CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 3

29-31 October 2019, the Hague, the Netherlands 17 September 2019

The Synthetic Data Challenge

Jennifer Taub and Mark Elliot (University of Manchester, United Kingdom)

jennifer.taub@manchester.ac.uk, mark.elliot@manchester.ac.uk

Abstract and Paper

Data synthesis is an alternative to controlling confidentiality risk through access restriction and traditional statistical disclosure control (SDC) methods. There are several different techniques used to produce synthetic data; the goal of all of them is to produce useful data with very low disclosure risk. An important and current research goal is the development of credible measures of both the risk and utility of synthetic data. Without such credible measures, it is difficult for analysts and data owners alike to trust synthetic data or to know which data synthesis methods work best (for their data). As part of the Isaac Newton Institute programme on Data Linkage and Anonymisation, we ran a challenge to test various synthetic data generation methods against one another. Four different research teams produced a total of eight synthetic versions of the same dataset using different synthetic data production methods. The synthetic datasets were then put through a battery of tests for both data utility and disclosure risk. The challenge study has produced significant insights not only about the synthetic data generation techniques, but also about the effectiveness of different measures designed to capture the utility/risk characteristics of synthetic data.

Creating the Best Risk-Utility Profile: The Synthetic Data Challenge

Jennifer Taub*, Mark Elliot*, Gillian Raab**, Anne-Sophie Charest***, Cong Chen****, Christine M. O'Keefe*****, Michelle Pistner Nixon******, Joshua Snoke*******, Aleksandra Slavković******

* Cathie Marsh Institute, The University of Manchester, Manchester, UK, {jennifer.taub, mark.elliot}@manchester.ac.uk

** Scottish Centre for Administrative Data Researach, University of Edinburgh, Edinburgh, UK, gillian.raab@ed.ac.uk

*** Departement de mathématiques et de statistique, Université Laval, Québec, Canada.

**** Public Health England, UK, Health Data Insight CIC, Cambridge, UK

***** CSIRO, GPO Box 1700, Canberra ACT 2601 AUSTRALIA (until 5 Oct 2018).

****** Department of Statistics, The Pennsylvania State University, University Park, PA, USA, {pistner, sesa}@psu.edu

******* RAND Corporation, Pittsburgh, PA, USA, jsnoke@rand.org

Abstract. Data synthesis is an alternative to traditional statistical disclosure control (SDC) methods or access restrictions for controlling the confidentiality risk arising from data sharing. There are several different techniques used to produce synthetic data; the goal of all of them is to produce useful data with very low disclosure risk. An important and current research goal is the development of credible measures of both the risk and utility of synthetic data. Without such credible measures, it is difficult for either analysts or data owners to trust synthetic data or to know which data synthesis methods work best (for their data). As part of the Isaac Newton Institute programme on Data Linkage and Anonymisation in 2016, we ran a challenge to test various synthetic data generation methods against one another. Four different research teams produced eight synthetic versions of the same dataset using different synthetic data production methods. The synthetic datasets were then put through a battery of tests for both data utility and disclosure risk. The challenge study has produced significant insights not only about the synthetic data generation techniques, but also about the effectiveness of different measures designed to capture the utility/risk characteristics of synthetic data.

1 Introduction

Data synthesis is an alternative to using traditional statistical disclosure control (SDC) methods or access restrictions to control the confidentiality risk arising from data sharing. Rubin (1993) first introduced the idea of synthetic data as an alternative to traditional SDC techniques. Since Rubin's original conception, a large literature has developed on methods for synthetic data generation, much of which is reviewed in the monograph by Dreschler (Dreschler, 2011) and implemented in software (Nowok et al. 2016, Raab et al., 2016). Most methods involve generating synthetic data from models of conditional distributions, within the original data. Many alternatives for modelling the distributions are available including CART (Reiter, 2005), random forests (Caiola and Reiter, 2010), and support vector machines (Drechsler, 2010). Additional features such as the predictors used for the conditional model, the order in which the conditional distributions are defined and whether stratification is used (Raab et al, 2017) can have an impact on both the utility and disclosure risk of the synthetic data.

As an outcome of the Isaac Newton Institute programme on Data Linkage and Anonymisation[1] in 2016, a group formed with the aim of exploring the utility and disclosure risk of synthetic data. The group decided to run a challenge amongst themselves to test various synthetic data generation methods against one another. The challenge was administered by the two lead authors of this paper.

In this paper, we report on the results of the challenge. We will describe the ways that the participants constructed their synthetic datasets and how that affected the synthetic datasets in terms of both risk and utility. We will also examine the trade off between utility and risk and which methods for synthesis optimise that balance. In section 2, we describe the datasets and the methods for synthesis. In section 3 we describe the methods being used to measure utility and risk. In section 4, we report the results for both utility and risk and discuss different ways of calculating a combined score. Section 5 presents a discussion of how the different synthesis methods affected their risk-utility scores. Section 6 concludes the paper.

2 The Data

2.1 The Original Dataset

The original dataset is derived from the Scottish historical census of 1901. It is a sample of 82,851 households in the Edinburgh region. The dataset contains 24 variables: 20 observed, 3 derived, and a unique identifier (See Appendix A for a full description of variables).[2]

[1] https://www.newton.ac.uk/event/dla [accessed 18th October 2017]

[2] The dataset can be found at Gillian Raab's page labeled as "Data for first challenge November 2016 at https://www.geos.ed.ac.uk/homes/graab [accessed 19 October 2017]

2.2 Synthesis Methods

Four different research teams (from Canada, the UK, and the USA) produced a total of eight synthetic versions of the same (original) dataset using different synthetic data production methods. In an evaluation stage, they were put through a battery of tests for both data utility and disclosure risk to determine which dataset had the best risk and utility profile.

The teams used a variety of different methods including CART, machine learning techniques, sampling, and quantile regression. Two of the teams produced a single dataset, with one team producing two different datasets using two different methods (Snoke, Pistner, and Slavković). While another team (Chen) created multiple synthetic dataset using one method. Here is a brief overview of the methods used. More detail on some of the methods can be found in Appendix C.

Team 1: Raab Raab used preprocessing as part of her synthesis process. She created an area level score based on a factor analysis of occupation, household occupancy, and people living on private means, which was used as a predictor for the synthesis. This score was defined at the level of the enumeration district so that the geographic structure of the data was maintained and the the score was used to predict the characteristic of families within the enumeration districts. Variables with many small categories were grouped into larger categories. The data were split into 3 geographic strata, and the *synthpop* package was used to create the synthetic data. The CART method was used for most variables, but some variables had special methods: *hsize*, *roomsperp*, *perperoom* were calculated from other variables that were synthesised before them and *Occlab2*, *occlab3* and *ctry_bth* were generated as random samples from the sub-group of the class within-which each was nested. The order of the variables was specified with the area level score at the start and a predictor matrix was used that selected the covariates to be used for each conditional distribution. Some additional smoothing of the upper parts of the distributions of *hsize, totrooms, bfamgteq15* and *nfamlt15* was carried out to prevent apparent singling out of extreme values.

Team 2: Snoke, Pistner and Slavković - Sequential Synthesis Model Snoke et al. used a non-parametric CART model, where variables are synthesised one at a time with all previous synthesised variables used as predictors. (See Appendix C for the synthesis order that they used)

Like, Raab, they used some pre-processing. Due to the dataset including categorical variables with skewed counts, the variables were partitioned and multi-level synthesis was performed. For example with marital status they split it into two groups: (i) married and widowed (which made up 85% of the observations) and (ii) the other marital statuses. A CART model was then fitted first to predict whether observations fall into one group or the other. After this, CART models were fitted

within each group and new observations are drawn within the predicted groups. For their synthesis, they grouped any categorical variable if a third or less of the categories contain more than 75% of observations. Using these settings, the following categorical variables had no partitioning: *sex, employ*, and *occlab1*.

For each nested level, they fitted a separate CART model to draw synthetic values. For each of the nested levels, the grouping described in the previous sections was carried out if the thresholds were achieved. The three nested variables, which all contain very large numbers of categories, were omitted as predictors for each other even though all previously synthesised variables would normally be included as predictors. These choices were made for computational reasons, with the goal of best preserving relationships among the variables while lowering the run time, which is known to be an issue when using CARTs with categorical predictors that have large numbers of levels.

Team 2b: Pistner, Snoke, and Slavković - Quantile Syntehsis Model Pistner et al. used quantile regression to synthesise both continuous and categorical variables. They wrote an R function, *qr.synth* that allows for the generation of synthetic data via quantile regression and censored quantile regression using the R package *quantreg*.

For the several variables with missing values they used logistic regression to model the "missingness". If the missingness was coded as a value, then quantile regression was used to completely synthesise the variable. If coded as missing, then the corresponding value was coded as missing in the synthetic data set.

They, like Raab and Snoke et al., used variable partitioning. For some categorical variables, many of the observations fell into one distinct category. For these variables, a hierarchal structure was employed. First, the factor levels were binned into two groups. For example, for disability, these groups would be None and One+. Then, the observations were split between these two levels using logistic regression. Quantile regression was then used to further separate the two major categories as needed. Variable partitioning was used for several variables, including marital status, country of birth, disability, and inactive. For each of the nested levels, they fit a separate quantile regression model to draw synthetic values. As with Snoke et al., the purpose of the partioning was to reduce the computational run time.

Variables are synthesised one at a time with all previous synthesised variables used as predictors. The synthesis order is in Appendix C. A further description of how quantile regression works is also included in Appendix C.

Team 3: Charest Charest generated the data intuitively, without complex modelling and statistical tests. For each of the variables in the dataset, synthetic values were obtained by randomly sampling with replacement from the original values for that variable. Some variables were sampled as a group to ensure consistency: parish

with enumeration districts, all three occupation variables together, employment with inactivity and all seven variables describing the number of individuals living in the house together. To better maintain conditional distributions, some of the variables were sampled within categories delimited by other variables. More specifically, age was sampled within sex (2 sub-populations), marital status was sampled within sex and age (18 sub-populations), occupational classification variables, employment and inactivity were sampled within sex and indicator of at least some disability (4 sub-populations) and the number of rooms was sampled according to the number of servants and the total size of the household (5 sub-populations: no servants with 1 or 2 people total, no servants with more than 2 people total, 1 or 2 servants with less than 2 people total, 1 or 2 servants with more than 2 people, more than 2 servants). Finally, household size, persons per room and rooms per person were derived from the other variables.

Team 4: Chen Chen manipulated the data in Matlab using code [3] written based on the statistics and machine learning toolbox. Functions were written to compute pairwise statistics between fields in the data for evidence of association by obtaining normalised chi-squared statistics, and construct a directed graph where vertices are fields and the two high-scoring edges from each vertex are included. This graph was generated for each stratum of the data and used to construct a bootstrap simulation of that stratum in the data, where each field is conditioned on its neighbourhood of fields with edges into it, which have already been simulated.

He verified that Household Size was a derived field, and that *pperroom* and *roomsperp* were also derived fields, and removed these and the Household ID field. The data were treated entirely as categorical data and simulated three times, firstly without stratification and then stratifying by either the inactive or *occlab1* field. The dataset that was generated by stratification on the inactive field was chosen and numerical fields were reformatted in post-processing to approximate the format of the input data, by resampling from the distributions corresponding to categories into they had been discretised.

3 Methods for Analysis- Data Utility and Disclosure Risk

This paper will examine both the data utility and the disclosure risk of the synethsised datasets. Given that there are many tests and methods for data utility, we will be using multiple utility tests, but just one measure for disclosure risk.

[3] Described briefly at https://simulacrum.healthdatainsight.org.uk/publications/ [accessed 13 September 2019]

3.1 Measuring Utility

We will use both broad and narrow measures to evaluate data utility. According to Karr et al (2006) narrow measures tend to be tailored to a certain kind of analysis and may, despite having high utility scores, allow for data to have low utility for other kinds of analyses. Broad utility measures may be "pretty good" for many analyses and "really good" for none. Therefore, we thought it important to include both and put together a battery of tests including descriptive statistics, regression models, multiple correspondence analyses, and propensity scores mean-squared errors.

3.1.1 Narrow Measures

To produce our narrow measures, we will attempt to think like a user. Taub et al. (2017) introduced the Purdam-Elliot methodology for synthetic data (Purdam and Elliot, 2007). This involves systematically repeating narrow measures. However, Taub et al. used the analyses from previously written papers to establish systematically what a real world user would be interested in; this approach is not possible with the 1901 Scottish census, for which there is currently a dearth of published output. Instead we endeavoured to imagine the kinds of analyses a data user would want to run and used this to inform our choice of narrow measures.

Frequency Tables and Cross-Tabulations We evaluated frequency tables and cross-tabs using the ratio of counts (ROC) and confidence interval overlap (CIO) metrics.

The ratio of counts is calculated by dividing the smaller number by the larger number. Thus, given two corresponding estimates (e.g. totals, proportions), where y^1_{orig} is the estimate from the original data and y^1_{synth} is the corresponding estimate from the synthetic data, the ROC is calculated as:

$$\frac{\min(y^1_{orig}, y^1_{synth})}{\max(y^1_{orig}, y^1_{synth})} \tag{1}$$

If $y^1_{orig} = y^1_{synth}$ then the ROC = 1. The ratio of estimates provides a value between 0 and 1. For each frequency table the per cell ratios of counts were averaged to give an overall ratio of counts..

The confidence interval overlap (CIO), is calculated as:

$$J_k = \frac{1}{2}\left(\frac{U_{,k} - L_{,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{,k} - L_{,k}}{U_{syn,k} - L_{syn,k}}\right) \tag{2}$$

where $U_{,k}$ and $L_{,k}$, denote the respective upper and lower bounds of the intersection of the confidence intervals from both the original and synthetic data for estimate k, $U_{org,k}$ and $L_{org,k}$ represent the upper and lower bounds of the original data, and $U_{syn,k}$ and $L_{syn,k}$ of the synthetic data. We, also, used the CIO to evaluate the means of some of the continuous variables.

Regression Models We created two different sets of regression models; (1) OLS regression where y= *pperroom*(people per room). We used *pperroom* as the response variable; this variable is created by dividing household size by number of rooms; although its scale status is certainly debatable, we treated this as continuous for this purpose in the interests of increasing the range of models tested. (2) Logistic regression using the presence of children as the response (where 1= children under 15 are present in the house and 0=no children under 15 are present). The Model is as follows: $Y \sim age + female + single + absent + spouse + widowed + employer + worker$ Descriptions of the dummy variables used are in Appendix B. The CIO was also calculated for the regression models using Equation (2).

3.1.2 Broad Measures

We used two different broad measures; the Propensity Score (pMSE) and a multiple correspondence analysis (MCA). The pMSE (Woo et al, 2009) is a traditional measure of data utility from the SDC community. While the MCA is a new intuitive approach in terms of evaluating utility for synthetic or SDC data.

Propensity Score A propensity score is the probability of being assigned to a treatment. Woo et al (2009) describe the procedure as follows: first, merge the original and masked data sets, adding a variable T, where T= 1 if it is the masked data set and T= 0 for the original dataset. Second, for each record in the combined dataset, compute the probability of being in the masked dataset, which is the propensity score. Third, compare the distributions of propensity scores in the original and masked data. To summarise the propensity score, Woo et al calculate:

$$pMSE = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2 \tag{3}$$

where N i the number of records in the merged dataset and $\hat{p}_i$ is the estimated propensity score for unit i and c is the proportion of data in the combined datasets that is masked data (which in many instances is $\frac{1}{2}$). Snoke et al (2018) have investigated the use of the pMSE to evaluate the utility of synthetic data and suggest a standardised score derived from it. In Table 1 we present the raw pMSE score.[4]. To create the pMSE scores we used most of the variables. However, given that number of rooms in house and number of people, and other variables counting the number of people, rooms per people were encapsulated in the variables people per room, we only used that variable. In total our logistic regression $k = 131$ parameters with no interaction variables or quadratic terms, thus effectively limiting the evaluation to univariate comparisons. An approach using a CART model to evaluate the utility as discussed in Snoke et al (2018) might have been more discriminating.

[4]Appendix E Table 9 presents the standardised scores.

MCA Multiple correspondence analysis (MCA) is a procedure wherein nominal variables are mapped in Euclidean space. It is the counterpart of principal component analysis for categorical data.

We used a MCA to capture the underlying structure of the datasets. We mapped both the original dataset and the synthetic dataset using MCA. Then to compare the two maps, we used a Euclidean distance metric. The Euclidean distance for two points is calculated as:

$$d = \sqrt{(x_o - x_s)^2 + (y_o - y_s)^2} \tag{4}$$

Wherein for any given nominal value, (x_o, y_o) represent its coordinates in the Euclidean for the original dataset and (x_s, y_s) represent the same coordinates for its synthetic counterpoint.

Given that MCA does not follow the 0-1 scale employed by other utility measures, we calculated a separate utility scale.Given that all averages fall into a 0 to 2 scale with the lowest performing the best, we applied the following equation to calculate a utility score for the MCA.

$$MCA_{utility} = 1 - MCA_{avg}/2 \tag{5}$$

The full results for the MCA can be see in Appendix E table 10.

3.2 Measuring Disclosure Risk using TCAP

Elliot (2014) and Taub et al (2018) introduced a measure for disclosure risk of synthetic data called *Differential Correct Attribution Probability* (DCAP), which consists of a Correct Attribution Probability (CAP) score. In this paper we will be using an adaptation of the CAP scores used in Taub et al. This adaptation we will be calling *Targeted Correct Attribution Probability* (TCAP).

The TCAP method in full is based on an intruder scenario in which two data owners produce a linked dataset (using a trusted third party), which is then synthesised and the synthetic data published. The adversary is one of the data owners who attempts to use the synthetic data to make inferences about the others' dataset. This is obviously a fairly strong attack.

More modestly, at the individual record level the adversary is somebody who has partial knowledge about a particular population unit (including the values for some of the variables in the dataset - the key - and knowledge that the population unit is in the original dataset) and wishes to infer the value of sensitive variable (the target) for that population unit for the original data.

The adversary views the synthetic dataset surmises that key equivalence class with low l-diversity on the target are most at risk. Here we assume that the adversary will focus on records which are in equivalence class which has l-diversity of 1 on the target and attempts to match them to their data. The TCAP metric is then the

probability that those matched records yield a correct value for the target variable (ie that the adversary makes a correct attribution inference).

TCAP is calculated as follows: We define d_o as the original data and K_o and T_o as vectors for the key and target information

$$d_o = \{K_o, T_o\} \tag{6}$$

Likewise, d_s is the synthetic dataset.

$$d_s = \{K_s, T_s\} \tag{7}$$

We then calculate the Within Equivalence Class Attribution Probability (WEAP) score for the synthetic dataset. The WEAP score for the record indexed j is the empirical probability of its target variables given its key variables,

$$WEAP_{s,j} = Pr(T_{s,j}|K_{s,j}) = \frac{\sum_{i=1}^{n}[T_{s,i} = T_{s,j}, K_{s,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{s,i} = K_{s,j}]} \tag{8}$$

where the square brackets are Iverson brackets, n is the number of records, and and K and T as vectors for the key and target information. Then using the WEAP score the synthetic dataset will be reduced to records with a WEAP score that is 1[5].
The TCAP for record j based on a corresponding original dataset d_o is the same empirical, conditional probability but derived from d_o,

$$TCAP_{o,j} = Pr(T_{s,j}|K_{s,j})_o = \frac{\sum_{i=1}^{n}[T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{o,i} = K_{s,j}]} \tag{9}$$

For any record in the synthetic dataset for which there is no corresponding record in the original dataset with the same key variable values, the denominator in Equation 4 will be zero and the TCAP is therefore undefined. If the TCAP score is 0 then the synthetic dataset carries little risk of disclosure; if the dataset has an TCAP score close to 1, then for most of the riskier records disclosure is possible. We will be using four different size keys of three to six variables (Shown in Appendix D).

4 Results

4.1 Utility

Table 1 shows the narrow measures being averaged together. (Summaries of the individual analyses can be found in Appendix E). Table 1 shows the Raab dataset has the best utility, with the Pistner et al dataset scoring the lowest. It is clear by looking at Table 1 that the Raab dataset performed so well due to its high performance with the regression models and the MCA. When only considering the frequency tables and means, all datasets, with the exception of the Pistner et al dataset, performed equally well.

[5]This choice is a simplifying assumption for our scenario and other values for the WEAP score are feasible. However, for the comparative purposes these are broadly monotonic

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4	Source Tables
Avg freq tables	0.863	0.884	0.528	0.913	0.899	0.882	0.931	0.892	4 &5
Avg cross- tabs	0.78	0.621	0.347	0.758	0.799	0.676	0.748	0.73	6
Avg CIO of means	0.88	0.505	0.145	0.676	0.632	0.575	0.706	0.62	7
Avg regression models	0.480	0.429	0.0303	0	0.180	0.221	0.2467	0.255	8
1-4*pMSE	0.999	0.879	0.831	0.999	0.863	0.878	0.879	0.879	9
MCA score	0.816	0.407	0.5425	0.197	0.611	0.4685	0.4155	0.856	10
Mean	0.803	0.621	0.404	0.591	0.664	0.617	0.654	0.705	

Table 1: Overall Utility Score

4.2 Disclosure Risk

Table 2 shows the mean TCAP score for each key. The Pistner dataset performed best since it had the lowest risk score. One other thing to note beyond the average risk scores, is how the risk scores interact with the keys. Ideally as the key becomes smaller, the TCAP score will become smaller. Since only the synthetic records with a CAP score of 1 are compared to the original, these would be the more sensitive records and therefore they should be less likely to match onto the original. As shown in Table 2 this is the case for the Pistner dataset wherein the TCAP decreases with key size. However, for many of the other datasets the TCAP size does not uniformly decrease, though it may for some variables.

Target	Key	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Employment	6	0.734	0.747	0.667	0.673	0.727	0.701	0.745	0.738
	5	0.728	0.750	0.674	0.668	0.717	0.701	0.709	0.704
	4	0.731	0.724	0.648	0.635	0.734	0.727	0.681	0.729
	3	0.782	0.775	0.554	0.597	0.550	0.847	0.667	0.791
Occupation	6	0.196	0.246	0.121	0.127	0.214	0.233	0.260	0.276
	5	0.207	0.256	0.154	0.117	0.174	0.254	0.269	0.265
	4	0.196	0.266	0.153	0.088	0.195	0.277	0.284	0.238
	3	0.038	0.400	0.074	0.179	0.250	0.519	0.321	0.372
Household Size	6	0.284	0.257	0.173	0.214	0.233	0.247	0.256	0.228
	5	0.278	0.276	0.143	0.251	0.218	0.226	0.273	0.251
	4	0.272	0.231	0.073	0.161	0.176	0.168	0.219	0.188
	3	0.3	0.186	0	0.091	0.200	0.077	0.234	0.175
Average		0.396	0.426	0.286	0.317	0.366	0.415	0.410	0.413

Table 2: Mean TCAP Scores

4.3 Total Score

We approached the total Risk-Utility score in four different ways. The first approach was to subtract the risk score from the utility. If the number was positive than we could informally say that the dataset had more utility than risk.The second approach is to take the minimum of the utility score and the inverse risk score. The third approach is to multiply together the utility score and the inverse risk. The final

approach is to take the geometric mean of the utility score and the inverse risk score.

As shown in Table 3 the Raab dataset had the highest score for all approaches, driven by its very high utility. The Pistner dataset performed worst, despite performing best on the risk assessment, due to having the lowest utility score. However, it should be noted that all datasets had higher utility scores than risk scores. Figure 1 shows a scatter plot mapping the risk-utility scores, creating an R-U map as suggested by Duncan and Stokes (2004).

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4	Avg
Utility score	0.803	0.621	0.404	0.591	0.664	0.617	0.654	0.705	0.632
TCAP (Risk) score	0.396	0.426	0.286	0.317	0.366	0.415	0.41	0.413	0.379
Utility - TCAP	0.407	0.195	0.118	0.274	0.298	0.202	0.244	0.292	0.254
Min(Utility, Inverse TCAP)	0.604	0.574	0.404	0.591	0.634	0.585	0.59	0.587	0.571
Utility * Inverse TCAP	0.485	0.356	0.288	0.403	0.421	0.361	0.386	0.414	0.441
Geometric mean(Utility, Inverse TCAP)	0.696	0.597	0.537	0.634	0.649	0.601	0.621	0.643	0.622

Table 3: Overall Risk Utility Scores

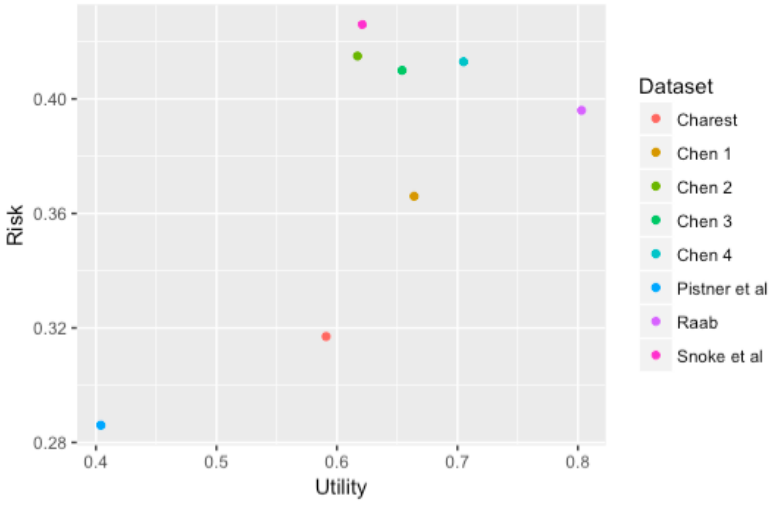

Figure 1: The Risk- Utility Map of the Different Synthetic Datasets

5 Discussion

Overall, the findings show that as utility increases, risk increases (see Figure 1). These findings also give us insight into how different types of synthetic data interact in terms of risk and utility.

5.1 Comparisons of the Synthetic Datasets

Both Raab and Snoke et al's datasets were synthesised using a CART method and *synthpop*, which led to high utility scores. However, the Raab dataset outperforms the Snoke et al dataset in terms of both utility and risk. The Raab synthesis was more customised to the specific features of the data. It used an area-level score (equivalent to modern area-level deprivation measures) as a predictor that was subsequently dropped from the data set. The predictors used in each conditional distribution were defined. These choices and the stratification by region were based on her knowledge of the social structure of the area at that time. Snoke et al. also customised their synthesis, but focused mainly on the computational problems posed by variables with many small categories. Additionally, the two teams used different synthesising orders.(See Appendix C for more details). It may be that the customisation of the synthesis used by Raab helped maintain the utility while not greatly affecting the risk score.

Pistner et al took extra measures to reduce the risk score of the synthetic dataset. Quantile regression as described in Pistner et al (2018) is designed to lower the disclosure risk. This calculation paid off, given that not only does the Pistner et al dataset have the lowest risk scores, the risk scores also reduce as the key size reduces (as discussed in Section 4.2). However, the Pistner dataset also had the lowest utility scores. One interesting thing to note is that the Pistner et al and the Snoke et al datasets have very similar synthesising orders (as shown in Appendix C). This perhaps indicates that the method of synthesis may affect the datasets risk-utility profile independent of the synthesising order.

The Charest dataset, like the Pistner et al dataset, tended towards lower risk scores. The one interesting thing to note is that the Charest dataset performed remarkably well on the pMSE. The explanation to this may lie with how the pMSE was calculated. We speculate that the Charest dataset may have performed so well because we did not use interaction variables when calculating the pMSE score. As shown in Section 4.1, the Charest dataset did not perform as well on utility tests that involved interactions between variables, such as the regression models.

The Chen datasets are interesting in that all four were synthesised using the same process. However, as shown in Figure 1, the Chen 1 dataset has a different profile to the other three Chen datasets, showing that even within a synthesis method the resultant datasets can vary in profile. In future studies, it would be interesting to run multiple trials of all the synthesis methods applied, to ensure that the behaviour of any given datasets is not due to random effects.

5.2 Calculation of the Risk-Utility Score

The way the Risk-Utility Score was calculated may also have a bearing on how the different synthetic datasets were evaluated. One problem with all methods for calculating the overall risk-utility score is that although both the utility and

risk scores are scaled between 0 and 1, in terms of their interpretation, they are measuring different things. Hence, the fact that all synthetic datasets have higher utility scores than risk scores may not actually be informative. It may be helpful if future research is able to create risk and utility metrics that work in tandem with a single scale.

An alternative to creating a one-scale risk-utility metric, would be to approach the problem by determining what is a sufficient utility score or a sufficient risk score and then merely optimising the other score (as suggested by Duncan and Stokes, 2004). However, this solution also has its complications. What is an acceptable utility score is to a certain extent a matter of judgement about the types of analyses that will be performed on the dataset. Likewise, what an acceptable risk score is depends in part on the sensitivity of a given dataset. Additionally, despite being grounded, the risk and utility metrics are nevertheless abstractions of some latent construct and while we expect that the metrics that we have created will be monotonic with actual risk and utility, where the cut off is for high utility or low risk is difficult to objectivise.

6 Conclusion

This paper has presented and discussed a case study that evaluates and compares different synthetic datasets. More trials will be need to draw conclusive results on the best way to generate synthetic data. However, mixing matching some of the aspects that worked well for each given dataset might be an interesting avenue to explore.

We have pointed out the need for more work developing risk and utility measures into a single framework. Future work along these lines could also compare synthesised data to that to which more traditional SDC has been applied.

References

[1] Caiola, G. and Reiter, J. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. Transactions on Data Privacy, 3, pp.27-42.

[2] Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. Lecture Notes in Computer Science, 6344, pp.148-161.

[3] Drechsler, J. (2011). Synthetic Data Sets for Statistical Disclosure Control. New York: Springer Science+Business Media.

[4] Domingo-Ferrer, J. and Rebollo-Monedero, D. (2009). Measuring Risk and Utility of Anonymized Data Using Information Theory. In Proceedings of the 2009 EDBT/ICDT Workshops, pp.126-130.

[5] Duncan, G., Elliot, M. and Salazar-Gonzales, J. (2011). Statistical Confidentiality: Principles and Practice. 1st ed. Springer New York.

[6] Duncan, G. and Stokes, S. (2004). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding. Chance, 17(3), pp.16-20.

[7] Elliot, M. (2014). Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. [online] CMIST. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF%20final.pdf [Accessed 17 Mar. 2017].

[8] Karr, A., Kohnen, C., Oganian, A., Reiter, J. and Sanil, A. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60(3), pp.224-232.

[9] Nowok, B., Raab, G. and Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), pp.1-26.

[10] Pistner, M., Slavkovic, A., and Vilhuber, L. (2018). Synthetic Data via Quantile Regression for Heavy-Tailed and Heteroskedastic Data. J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126. pp.92-108

[11] Purdam, K. and Elliot, M. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environment and Planning A, 39, pp.1101-1118.

[12] Raab, G., Nowok, B. and Dibben, C. (2017) Guidelines for Producing Useful Synthetic Data Preprint available at: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/407/390 [Accessed 31 May. 2019].

[13] Raab, G., Nowok, B. and Dibben, C. (2016) Practical data synthesis for large samples. Journal of Privacy and confidentiality. Available at: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/407/390 [Accessed 31 May. 2019].

[14] Read, T. and N. Cressie (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York.

[15] Reiter, J. (2005). Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, 21, pp. 441-462.

[16] Rubin, D. B. (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9(2), 461-468.

[17] Snoke, J., Raab, G., Nowok, B., Dibben, C. and Slavkovic, A. (2018). General and specific utility measures for synthetic data. Journal of the Royal Statistical Society Series A (Statistics in Society).

[18] Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126. pp. 122-137

[19] Taub, J., Elliot, M., and Saukshaug, J. (2017) A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records. In proceedings of UNECE Statistical Data Confidentiality Work Session.

[20] Woo, M., Reiter, J., Oganian, A. and Karr, A. (2009). Global Measures of Data Utility for Microdata Masked of Disclosure Limitation. The Journal of Privacy and Confidentiality, 1(1), pp.111-124.

Appendices

A Description of Variables

Variable name	Class	No. of categories	Description
household	numeric	82851	Household number
parish	Factor	29	Parish
enum_dist	Character	888	Nested within enumeration district
sex	Factor	2	labels M, F
age	numeric	90	Numeric age in years calculated from age in years and months (for <1)
mar_stat	Factor	6	Marital status*
disability	Factor	7	Type of disability (grouped)*
occlab1	Factor	23	Historic classification of occupations see I-CeM Guide level 1*
occlab2	Factor	75	HISCO level 2 nested within level 1
occlab3	factor	738	HISCO level 3 nested within level 2
employ	factor	3	Whether employer (E) or worker (W) or blank
inactive	factor	9	Reason for inactivity *
ctry_bth	factor	78	Codes for country of birth see I-CeM guide Includes UNK for not known
hsize	Integer	26	Household size
nservants	integer	12	Number of servants
nboarders	Integer	12	Number of boarders
nlodgers	integer	12	Number of lodgers
nvisitors	integer	10	Number of visitors
nfamgteq15	integer	17	Numbers of relatives (including step-children in-laws etc.) aged 15 or over
nfamlt15	integer	16	Numbers of relatives (including step-children in-laws etc.) aged under 15
nkn	integer	9	Number with relationship not known
totrooms	numeric	46	Number of rooms
pperroom	Numeric	269	Persons per room
roomsperp	numeric	269	Rooms per person

B Description of variables derived for this study

Variable	Original Variable derived from	Description
Disabled	disability	0=no disability 1=disability present
Edinburgh	parish	1=Edinburgh 0= other parish
Servants	nservants	0= no servants 1= servants present
Boarders	nboarders	0= no boarders 1= boarders present
Lodgers	nolodgers	0= no lodgers 1= lodgers present
Visitors	nvisitors	0= no visitors 1= visitors present
Unknown relationship	nkn	0= no unknown relations 1= unknown relations present
Family over 15	nfamgteq15	0= no family over 15 1= family over 15 present
Family under 15	nfamlt15	0= no children under 15 1= children under 15 present
Female	Sex	0= male 1= female
Single	mar_stat	1= single 0= other relationship status
Spouse absent	mar_stat	1= spouse absent 0=other relationship status
Widowed	mar_stat	1=widowed 0=other relationship status
Employer	employ	1=employer 0= other work status
Worker	employ	1=worker 0=other work status
Age group- utility	age	0-19, 20-39, 40-59, 60-80, 80+
Age group- dcap	age	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+

C Additional Information about Synthesis Methods

C.1 Synthesis Order

Raab {parish, enumdist, area score, nfamgteq15, nfamlt15, age, sex, inactive, occlab1, occlab2, occlab3, mar_stat, newcob, ctry_bth, employ, disability, nboarders, nlodgers, nvisitors, nkn, nservants, hsize, totrooms}
Snoke et al: {age, nservants, nboarders, nlodgers, nvisitors, nfamgteq15, nfamlt15, nkn, totrooms, sex, employ, mar stat, disability, inactive, occlab1, parish, ctry_bth, occlab2, occlab3, enum_dist}
Pistner et al: {age, nservants, nboarders, nlodgers, nvisitors, nfamgteq15, nfamlt15, nkn, totrooms, sex, mar_stat, employ, inactive, disability, ctry_bth, parish, enumdist, occlab1, occlab2, occlab3}

C.2 Additional Information on Quantile Regression

Variables are synthesised one at a time with all previous synthesised variables used as predictors.

For continuous variables, the synthesis was conducted as follows. First, we partitioned (0,1] into 25 separate bins, i.e., (0, 0.04], (0.04, 0.08], ..., (0.96, 1]. Using these partitions, we fit 25 separate quantile regression models with equal to the midpoint of the respective bin. Then, for each row in the data set, we generated a random quantile according to a Unif(0,1) random variable. With each row, we found the corresponding bin and used the corresponding model to generate the synthetic data value. For the first variable that was synthesised, age, a bootstrap approach was used instead of one based off of quantile regression.

Synthesis of categorical variables was not as straightforward. For a given categorical variable, first it had to be recorded into numerical form. For simplicity, we recoded these variables such that the highest frequency level was now equal to one, the second highest frequency level was now equal to two, and so on. Quantile regression was then used to synthesise with these new values as responses. Next, the continuous predictions had to be recoded back to the original categorical levels. To do this, quantile cut-offs for each of these levels was calculated from the original codings. Then, the predicted value for each of these quantiles was calculated from the predictions and the data were binned according to these cutoffs.

D TCAP keys

The TCAP score will be calculated using the following key variables:

Key 6= sex, age group, marital status, parish, country of birth, presence child
Key 5= sex, age group, marital status, parish, presence child
Key 4= sex, age group, marital status, parish
Key 3= sex, marital status, parish

E Breakdown of Results

Table 4 shows the ROC and CIO for 8 different variables. While Table 5 shows the ROC and CIO for variables that were originally continuous, but due to their low number of counts were recoded as artificial binaries where >= 1 is 1 and 0 is 0.

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Marital Status	0.985	0.959	0.967	0.995	0.987	0.982	0.986	0.992
	0.877	*0.497*	*0.569*	*0.939*	*0.759*	*0.786*	*0.799*	*0.87*
Sex	0.997	0.996	0.996	0.999	0.999	0.997	0.998	0.998
	0.786	*0.757*	*0.73*	*0.94*	*0.913*	*0.825*	*0.854*	*0.885*
Employ	0.993	0.992	1	0.996	0.996	0.999	0.996	0.999
	0.804	*0.786*	*1*	*0.893*	*0.843*	*0.97*	*0.844*	*0.641*
Country of Birth	0.989	0.999	0.936	0.999	0.997	0.998	0.999	0.998
	0.382	*0.964*	*0*	*0.937*	*0.814*	*0.872*	*0.966*	*0.876*
Disabled	0.953	0.985	0.725	0.982	0.994	0.979	0.993	0.987
	0.598	*0.875*	*0*	*0.845*	*0.953*	*0.826*	*0.937*	*0.889*
Edinburgh	1	0.999	0.936	0.999	0.997	0.998	0.999	0.998
	1	*0.892*	*0.973*	*0.812*	*0.954*	*0.767*	*0.803*	*0.965*
Occupation	0.962	0.978	0.772	0.976	0.984	0.979	0.977	1
	0.669	*0.734*	*0.418*	*0.755*	*0.81*	*0.762*	*0.744*	*1*
Age group	0.98	0.976	0.58	0.985	0.79	0.978	0.984	0.98
	0.854	*0.648*	*0*	*0.832*	*0.834*	*0.862*	*0.874*	*0.847*
Mean	0.982	0.986	0.864	0.991	0.968	0.989	0.991	0.994
	0.746	*0.769*	*0.461*	*0.869*	*0.86*	*0.834*	*0.853*	*0.871*
Avg of ROC and CIO	0.864	0.8775	0.6625	0.93	0.914	0.9115	0.922	0.9325

Table 4: Mean Ratio of Counts and CIO (in italics) for Univariate Frequency Tables

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Servants	0.986	0.998	0.864	0.99	0.99	0.989	0.994	0.993
	0.653	*0.943*	*0*	*0.747*	*0.736*	*0.71*	*0.846*	*0.815*
Boarders	0.994	0.988	0.919	0.991	0.992	0.992	0.997	0.997
	0.815	*0.603*	*0*	*0.705*	*0.744*	*0.736*	*0.901*	*0.925*
Lodgers	0.993	0.995	0.966	0.992	0.992	0.995	0.996	0.99
	0.756	*0.819*	*0*	*0.742*	*0.728*	*0.832*	*0.875*	*0.672*
Visitors	0.978	0.984	0.498	0.964	0.991	0.951	0.997	0.96
	0.8	*0.713*	*0*	*0.909*	*0.819*	*0.845*	*0.787*	*0.845*
Unknown relationship	0.978	0.984	0.498	0.964	0.991	0.951	0.997	0.96
	0.806	*0.863*	*0*	*0.674*	*0.925*	*0.567*	*0.97*	*0.646*
Family over 15	0.794	0.917	0.5	1	0.885	0.833	0.955	0.812
	0.519	*0.848*	*0*	*1*	*0.777*	*0.644*	*0.922*	*0.58*
Family under 15	1	0.998	0.981	0.998	0.997	0.998	0.999	0.995
	0.979	*0.822*	*0*	*0.831*	*0.791*	*0.84*	*0.93*	*0.651*
Mean	0.96	0.981	0.747	0.986	0.977	0.958	0.991	0.958
	0.761	*0.802*	*0.143*	*0.803*	*0.789*	*0.739*	*0.89*	*0.733*
Avg of ROC and CIO	0.8605	0.8915	0.445	0.8945	0.883	0.8485	0.9405	0.8455

Table 5: The Ratio of Counts and CIO(in italics) for Univariate Frequency Tables of Household Member types

Table 6, like with tables 4 and 5 used both the ROC and CIO measurements but for cross-tabs consisting of variables that we thought a researcher might cross reference.

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Marital status	0.967	0.953	0.744	0.987	0.95	0.968	0.969	0.97
by sex	*0.762*	*0.602*	*0.21*	*0.858*	*0.715*	*0.786*	*0.805*	*0.774*
Age group	0.949	0.887	0.642	0.98	0.995	0.939	0.978	0.973
by sex	*0.707*	*0.274*	*0.161*	*0.847*	*0.901*	*0.436*	*0.57*	*0.64*
Marital status	0.882	0.953	0.13	0.859	0.679	0.85	0.933	0.802
by age group	*0.628*	*0.341*	*0.114*	*0.644*	*0.342*	*0.625*	*0.804*	*0.404*
Employ	0.861	0.794	0.61	0.716	0.974	0.749	0.791	0.912
by age group	*0.483*	*0.16*	*0.157*	*0.167*	*0.834*	*0.0487*	*0.133*	*0.362*
Mean	0.915	0.897	0.532	0.886	0.9	0.877	0.918	0.914
	0.645	*0.344*	*0.161*	*0.629*	*0.698*	*0.474*	*0.578*	*0.545*
average of roc and cio	0.78	0.621	0.347	0.758	0.799	0.676	0.748	0.73

Table 6: Ratio of Counts and CIO (in italics) for Bivariate Cross-tabulations

Table 7 shows the CIOs of the means and standard deviation of six count and continuous variables.

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Age	0.964	0.850	0	0.885	0.845	0.725	0.891	0.911
People per room	0.747	0.223	0	0	0	0	0	0
Household size	0.934	0.202	0	0.975	0.867	0.498	0.866	0.525
Number of Rooms	0.689	0.953	0.652	0.687	0.214	0.982	0.909	0.958
Family over 15	0.949	0.321	0	0.819	0.927	0.461	0.709	0.765
Family under 15	0.999	0.486	0.0714	0.689	0.939	0.785	0.891	0.566
Mean	0.880	0.505	0.120	0.676	0.632	0.575	0.706	0.620

Table 7: CIO for Count/Continuous Variables

Table 8 shows the mean CIO for the OLS regression models and logistic regression.

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
people per room-linear	0.366	0.365	0.0504	0	0.0888	0.115	0.247	0.294
family under 15-binary	0.593	0.493	0.0101	0	0.272	0.327	0.246	0.216
mean	0.480	0.429	0.0303	0	0.180	0.221	0.247	0.255

Table 8: CIO scores for Regression models

Table 9 shows the pMSE scores for the different datasets. Please note that only the pMSE score is shown in Table 1, not the standardised pMSE and pMSE ratio.

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
pMSE	0.000211	0.0303	0.0422	0.000211	0.0342	0.0305	0.0303	0.0302
Standardized pMSE	9.269	2478.909	3464.961	9.269	2807.305	2497.305	2485.409	2471.225
pMSE ratio	2.150	308.471	430.776	2.150	349.203	310.753	309.277	307.518

Table 9: pMSE Scores

Table 10 shows the Euclidean distance between the original and synthetic datasets.

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Married	0.0118	0.00768	0.0124	0.0235	0.0264	0.0136	0.0228	0.00694
Spouse absent	0.176	2.829	0.149	2.746	0.569	2.764	3.192	0.0599
Not known	2.142	2.305	5.621	3.249	4.281	0.199	2.000	2.791
Single	0.184	2.383	1.025	1.850	0.126	2.100	2.476	0.255
Widowed	0.0176	0.239	0.477	0.0550	0.0195	0.176	0.174	0.0572
F	0.0229	0.0694	0.0509	0.0215	0.0336	0.0261	0.0294	0.00588
M	0.00726	0.0220	0.0177	0.00779	0.0108	0.00754	0.00926	0.00123
Employer	0.152	4.033	0.620	3.146	0.818	2.403	3.974	0.224
Not working	0.0114	0.135	0.0906	0.117	0.0431	2.828	0.160	0.0230
Worker	0.0143	0.360	0.111	0.270	0.109	0.367	0.333	0.0182
Edinburgh	0.125	0.860	0.663	0.562	0.367	0.796	0.783	0.0544
Other Parish	0.202	1.385	1.066	0.901	0.590	1.275	1.254	0.0881
Able	0.00653	0.00631	0.00655	0.0330	0.0167	0.00434	0.00524	0.00289
Disabled	2.409	1.959	1.817	9.849	5.257	1.483	1.632	0.881
Scotland	0.0603	0.378	0.438	0.456	0.0263	0.409	0.423	0.0239
Other Country	0.340	2.000	2.473	2.410	0.146	2.159	2.236	0.121
Mean	0.368	1.186	0.915	1.606	0.778	1.063	1.169	0.288
utility score	0.816	0.407	0.5425	0.197	0.611	0.4685	0.4155	0.856

Table 10: Euclidean distance between MCA for Original and Synthetic Datasets