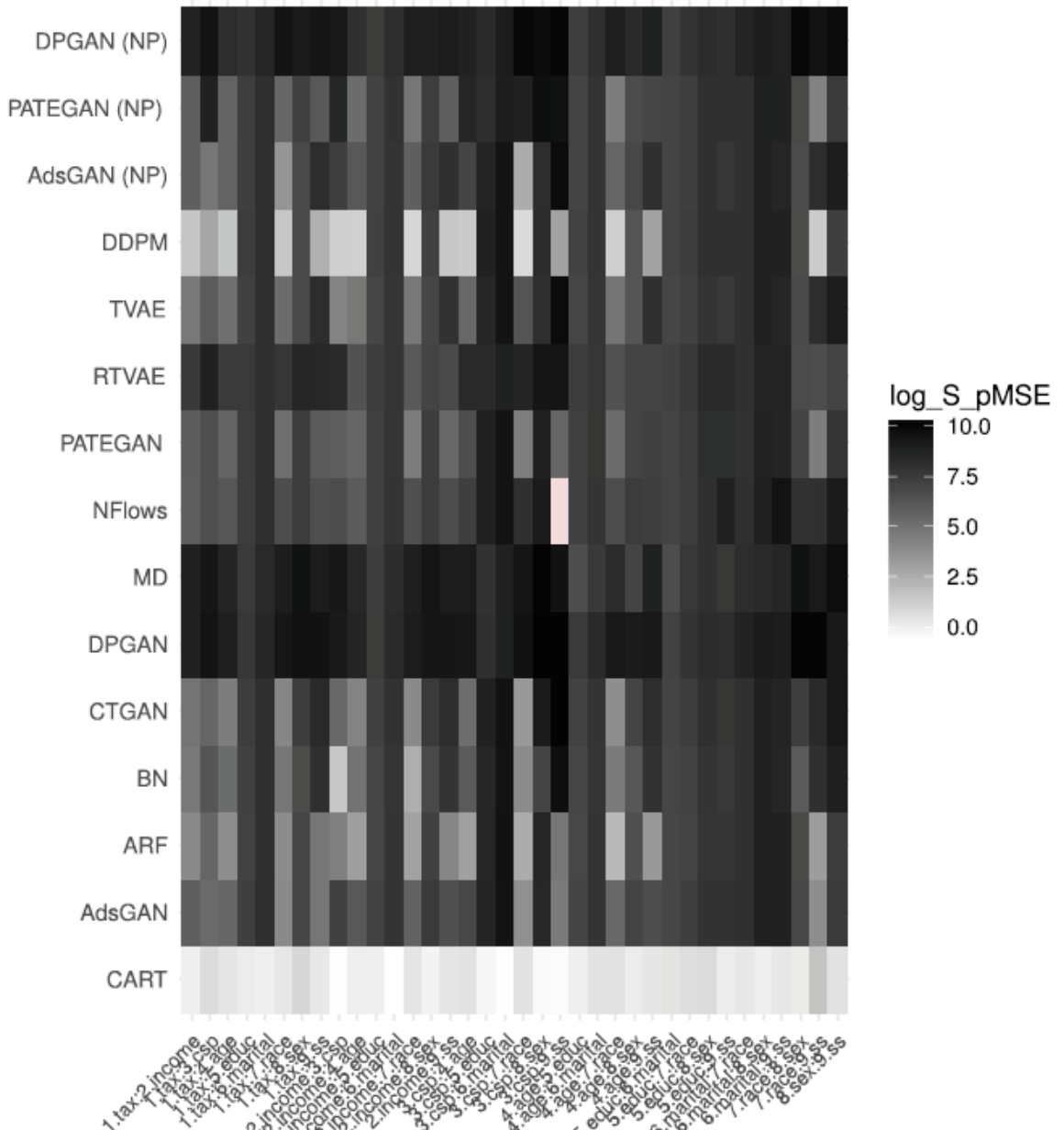
- GANs
- VAEs
- SVMs

Lots of software available to implement these methods, both open-source and commercial

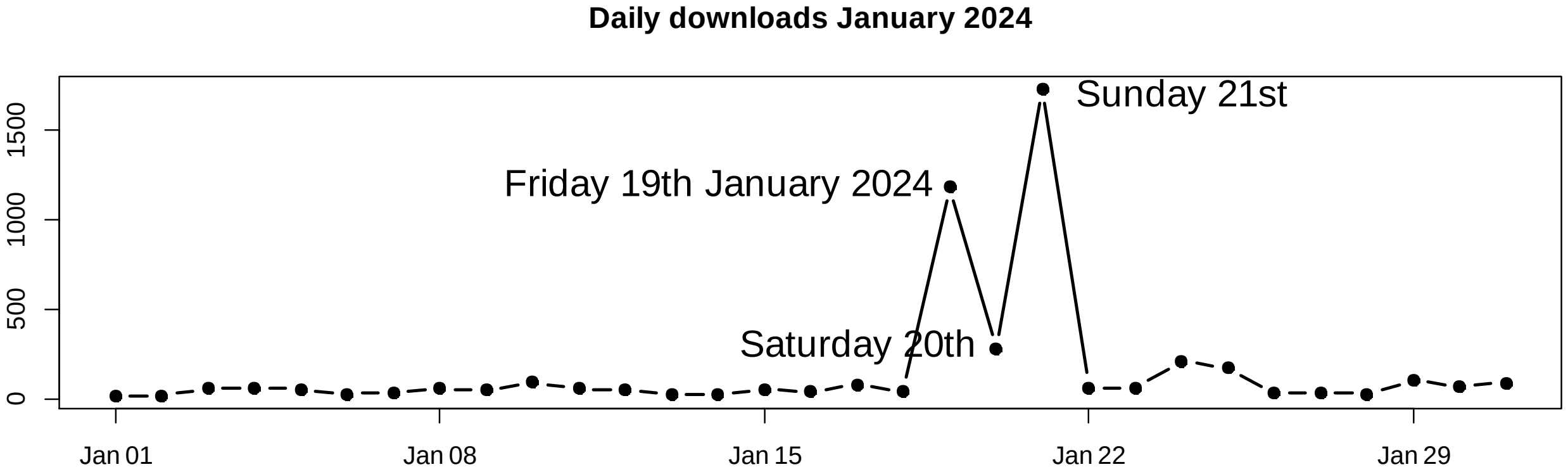


"As a benchmark we used the CART-based synthesizer as implemented in the **synthpop** package for R.

Our simulations confirmed findings of previous papers that CART methods tend to offer much better analytical validity than other synthesizers."



# Why so many downloads in January 2024?



# Searching for “synthpop” on Google