INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

UNDERSTANDING THE TRADE-OFF BETWEEN UTILITY AND RISK IN CART BASED MODELS USING SIMULATION DATA

Berlin,
7-8. Oktober, 2024

Jonathan Latner, PhD
Dr. Marcel Neuenhoeffer
Prof. Dr. Jörg Drechsler

SECTION 1: INTRODUCTION

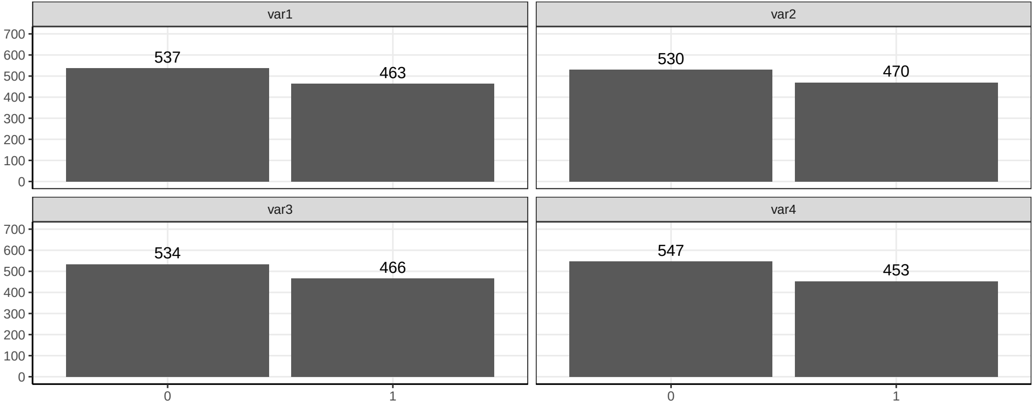
DATA

From Reiter et al., 2014

“We use a simple simulation scenario that illustrates many of the main issues: protecting a 2^4 binary table with fully synthetic data. For $i = 1, \ldots, 1000 = n$, let $y_i = (y_{1i}, y_{2i}, y_{3i}, y_{4i})$ comprise four binary variables. Let each of the $K = 16$ possible combinations be denoted c_k, where $k = 1, \ldots, 16$. Let $c_{16} = (0, 0, 0, 0)$, and let $C_{-16} = (c_1, \ldots, c_{15})$. We generate an observed dataset D as follows. For $i = 1, ..., n - 1 = 999$, sample y_i from a multinomial distribution such that $p(y_i = c_k) = 1/15$ for all $c_k \in C - 16$. Set $y_{1000} = c_{16}$. Since we do full synthesis, $X = \theta$”

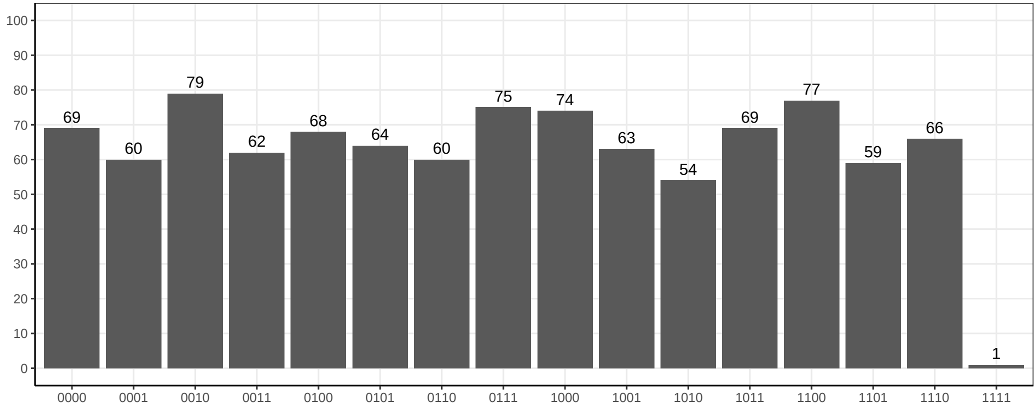
VARIABLE FREQUENCY

Figure 1



HISTOGRAM

Figure 2



SYNTHPOP

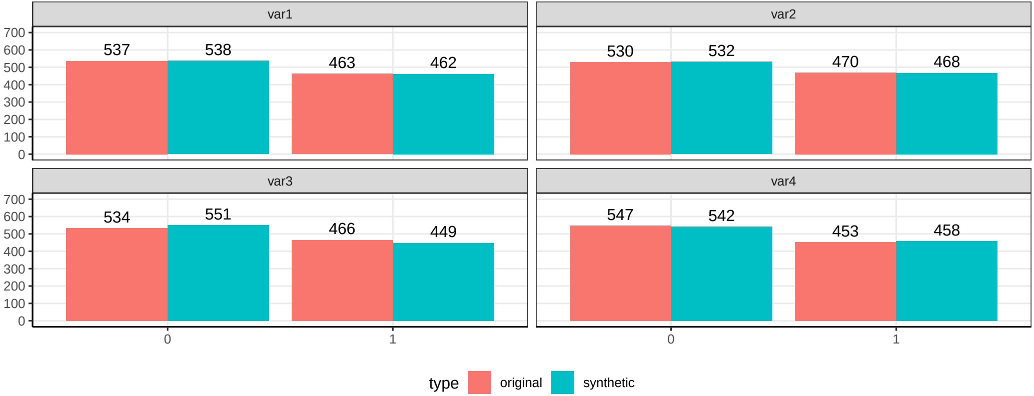
```
> sds <- syn(df_ods, m=1)
Warning: In your synthesis there are numeric variables with 5 or fewer levels: var1, var2, var3, var4.
Consider changing them to factors. You can do it using parameter 'minnumlevels'.

Synthesis
-----------
 var1 var2 var3 var4
```

notice the "Warning". It means that the variables are being synthesized as numerical values (0/1), and Synthpop is suggesting they should be synthesized as categorical values (“0”/“1”)

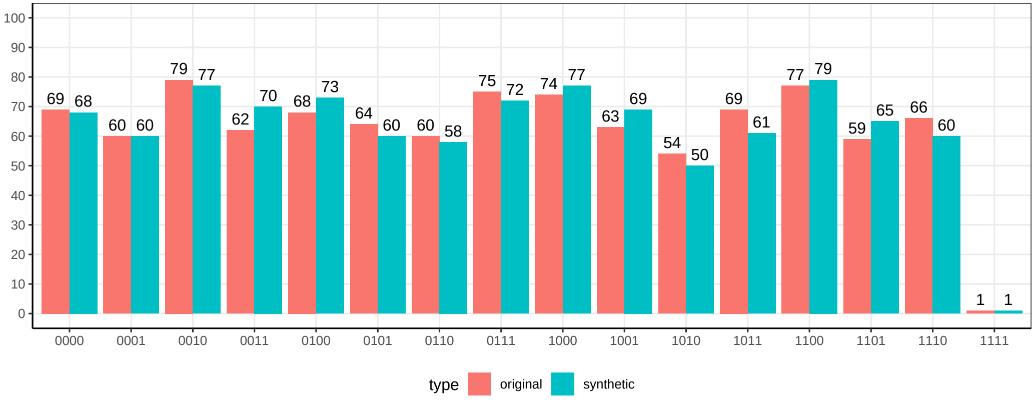
COMPARE FREQUENCY (NUMERICAL)

Figure 3



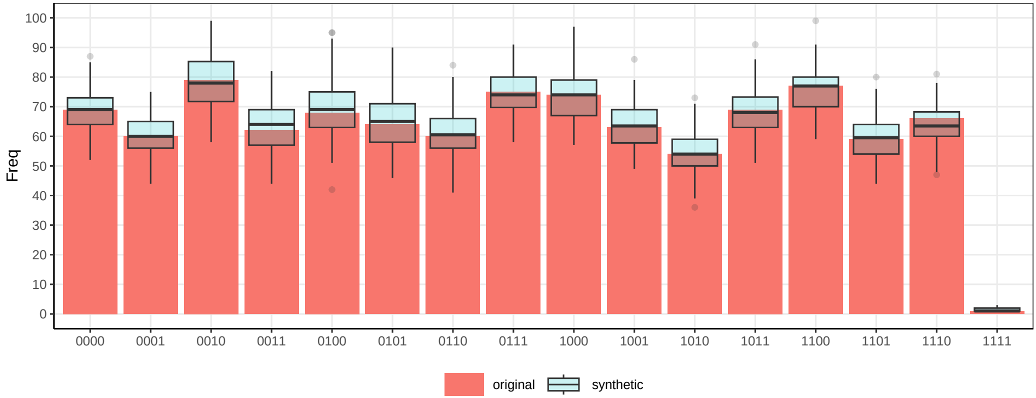
COMPARE HISTOGRAM (NUMERICAL)

Figure 4



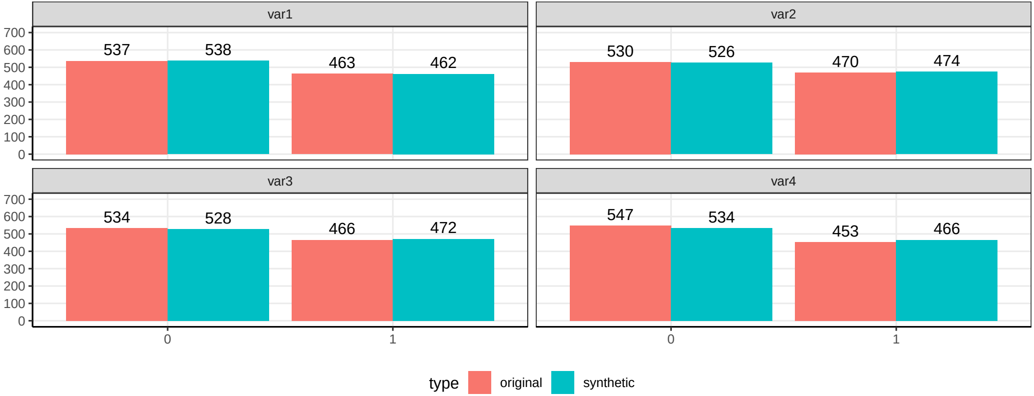
COMPARE HISTOGRAM (NUMERICAL) X 100 SYNTHETIC DATASETS

Figure 5



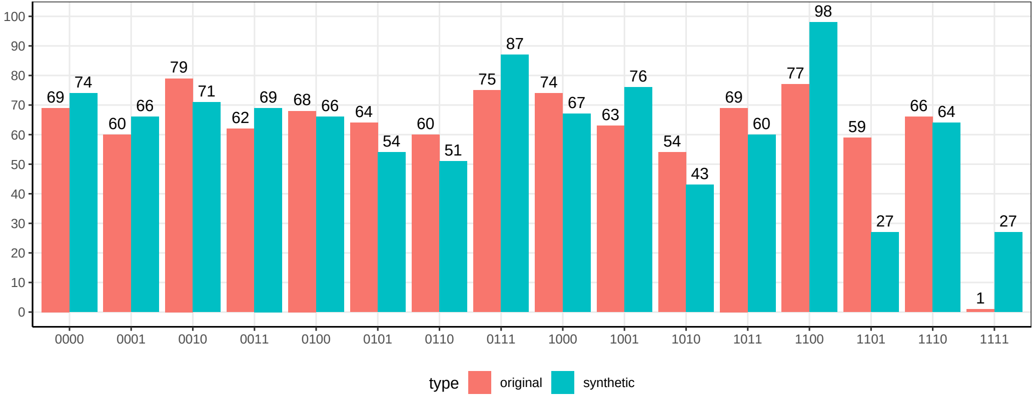
COMPARE FREQUENCY (CATEGORICAL)

Figure 6



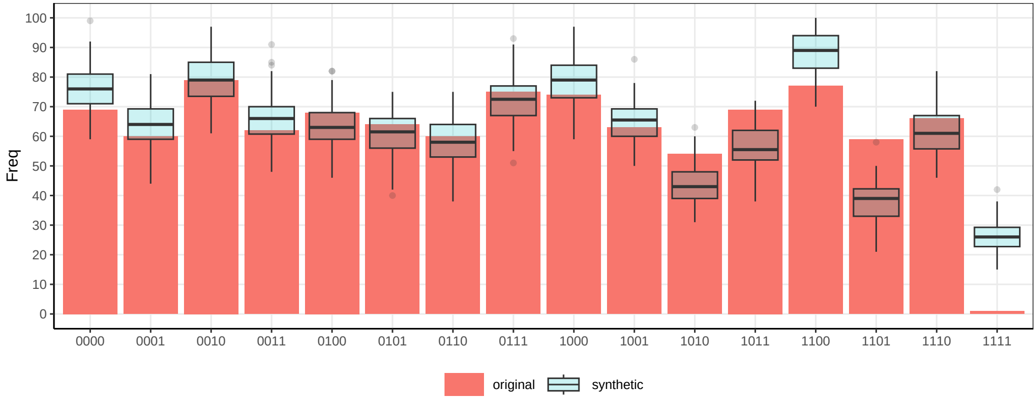
COMPARE HISTOGRAM (CATEGORICAL)

Figure 7



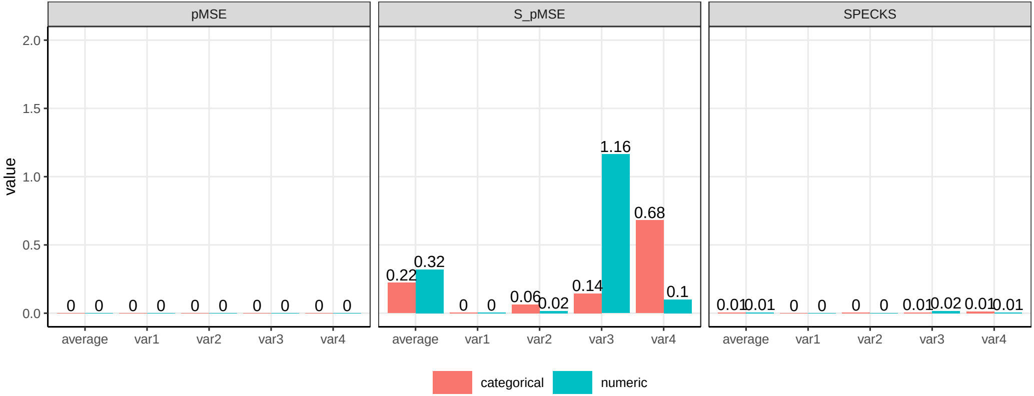
COMPARE HISTOGRAM (CATEGORICAL) X 100 SYNTHETIC DATASETS

Figure 8



COMPARING UTILITY MEASURES

Figure 9: Utility measures close to 0, i.e. high utility

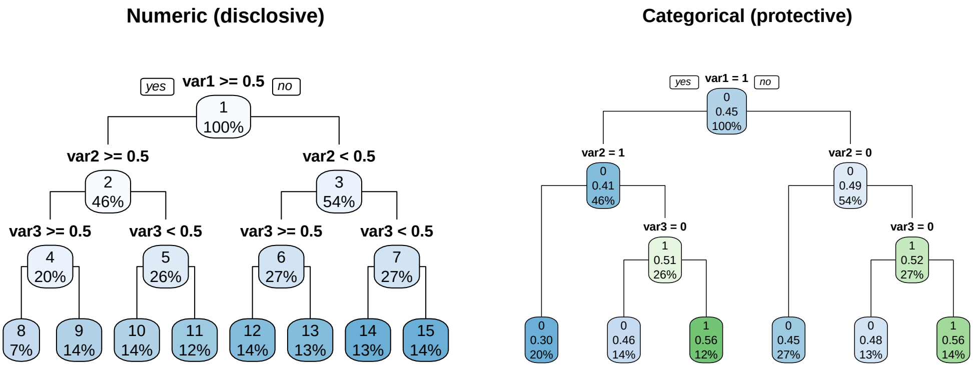


COMPARING PRIVACY MEASURES

all privacy measures close to 0, i.e. low privacy risk

HOW DO WE EXPLAIN THIS?

Figure 10

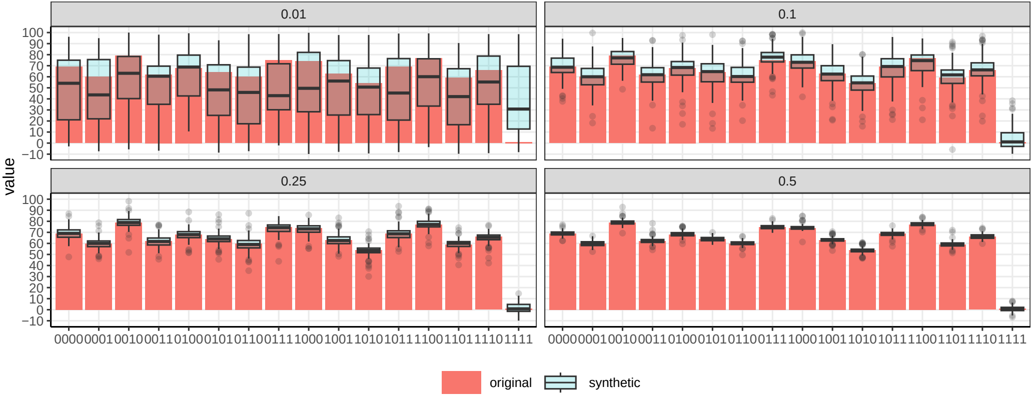


Numeric (disclosive)

Categorical (protective)

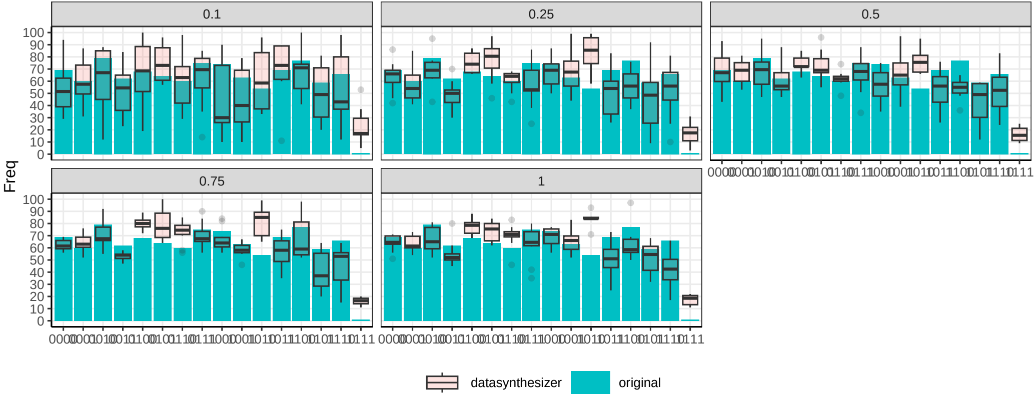
HISTOGRAM WITH DIFFERENTIAL PRIVACY X 100 SIMULATIONS

Figure 11



HISTOGRAM WITH DP (DATASYNTHESIZER) X 100 SIMULATIONS

Figure 12


Freq
datasynthesizer
original