Do samples taken from a synthetic microdata population replicate the relationship between samples taken from an original population?

MARK ELLIOT, CLAIRE LITTLE, RICHARD ALLMENDINGER

UNIVERSITY OF MANCHESTER

MANCHESTER
1824
The University of Manchester

Introduction

Is the relationship between:

- a population dataset and samples drawn from it

replicated by

- a synthetic version of the same population and samples drawn from it?

Population data usually unavailable - if synthetic samples can mimic this relationship, it would be useful

Extends previous work (Little et al., 2022) using samples to determine the sample equivalence of synthetic data to the original dataset

- (to be able to say, for example, “the synthetic dataset has utility equivalent to a 10% original sample and risk equivalent to a 5% original sample”)

Study Design - Data

UK 1991 Census microdata (University of Manchester, 2023) is used to represent the population

- subsetted on geographical region (West Midlands)
- 104267 records
- 15 variables (13 categorical, 2 numerical)

Area	Age	Country of birth	Economic group	Ethnic group	Family type	Hours worked	Long term illness	Marital status	Num qualifications	Relationship	Sex	Social class	Transport to work	Housing tenure
Sandwell	7	England	NA	Bangladeshi	Married dep. Children	NA	No	Single	None	Child	M	NA	NA	Own outright
Coventry	40	England	Employee FT	White	NA	50	No	Married	None	NA	F	Manag. tech	Car	NA
Walsall	70	England	Retired	White	Married no children	39	Yes	Married	None	Household head	M	Part skilled	NA	Own buying

Study Design

synthpop (Nowok et al. 2016) used to generate synthetic data

- Default parameters
- Visit sequence ordered by ascending number of categories, with numerical variables first

Data samples were drawn randomly without replacement

Various sample fractions

- 0.1%, 0.25%, 0.5%, 1%, 2%, 3%, 4%, 5%, 10%, 20%, …, 80%, 90%, 95%, 96%, 97%, 98%, 99%
  - 22 overall
- n = 100 samples randomly drawn for each sample fraction
- 2200 samples

Study Design – Metrics

Disclosure Risk

- For synthetic data reidentification risk not meaningful
- Attribution is possible
- Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
  - Probability that an intruder makes a correct attribution inference about a particular target variable, given partial knowledge (key variables)
- We use marginal TCAP score
  - Calculate baseline – probability of intruder being correct if they drew randomly from univariate distribution of target variable
  - Scale TCAP score between baseline and 1
  - marginal TCAP indicates risk above the baseline
    - Value between -x and 1, where a higher value indicates greater risk

Study Design – Metrics

Utility

- Confidence Interval Overlap (CIO) (Karr et al., 2006)
  - Logistic regressions performed on synthetic and original data (using same target/predictors for each)
  - Regression coefficients are compared
  - Score between 0 (no overlap) and 1
- Ratio of Counts/Estimates (ROC)
  - For univariate and bivariate cross-tabulations
  - Compares proportion of synthetic and original data estimates by taking the ratio
  - Score between 0 and 1
- Overall utility score
  - Mean of CIO, ROC univariate and ROC bivariate
    - Value between 0 and 1, where a higher value indicates greater utility

Study Design – Metrics

Risk-Utility comparison

- R-U confidentiality map (developed by Duncan et al. 2004)
- Plots utility against risk (TCAP) score
- Ideally disclosure risk is minimised, utility is maximised

Synthetic / Sample data

- Utility and risk metrics calculated in the same way for samples of original data as for samples of synthetic data
  - By comparing against the dataset that the samples were drawn from
- Allows comparison on R-U map

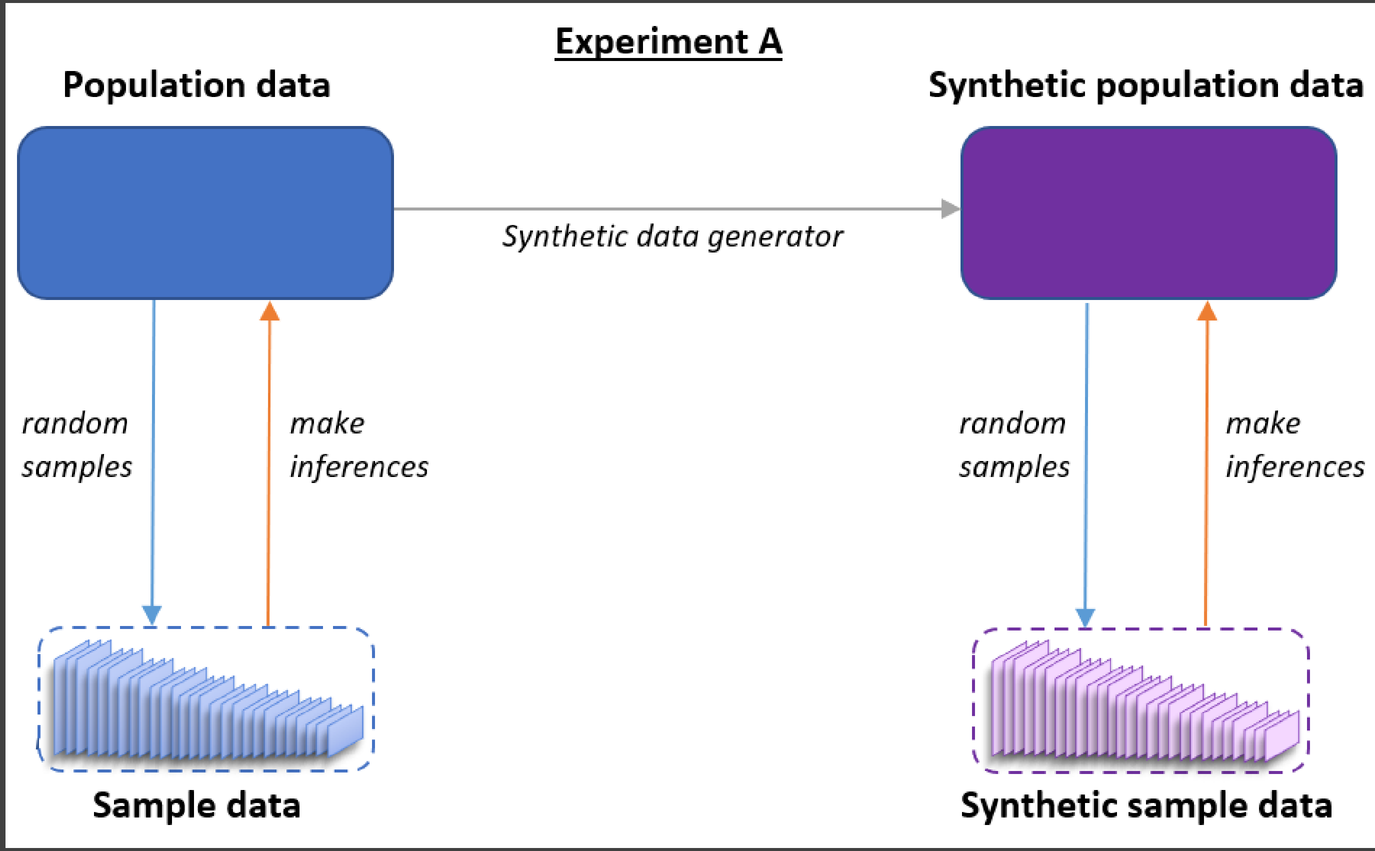
Results - Experiment A

A synthetic population was generated from the original population

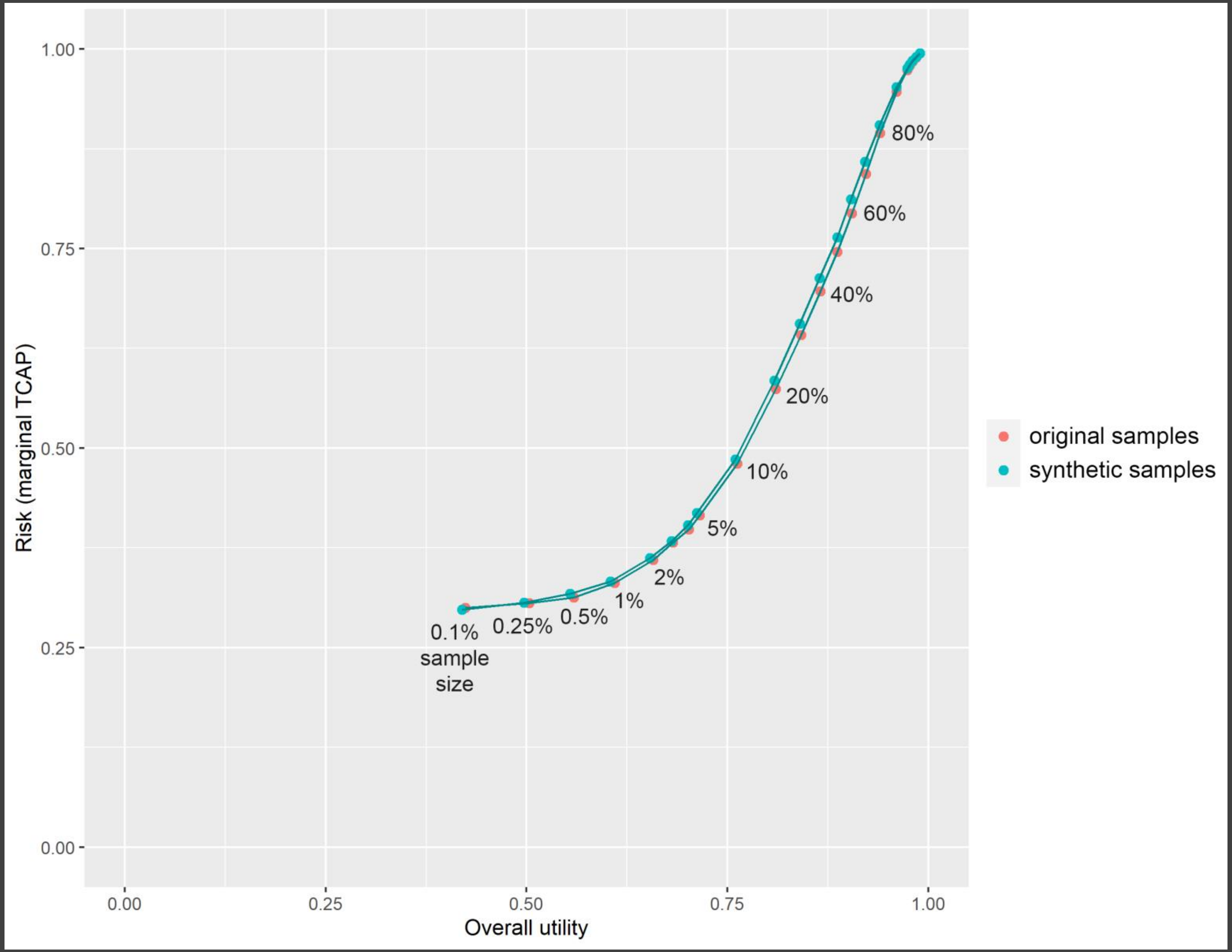
Random samples taken from both populations

Risk and utility calculated for each sample compared to the population it was sampled from

Results compared

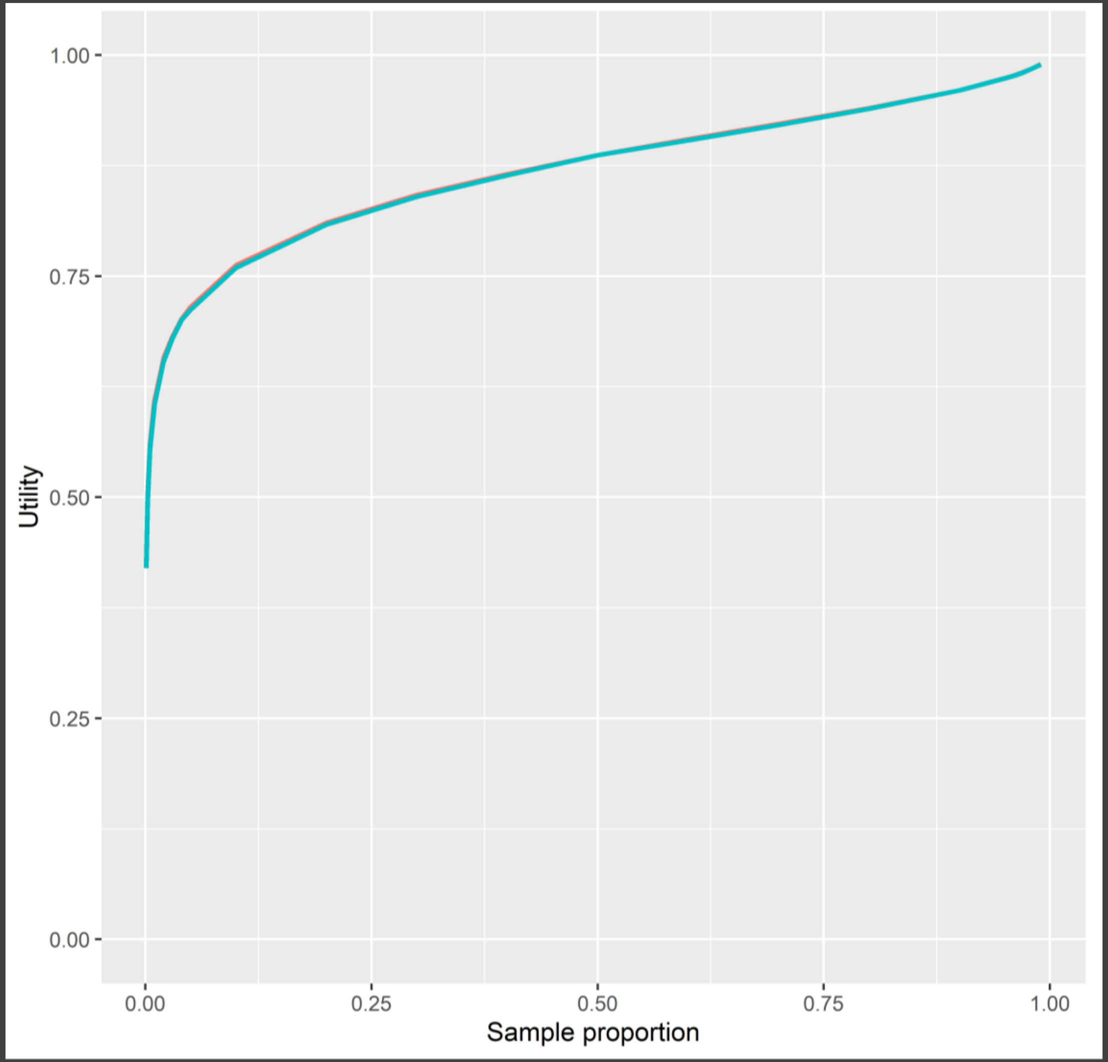


Experiment A: Risk-Utility map showing the original samples and synthetic samples

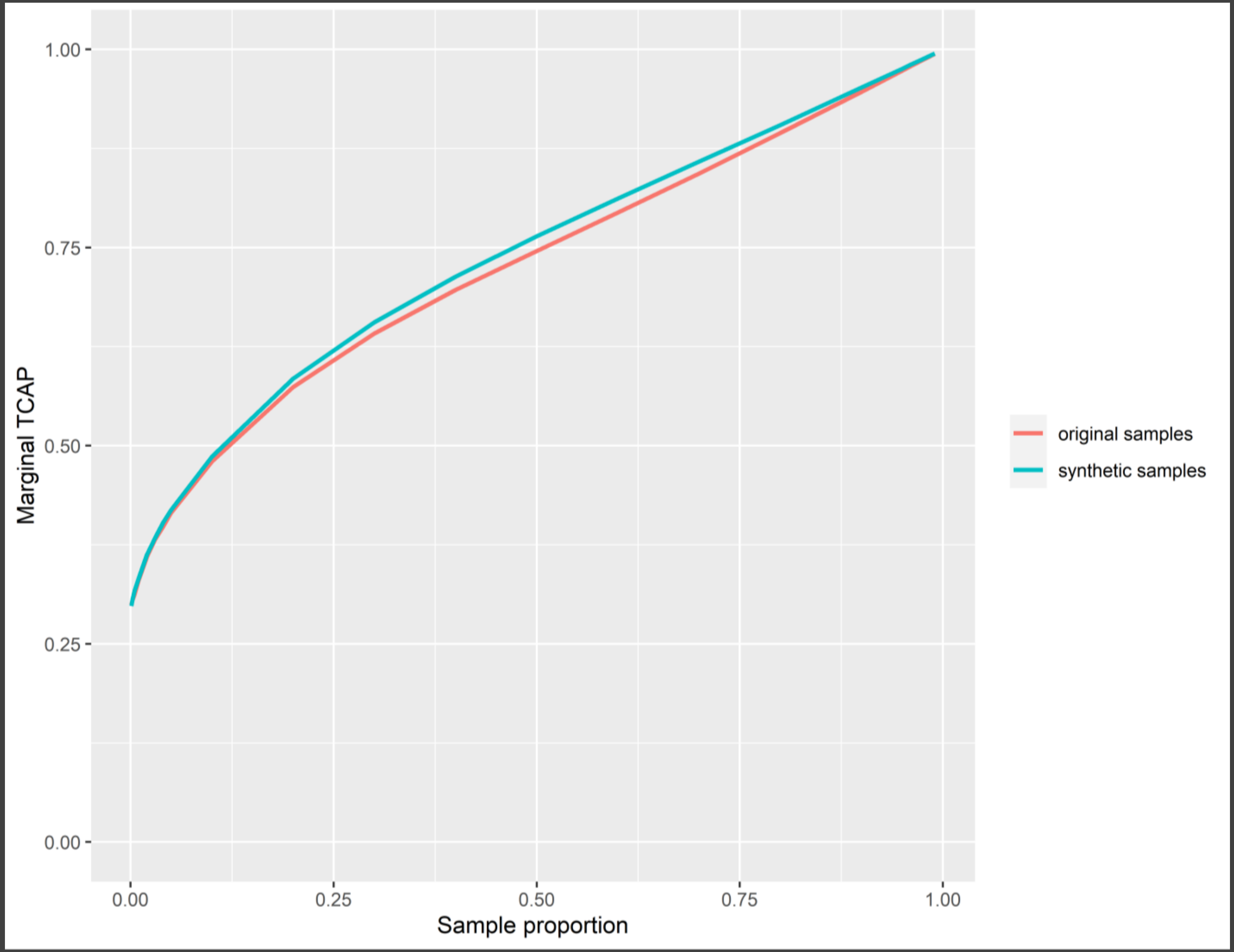


Experiment A: Individual plots showing the original samples and synthetic samples for:

Utility

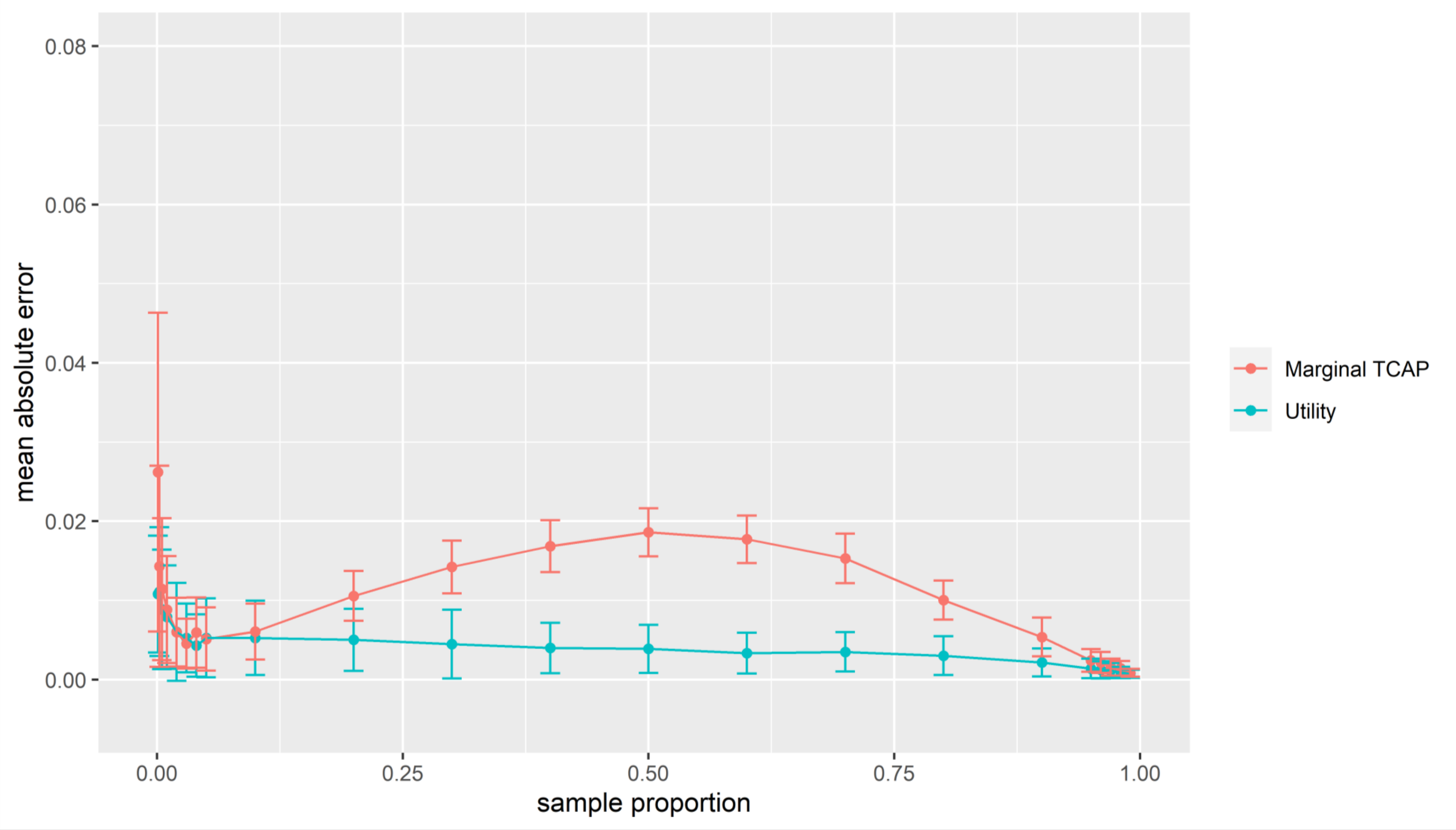


Risk (Marginal TCAP)



Mean Absolute Error of the utility and marginal TCAP for each synthetic sample size

(calculated against the original samples, error bars show +- 1 standard deviation)



Results - Experiment B

UK 1991 Census data represents the population

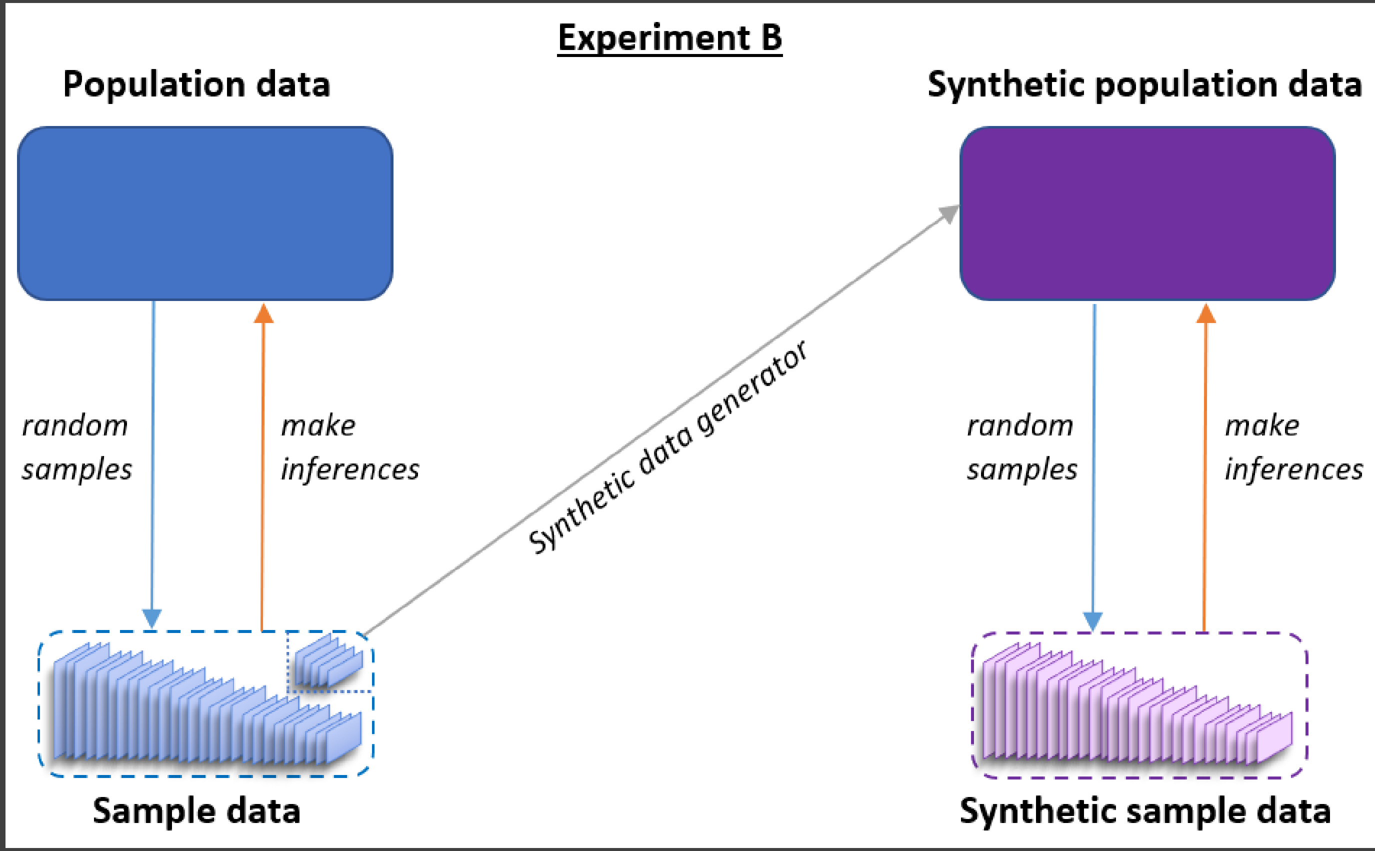
Take samples from the population (1%, 2%, 3%, 4%, 5%)

Generate synthetic populations from the samples

Random samples taken from original and synthetic populations

Risk and utility calculated for each sample compared to the population it was sampled from

Results compared



Experiment B

Synthetic population generated from smaller samples

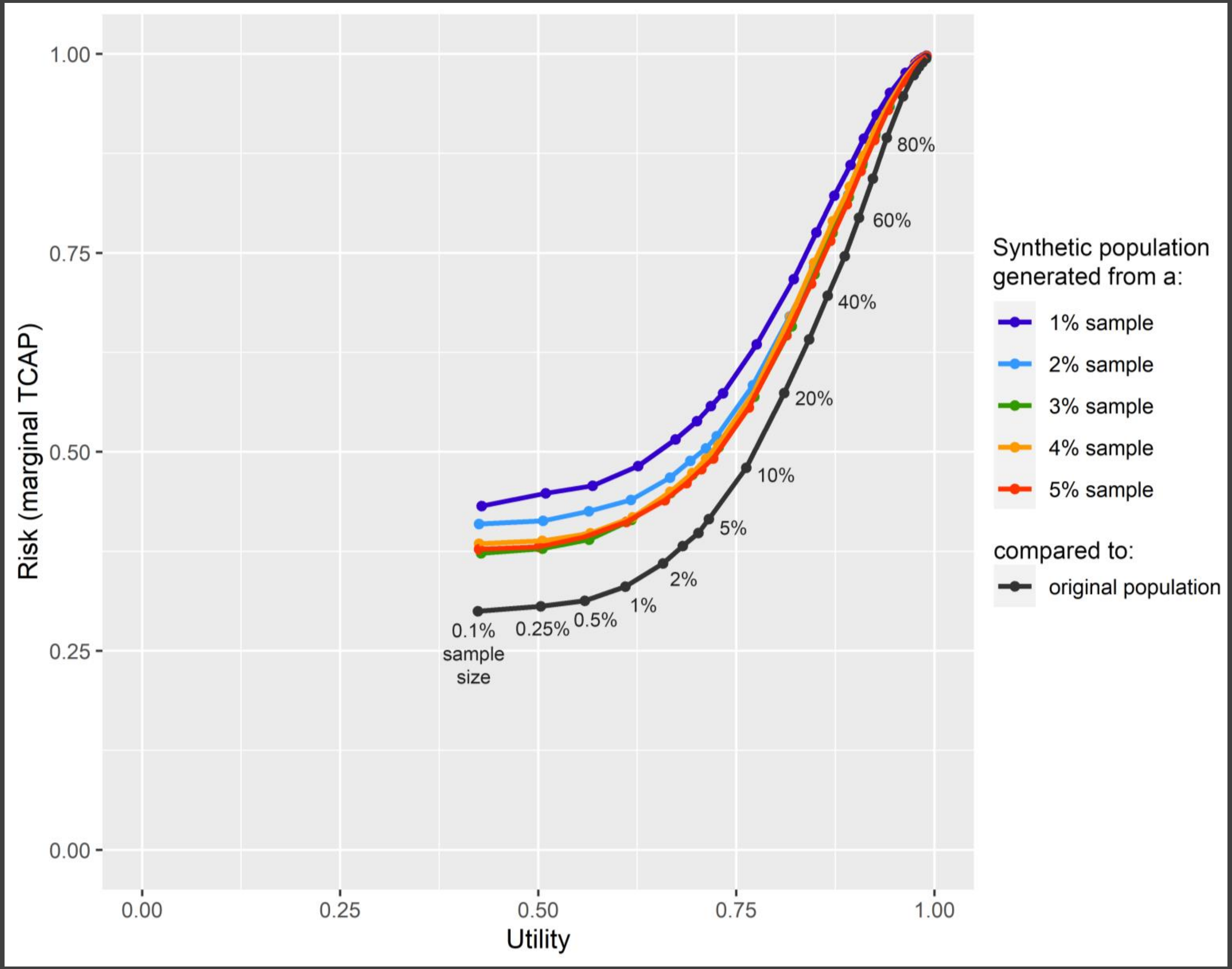
- A more likely scenario

Process:

- Take samples from the original population
  - 1%, 2%, 3%, 4%, 5%
- From each sample, a synthetic dataset the same size as the population (n=104267) was generated
  - Utility increases with sample size
  - TCAP differs

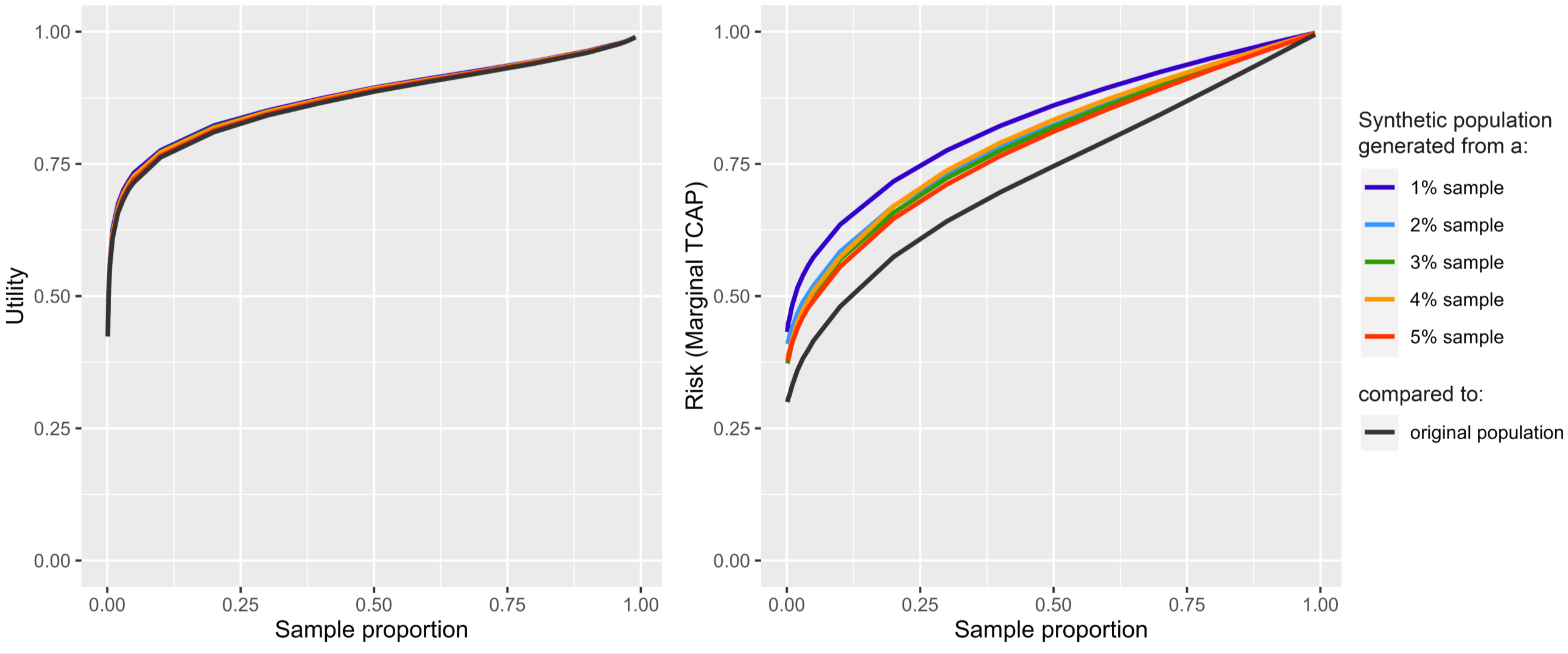
Synthetic population generated from a:	Utility	Marginal TCAP
1% sample	0.539	0.407
2% sample	0.585	0.351
3% sample	0.591	0.370
4% sample	0.616	0.409
5% sample	0.643	0.423

Risk-Utility map contrasting the results for samples drawn from synthetic populations to those drawn from original population



Individual plots contrasting the results for samples drawn from synthetic populations to samples drawn from the original population, for:

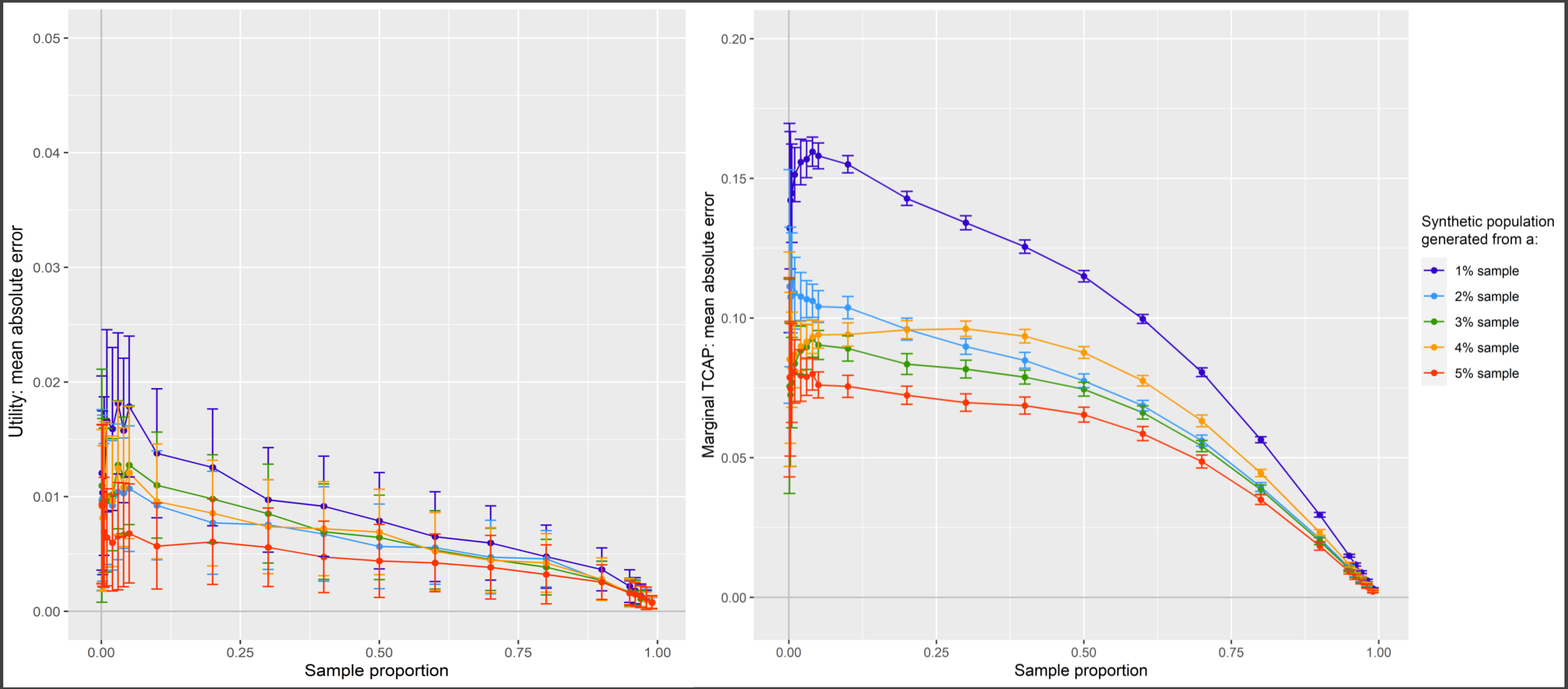
Utility　　　　　　　　　　　　　　Risk (Marginal TCAP)



Mean Absolute Error of the utility and marginal TCAP for each synthetic sample size
(calculated against the original samples, error bars show +- 1 standard deviation)

Utility

Risk (marginal TCAP)

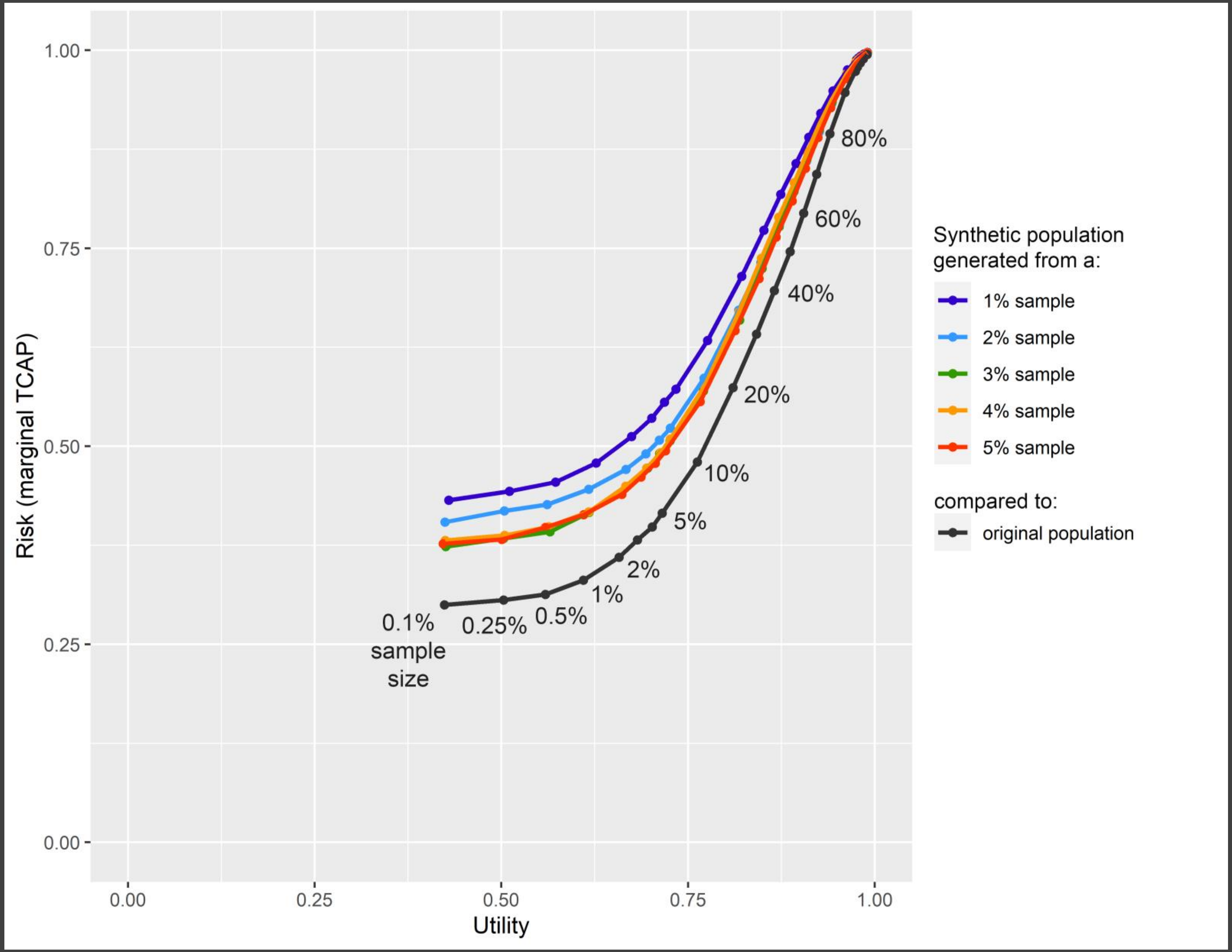


An aside:

Risk-Utility map contrasting the results for samples drawn from synthetic populations to those drawn from original population…

where the synthetic population also contains the original sample used to generate it

- very little difference whether or not the original sample is included



Observations

Experiment A → Synthetic population generated from original population

- Relationship between synthetic samples and the synthetic population follows closely the relationship between original samples and the original population

Experiment B → Synthetic populations generated from samples drawn from original population

- Overall relationship similar to original populations results (similar curve on the RU map)
- But the smaller the original sample (used to generate the synthetic population) the more the risk is overestimated
- Utility similar no matter the original sample size

Caveats

Experiments conducted on samples of Census microdata
- May not generalise to full population data

Only one data synthesis method used
- Synthpop – which tends to create high utility (but also higher risk) synthetic data

Only one dataset used
- It may be useful to repeat this on other datasets

Underestimation of the risk of samples, relative to synthetic data
- Whilst synthetic data should not contain re-identification risk, sample data does

Risk measure uses a response knowledge attribution disclosure
- OK for Census data, but presence detection may be a significant risk in other data

Different risk and utility metrics may produce different results

Future Work

Run experiments on full population data

Use different data synthesis methods

Use different datasets

Assess other utility measures

Assess other disclosure control methods

References

Nowok, B., Raab, G.M. and Dibben, C., 2016. synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, *74*(1), pp.1-26.

Little, C., Elliot, M. & Allmendinger, R., 2022, Comparing the Utility and Disclosure Risk of Synthetic Data with Samples of Microdata. In *Privacy in Statistical Databases: International Conference, PSD 2022, Paris, France, September 21–23, 2022, Proceedings.* Lecture Notes in Computer Science vol. 13463 LNCS, Springer Nature, Cham, Switzerland, pp. 234-249. https://doi.org/10.1007/978-3-031-13945-1_17

University of Manchester, Cathie Marsh Centre for Census and Survey Research, Office for National Statistics, Census Division. (2023). *Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs)*. [data collection]. UK Data Service. SN: 7210, DOI: http://doi.org/10.5255/UKDA-SN-7210-1

Taub, J., Elliot, M., Raab, G., Charest, A., Chen, C., O'Keefe, C. M., Nixon, M. P., Snoke, J., Slavkovic, A., 2019. The synthetic data challenge. Joint UNECE/Eurostat Work Session on Statistical Data Condentiality. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf

Duncan, G.T., Keller-McNulty, S.A. and Stokes, S.L., 2004. Database security and confidentiality: examining disclosure risk vs. data utility through the RU confidentiality map.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A framework for evaluating the utility of data altered to protect confidentiality. Am. Stat. **60**(3), 224–232 (2006).