UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Confidentiality

26–28 September 2023, Wiesbaden

ASSESSING THE UTILITY OF SYNTHETIC DATA: A DENSITY RATIO PERSPECTIVE

Thom Benjamin Volker (Utrecht University, the Netherlands; Statistics Netherlands, the Netherlands)
Erik-Jan van Kesteren (Utrecht University, the Netherlands)
t.b.volker@uu.nl, e.vankesteren1@uu.nl

Abstract

High quality synthetic data can be a solution to overcome disclosure risks that arise when disseminating research data to the public. However, for inferential purposes, the usefulness of the synthetic data largely depends on whether the synthetic data model approximates the distribution of the observed data close enough. Often, multiple candidate models are considered and improvements are made iteratively. Evaluating the utility of the synthetic data after the model building steps is a crucial but complicated endeavor, and although many methods exist, their results may tell an incomplete story. We propose to evaluate the utility of synthetic data using density ratio estimation techniques, which allows to embed both general and specific utility measures into a common framework. Using techniques from the density ratio estimation field, we show how an interpretable global utility measure can be obtained from the ratio of the observed and synthetic data densities that can even be used to express the utility of individual (synthetic) data points. Applying these techniques on (an approximation to the) posterior distributions of parameters gives rise to analysis-specific utility measures. Using empirical examples, we show that framing utility from a density ratio perspective improves on existing global utility measures. Lastly, we implemented the procedure along with existing utility measures in an R-package.

Introduction

This will be the introduction.

Introduction

Here is one paragraph of the introduction.

Here is a second paragraph of the introduction.

Primary heading

Secondary heading

Here is a paragraph.

Secondary heading

In this paragraph, we cite this book, **?**, but Figure~1 is not taken from that book. Another paper contains some formulas (see **?**). Also,

1. Xxxxxxxx
   - Xxxxx
     a. Yyyyy
   - Xxxxx
2. Xxxxxxxxxxxxxxxxxx
3. Xxxxxxxxxxxxxxxxxx

Primary heading

Secondary heading

This paragraph is a pretext to include a table:

Team	P	W	D	L	F	A	Pts
Manchester United	6	4	0	2	10	5	12
Celtic	6	3	0	3	8	9	9
Benfica	6	2	1	3	7	8	7
FC Copenhagen	6	2	1	2	5	8	7

Another paragraph with high-flying content, maybe refering to Section~. We forgot to mention that Figure~1 has been produced with R.

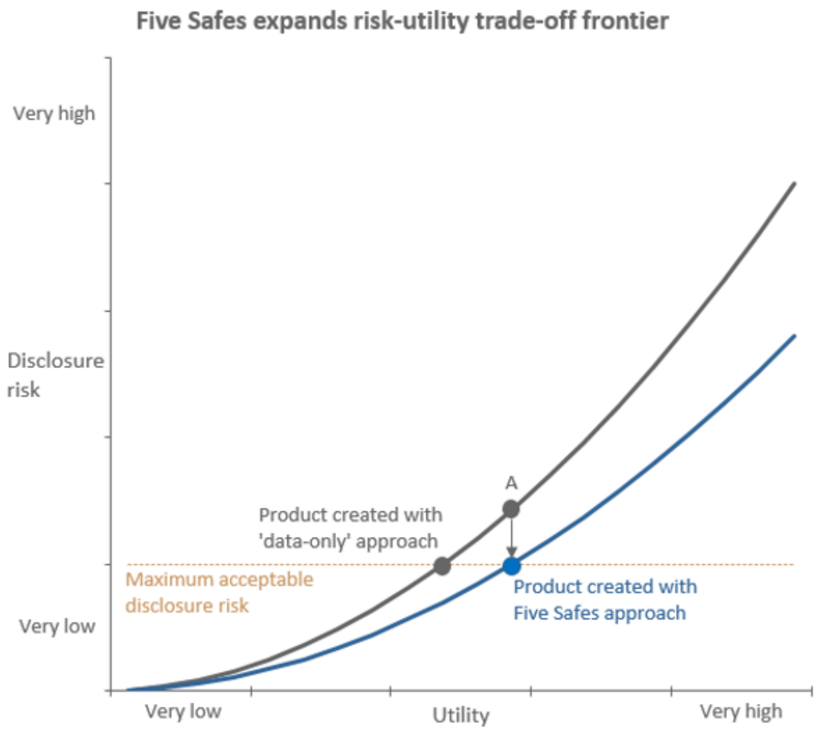

Figure 1. The Five Safes expands the R-U trade-off frontier. For each level of utility, we can achieve lower disclosure risk.
See https://www.abs.gov.au/statistics/research/confidentiality-abs-microdata