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

BALANCING DATA UTILITY AND PRIVACY: EVALUATING COMPUTER SCIENCE AND STATISTICAL APPROACHES TO CREATING SYNTHETIC DATA

Statistische Woche
Dortmund, Deutschland
11. September, 2023

Jonathan Latner, PhD
Prof. Dr. Jörg Drechsler

SECTIONS

1. Introduction
2. Data
3. Methods
4. Results
5. Conclusion

Section 1: Introduction

OVERVIEW

- **Definition:** Synthetic data are data that mimic the characteristics of original or ‘real’ data.
- **Goal:** High utility + privacy protection = more knowledge
  - Utility: any analysis using synthetic data should provide approximately the same answers as analysis using the original data.
  - Privacy: Synthetic data look like real data, but without any of the threats to privacy contained in real data (theoretically).
  - More knowledge is created if more data can be released with no (or little) risk to privacy, and accessed more easily.
- **The problem:** Trade off between utility and privacy
- Different approaches for the generation of synthetic data have been developed.
  - In statistics: formal guarantee of statistical utility, but no formal guarantee for privacy
  - In computer science: formal guarantee of privacy, but no guarantee for statistical utility
- Hard to evaluate which approach is correct or even better
  - No one, single definition of utility or privacy protection
  - Different approaches (data packages) solve different data problems

THE GOAL

- Evaluate (compare and contrast) how different approaches to the creation of synthetic data balance the twin goals of data utility and data privacy.
- For now, we focus on utility (next steps: privacy)
- In this talk, we will present some first results from this project.

LITERATURE REVIEW

- Not many papers that compare and contrast
- Most papers are often written by the package authors
  - In these papers, their package is often the 'winner'
  - May be biased, but not necessarily wrong
  - Different data packages solve different data problems
  - Different data packages have different strengths and weaknesses
- Few 'independent' papers
  - Little et al., 2021/2023 (*W*orking paper)
    - They don't tune the packages (defined later)
    - Low dimensional data/majority categorical data
  - Dankar and Ibrahim (2021)
    - 15 data sets (categorical, continuous, mixed)
    - Low dimensional data

PREVIEW RESULTS

- 3 data packages (CTGAN, Datasynthesizer, Synthpop)
- 3 types of data
- Focus on utility (next step: privacy)
- Synthpop is the ‘winner’ (similar to both Little et al., 2021/2023 and Dankar and Ibrahim, 2021)
- Results are not surprising
  - Synthpop emphasizes utility over privacy protection
  - Packages are evaluated on data with low dimensionality (observations/variables), where Synthpop performs better

Section 2: Data

3 DATA SETS

- 2 Simulated data sets - Examine differences in a controlled environment
  1. Simulated categorical data
     - 1.000 observations
     - 4 bivariate categorical variables (‘Y’, ‘N’)
  2. Simulated continuous data
     - 1.000 observations
     - 3 continuous variables (‘income’, ‘wealth’, ‘age’)
- 1 Real data set
  - UK 1991: Individual Sample of Anonymised Records (SAR) for the British Census, subsetted on the region of West Midlands
  - 20% sample ($\approx$ 20.000)
  - 12 variables: 1 numerical and 11 categorical, includes missing values
  - Benchmark our results to Little et al., 2021/2023:

Section 3: Methods

COMPARE 3 PACKAGES FOR CREATING SYNTHETIC DATA

- We choose these three because they are commonly compared (Little et al., 2021/2023; Dankar and Ibrahim, 2021)

  1. CTGAN (Conditional Tabular Generative Adversarial Network) in Synthetic Data Vault (SDV) package (Patki et al., 2016)
  2. Datasynthesizer (Ping et al., 2017)
  3. Synthpop (Nowak et al., 2016)

Table 1: Comparison of data synthesis packages

Variable/data	CTGAN (GANs)	Datasynthesizer (PrivBayes)	Synthpop (CART)
Continuous variables	✓		✓
Categorical variables			✓
Mixed data		✓	✓
Privacy protection	✓*	✓	✓†
High dimensional datasets	✓	✓	

* In theory, GANs (Synthpop) offer no formal privacy protection. In reality, one can adjust the training procedure of the discriminator to satisfy a formal guarantee (Beaulieu-Jones, et al., 2019; Neunhoffer, et al., 2021).

† Parameters can be used to adjust privacy.

MEASURING UTILITY (2 'SPECIFIC' MEASURES)

1. Descriptive (non-parametric): Ratio of estimates (ROE)*
   - The average difference between the values (i.e. proportion/total) of a given categorical variable (or binned continuous variable) between synthetic and original data
   - Higher is better utility. Max = 1, min = 0.
   - The ROE is calculated over univariate and bivariate values of a given variable(s).
2. Parametric: Difference/overlap in the estimate/confidence interval†
   - Apply the same regression model to original and synthetic data
     - Research choice: Theoretical/athoeretical models
   - Standardized difference in the β - The average difference between each point estimate. Lower is better utility (closer to 0).
   - Confidence interval overlap (CIO) - The percent overlap in the 95% confidence interval. Higher is better utility. Max = 1, negative value indicating no overlap (here, negative = 0).
3. Others ('Universal' measures: next steps)

* https://github.com/clairelittle/psd2022-comparing-utility-risk/blob/main/code/ROC_Ratio_of_Counts_Estimates.R
† https://github.com/clairelittle/psd2022-comparing-utility-risk/blob/main/code/CIO_Confidence_Interval_Overlap.R

TUNING

- **Definition:** Adjusting the data packages to create synthetic data that are more representative of the original data
- Data packages can be tuned to various levels using hyperparameters
  - Authors of data packages state that tuning the packages are important
  - One should not simply use the default values
- Tuning is a time consuming process (described later)
  - Not 100% clear on how to best tune the data
  - No one, single measure for utility or privacy protection
- For each package (Datasynthesizer, CTGAN, Synthpop), estimate the effect of each categorical hyperparameter value (h_j) on each utility measure (y_i) (ROE_u, ROE_b, Std. Diff, and CIO).

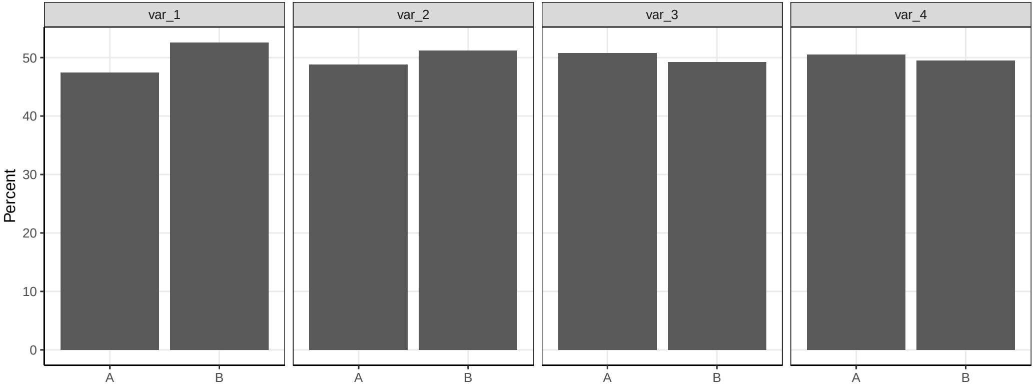
$$y_i = \beta_0 + \sum_{h=1,j=1}^{H,J} \beta_{h,j} x_{h,j,i} + \epsilon \tag{1}$$

Section 4: Results

Section 4a): Simulated categorical data

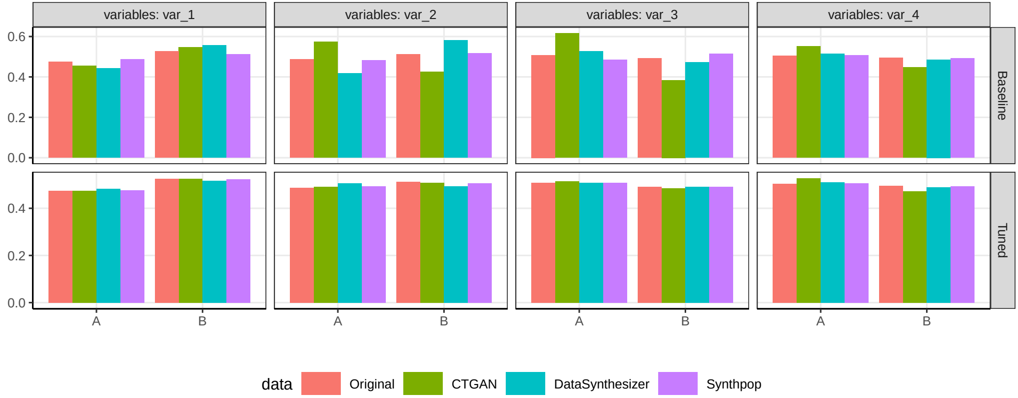
DESCRIPTIVE STATISTICS - CATEGORICAL DATA

Figure 1:



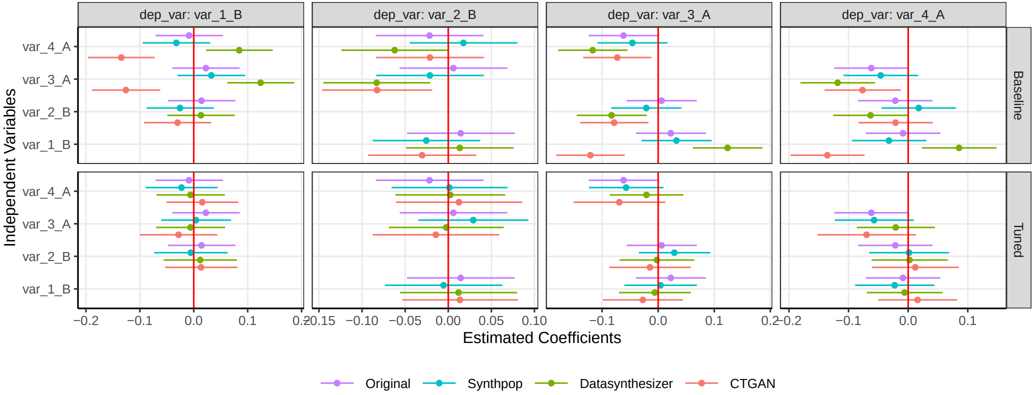
COMPARE FREQUENCY COUNTS BETWEEN BASELINE AND TUNED

Figure 2:



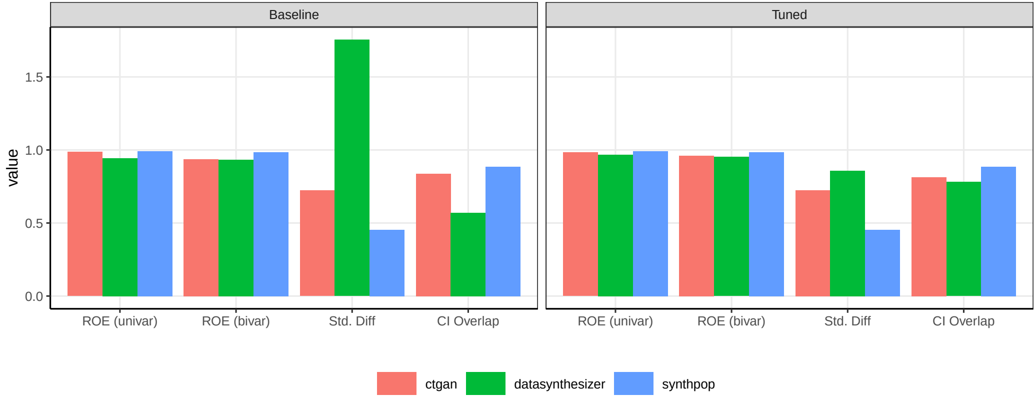
COMPARE REGRESSION OUTPUT BETWEEN BASELINE AND TUNED

Figure 3:



MEASURING UTILITY FOR SIMULATED CATEGORICAL DATA

Figure 4:



SUMMARY: SYNTHPOP HAS HIGHEST UTILITY

Table 2: Comparison of Results

Data	ROE univar	ROE bivar	Std. Diff	CIO
Simulated categorical variables	=	=	SP	SP

- CTGAN/Datasynthesizer requires tuning (not Synthpop)
- Unexpectedly, CTGAN performs 2nd best for categorical data

Section 4: Results
Section 4b): Simulated continuous data

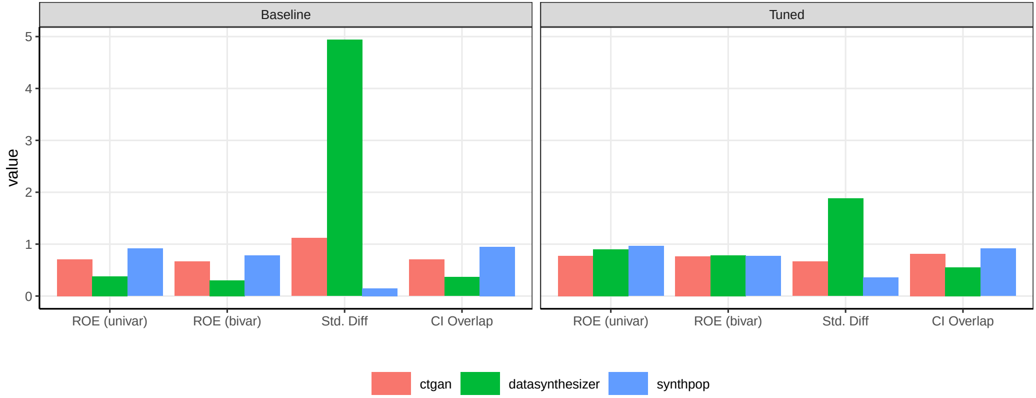
DESCRIPTIVE STATISTICS - CONTINUOUS VARIABLES

Table 3:

number	variable	min	max	mean	std	median
1	age	16.00	94.00	55.65	22.56	57.00
2	income	624.00	349355.00	35721.10	41455.86	22421.50
3	wealth	-19560.00	89975356.00	659520.71	3165853.95	127616.00

MEASURING UTILITY FOR SIMULATED CONTINUOUS DATA

Figure 5:



SUMMARY: SYNTHPOP HAS HIGHEST LEVELS OF UTILITY

Table 4: Comparison of Results

Data	ROE univar	ROE bivar	Std. Diff	CIO
Simulated continuous variables	SP	=	SP	SP

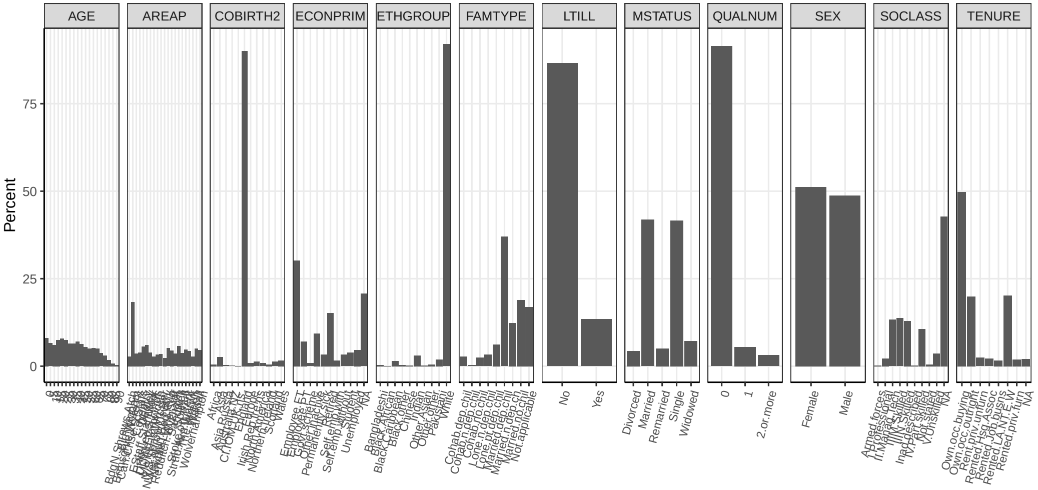
- CTGAN has similar level of CIO but higher Std. Diff

Section 4: Results

Section 4c): Individual Sample of Anonymised Records (SAR) for the British Census, subsetted on the region of West Midlands (UK 1991)

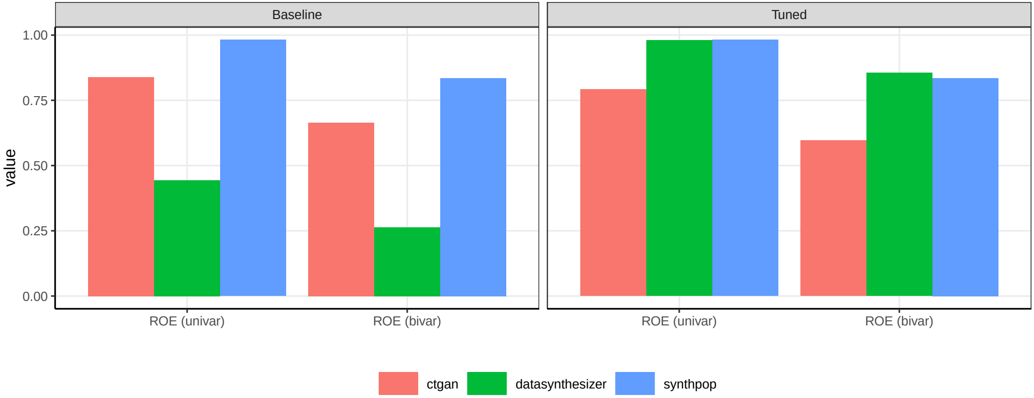
DESCRIPTIVE STATISTICS - UK 1991

Figure 6:



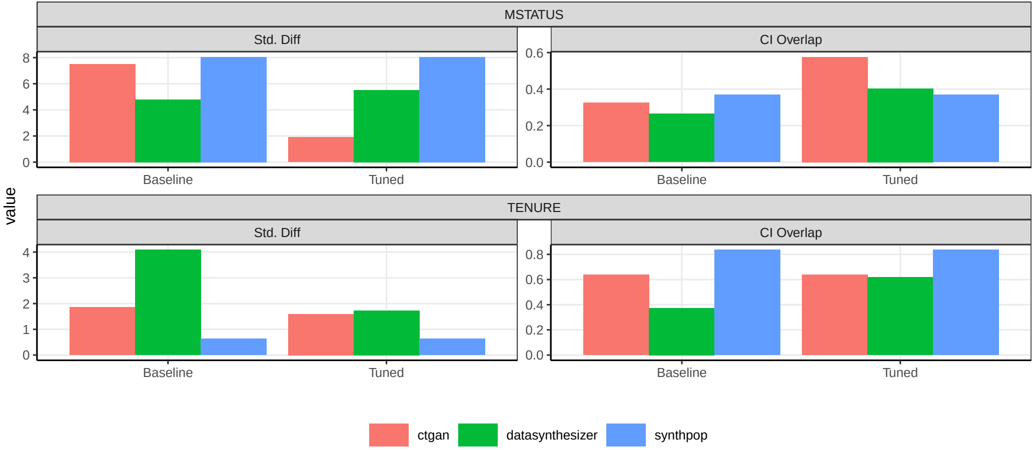
MEASURING UTILITY

Figure 7: Ratio of estimates



MEASURING UTILITY

Figure 8: Confidence interval overlap



COMPARING DURATION TO CREATE 1 SYNTHETIC DATA SET ($\times 5$)

Table 5: UK 1991 data, 12 variables (1 continuous), and $\approx$ 20.000 observations

synthesizer	baseline	tuned
Synthpop	1 Min 48.0 Sec	1 Min 23.0 Sec
Datasynthesizer	2 Min 0.0 Sec	0 Min 54.0 Sec
CTGAN	6 Min 30.0 Sec	41 Min 9.0 Sec

SUMMARY: THINGS BECOME MORE COMPLICATED IN ‘REAL’ DATA

Table 6: Comparison of Results

Data	ROE univar	ROE bivar	Std. Diff	CIO
UK1991	DS/SP	DS/SP		
DV: MSTATUS			CTGAN	CTGAN
DV: TENURE			SP	SP

Section 5: Conclusion

SUMMARY

Table 7: Comparison of Results

Data	ROE univar	ROE bivar	Std. Diff	CIO
Simulated categorical variables	=	=	SP	SP
Simulated continuous variables	SP	=	SP	SP
UK1991	DS/SP	DS/SP		
DV: MSTATUS			CTGAN	CTGAN
DV: TENURE			SP	SP

SUMMARY OF RESULTS

- Results a reflection of low dimensional data and focus on data utility
- Main message: Synthpop is the ‘winner’
  - However, it is not always the ‘best’
  - Easy - little tuning required
  - Fastest (data dimensionality)
  - No privacy protection†
- Datasynthesizer/CTGAN
  - Requires tuning
- CTGAN
  - Slowest

QUESTIONS

- Where is Synthpop not right?
- Where is CTGAN right? Is it worth it?
- What are the right utility measures to use and when do we use them?

NEXT STEPS

- Privacy protection
- High dimensional data
- Assumption is that CTGAN/Datasynthesizer will perform better

Thank you