UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

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

Mark Elliot, Claire Little and Richard Allmendinger (University of Manchester)
mark.elliot@manchester.ac.uk

Abstract

Assessment of disclosure risk in sample surveys by data controllers who don't have access to the population data are constrained by verifiability challenges. A sample unique may not be population uniques. Statistics generated at the sample level may not carry over to the population level. Privacy models such as k-anonymity simply may not make sense when applied to sample data (or only make sense for some scenarios) This study aims to understand whether samples generated from a synthetic population present the same relationship, in terms of risk and utility, to the synthetic population, as samples generated from the original population. Note that this is a very different question from the more general questions about the utility of synthetic data which compares the synthetic and original data. Here we are comparing two relationships. This opens the possibility of being able to test and set parameters for models of risk assessment to be applied to real data using synthetic data.

1 Introduction

This document explores whether the relationship between a population dataset and samples drawn from it is maintained when the samples are drawn from (and compared to) a synthetic version of the same population. This extends the work of Little et al. (2022), where samples were used to determine the sample equivalence of synthetic data to the original dataset (for example, to be able to say "the synthetic dataset has utility equivalent to a 10% original data sample and risk equivalent to a 5% sample"). In real-life scenarios the population data may not be available, so if synthetic samples were able to mimic this relationship, it would be useful.

As visualised below, two scenarios are explored: Experiment A (Figure 1), where we do not have access to the original population data but have a synthetic dataset generated from it that is the same size as the original population; and Experiment B (Figure 2), where we have a sample of the original population dataset and from that create a larger synthetic population. An extension to Experiment B (named B2) is to include the original sample within the synthetic population.

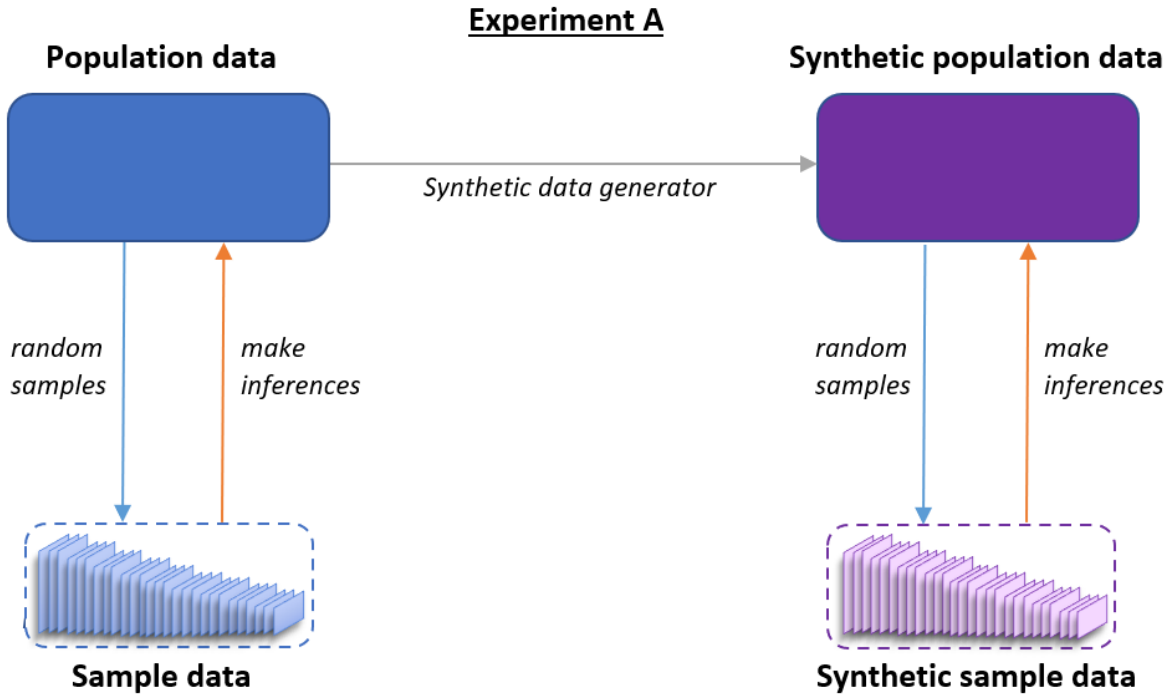


Figure 1: Diagram of data relationships for Experiment A

Experiments were performed using the UK 1991 Census dataset (although it may make sense to repeat these experiments on other Census datasets in the future). The synthetic data was generated using Synthpop (Nowok et al., 2016). This was selected because in previous experiments it produced data with the highest utility compared to other methods (although it should be noted this came with higher disclosure risk). It may make sense to also experiment with other methods in the future.

The next section introduces the dataset and data/sample generation approach adopted in this study. Section 3 describes the risk and utility measures used, and Section 4 presents an analysis of Experiment A and B. Finally, Section 5 concludes the paper and discusses areas for future research.

2 Data

2.1 UK 1991 Census

A subset of the UK 1991 Individual Sample of Anonymised Records for Great Britain (SARs) was used to simulate a population. The SARs data was downloaded from the UK Data Service on

29/05/21.[1] This consists of a 2% sample of the population of Great Britain (excluding Northern Ireland), with 1,116,181 individual records and 67 attributes. The dataset includes children and adults and contains information on topics such as age, gender, ethnicity, employment, and housing. To reduce the computational load the data was subsetted on geographical region (the REGIONP attribute); there are 12 regions, and the West Midlands was randomly selected for use in this study. Details of each of the variables are contained in Appendix A. The subset consisted of 104,267 records (9.34% of overall sample) and fifteen variables (thirteen categorical, two numeric). This subset will be henceforth referred to as the *original population.*

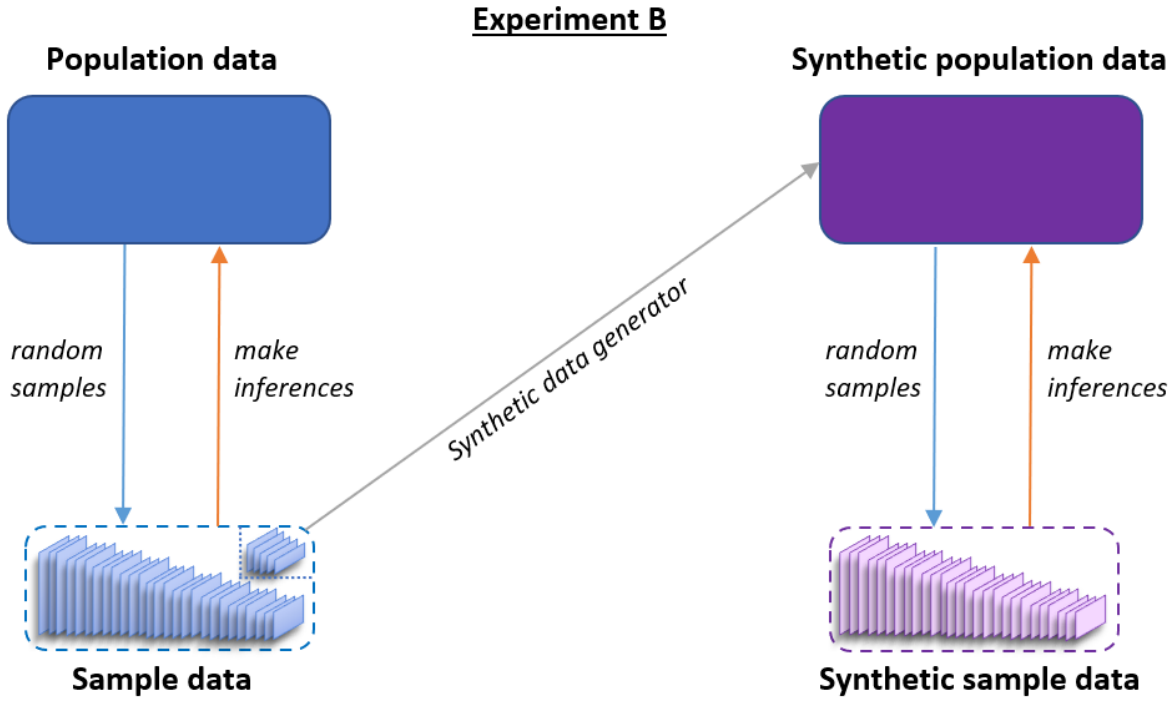


Figure 2: Data relationships for Experiment B

2.2 Synthetic Data Generation

Synthpop, developed by Nowok et al. (2016), was used to generate the synthetic data. Synthetic data the same size as the original population (104,267 records) was generated. Default parameters were used, with the visit sequence ordered with numerical variables first, followed by categorical variables with least number of categories to most (with ties decided alphabetically). That gave a visit sequence of: AGE, HOURS, LTILL, SEX, QUALNUM, MSTATUS, TENURE, RELAT, FAMTYPE, SOCLASS, ECONPRIM, ETHGROUP, TRANWORK, AREAP, COBIRTH.

2.3 Sample Generation

Random samples of sizes 99%, 98%, 97%, 96%, 95%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.25%, 0.1% were drawn (without replacement) from both the original and synthetic populations. For each sample size 100 samples were drawn. This follows the framework developed in earlier experiments (as reported in Little et al., 2022).

3 Risk and Utility Measures

For calculating the associated risk and utility the sample datasets were measured against the population dataset. That is, the synthetic samples were measured against the synthetic population they

[1] Study Number 7210 (Office for National Statistics, Census Division, University of Manchester, Cathie Marsh Centre for Census and Survey Research 2013).

were sampled from, and the original samples were measured against the original population that they were sampled from. Risk-Utility (R-U) maps, as developed by Duncan et al. (2004), were used to visualise the trade-off between risk and utility.

3.1 TCAP for disclosure Risk

Elliot (2014) and Taub et al. (2018) introduced a measure for the disclosure risk of synthetic data called the Correct Attribution Probability (CAP) score. The disclosure risk is calculated using an adaptation used in Taub et al. (2019) called the Targeted Correct Attribution Probability (TCAP). TCAP is based on a scenario whereby an intruder has partial knowledge about a particular individual. Specifically, they know (i) the values for some of the variables in the dataset (the keys) and (ii) that the individual is in the original dataset. We assume that the intruder wishes to infer the value of a sensitive variable (the target) for that individual. The TCAP metric is then the probability that those matched records yield a correct value for the target variable (i.e., that the adversary makes a correct attribution inference).

Three target variables, and corresponding key variables were identified from the UK Census data. For each target, the TCAP score was calculated using sets of 3, 4, 5 and 6 keys. The overall mean of the TCAP scores (for each of the target and key combinations) was calculated as the overall disclosure risk score.

The TCAP statistic has a value between 0 and 1; a low value indicates that the synthetic dataset carries little risk of disclosure whereas a score close to 1 indicates a higher risk. A baseline value can be calculated (the usual one being the probability of the intruder being correct if they drew randomly from the univariate distribution of the target variable) and then the TCAP score is rescaled so that the baseline equals zero.[2] We refer to the rescaled TCAP value as the marginal TCAP, i.e., it is the increase in risk above the baseline. Rescaling is performed by subtracting the baseline from the TCAP score and then dividing by 1 minus the baseline. For all experiments the targets were:

- LTILL : baseline = 0.774
- FAMTYPE : baseline = 0.223
- TENURE : baseline = 0.329

With a mean baseline of 0.442. The keys for each were:

- 6 keys: AREAP, AGE, SEX, MSTATUS, ETHGROUP, ECONPRIM
- 5 keys: AREAP, AGE, SEX, MSTATUS, ETHGROUP
- 4 keys: AREAP, AGE, SEX, MSTATUS
- 3 keys: AREAP, AGE, SEX

3.2 Utility

Following previous work (Little et al. 2022) the mean of the Ratio of Counts (ROC) and Confidence Interval Overlap (CIO) was calculated as the overall utility score. This was to provide a more complete view of the utility, rather than just using a single measure.

3.2.1 Ratio of Counts (ROC)

The Ratio of Counts (ROC) was calculated for univariate and bivariate cross tabulations of the data. This is calculated by taking the ratio of the synthetic and original data estimates (where the smaller is divided by the larger one). Thus, given two corresponding estimates (for example, the number of records with SEX = female in the original dataset, compared to the number in the synthetic dataset), where y_{orig} is the estimate from the original data and y_{synth} is the corresponding estimate from the synthetic data, the ROC is calculated as:

$$\mathrm{ROC} = \frac{\min(y_{orig}, y_{synth})}{\max(y_{orig}, y_{synth})}$$

[2] This does create the possibility of a synthetic dataset receiving a negative TCAP score (which can still be plotted on the R-U map) but that simply indicates a risk level below that of the baseline and will only occur in degenerate cases.

If $y_{orig} = y_{synth}$ then the ROC = 1. Where the original and synthetic (or sample) datasets are of different sizes (as is the case when calculating the ROC for the various sample datasets) the proportion, rather than the count can be used. The ROC was calculated over univariate and bivariate cross-tabulations of the data and takes a value between 0 and 1. For each variable the ROC was averaged across categories to give an overall score.

3.2.2 Confidence Interval Overlap (CIO)

To calculate the CIO (using 95% confidence intervals), the coefficients from regression models built on the original and synthetic datasets are used. The CIO, proposed by Karr et al. (2006), is defined as:

$$CIO = \frac{1}{2}\{\frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_o - l_o} + \frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_s - l_s}\}$$

where u_o, l_o and u_s, l_s denote the respective upper and lower bounds of the confidence intervals for the original and synthetic/sample data. This can be summarised by the average across all regression coefficients, with a higher CIO indicating greater utility (maximum value is 1 and a negative value indicating no overlap).

For each synthetic (or sample) dataset two logistic regressions were performed, and the CIO (between the same regression on the original data) for each was calculated. The mean CIO over all coefficients was used (where a negative overlap was equivalent to no overlap and therefore set to zero). The mean of the two CIOs was then calculated as the overall score.

The target variables were marital status (MSTATUS) and housing tenure (TENURE), and they were converted into a binary attribute: for marital status this was married (or living as married) and anything else; and for tenure this was whether an individual owns their property (or lives in property that is owned by a family member), and anything else. Eight variables were used as predictors, using more would seem to overcomplicate the models. The predictors were: AGE, ECONPRIM, ETHGROUP, LTILL, QUALNUM, SEX, SOCLASS, and TENURE or MSTATUS (whichever was not the target).

4 Results

4.1 Experiment A

The scenario where we do not have access to the original/population data but have a synthetic dataset the same size created from it. This explores using a synthetic dataset to model the relationship between samples and population data. To be clear, throughout this section, the original dataset (the UK 1991 sample, n=104,267) is referred to as the *original population*, and the synthetic dataset created from this is referred to as the *synthetic population*. The samples are referred to as *original samples* and *synthetic samples*.

The synthetic population was created (using Synthpop) from the original population. The synthetic population had utility = 0.7596 and Marginal TCAP = 0.7228 (to 4dp) compared to the original. Samples were drawn from the synthetic population to determine if the results follow the same patterns as samples drawn from the original population. The same sample sizes were used as in previous experiments (0.1%, 0.25%, …, 99%, see Little et al., 2022).

The utility and TCAP scores for each sample size were calculated by measuring against the 100% synthetic population dataset, not the original population since this would not be available in this scenario. The baseline TCAP scores (used for calculating Marginal TCAP) were calculated from the 100% synthetic population, and these vary slightly from the original population:

- Original TCAP baseline = 0.442
- Synthetic TCAP baseline = 0.441

For each sample size 100 datasets were drawn, and the results are the mean of the 100. The risk and utility of the synthetic samples were contrasted with the equivalent results from the original samples. Tables with the mean utility and TCAP scores for each sample size, and the standard deviation (all values less than 0.04) are contained in Appendix B. Figure 3 displays the R-U map for the original sample data at each sample size, together with the results for the synthetic sample data. The plot and tables indicate that the relationship in terms of (risk and utility) between synthetic samples and the

synthetic population follows closely to the relationship between the original samples and original population. However, the synthetic samples have moderately higher risk (particularly around the 50% sample size) and moderately lower utility.

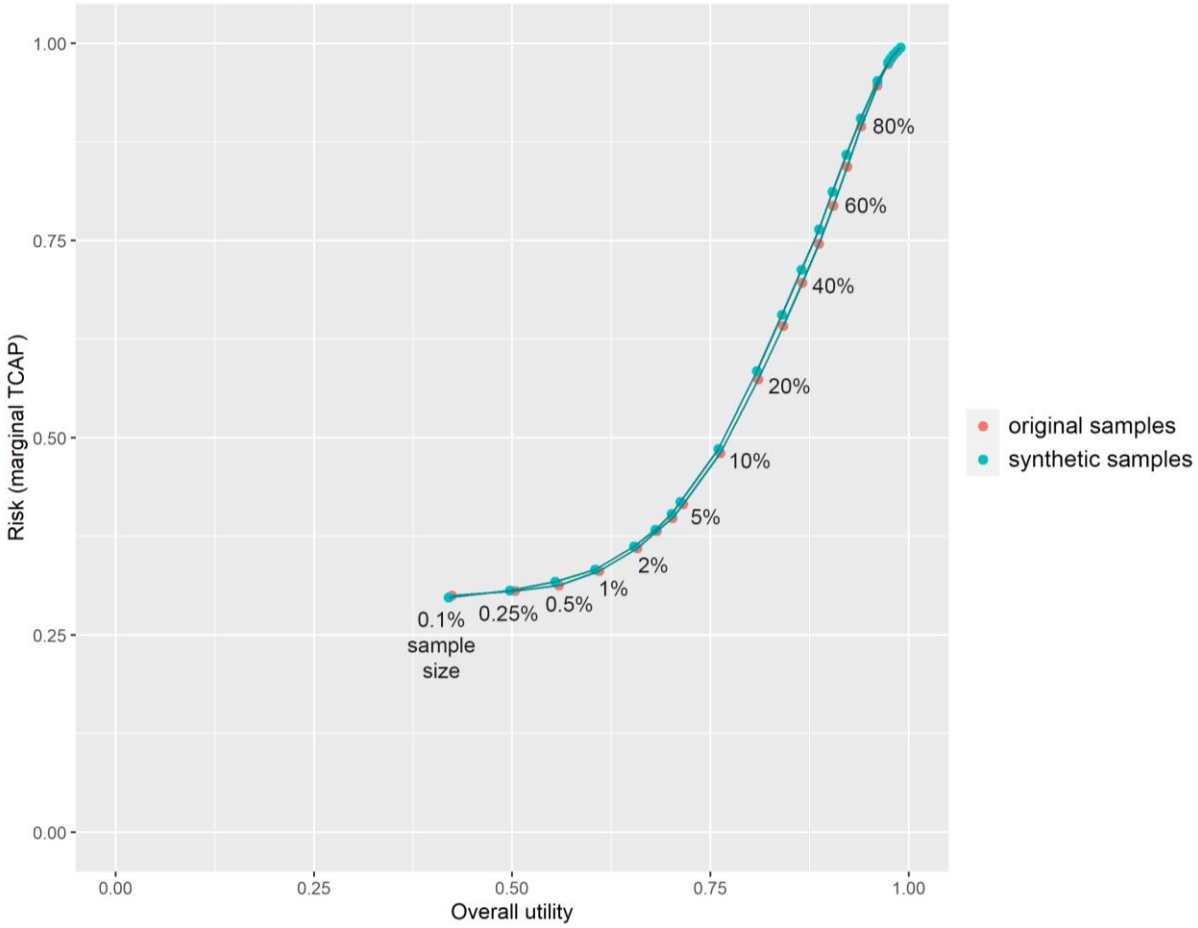


Figure 3: R-U map showing the original samples and the synthetic samples (mean of n=100) in experiment A.

Appendix B contains a table with the mean absolute error (MAE) and standard deviation (SD) of the synthetic utility and TCAP values (when calculated against the original samples), for each sample size. Figure 4 illustrates the values in the table, displaying the MAE of the utility and TCAP scores. It highlights that the MAE in terms of utility is low and generally decreases as sample size increases, whereas whilst the MAE for the TCAP is also low it displays an interesting curve around the 50% point and then decreases beyond that as sample size increases.

4.2 Experiment B

This scenario where the original (UK 1991 Census sample, n=104,267) dataset represents the population, then:

- take smaller samples from the original population (1%, 2%, 3%, 4%, 5%)
- generate synthetic populations (the same size as the original population) from the smaller samples
- then draw multiple samples of different sizes from each synthetic population
- calculate the risk and utility of the samples and contrast with original population samples

This is perhaps the more likely scenario (compared to Experiment A) since we do not usually have access to the population data – it is more likely a small sample will be provided, and we can then use this to generate a synthetic population. From this synthetic population samples can be drawn and the resulting utility and risk of these can be compared to the equivalent results from the original population samples.

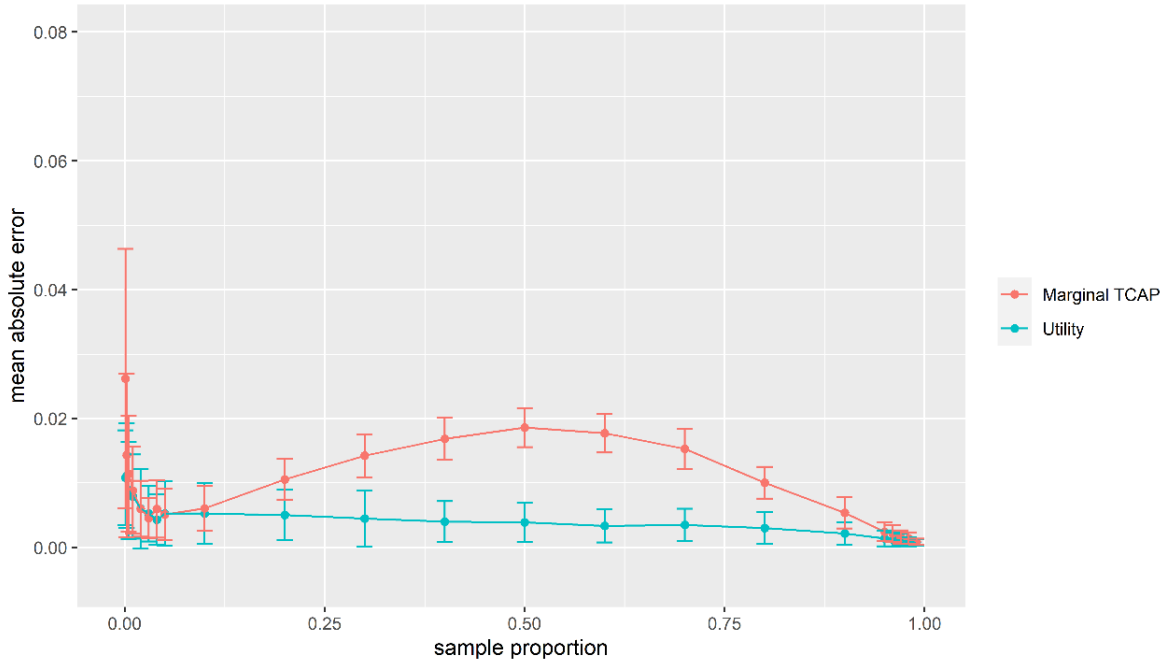


Figure 4: Mean Absolute Error of the utility and marginal TCAP for each synthetic sample size (calculated against the original samples), with error bars shows +-1 standard deviation.

To calculate the utility and risk, the synthetic samples are measured against the synthetic population they were drawn from. They are not measured against the original population as that data would not be available.

4.2.1 Samples to generate the synthetic populations

Sample sizes of 1%, 2%, … 5% were drawn from the population data, Table 1 lists the number of records in each sample. Note that only 1 sample was (randomly) drawn for each size, this is because emanating from each of these individual samples were hundreds of datasets, therefore, to keep complexity down only one of each size was drawn initially.

Table 1: Number of records for each sample size

Sample size	1%	2%	3%	4%	5%
Number of records	1042	2085	3128	4170	5213

Synthpop was used to generate a synthetic population from each sample, using default parameters (and with the visit sequence as detailed in Section 2.2). One synthetic population the same size as the original population (104,267) was generated for each sample; therefore 5 synthetic populations were produced. Table 2 indicates the utility and risk values for each synthetic population measured against the original population. It highlights that (even with these small sample sizes), the utility of a population generated from a smaller sample is lower than the utility of a population generated from a larger sample, as might be expected. The risk (TCAP) exhibits a different pattern, and it is notable that the TCAP score for the synthetic population generated from a 1% sample is higher than that for the 2% and 3% sample populations.

For each of these five synthetic populations, random samples the same size as used in previous experiments (0.1%, 0.25%, …, 99%, see Little et al., 2022) were drawn (without replacement). For each sample size 100 samples were drawn.

Table 2: Utility and risk scores for each synthetic population, to 3dp

Synthetic population generated from a:	Utility	TCAP	Marginal TCAP
1% sample	0.539	0.669	0.407
2% sample	0.585	0.638	0.351
3% sample	0.591	0.648	0.370
4% sample	0.616	0.670	0.409
5% sample	0.643	0.678	0.423

4.2.2 Utility and Risk

Appendix C contains tables with the results for utility and Appendix D for TCAP. To calculate the utility and TCAP the synthetic samples are measured against the synthetic population they were drawn from (they are not compared against the original population as that data would not be available). **Error! Reference source not found.** plots (in the left panel) the utility for each of the synthetic populations at different sample sizes, with the original population plotted for comparison. The plot highlights that, regardless of the synthetic population origin (whether it was generated from a 1% sample of the original population or a 5% sample) the relationship between the utility and the sample proportion is similar.

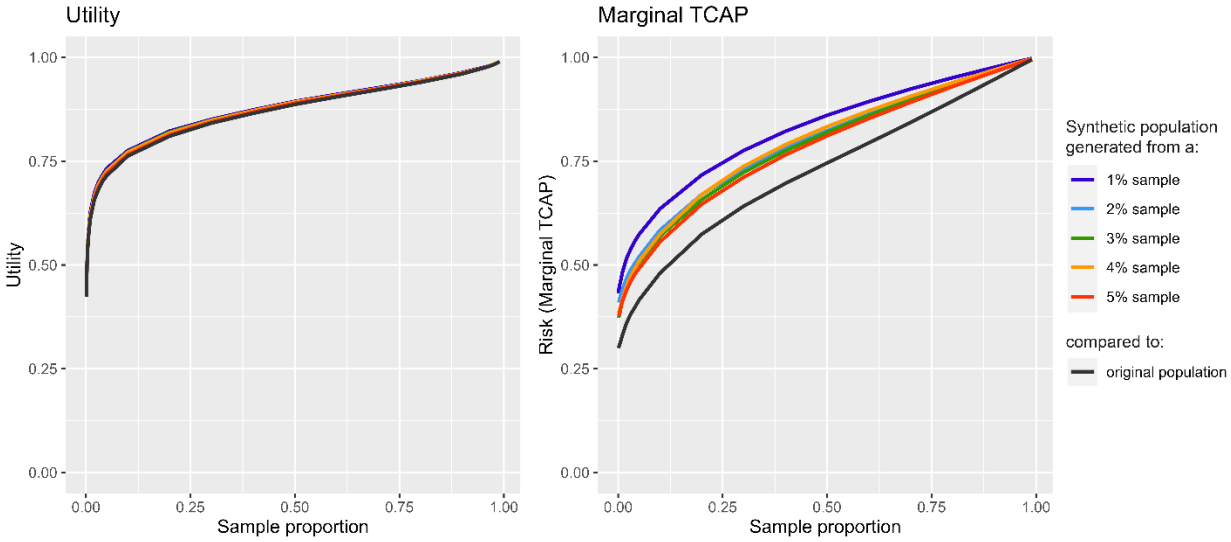


Figure 5:The utility (left) and marginal TCAP (right) for samples drawn from the synthetic populations, contrasted with samples from the original population, in experiment B

The panel on the right in Figure 5 displays the marginal TCAP results for each synthetic population. This illustrates that, whilst they all follow a similar curve, the synthetic samples all overestimate the TCAP compared to the original samples - the samples taken from the synthetic population generated from a 1% sample of the original population particularly so.

The R-U map (plotting the utility against the marginal TCAP) can be visualised for each synthetic population. Figure 6 plots them all in one plot, alongside the original population results. Whilst they all follow a similar pattern, the results from synthetic populations generated from smaller original samples tend to have higher TCAP values than those generated from larger samples.

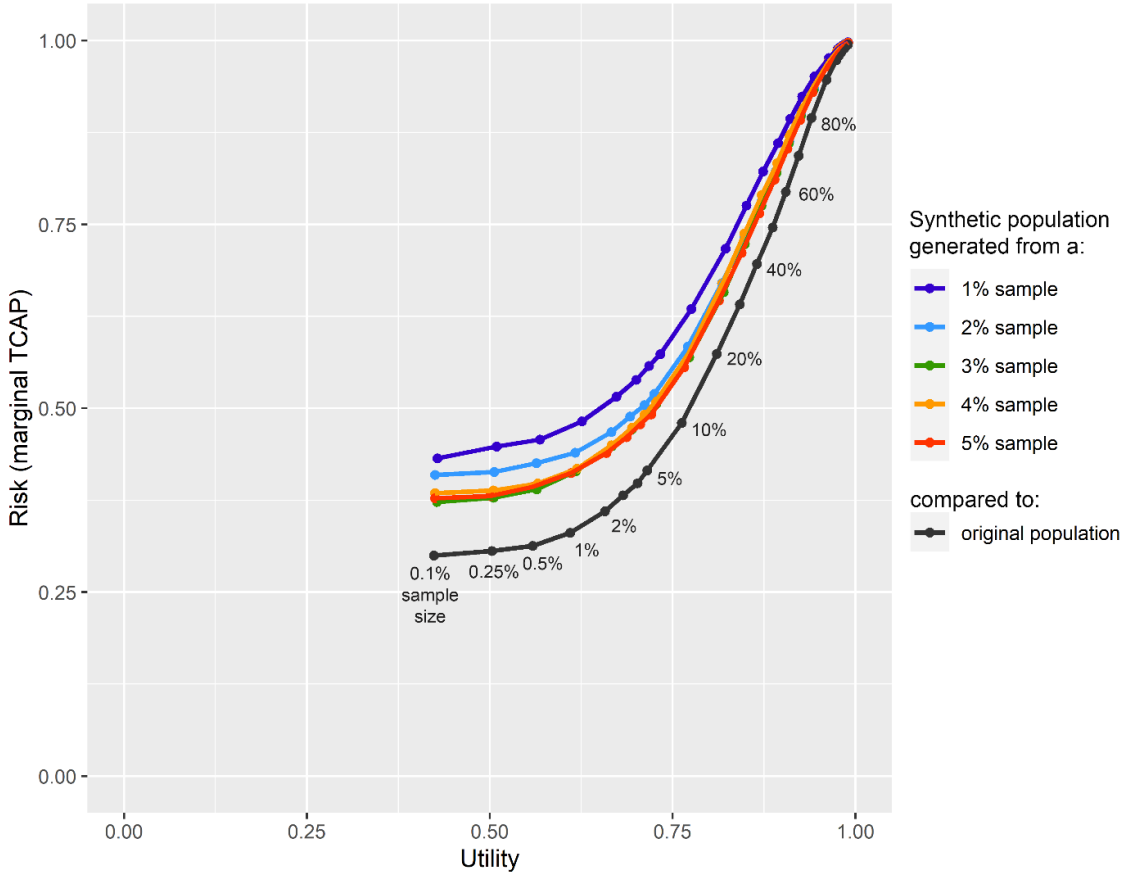


Figure 6: R-U map contrasting the results for samples generated from synthetic populations to the original population (with sample sizes labelled) in experiment B.

Plots and tables of the MAE (and standard deviation) are in Appendix C (utility) and Appendix D (marginal TCAP). The marginal TCAP plot indicates that the overall pattern of the MAE fluctuates at lower sample sizes and then generally decreases as the sample size gets larger. The samples from the synthetic population generated from a 1% sample of the original data have higher MAE than those generated from larger samples. The samples generated from a 2%, 3% and 4% synthetic population exhibit unusual behaviour in that they are not in the order one might expect, this is likely due to variation in the samples for the TCAP key and target variables.

5 Final Thoughts

The results show that, at least in terms of the risk and utility of samples drawn from a synthetic population, the relationship is similar to the results obtained by drawing samples from the original population. For Experiment A, which used a synthetic population generated directly from the original population, the relationship between the synthetic samples and the synthetic population follows closely the relationship between the original samples and the original population; the lines on the R-U map were very close together when compared.

For Experiment B, which is perhaps a more likely scenario (since we do not usually have access to the population data), synthetic populations were generated from samples (of varying sizes) drawn from the original population. For each synthetic population samples were drawn, and the risk and utility calculated, with the results compared (in terms of risk and utility) to the results of samples drawn from the original population. For each of the synthetic populations, the overall relationship, in terms of the curve on the R-U map, is similar to the original population results. However, each of the synthetic populations had higher risk (TCAP), pushing the curve upwards; and as the sample that the synthetic population was generated from gets smaller the curve moves further away from the original population curve.

Further work on this might involve using a different data synthesizer – Synthpop was selected because it generally produces data of high utility (and therefore higher risk) – but it may make sense to perform these experiments with synthetic data of lower utility/risk to determine whether the results replicate. It is also possible that using different risk and utility metrics may produce different results. Repeating the experiments with different datasets may also make sense. As in previous work, we have used a sample to represent the population data, so a further extension would be to access population data and repeat these experiments.

6 References

Duncan, G.T., Keller-McNulty, S.A. and Stokes, S.L. (2004). Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map.

Elliot, M. (2014). Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. [online]. Available from: https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02 -Report on disclosure risk analysis of synthpop synthetic versions of LCF_ final.pdf.

Karr, A.F. et al. (2006). A framework for evaluating the utility of data altered to protect confidentiality. American Statistician, 60(3), pp.224–232.

Little, C., Elliot, M. and Allmendinger, R. (2022). Comparing the Utility and Disclosure Risk of Synthetic Data with Samples of Microdata. In Privacy in Statistical Databases. PSD 2022. Springer International Publishing, pp. 234–249. [online]. Available from: https://doi.org/10.1007/978-3-031-13945-1_17.

Nowok, B., Raab, G.M. and Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software, 74(11).

Office for National Statistics, Census Division, University of Manchester, Cathie Marsh Centre for Census and Survey Research. (2013). Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs). UK Data Service. [online]. Available from: http://doi.org/10.5255/UKDA-SN-7210-1 [Accessed May 29, 2021].

Taub, J. et al. (2019). Creating the Best Risk-Utility Profile: The Synthetic Data Challenge. In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

Taub, J. et al. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. In Privacy in Statistical Databases. pp. 122–137. [online]. Available from: http://dx.doi.org/10.1007/978-3-319-99771-1_9.

7 Appendix A

The UK 1991 Census dataset sample, 104267 records and 15 variables:

Variable Name	Description	Number of Values	Number of missing
AREAP	Individual SAR area, e.g., Birmingham, Solihull	21	0
AGE	Age Range: 0 - 95	94	0
COBIRTH	Country of birth	42	0
ECONPRIM	Primary economic position, e.g., Employee FT, Student, Retired Note: omits individuals < 16	10	21467 (20.6%)
ETHGROUP	Ethnic group e.g., White, Black Caribbean	10	0
FAMTYPE	Family type e.g., Married no children, Cohabiting with children Note: n/a for individuals in communal establishments or with no family	9	0
HOURS	Number of hours worked weekly Range: 1-81 Note: excludes individuals aged <=16 and those who have not worked in previous ten years	72	46979 (45.1%)
LTILL	Limiting long-term illness. Two categories: Yes or no	2	0
MSTATUS	Marital status e.g., Single, married, divorced Note: individuals < 16 are categorised as ‘single’	5	0
QUALNUM	Number of higher educational qualifications Three categories: 0, 1 or 2+ Note: individuals < 18 have a “0”	3	0
RELAT	Relationship to household head e.g., Head, spouse, daughter	8	2113 (2.0%)
SEX	Sex Two categories: Male or female	2	0
SOCLASS	Social class (based on occupation) e.g., Professional, skilled Note: omits individuals < 16, & those not in paid work in last 10 years	9	44537 (42.7%)
TENURE	Tenure of household space e.g., Owner occupied outright, rented privately Note: omits individuals not in a household	7	2113 (2.0%)
TRANWORK	Mode of transport to work e.g., Bus, on foot Note: omits individuals not in employment in the week before Census	11	59249 (56.8%)

8 Appendix B

Experiment A: the mean utility and TCAP scores for each synthetic sample size (to 3dp, n=100), contrasted with the mean utility and TCAP of samples taken from the original population

	Overall utility		TCAP (3 targets)		Marginal TCAP (3 targets)	
Sample size	**Original**	**Synthetic**	**Original**	**Synthetic**	**Original**	**Synthetic**
0.1%	0.424	0.420	0.609	0.607	0.300	0.298
0.25%	0.503	0.497	0.613	0.612	0.306	0.306
0.5%	0.559	0.554	0.617	0.618	0.313	0.317
1%	0.610	0.605	0.627	0.627	0.331	0.333
2%	0.657	0.653	0.643	0.643	0.360	0.362
3%	0.682	0.680	0.655	0.655	0.382	0.384
4%	0.702	0.701	0.664	0.666	0.398	0.403
5%	0.715	0.712	0.674	0.675	0.416	0.419
10%	0.762	0.760	0.710	0.713	0.480	0.486
20%	0.810	0.808	0.762	0.768	0.574	0.585
30%	0.842	0.840	0.800	0.807	0.641	0.656
40%	0.865	0.864	0.831	0.840	0.696	0.713
50%	0.887	0.887	0.858	0.868	0.746	0.764
60%	0.905	0.904	0.885	0.895	0.794	0.812
70%	0.922	0.921	0.913	0.921	0.843	0.859
80%	0.940	0.939	0.941	0.947	0.895	0.905
90%	0.960	0.960	0.970	0.973	0.947	0.952
95%	0.974	0.974	0.985	0.986	0.974	0.976
96%	0.977	0.977	0.988	0.989	0.979	0.981
97%	0.980	0.980	0.991	0.992	0.984	0.985
98%	0.985	0.985	0.994	0.995	0.989	0.990
99%	0.990	0.990	0.997	0.997	0.995	0.995

Experiment A, the standard deviation to 4dp (n=100) of the utility and TCAP scores for the original and synthetic data samples

	Overall utility		TCAP (3 targets)		Marginal TCAP (3 targets)	
Sample size	**Original**	**Synthetic**	**Original**	**Synthetic**	**Original**	**Synthetic**
0.1%	0.0106	0.0125	0.0192	0.0185	0.0344	0.0331
0.25%	0.0114	0.0122	0.0108	0.0107	0.0193	0.0192
0.5%	0.0101	0.0108	0.0077	0.0078	0.0138	0.0139
1%	0.0078	0.0089	0.0061	0.0062	0.0109	0.0110
2%	0.0064	0.0076	0.0044	0.0039	0.0080	0.0070
3%	0.0066	0.0066	0.0034	0.0030	0.0061	0.0053
4%	0.0060	0.0057	0.0029	0.0031	0.0052	0.0055
5%	0.0068	0.0066	0.0028	0.0031	0.0050	0.0056
10%	0.0054	0.0065	0.0024	0.0022	0.0042	0.0039
20%	0.0059	0.0061	0.0021	0.0018	0.0037	0.0032
30%	0.0049	0.0060	0.0019	0.0019	0.0035	0.0033
40%	0.0067	0.0050	0.0016	0.0018	0.0028	0.0033
50%	0.0048	0.0049	0.0022	0.0017	0.0039	0.0030
60%	0.0045	0.0041	0.0021	0.0017	0.0037	0.0030
70%	0.0041	0.0041	0.0018	0.0017	0.0032	0.0031
80%	0.0036	0.0038	0.0021	0.0014	0.0038	0.0025
90%	0.0027	0.0028	0.0017	0.0014	0.0030	0.0025
95%	0.0019	0.0018	0.0013	0.0010	0.0024	0.0017
96%	0.0019	0.0016	0.0012	0.0010	0.0021	0.0018
97%	0.0016	0.0015	0.0011	0.0008	0.0020	0.0014
98%	0.0011	0.0011	0.0009	0.0008	0.0016	0.0014
99%	0.0010	0.0009	0.0005	0.0005	0.0010	0.0009

Experiment A: Mean Absolute Error (n=100) and standard deviation to 4dp of the utility and TCAP values of synthetic samples compared to the original samples

	Overall utility		**TCAP (3 targets)**		**Marginal TCAP (3 targets)**	
Sample size	**MAE**	**SD**	**MAE**	**SD**	**MAE**	**SD**
0.1%	0.0108	0.0074	0.0147	0.0113	0.0262	0.0201
0.25%	0.0111	0.0081	0.0080	0.0071	0.0143	0.0127
0.5%	0.0088	0.0075	0.0062	0.0049	0.0114	0.0090
1%	0.0079	0.0065	0.0048	0.0038	0.0088	0.0068
2%	0.0060	0.0062	0.0032	0.0023	0.0060	0.0043
3%	0.0052	0.0043	0.0024	0.0017	0.0045	0.0031
4%	0.0043	0.0039	0.0029	0.0022	0.0059	0.0044
5%	0.0053	0.0050	0.0026	0.0021	0.0051	0.0040
10%	0.0053	0.0047	0.0029	0.0018	0.0061	0.0035
20%	0.0050	0.0039	0.0054	0.0018	0.0106	0.0032
30%	0.0045	0.0044	0.0075	0.0019	0.0142	0.0033
40%	0.0040	0.0032	0.0091	0.0018	0.0169	0.0033
50%	0.0039	0.0030	0.0101	0.0017	0.0186	0.0030
60%	0.0034	0.0026	0.0097	0.0017	0.0177	0.0030
70%	0.0035	0.0025	0.0084	0.0017	0.0153	0.0031
80%	0.0030	0.0025	0.0055	0.0014	0.0100	0.0025
90%	0.0022	0.0018	0.0029	0.0014	0.0054	0.0024
95%	0.0014	0.0012	0.0013	0.0008	0.0024	0.0014
96%	0.0012	0.0011	0.0012	0.0007	0.0022	0.0013
97%	0.0012	0.0009	0.0009	0.0006	0.0016	0.0011
98%	0.0009	0.0007	0.0008	0.0005	0.0014	0.0009
99%	0.0007	0.0005	0.0005	0.0003	0.0009	0.0005

9 Appendix C

Experiment B: Mean utility of original samples and synthetic samples, by sample size to 3dp. This is the mean utility (across 100 samples) of each sample size (the rows) for each of the synthetic populations (columns).

Sample size	Original Population	Synthetic population generated from:				
		1% sample	2% sample	3% sample	4% sample	5% sample
0.1%	0.424	0.429	0.425	0.428	0.425	0.425
0.25%	0.503	0.509	0.506	0.505	0.505	0.500
0.5%	0.559	0.569	0.564	0.564	0.566	0.558
1%	0.610	0.626	0.617	0.618	0.619	0.611
2%	0.657	0.673	0.666	0.667	0.666	0.660
3%	0.682	0.700	0.692	0.694	0.694	0.687
4%	0.702	0.718	0.712	0.714	0.711	0.706
5%	0.715	0.733	0.725	0.728	0.727	0.721
10%	0.762	0.776	0.771	0.773	0.772	0.766
20%	0.810	0.823	0.817	0.820	0.818	0.813
30%	0.842	0.851	0.848	0.849	0.848	0.844
40%	0.865	0.874	0.871	0.871	0.872	0.868
50%	0.887	0.894	0.891	0.892	0.893	0.890
60%	0.905	0.911	0.909	0.909	0.909	0.907
70%	0.922	0.927	0.926	0.925	0.925	0.924
80%	0.940	0.944	0.944	0.943	0.943	0.941
90%	0.960	0.963	0.962	0.962	0.962	0.961
95%	0.974	0.976	0.974	0.975	0.975	0.974
96%	0.977	0.978	0.978	0.978	0.978	0.977
97%	0.980	0.981	0.981	0.981	0.981	0.980
98%	0.985	0.985	0.985	0.985	0.985	0.984
99%	0.990	0.990	0.990	0.990	0.990	0.990

Experiment B: the standard deviation to 4dp (n=100) of the utility for samples taken from the original population, and the five synthetic populations

Sample size	Original Population	Synthetic population generated from:				
		1% sample	2% sample	3% sample	4% sample	5% sample
0.1%	0.0106	0.0140	0.0125	0.0145	0.0113	0.0116
0.25%	0.0114	0.0110	0.0120	0.0119	0.0102	0.0113
0.5%	0.0101	0.0095	0.0092	0.0102	0.0097	0.0084
1%	0.0078	0.0086	0.0087	0.0085	0.0075	0.0076
2%	0.0064	0.0073	0.0066	0.0061	0.0066	0.0070
3%	0.0066	0.0062	0.0069	0.0070	0.0066	0.0064
4%	0.0060	0.0064	0.0059	0.0057	0.0073	0.0071
5%	0.0068	0.0062	0.0067	0.0058	0.0064	0.0057
10%	0.0054	0.0071	0.0063	0.0062	0.0056	0.0060
20%	0.0059	0.0056	0.0057	0.0047	0.0060	0.0064
30%	0.0049	0.0058	0.0056	0.0064	0.0055	0.0060
40%	0.0067	0.0059	0.0051	0.0053	0.0053	0.0048
50%	0.0048	0.0050	0.0052	0.0052	0.0049	0.0047
60%	0.0045	0.0045	0.0046	0.0051	0.0046	0.0045
70%	0.0041	0.0046	0.0046	0.0044	0.0044	0.0043
80%	0.0036	0.0035	0.0035	0.0034	0.0037	0.0038
90%	0.0027	0.0025	0.0024	0.0025	0.0026	0.0028
95%	0.0019	0.0020	0.0020	0.0019	0.0019	0.0019
96%	0.0019	0.0016	0.0018	0.0017	0.0017	0.0018
97%	0.0016	0.0014	0.0014	0.0012	0.0014	0.0016
98%	0.0011	0.0012	0.0014	0.0012	0.0011	0.0015
99%	0.0010	0.0009	0.0009	0.0009	0.0009	0.0010

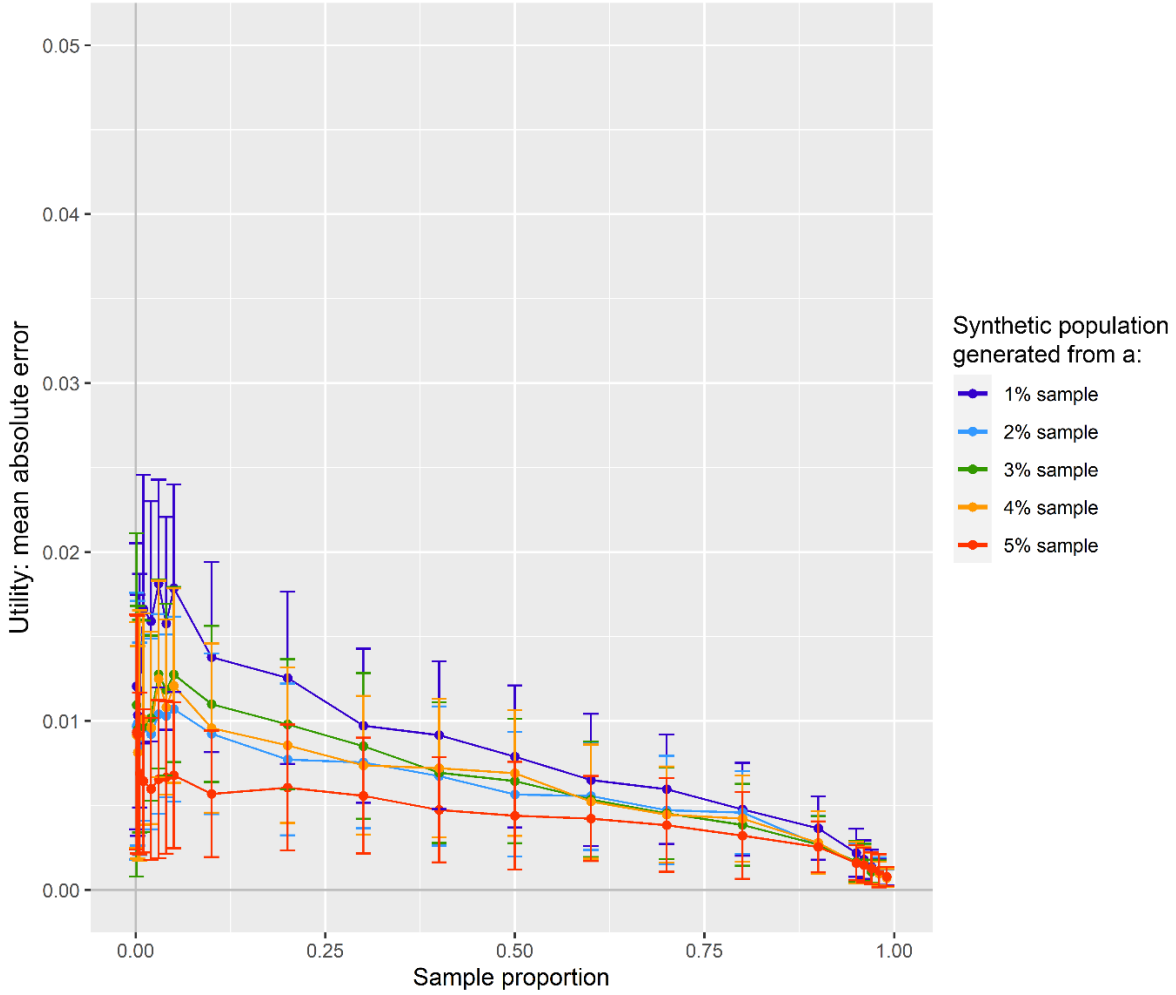
Experiment B: The MAE to 4dp (n=100) between the utility of the original population samples and each of the synthetic population samples

Sample size	Synthetic population generated from:				
	1% sample	2% sample	3% sample	4% sample	5% sample
0.1%	0.0121	0.0097	0.0109	0.0092	0.0093
0.25%	0.0103	0.0099	0.0095	0.0081	0.0092
0.5%	0.0118	0.0084	0.0097	0.0091	0.0069
1%	0.0166	0.0096	0.0097	0.0101	0.0065
2%	0.0159	0.0092	0.0102	0.0096	0.0060
3%	0.0181	0.0104	0.0128	0.0125	0.0066
4%	0.0158	0.0103	0.0119	0.0108	0.0067
5%	0.0179	0.0107	0.0128	0.0121	0.0068
10%	0.0138	0.0092	0.0110	0.0096	0.0057
20%	0.0125	0.0077	0.0098	0.0086	0.0061
30%	0.0097	0.0076	0.0085	0.0074	0.0056
40%	0.0092	0.0067	0.0069	0.0072	0.0047
50%	0.0079	0.0057	0.0065	0.0069	0.0044
60%	0.0065	0.0056	0.0054	0.0052	0.0042
70%	0.0060	0.0047	0.0045	0.0045	0.0038
80%	0.0048	0.0046	0.0039	0.0042	0.0032
90%	0.0037	0.0027	0.0027	0.0028	0.0025
95%	0.0022	0.0016	0.0017	0.0016	0.0016
96%	0.0018	0.0016	0.0016	0.0015	0.0015
97%	0.0014	0.0013	0.0011	0.0013	0.0013
98%	0.0011	0.0011	0.0010	0.0010	0.0011
99%	0.0008	0.0007	0.0008	0.0007	0.0008

Experiment B: the standard deviation for the MAE of the utility, to 4dp

	Synthetic population generated from:				
Sample size	1% sample	2% sample	3% sample	4% sample	5% sample
0.1%	0.0085	0.0079	0.0102	0.0067	0.0069
0.25%	0.0071	0.0072	0.0073	0.0063	0.0070
0.5%	0.0069	0.0063	0.0063	0.0074	0.0048
1%	0.0079	0.0055	0.0063	0.0063	0.0042
2%	0.0071	0.0056	0.0049	0.0057	0.0042
3%	0.0062	0.0059	0.0056	0.0058	0.0047
4%	0.0063	0.0048	0.0051	0.0052	0.0045
5%	0.0061	0.0055	0.0052	0.0058	0.0043
10%	0.0056	0.0047	0.0046	0.0050	0.0037
20%	0.0051	0.0045	0.0039	0.0046	0.0037
30%	0.0046	0.0039	0.0043	0.0041	0.0034
40%	0.0044	0.0041	0.0042	0.0041	0.0031
50%	0.0042	0.0037	0.0037	0.0037	0.0032
60%	0.0039	0.0032	0.0034	0.0034	0.0025
70%	0.0032	0.0032	0.0027	0.0028	0.0028
80%	0.0027	0.0025	0.0024	0.0026	0.0026
90%	0.0019	0.0017	0.0017	0.0019	0.0015
95%	0.0014	0.0012	0.0012	0.0012	0.0010
96%	0.0011	0.0011	0.0011	0.0011	0.0010
97%	0.0010	0.0009	0.0007	0.0009	0.0010
98%	0.0008	0.0008	0.0007	0.0007	0.0010
99%	0.0005	0.0005	0.0005	0.0005	0.0006

Experiment B: the MAE for the utility by sample proportion, for each synthetic population, with error bars indicating +- 1 standard deviation



Appendix D

Experiment B: the mean (n=100) Marginal TCAP values from each of the synthetic populations, and the original population, to 3dp.

Sample size	Original Population	Synthetic Population (1%)	Synthetic Population (2%)	Synthetic Population (3%)	Synthetic Population (4%)	Synthetic Population (5%)
0.1%	0.300	0.432	0.410	0.373	0.385	0.378
0.25%	0.306	0.448	0.414	0.378	0.388	0.381
0.5%	0.313	0.458	0.425	0.390	0.398	0.393
1%	0.331	0.482	0.440	0.415	0.418	0.412
2%	0.360	0.516	0.468	0.448	0.450	0.439
3%	0.382	0.539	0.489	0.472	0.473	0.461
4%	0.398	0.558	0.504	0.491	0.491	0.478
5%	0.416	0.574	0.520	0.506	0.510	0.492
10%	0.480	0.635	0.584	0.569	0.574	0.556
20%	0.574	0.717	0.670	0.658	0.670	0.646
30%	0.641	0.776	0.731	0.723	0.738	0.711
40%	0.696	0.822	0.781	0.775	0.790	0.765
50%	0.746	0.861	0.823	0.820	0.833	0.811
60%	0.794	0.894	0.863	0.860	0.872	0.853
70%	0.843	0.924	0.899	0.898	0.907	0.892
80%	0.895	0.951	0.934	0.933	0.939	0.930
90%	0.947	0.976	0.968	0.967	0.970	0.965
95%	0.974	0.988	0.984	0.984	0.985	0.983
96%	0.979	0.991	0.987	0.987	0.988	0.986
97%	0.984	0.993	0.991	0.990	0.991	0.990
98%	0.989	0.995	0.994	0.993	0.994	0.993
99%	0.995	0.998	0.997	0.997	0.997	0.997

Experiment B: The standard deviation (to 4dp) of the marginal TCAP scores for samples from each of the synthetic populations. The original population results are included for comparison.

Sample size	Original Population	Synthetic Population (1%)	Synthetic Population (2%)	Synthetic Population (3%)	Synthetic Population (4%)	Synthetic Population (5%)
0.1%	0.0344	0.0374	0.0456	0.0429	0.0389	0.0374
0.25%	0.0193	0.0246	0.0250	0.0256	0.0270	0.0241
0.5%	0.0138	0.0176	0.0179	0.0162	0.0170	0.0177
1%	0.0109	0.0097	0.0128	0.0135	0.0118	0.0114
2%	0.0080	0.0082	0.0086	0.0086	0.0088	0.0092
3%	0.0061	0.0066	0.0066	0.0080	0.0062	0.0067
4%	0.0052	0.0052	0.0059	0.0067	0.0059	0.0058
5%	0.0050	0.0046	0.0057	0.0052	0.0048	0.0047
10%	0.0042	0.0031	0.0040	0.0045	0.0042	0.0040
20%	0.0037	0.0025	0.0039	0.0037	0.0032	0.0032
30%	0.0035	0.0025	0.0028	0.0032	0.0027	0.0031
40%	0.0028	0.0024	0.0028	0.0025	0.0024	0.0031
50%	0.0039	0.0021	0.0025	0.0025	0.0021	0.0027
60%	0.0037	0.0016	0.0019	0.0024	0.0019	0.0025
70%	0.0032	0.0016	0.0020	0.0020	0.0021	0.0023
80%	0.0038	0.0011	0.0016	0.0017	0.0013	0.0018
90%	0.0030	0.0008	0.0012	0.0013	0.0010	0.0015
95%	0.0024	0.0005	0.0009	0.0009	0.0008	0.0007
96%	0.0021	0.0005	0.0008	0.0008	0.0007	0.0009
97%	0.0020	0.0005	0.0008	0.0008	0.0006	0.0007
98%	0.0016	0.0004	0.0006	0.0007	0.0005	0.0007
99%	0.0010	0.0003	0.0004	0.0004	0.0003	0.0005

Experiment B: the MAE of the marginal TCAP for each synthetic population by sample proportion, with error bars indicating +- 1 standard deviation