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

AniGeD

BUYER BEWARE: 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

OVERVIEW

- It is well established that there is a trade-off between utility and privacy when generating synthetic data
- Utility in CART based synthesizers is high (Little et al., 2022; Danker and Ibrahim, 2021)
- Are CART based synthesizers actually preserving privacy? If so, how?
- Using simulation data (Reiter et al., 2014), results suggest synthetic data from CART models are disclosive
- Disclosive in ways that are not observable using traditional privacy measures

WHATS THE GOAL OF SYNTHETIC DATA?

- Synthetic data can accelerate development by replacing sensitive values with synthetic ones with minimal distortion of the statistical information contained in the original data set. (Jordan et al., 2022; Nowak et al., 2016)
- Low disclosure risk (R)
- High data utility (U)
- Visualize the trade-off using the R-U confidentiality map (Duncan et al., 2004)

WHATS THE PROBLEM?

- High data utility – It must be similar to and different from the original data.
  - At the extreme, if the goal is high utility, why not just release the original data?
- Low disclosure risk – Synthetic data is not automatically private.
  - At the extreme, if the goal is low privacy risk, why should we release any data?
- Many measures of utility and privacy exist
  - Therefore, its not clear if data have high utility or low risk
  - 2 problems
    - More specifically, how can we map R-U trade-off if there are multiple measures of both?
    - More generally, how do we know if the data have high levels of utility and low levels of privacy?

WHAT DO WE KNOW?

- Reiter (2005) suggested using sequential modeling with Classification and Regression Trees (CART).
- Utility
  - Drechsler and Reiter (2011) found that CART models offered the best results in terms of preserving the information from the original data.
  - Other comparisons also found CART is superior (Little et al., 2022; Danker and Ibrahim, 2021)
- Privacy
  - Some evidence also suggests CART is superior (Little et al., 2022)
  - However, other evidence indicates that CART-based synthesis simply replicates most of the original records (Manrique-Vallier and Hu, 2018)

HOW DOES SEQUENTIAL MODELING WITH CART? (NOWAK ET AL., 2022)

- Consider as an example a default synthesis, i.e. synthesis with all values of all variables $(Y_1, Y_2, \ldots, Y_p)$ to be replaced.
- The first variable to be synthesised (Y_1) cannot have any predictors and therefore its synthetic values are generated by random sampling with replacement from its observed values.
- The second variable to be synthesized (Y_2) is generated using the fitted model and the synthesised values of (Y_1).
- The third variable to be synthesized (Y_3) is generated using the fitted model and the synthesized values of Y_1 and (Y_2)
- The distribution of the last variable (Y_p) will be conditional on all other variables.

EXAMPLE

Simulation data with 1.000 observations, 4 columns (var1, var2, var3, var4), and each column is random ("0"/"1")

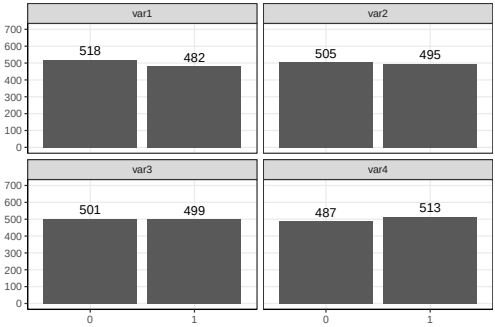


Figure 1: Frequency

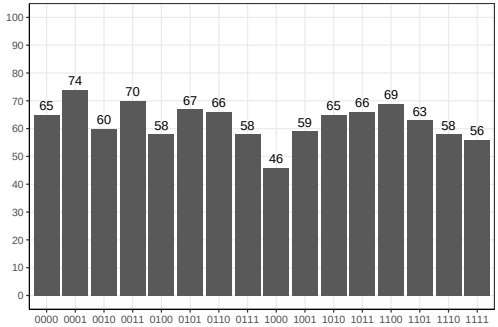


Figure 2: Histogram

COMPARE ORIGINAL TO SYNTHETIC DATA GENERATED BY CART

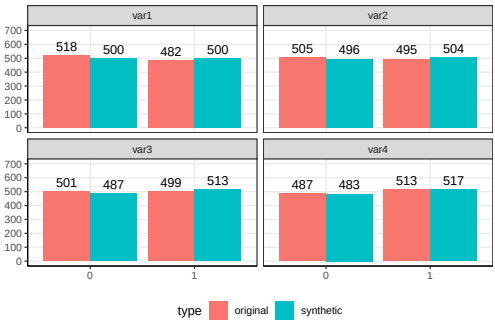


Figure 3: Frequency

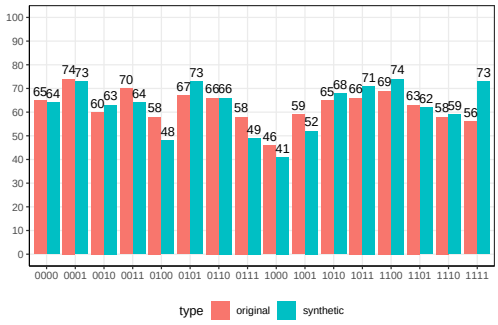


Figure 4: Histogram

SIMULATE DATA WITH A UNIQUE RECORD

Borrowing from Reiter et al. (2014), we set the first 999 observations to be a random sample from a multinomial distribution for all combinations of $var1(0,1)$, $var2(0,1)$, $var3(0,1)$, $var3(0,1)$ except (var1=1,var2=1,var3=1,var4=1), which we set to be the 1000th observation.

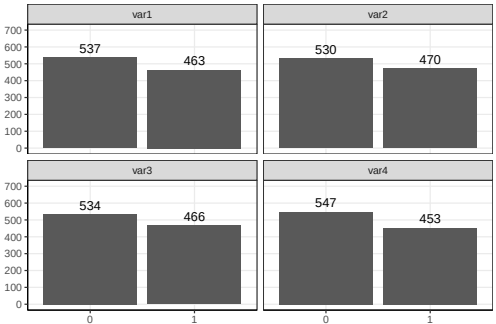


Figure 5: Frequency

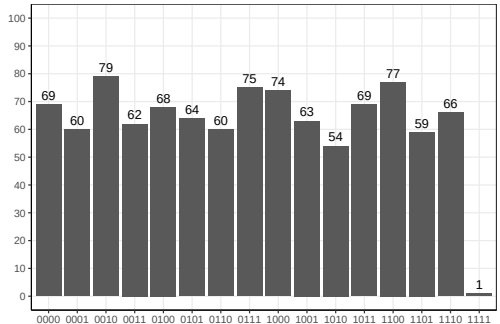


Figure 6: Histogram

SIMULATE DATA WITH A UNIQUE RECORD

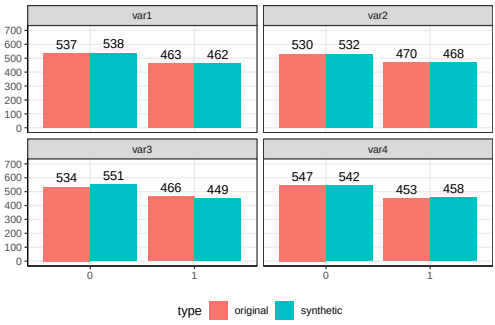


Figure 7: Frequency

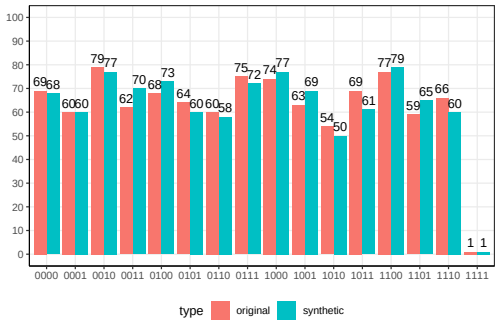
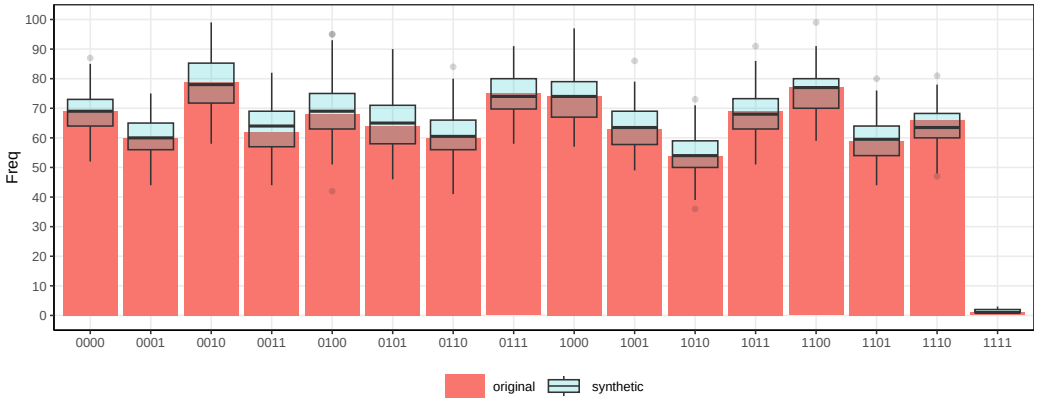


Figure 8: Histogram

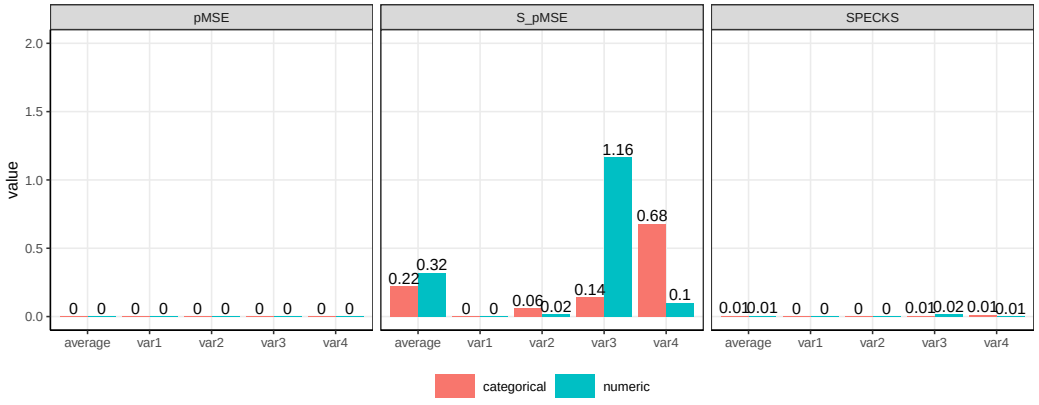
COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS

Figure 9



COMPARING UTILITY MEASURES

Figure 10: Utility measures close to 0, i.e. high utility



COMPARING PRIVACY MEASURES

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