Presentation at the UNECE Work Session on Statistical Data Confidentiality
December 2021

GENERATIVE ADVERSARIAL NETWORKS FOR SYNTHETIC DATA GENERATION: A COMPARATIVE STUDY

Claire Little, Mark Elliot, Richard Allmendinger, Sahel Shariati Samani

Centre for Digital Trust and Society

University of Manchester

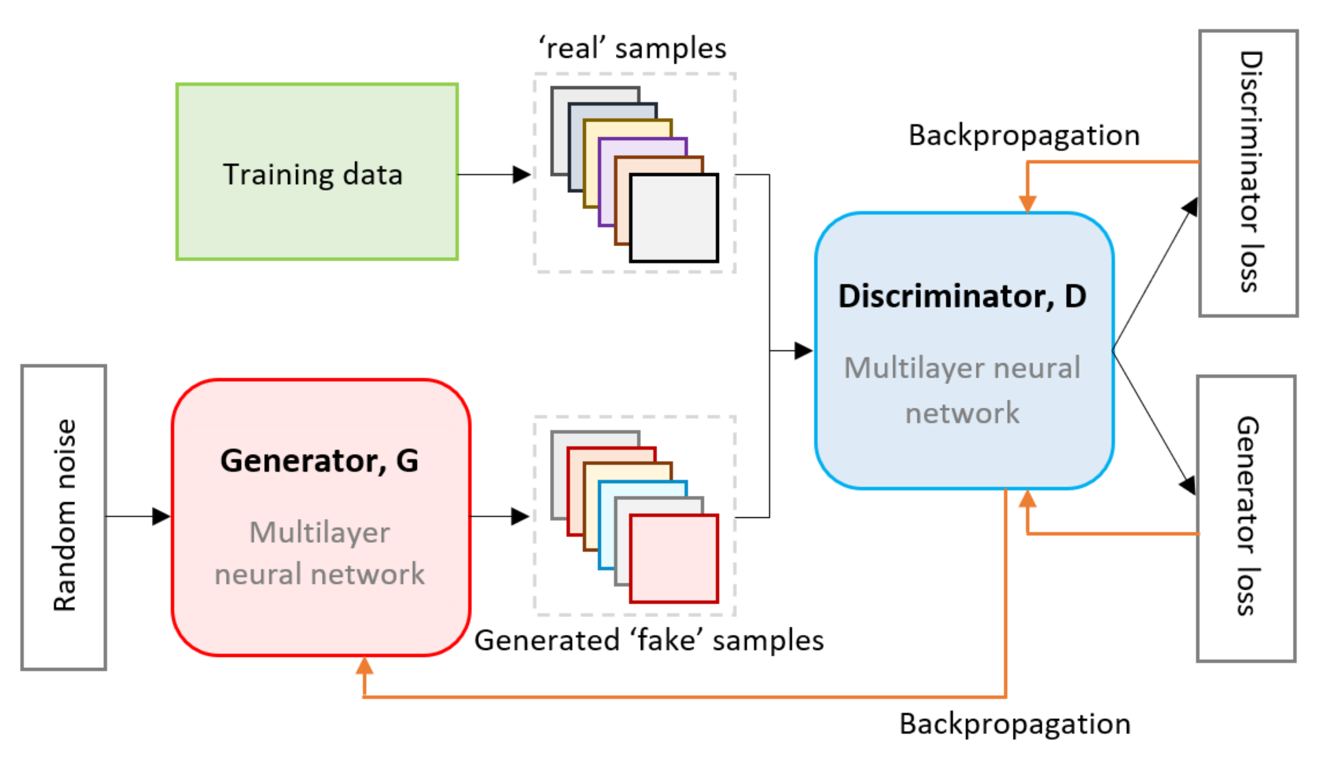
INTRODUCTION

- Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are gaining increasing attention as a means of synthesising data
- GANs have so far been used predominantly for image generation
- Less research into structured microdata synthesis
  - e.g. synthesising census or social survey data
- We compare two GANs with two statistical methods:
  - generate synthetic census data
  - perform analysis using disclosure risk and utility metrics

Synthetic images produced by NVIDIA's Style-Based GAN (Karras et al, 2019)

GENERATIVE ADVERSARIAL NETWORKS (GANS)

- Composed of two neural networks
  - Generator, *G*
  - Discriminator, *D*
- Discriminator aims to determine whether a sample of data is from the real distribution or whether it was generated by *G*
- Generator creates data samples in order to fool the discriminator
  - Generator never sees the original data and learns only from error
- Performance improves over time
- Success if the discriminator cannot determine fake from real data



Example of GAN architecture

STUDY DESIGN

- Census data
  - 1991 Individual Sample of Anonymised Records (SAR) for the British Census (ONS 2013), a 2% sample (1,116,181 records) including adults and children
  - We subsetted one geographical region (n=104,267, 9.34% of total)
  - Twelve variables used (11 categorical, 1 numerical)

Area	Age	Sex	Marital Status	Economic group	Ethnic group	Birth country	Tenure	Social class	Long term ill	Num quals	Family type
Birmingham	28	F	Single	Employed ft	White	England	Rent LA	Skilled	No	one	Lone no dep. child
Walsall	10	M	Single	NA	Indian	England	Rent private	NA	No	none	Married dep. child
Dudley	78	M	Married	Retired	White	Scotland	Own outright	NA	Yes	none	Married no dep child

STUDY DESIGN

- Synthesis Methods
  - Statistical
    - Synthpop (Nowok et al. 2016) – CART based
    - DataSynthesizer (Ping et al. 2017, Zhang et al., 2017) – uses Bayesian networks
  - GAN
    - CTGAN (Xu et al. 2019)
    - TableGAN (Park et al. 2018)

All methods used default parameters and generated synthetic data the same size as original dataset (n=104,267)

STUDY DESIGN

- Metrics
  - Disclosure risk
    - Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
    - Provides a score between 0 and 1
      - Higher value implies higher risk
  - Utility
    - Propensity mean squared error (pMSE) (Snoke et al. 2018, Woo et al. 2009)
    - Confidence interval overlap (CIO)
    - Ratio of estimates (ROE)
- Risk-Utility comparison
    - R-U confidentiality map (developed by Duncan et al. 2004)
      - plots overall utility score against TCAP (risk) score
    - Ideally disclosure risk is minimised and utility is maximised

STUDY DESIGN

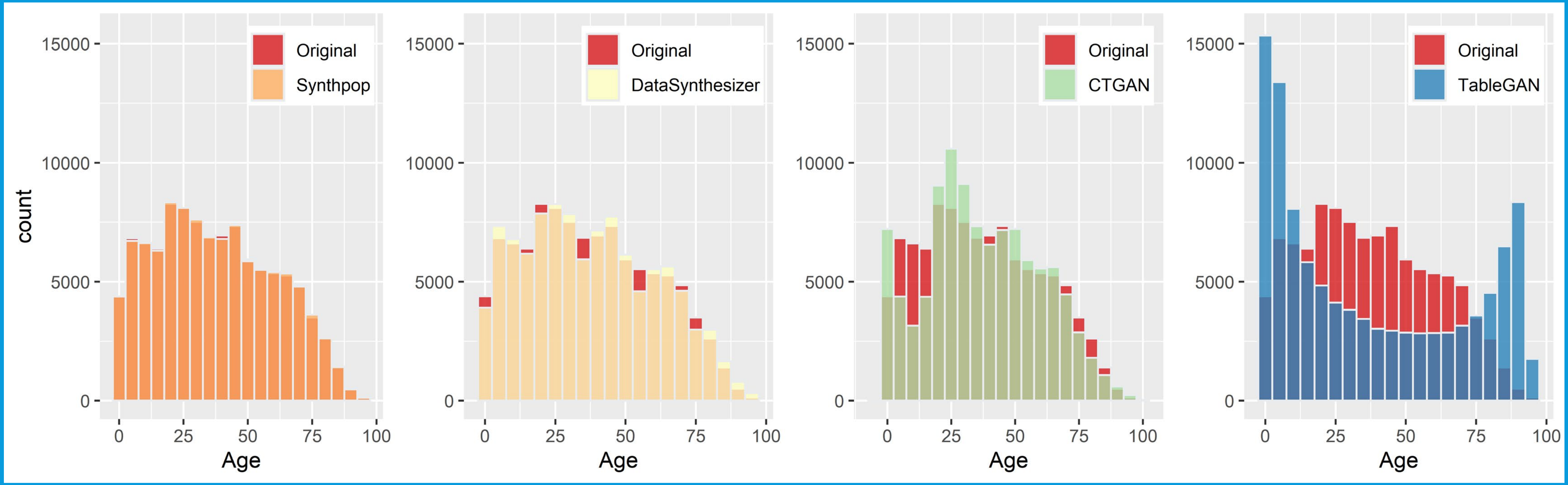
- Metrics
  - Disclosure risk
    - Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
    - Provides a score between 0 and 1
      - Higher value implies higher risk
  - Utility
    - Propensity mean squared error (pMSE) (Snoke et al. 2018, Woo et al. 2009)
    - Confidence interval overlap (CIO)
    - Ratio of estimates (ROE)
- Risk-Utility comparison
  - R-U confidentiality map (developed by Duncan et al. 2004)
    - plots overall utility score against TCAP (risk) score
  - Ideally disclosure risk is minimised and utility is maximised

STUDY DESIGN

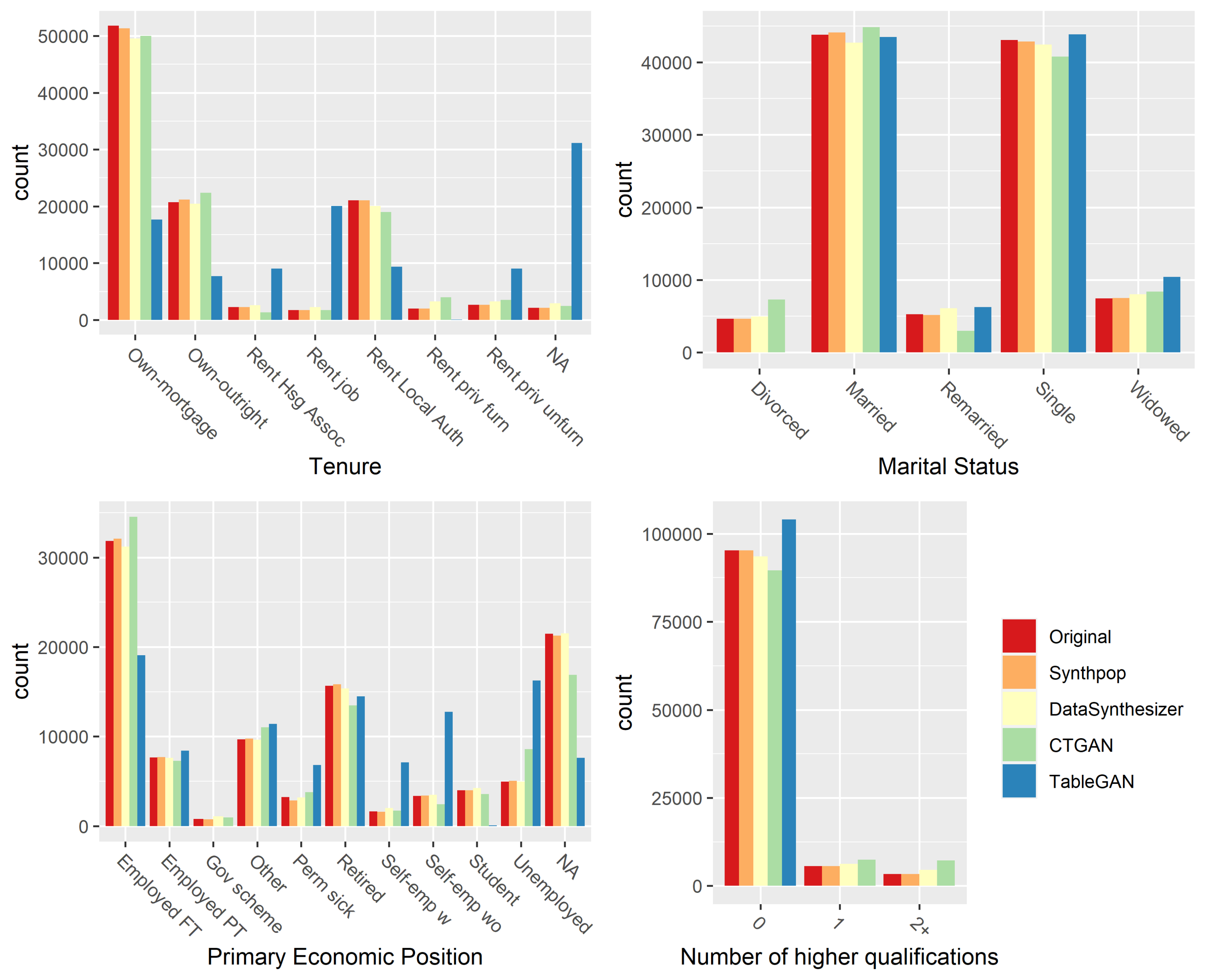
- Metrics
  - Disclosure risk
    - Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
    - Provides a score between 0 and 1
      - Higher value implies higher risk
  - Utility
    - Propensity mean squared error (pMSE) (Snoke et al. 2018, Woo et al. 2009)
    - Confidence interval overlap (CIO)
    - Ratio of estimates (ROE)
- Risk-Utility comparison
    - R-U confidentiality map (developed by Duncan et al. 2004)
      - plots overall utility score against TCAP (risk) score
    - Ideally disclosure risk is minimised and utility is maximised

RESULTS

Histograms comparing original data with synthetic data for age



Synthpop closely matched the age distribution whilst TableGAN struggled



Bar graphs comparing original to synthetic data

Data produced by Synthpop and DataSynthesizer had similar counts to the original data. TableGAN did not manage to identify all categories

RESULTS

The basket of utility metrics

Metric	Synthpop	DataSynthesizer	CTGAN	TableGAN
pMSE	**0.00015**	0.01438	0.03162	0.17529
1 - (4 x pMSE)	**0.9994**	0.9425	0.8735	0.2988
ROE univariate (mean)	**0.981**	0.821	0.743	0.499
ROE bivariate (mean)	**0.847**	0.616	0.587	0.255
CI Overlap (mean)	**0.506**	0.365	0.410	-
Overall utility	**0.833**	0.686	0.653	0.351

Synthpop had optimal results for all metrics

RESULTS

TCAP scores for the synthetic methods, four key sizes

Target	Key	Synthpop	DataSynthesizer	CTGAN	TableGAN	Baseline
LTILL	6	0.935	0.929	0.912	0.911	0.774
	5	0.897	0.898	0.891	0.907	
	4	0.894	0.899	0.889	0.907	
	3	0.936	0.951	0.931	0.901	
FAMTYPE	6	0.709	0.623	0.598	0.301	0.223
	5	0.725	0.658	0.639	0.384	
	4	0.736	0.654	0.651	0.416	
	3	0.809	0.608	0.648	0.420	
TENURE	6	0.596	0.429	0.490	0.217	0.329
	5	0.504	0.376	0.453	0.336	
	4	0.500	0.350	0.447	0.341	
	3	0.496	0.353	0.482	0.279	
Average		**0.728**	**0.644**	**0.669**	**0.527**	0.442

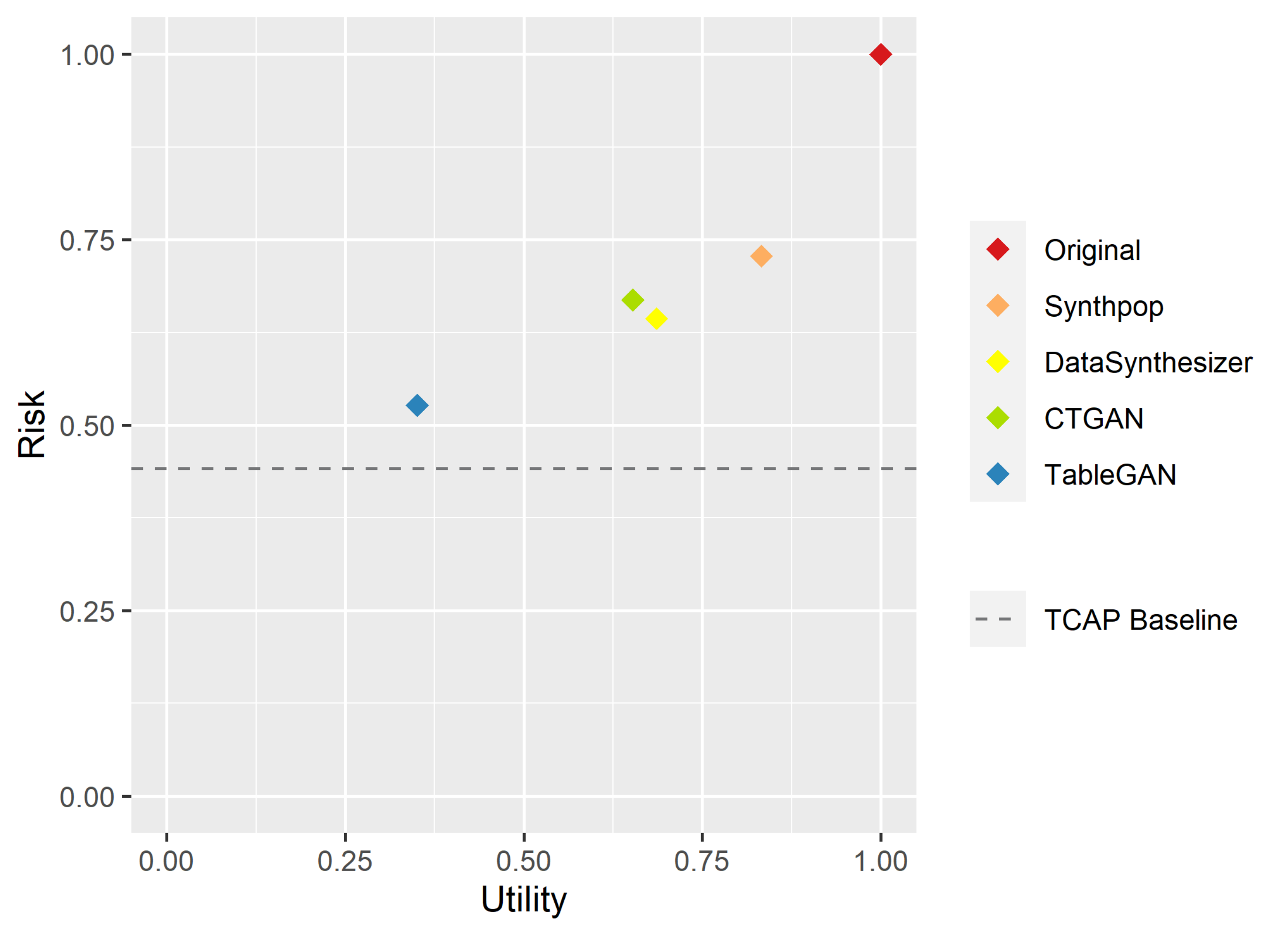
Synthpop had highest disclosure risk, TableGAN had the lowest

RESULTS

RU Confidentiality map and table of results

	Utility (overall)	Risk (TCAP)
Synthpop	0.833	0.728
DataSynthesizer	0.686	0.644
CTGAN	0.653	0.669
TableGAN	0.351	0.527

Risk-Utility relationship appears to approximately follow a straight line – excluding the original data



CONCLUSIONS

- Trade-off between utility and disclosure risk appears to fall on a relatively straight line
- Synthpop showed both highest utility and disclosure risk
- TableGAN had lowest disclosure risk but with unacceptably low data quality

- Methods only tested on a single dataset
- Methods only tested on a subset of records
- Bucket of analyses for the utility tests needs expanding
- Default parameters used for each method

FUTURE WORK

- Much wider range of tests examining effects of parameter changes on the RU map
- Investigating other GAN architectures
- Investigating whether any method can effectively optimise both risk and utility
- Testing on larger datasets (number of variables and cases) and determining scalability of methods

THANK YOU FOR LISTENING!