


Utility Assessment of Synthetic Data Generation Methods

Md Sakib Nizam Khan¹ , Niklas Reje¹, and Sonja Buchegger¹

KTH Royal Institute of Technology, Stockholm, Sweden
{msnkhan , nreje, buc}@kth.se

Abstract. Big data analysis poses the dual problem of privacy preservation and utility, i.e., how accurate data analyses remain after transforming original data in order to protect the privacy of the individuals that the data is about - and whether they are accurate enough to be meaningful. In this paper, we thus investigate across several datasets whether different methods of generating fully synthetic data vary in their utility a priori (when the specific analyses to be performed on the data are not known yet), how closely their results conform to analyses on original data a posteriori, and whether these two effects are correlated. We find some methods (decision-tree based) to perform better than others across the board, sizeable effects of some choices of imputation parameters (notably number of released datasets), no correlation between broad utility metrics and analysis accuracy, and varying correlations for narrow metrics. We did get promising findings for classification tasks when using synthetic data for training machine-learning models, which we consider worth exploring further also in terms of mitigating privacy attacks against ML models such as membership inference and model inversion.

Keywords: Synthetic Data · Utility · Metrics · Analysis · Correlation.

1 Introduction

In this era of big data and artificial intelligence, technologies are becoming increasingly dependent on data processing and analysis. Since a fair share of the data used by these technologies is the privacy-sensitive data of individuals, there is a growing need for methods that facilitate privacy-preserving data analysis and sharing. Differential privacy [11] and k-anonymity-related [32] mechanisms focus on providing privacy guarantees for unlinkability, at the cost of utility due to the noise additions required by the mechanisms.

A promising alternative is to generate synthetic data from the original data to use for analysis instead. Firstly, with synthetic data, there is no given direct linkability from records to individuals as the records are not real. Secondly, given the large variety of ways to generate synthetic data, there is potential for better utility to arrive at an acceptable level of accuracy for a given analysis.

The main idea of statistical methods for synthetic data generation is imputation [30] which is originally developed for replacing missing data in survey

results. There are different choices of parameters related to imputation such as which other variables to choose to impute one variable, the order of variables chosen for imputation, etc. which can impact the utility of the generated data. Similarly, depending on the dataset the choice of the synthesizer can have a large impact on the utility of the synthetic data since some synthesizers perform better for numerical variables and some perform better with categorical variables. There has been a lot of research on synthetic data utility in recent years however, there still exists some research gaps concerning the impact of different choices during the synthetic data generation process on the utility. Further investigation is also required to find out if there is any correlation between the general utility metrics commonly used to evaluate synthetic data and how well the synthetic data performs for a given analysis.

In this work, we thus investigate the utility provided by different synthetic data generation techniques and imputation parameters, on separate but similar datasets (Adult and Polish, both census-type data) and on a dataset with different characteristics (Avila). First, we determine the effects on utility as measured by various metrics for the similarity between the original and the synthetic data and, second, by comparing the similarity of results from analyses performed on the original versus on the synthetic data. Third, we investigate to what extent the first can predict the second when the analysis is not known beforehand. In summary, our main contributions are:

- A comprehensive evaluation of synthetic data utility using three publicly available datasets.
- Identification of the individual impact of the choice of variables during imputation, imputation order, sampling, and number of datasets on the utility of synthetic data.
- A comprehensive study on the correlation between different utility metrics and how well the synthetic data performs for any given analysis.

Organization. The rest of the paper is organized as follows. In Section 2, we discuss the related work, followed by an overview on the synthetic data generation process and the utility metrics in Section 3. We then present our experimental results in Section 4, followed by our discussion and concluding remarks in Section 5.

2 Related Work

There have been various research works concerning synthetic data generation and the utility provided by them. The related works that we have distilled from the literature mostly focus on the comparison between imputation methods (single and multiple imputation [33], hierarchical Bayes and conventional generalized linear imputation models [13]), evaluation of different mechanisms and tools for synthetic data generation (fully and partially synthetic data [2], tools for regression [15] and classification [14] tasks), evaluation of the usefulness of synthetic

data [8,12,19,20,34], and what utility metrics to use for comparing original and synthetic data [3,31]. In this section, we review some of these works in detail.

Among the works on imputation methods, Taub et al. [33] evaluated the real word analyses replicability of singly and multiply imputed CART generated synthetic data using Purdam and Elliot’s [24] methodology. To evaluate the impact of disclosure control on the analysis outcomes, Purdam and Elliot [24] replicated the published analysis on datasets using the disclosure controlled versions of the same datasets. Based on this methodology, Taub et al. [33] replicated 9 different sets of analyses involving 28 different tests and models. According to the findings of the authors, multiply imputation performed better for some analyses, but not for all, and depending on the complicity of the analysis, single imputation can be useful in some scenarios. The authors also investigated the relationship between the utility metrics and how well a synthetic dataset performs for a given analysis and found that there is no clear relationship.

To investigate the usefulness of synthetic data, Nowok [20] evaluated the performance of non-parametric tree-based synthetic data generation methods (Classification and Regression Trees (CART), bagging, and random forests) using synthpop. Besides general utility measures, the authors used some hypothetical analyses to evaluate analysis specific utility of the synthetic data. From the empirical evaluation, the authors conclude that it is possible to produce useful completely synthetic data using automated methods. Similarly, Drechsler et al. [8] replicated an already published analysis of a dataset using synthetic data and found that the regression coefficients of the synthetic and the original were almost identical and concluded that the authors of the published analysis would have drawn the same conclusion using the synthetic dataset. Lee et al. [19] used univariate, bivariate, and linear regression-based exploratory data analysis to evaluate the utility of synthetic data generated using CART models and found that for univariate analysis synthetic data results match the original data, whereas for bivariate and linear regression it was not true. According to the authors, the reason for this could be that the CART model underestimates the strong correlation between the variables in the dataset.

Among the works on utility metrics, Dankar et al. [3], first classified the available utility metrics for synthetic data comparison into different categories based on the measure they attempt to preserve. Then the authors chose one metric from each category depending on popularity and consistency and used them to compare the utility of four data synthesizers (i.e Data-Synthesizer (DS), Synthetic Data Vault (SDV), Synthpop Parametric, Synthpop Non-parametric). According to the authors, their experimental results show that the Synthpop Non-parametric is the best performing synthesizer overall and provides the best average values across all metrics as well as the best (overall) stability and consistency. Snoke et al. [31] investigated how general utility compares to the specific utility for synthetic data and also presented two contrasting example evaluations. The authors conclude that general utility measures can be used to tailor the methods that are used to synthesize datasets for specific purposes whereas

specific utility measures can be used for reassurance after performing any standard exploratory analysis on the synthetic data that the data was not misleading.

Even though there have been a lot of studies on different aspects of synthetic data generation, there is still a lack of understanding regarding how choice and order of variables during imputation, proper and non-proper synthesis, number of datasets, etc. affect the utility of generated synthetic data. Similarly, there is a lack of concrete work on the correlation between commonly used utility metrics for synthetic data and the performance of synthetic data in any given analysis.

3 Synthetic Data Generation

Figure 1 provides an overview of synthetic generation process and its two main components, Data Synthesis and Evaluation.

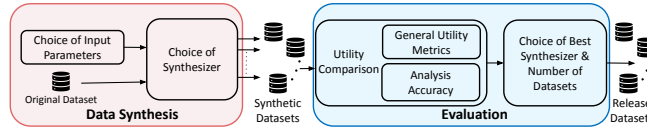


Fig. 1: Synthetic Data Generation Process

3.1 Data Synthesis

Choice of Input Parameters. The goal of the data synthesis component is to generate synthetic datasets based on the original dataset. The synthesis process requires a set of input parameters related to imputation.

In statistics, imputation is the process of replacing missing data with substitute values drawn from similar records. It was originally developed for solving the problem of missing values in surveys. Rubin [29] developed an approach called multiple imputation and later on [30] proposed to use it for generating fully synthetic datasets. The basic idea proposed by him is to treat the population not selected in a sample as missing data and then use multiple imputation to create synthetic datasets. For multiple imputations, two general approaches have emerged over the years: the joint modeling approach and the fully conditional specification (FCS) approach [5]. With the FCS approach variables are replaced in a specific order whereas with the joint modeling approach multiple variables are replaced together. More details about these approaches can be found in [5]. Here, we present the three main parameter choices for imputation.

Imputation Order. For the joint modeling approach, the values in each record are generated at once whereas for FCS a particular order needs to be selected to impute the variables. The order that seems most logical is usually chosen. There are also other approaches [27] for choosing the order. Nonetheless, the order needs to be chosen based on the relations between the variables and

also the computational complexity. Since the order impacts which relationships are kept between the variables, it can also impact the utility of synthetic data

Simple & Selective Imputation. For the FCS approach, it is possible to select only a subset of variables for imputation of another variable (we term this approach Selective) instead of using all other variables (which we call Simple). The Selective approach can be taken in an effort to reduce computation time or to preserve the relationships between variables that are most essential. However, the downside is that the correlations that are not chosen can get lost, and knowing which relationships to preserve is also not straightforward.

Proper & Non-proper Imputation. In Synthpop, one can chose to use bootstrap samples from the original dataset (Proper) or the entire original dataset (Non-Proper) while generating imputation models.

Choice of Synthesizers. There are several methods developed over the years for generating synthetic data using multiple imputation. We use three of the most widely used methods: Parametric and Decision Trees based on the FCS approach and Saturated model based on the joint modeling approach.

Parametric. In parametric method, the first variable is chosen together with a suitable regression model. A subset of the variables are then fitted to it and beta-coefficients (β) and variances (σ^2) are estimated. Those are then used together with Y_{-j} in some distribution like: $N(\beta Y_{-j}, \sigma^2)$ to draw synthetic values from. When that is done the second variable is picked and a regression is made to fit the other variables to it but this time the new synthetic values are used instead. The parametric method uses different kind of regression models depending on what type of data values are meant to be synthesized with linear regression used for numerical variables and logistic regression used for categorical variables.

Decision Trees. Decision Trees or as they are commonly referred to in machine learning Classification and Regression Trees (Cart) is a non-parametric alternative to the Parametric methodology [22]. Using decision trees to generate synthetic data was introduced in [27] and was based on a previous work on replacing missing values, just like Parametric. A decision tree can be used to represent a multivariate conditional distribution so instead of trying to fit a regression model on the variable, a tree is created for the imputed variable which is split on the subset of variables.

A decision tree can be visualized as a flow chart where at each decision, the tree is split i.e. a test is made against the value of some variable/s for which path to take. After going through the tests down the tree, one ends up at a leaf which holds some number of values for the variable we are imputing. To create such a tree, it starts with all values for the variable as one leaf each with splits based on the values of the other variables. The tree is then pruned so that some leaves are combined and the remaining leaves hold a number of different values. After going down the tree to one leaf, one value is sampled from the values within the leaf.

Saturated Model. The Saturated Model approach works by fitting a model that perfectly reproduces the data, to a cross-tabulated table of all the categorical variables [18,23]. Certain combinations in the cross-tabulation can be manually excluded from being used for imputation by giving them a zero probability in the multinomial distribution. Synthetic values are then imputed by sampling from the multinomial distribution where the parameters come from the probabilities of each variable combination. Since this method only works with categorical variables, a parametric method or decision tree is needed in combination for the numerical variables.

3.2 Evaluation

For evaluating the utility of synthetic datasets a common approach found in the literature is to use different utility metrics. Besides this, some studies have also used the accuracy of analysis to measure how well synthetic data imitate the original data for any given analysis. In this section, we briefly discuss these two comparison approaches.

General Utility Metrics. The existing works on synthetic data use different metrics to estimate the utility provided by such data. However, there is no such metric that can individually capture all the different aspects of utility concerning synthetic data. Hence, a combination of different metrics is needed to compare the utility provided by synthetic data and the original data. There are two categories [9,16] of utility measures which are commonly termed broad measures and narrow measures of utility. The broad measures mostly compare the utility between the entire distributions of synthetic and original data using some statistical distance measures such as Kullback-Leibler (KL) divergence [16] score. Narrow measures on the other hand tend to compare specific models (e.g., regression, point estimation, etc.) between the synthetic and original dataset and are more widely used in the literature for the utility comparison of synthetic data. Confidence Interval Overlap (CIO) developed by Karr et al. [16] is a widely used [7,9,33] narrow measure of utility for synthetic data. In this work, we use the both CIO and KL divergence as the general utility metrics.

Confidence Interval Overlap (CIO). To measure the utility of synthetic data based on confidence interval overlap (CIO), we use a combination of mean point estimators and regression fit coefficient estimation. We calculate the 95% Confidence Interval (CI) for the mean point estimators and the regression fits coefficients estimations and the real coefficients from the original data. We then calculate CIO according to the following equation which is proposed by Karr et al. in [16].

$$CIO = \max(0.5 \cdot \left(\frac{\min(U_o, U_s) - \max(L_o, L_s)}{U_o - L_s} + \frac{\min(U_o, U_s) - \max(L_o, L_s)}{U_s - L_o} \right), 0) \quad (1)$$

Where L and U are the lower and upper bounds for the synthetic (s) and original (o) CI. We have a lower bound on CIO because it can otherwise grow to large negative numbers if the CI do not overlap. For the regression fit and mean point

estimation, we calculate the average CI overlap for all the coefficients of each regression fit and estimation. We then take the average over all the tests for each m and k where m is number of dataset to release and k is the number tests performed. These averages is then further averaged over k to get a summary for a specific synthesizer combination and m .

For the regression fit and mean point estimation, we calculate the average CIO for all the coefficients of each regression fit and estimation. We then take the average over all the tests for each m and k where m is number of datasets to release and k is the number tests performed (i.e., the number of different sets of m datasets). These averages are then further averaged over k to get a summary for a specific synthesizer combination and m . For the point estimates and regression fits we need to estimate the variance. There are different equations for variance estimation based on combining rules for imputation proposed by Raghunathan et al. [26]. For more details about the imputation combining rules and variance estimation see [6, 33]. In this work, we use the equation proposed by Raab et al. [25] for the variance estimation. The reason for choosing the equation is that it is not dependent on the number of datasets and thus valid even if there is only one synthetic dataset. The equation is as follows

$$T_s = (1 + \frac{1}{m})\bar{v}_m$$

Kullback-Leibler Divergence. Kullback-Leibler Divergence is used to measure how close two probability distributions are, based on information entropy and is a global measure of data utility [35]. We measure the KL divergence score of each variable individually and then take the average over all the variables for the comparison between synthesizers. KL divergence for discrete distributions are defined as follows:

$$D_{KL}(P|Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

P and Q is then the distribution of a variable from the original dataset and the distribution of the same variable from a synthetic dataset though it does not matter which is P is which is Q as long as it is consistent. It is also possible to calculate both the alternatives and then use the average of the two. For categorical variables, X is just the different values while for the numerical variables, X is a quantization with the upper and lower bounds being the largest and smallest value from the original and the synthetic datasets with a step-size of suitable magnitude.

Average Percentage Overlap (APO) above 90%. A CIO of 90% or higher has been regarded as good. With the APO measure, we can have a measure of confidence in the utility beyond the average CIO by calculating the ratio of the CIOs over 90% to all CIOs.

Analysis Accuracy The problem with broad and narrow measures is that none of them is individually sufficient for measuring the utility of synthetic data. They

both have certain weaknesses. For instance, narrow measures perform really well for certain analyses but are unable to provide any insights on the overall utility of the dataset whereas broad measures perform well for many analyses but really well for none [16]. To overcome these issues, multiple studies use either already published analysis [24, 33] on original data or ad hoc analysis [18, 20] to measure the utility provided by synthetic data. Since this approach uses multiple analyses for the utility comparison, it avoids the possibility of reporting unreasonably high or low utility measures based on a single measure which is the case for broad and narrow measures. In this work, for measuring the utility, we use machine learning-based classification accuracy along with the accuracy of multiple ad hoc analyses which compare the multivariate relationship and univariate distributions between synthetic and original data.

4 Experimental Results

4.1 Datasets and Experimental Environment

In this work, for synthetic data generation, we use the R package `synthpop` developed by Nowok et al. [21]. For the datasets, we use three publicly available ones, *i.e.*, Polish quality of life dataset (SD2011) [1] from `synthpop` example datasets, Adult dataset [17], and Avila dataset [4] from UCI Machine Learning Repository [10].

- **Polish Dataset:** The `synthpop` R package comes with one example dataset which is a Polish census on the quality of life in Poland [1]. We use a modified version of the Polish dataset where 14 out of the 35 variables are kept to simplify the computation. The dataset contains records with missing values which are replaced with similar values using random sampling. The total number of records is 5000 for the polish dataset.
- **Adult Dataset:** The Adult dataset [17] is a popular machine learning census dataset from UCI Repository intended for predicting whether income exceeds 50K/yr based on census data. The dataset contains 15 variables where we removed numerical education level which is a redundant version of the categorical education variable. We also modify the levels of the native country variable from 44 countries to 7 coarser ones, a change that was required for most of the synthesizers to work. To simplify computation, we use the first 10000 records of the dataset.
- **Avila Dataset:** The Avila dataset [4] is also a machine learning dataset from UCI Repository intended for predicting the copyist based on the patterns of segments of a Latin bible. All the variables except the author class are continuous in Avila dataset. The dataset has no missing values and all 11 variables and all 10430 records are used.

While both the Polish and the Adult datasets are standard census surveys, Avila is a completely different type that is not commonly used in existing works on synthetic data generation. We choose Avila because we want to see how well the methods which are mostly developed for census or similar datasets perform for a completely different dataset.

Table 1: Naming Convention

Synthesizer	Symbol	Imputation Order	Name	Suffix	Synthesis Method	Name	Suffix
Standard Parametric (SAP)	P		Original Ordering	No Suffix			
Decision Tree / Classification & Regression Tree (CART)	D		Opposite Ordering	O			
Saturated Model (Catall) with Predictive Mean Matching (CAP)	CP		Own Ordering	V		Non-Proper	No Suffix
Saturated Model (Catall) with CART (CAC)	CC		Largest Categorical Variables First	H		Proper	T
Sample	S		Largest Categorical Variables Last	L			

For the experiments, we generated multiple synthetic datasets for Standard Parametric (SAP or P), Cart (D), Catall (i.e., saturated model) with PMM (CAP or CP), and Catall with Cart (CAC or CC). For parametric (P) and decision trees (D), we generated synthetic datasets with 3 different orders, a total of 6 different datasets. Next, for CAP, CAC, and the 6 datasets for P and D we generated Proper and Non-Proper combinations which sum up to 16 datasets. Additionally, for baseline comparison, we created a fifth synthesizer, Sample, which just samples values from the original dataset for each variable without attempting to retain any relationships between the variables. Thus, for each of the 3 original datasets, we generate 17 dataset collections. Table 1 shows how we symbolize each method and parameter in this work. We use the naming convention shown in the table for the rest of this paper. Our code for all the experiments and an extended version of the paper containing more detailed explanations of the experimental results are available on GitHub¹.

4.2 General Utility Metrics

Confidence Interval Overlap (CIO). To evaluate the utility of synthetic datasets using confidence interval overlap (CIO), we perform mean point estimation and regression fitting. The mean point estimations are performed on the numerical variables of each dataset. In the case of regression fitting, we perform 20 regression fits on the Polish dataset and 24 regression fits each on the Adult and Avila datasets. The details of regression fits for each dataset can be found in [28].

Mean Point Estimation. Table 2 shows the APO above 90% of mean point estimations for all three datasets. In table 2a, for polish dataset, we can immediately see that the mean point estimation for $m = 1, 2$ have some great reduction in utility compared to the higher m for some of the synthesizer combinations. This is due to the variance estimator overestimating the variance, which is reduced with higher m . We can also see that for all synthesizers at $m = 3$ all of the variables have an above 90 ratio. The sample method also performs well here which is expected since it should be able to preserve the univariate distribution of each variable and that is the only thing we test here.

¹ <https://github.com/sakib570/synthetic-data-utility>

Table 2: APO above 90% for Mean Point Estimations

m	S	P	D	CP	CC
1	0.00	0.08	0.00	0.00	0.00
2	0.90	0.99	0.85	0.86	0.86
3	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00
100	1.00	1.00	1.00	1.00	1.00

(a) Polish Dataset

m	S	P	D	CP	CC
1	0.01	0.00	0.00	0.01	0.00
2	0.90	0.50	0.86	0.83	0.82
3	1.00	0.50	0.99	1.00	0.98
4	1.00	0.50	1.00	1.00	1.00
5	1.00	0.50	1.00	1.00	1.00
6	1.00	0.50	1.00	1.00	1.00
7	1.00	0.50	1.00	1.00	1.00
8	1.00	- 1.00	1.00	1.00	1.00

(b) Adult Dataset

m	S	P	D	CP	CC
1	0.05	0.08	0.06	0.14	0.05
2	0.80	0.51	0.80	0.92	0.78
3	0.90	0.68	0.91	0.93	0.90
5	0.95	0.75	0.96	0.92	0.95
10	0.98	0.82	0.99	0.91	0.98
20	1.00	0.86	1.00	0.90	1.00
50	1.00	0.89	1.00	0.89	1.00
100	1.00	0.90	1.00	0.90	1.00

(c) Avila Dataset

For the Adult dataset, the mean point estimation results are mostly similar to Polish with some small differences. In Table 2b, we see the point estimation results for Adult where for smaller m APO above 90% on an average is not achieved for all the synthesizers. However, in the case of Adult, two of the coefficients for P do not have an overlap above 90% regardless of m . The reason can be that the two concerned variables contain a lot of zero values.

The point estimation for Avila dataset has some interesting results. As shown in Table 2c, in the case of Avila, for P and CP there is always at least one variable that is below 90%. D and CC instead have all variables above 90% for large m but for small m the ratio is similar as seen previously for Polish and Adult datasets.

Findings. The mean point estimation results for all three datasets show that at least $m = 3$ or higher is required for all variables to reach APO above 90%. In terms of synthesizer, though CP performed well for Adult and Polish, it struggled for Avila whereas D and CC performed well for all three datasets. P struggled for both Adult and Avila and was unable to reach APO above 90% regardless of the number of datasets released.

Regression Fit. For regression fit, we calculate both the average CIO and the APO above 90% over all the regression fits for each dataset. Figure 2 shows the average CIO for the regression fits.

As we can see from the figure, for the Polish dataset (Figure 2a), CC is the best synthesizer and the only one to get above 90% but CP and D are not that far behind. The Sample method performs far worse than the others which confirms that the other synthesizers have been able to retain relationships between the variables. We can also see that there is a steady increase in overlap as m increases, though there appears to be a somewhat diminishing return with the largest increases being below 10 to 20 datasets and the overlaps above only raising the results a few percentage points at most. In terms of imputation order, the original ordering performs best and the opposite ordering is worse. Non-proper methods perform better overall than proper ones. One interesting observation to note here is that the proper versions are worse on average for small m , however, it approaches close to Non-Proper with higher m .

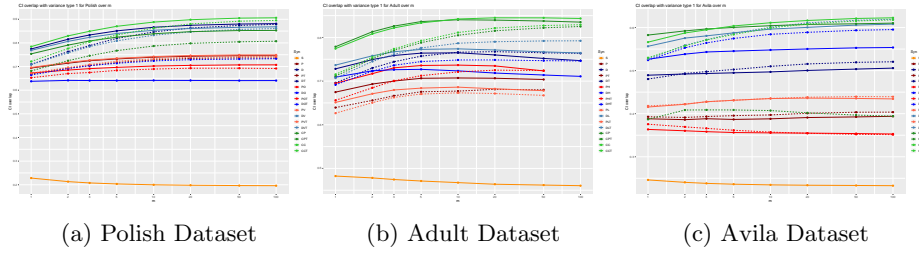


Fig. 2: CIO for Regression Fits

For the Adult dataset (Figure 2b), the results are very similar to the Polish dataset overall with some minor differences. In the case of Adult, CP performs better than D with about a 5 percentage point advantage, which is a bit more than the difference between these two in the case of Polish. CC is again best performing for higher m but the difference between CP and CC is not as large as it was for Polish. P is the worst once again if we do not include Sample. A key difference between the result from Polish and Adult is that the average CI overlap drops for many of the synthesizers as m increases, something which only happened for Sample in the case of Polish. Moreover, the results start to drop or plateau for a lot of the synthesizers after $m = 5$. Only CP and CC have some increase for the larger m which are the only ones that get above 80% average CI overlap. For imputation order, in the case of D, the largest categorical variable last (L) performs slightly better than the original ordering and much better than the largest categorical variable first (H) ordering and for P the H ordering performed better than the other two. In terms of synthesis method except for D for higher m non-proper performs better overall than proper.

In the case of Avila, the results differ from Polish and Adult. As shown in Figure 2c, the results are worse for all synthesizers with the best ones (CC and CP) barely able to reach 60% on average. While some do better than others, they are far from 90 which is in practice useless. As seen previously for the poorly performing synthesizers, with the increase of m the score either plateau very early or starts to decrease indicating that the synthesizers perform poorly with Avila. For Avila, in the case of D and P and own ordering (V) and proper was better than other orders and non-proper synthesis.

In Figure 3, we see the APO above 90% results of regression coefficients for all synthesizer combinations and datasets. The first thing to notice is that the lower m perform significantly worse than higher m for all three datasets. This has to do with how the variance estimation works with smaller m resulting in larger variances which here clearly did not match the real variance.

For Polish (Figure 3a), CC preforms far better than the other synthesizers and gets very close to an average of two thirds for high m . D and CP perform very similarly. P again is not as good as the other synthesizers. For Adult (Figure 3b), the synthesizer combination perform worse overall compare to the Polish dataset. CP with Proper imputation (CPT) is the best synthesizer which per-

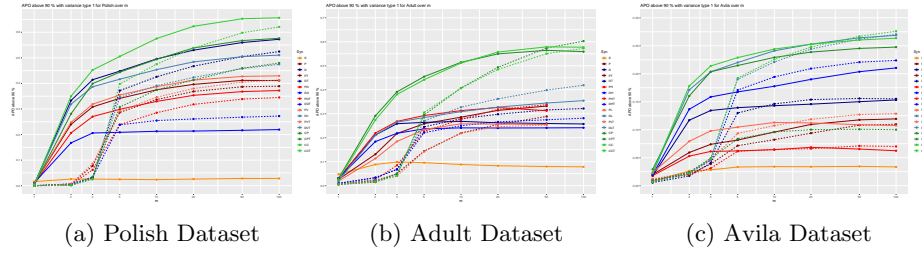


Fig. 3: APO above 90% for Regression Fits

forms slightly better than CC around $m \approx 60$ and is the only one that get above 60%. We also have some surprising results for P that it performs better than D for $m > 6$ which is not the case for average CI overlap.

For Avila (Figure 3c), all the synthesizers performed very poorly and were below 30%. CC performs better compared to other synthesizers. The fact that we do get some difference (i.e., relatively large difference for some synthesizers) between Sample and the other synthesizers indicates that there has been at least some preservation of the relationships between the variables. We can also see that the APO above 90% had a steady increase as the number of datasets to release increased. The increase of APO above 90% as m increases is also a lot more than for average CI overlap and is far more significant for higher m as well. For all three datasets for lower m (e.g. $m = \{1, 2, 3\}$) proper methods perform much worse than Non-Proper methods regardless of synthesizer combinations. In terms of imputation order, we see similar results as CI overlap.

Findings. Based on average CI overlap, CC is the best synthesizer by being the only one to reach above 90% for Polish and achieving the highest scores for both Adult and Avila. However, for Adult and Avila, none of the synthesizers can reach above 90% and in the case of Avila, the best one was able to 60%. CP and D are the next best ones for all three datasets. The APO 90% of also shows similar results as average CI overlap in terms of synthesizer performance where CC, CP, and D perform well for all three datasets. The major difference between average CI overlap and APO above 90% is the results for lower m where APO 90% shows that lower m (i.e., $m < 3$) perform very poorly utility wise compare to higher m which is not the case for average CI overlap.

Table 3: KL-Divergence Score

Dataset	metric	S	P	D	PT	DT	PO	DO	POT	DOT	PV	DV	PVT	DVT	CP	CPT	CC	CCT
Polish	Average	1.00	1.21	1.06	1.96	2.28	1.13	1.29	1.86	2.33	1.23	1.07	1.96	2.29	1.07	2.37	1.03	2.1
		S	P	D	PT	DT	PH	DH	PHT	DHT	PL	DL	PLT	DLT	CP	CPT	CC	CCT
Adult	Average	1.00	9.5	1.04	10.17	2.17	4.72	1.05	5.62	2.21	16.92	1.05	17.21	2.12	1.06	2.48	1.00	2.03
		S	P	D	PT	DT	PH	DH	PHT	DHT	PL	DL	PLT	DLT	CP	CPT	CC	CCT
Avila	Average	1.00	6.96	1.08	7.32	2.31	8.74	1.6	8.96	2.78	5.7	1.08	5.97	2.28	1.13	25.36	1.08	2.31
		S	P	D	PT	DT	PH	DH	PHT	DHT	PL	DL	PLT	DLT	CP	CPT	CC	CCT

KL-Divergence. For the KL-Divergence score (Table 3), we calculate the score for all variables and normalize them over the score of Sample and then take the average over all variables for each synthesizer for easier comparison.

For polish, D, CP, and CC perform well with minor differences between them and P is not also far off. In the case of the Adult dataset, CC is the best performing synthesizer. D and CP are also not far behind CC. However, for P the average score increased a lot because of the variables income, capital gain, and capital loss. Among the three variables, capital gain and loss have the most values as zero which can be a reason for higher divergence value. For the Avila dataset, we see that D and CC perform similarly. This is not surprising because CC in this case is just D with a different imputation order. In terms of imputation order, the original ordering performs better with few exceptions for all three datasets. Similarly, non-proper performs much better than proper in most cases.

Findings. The KL divergence results conform to the confidence interval overlap results that CC and D perform well for all three datasets and also CP is not far off.

4.3 Analysis Accuracy

Adopting the methodology used in [24, 33], we performed some analysis on both the synthetic data and the original data for the comparative evaluation. The idea is to see how well synthetic data perform on such analysis and whether it provides the same conclusion as the original data.

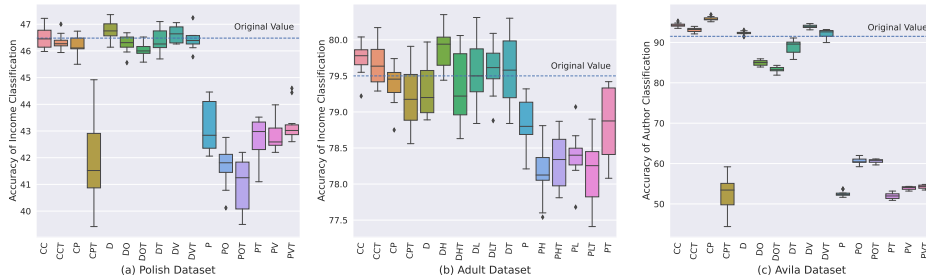


Fig. 4: Classification Accuracy

Classification Accuracy. The three datasets used in this work are publicly available datasets. Thus, we first look at what kind of analysis are commonly performed on the datasets. From the literature search, we find that Adult and Avila dataset are primarily used for classification tasks using machine learning. The classification task for Adult is to classify whether the income is above or

below 50K. Similarly, for Avila it is to classify the author based on the patterns. For Polish dataset we did not find any such common analysis. Nonetheless, since it is a census dataset similar to Adult and has a income variable, we perform a similar income classification on the dataset. For each original dataset, we train the 10 machine learning models using 10 different synthetic datasets and take the average accuracy score. Since we have the original dataset which can serve as the ground truth, we test the accuracy of each model using the original dataset. We also calculate the accuracy score of the original datasets.

Figure 4 shows the classification accuracy results for all three datasets. As shown in the figure, CC, CP and D perform well and P is consistently poor for all three datasets. The synthesizers that perform well have similar or even sometimes higher accuracy than the original dataset. Since, we test the synthetic model accuracy using the original dataset, it is certain that the models trained using the synthetic data can perform well on the real data. For the Polish dataset, as the accuracy of original data is poor, we performed further verification by looking at whether the original model and the synthetic model correctly and wrongly classify similar data or they differ. The investigation revealed that not only for Polish but for all three datasets the synthetic model and the original model classify similar records correctly and wrongly and only differ depending on the difference in accuracy score.

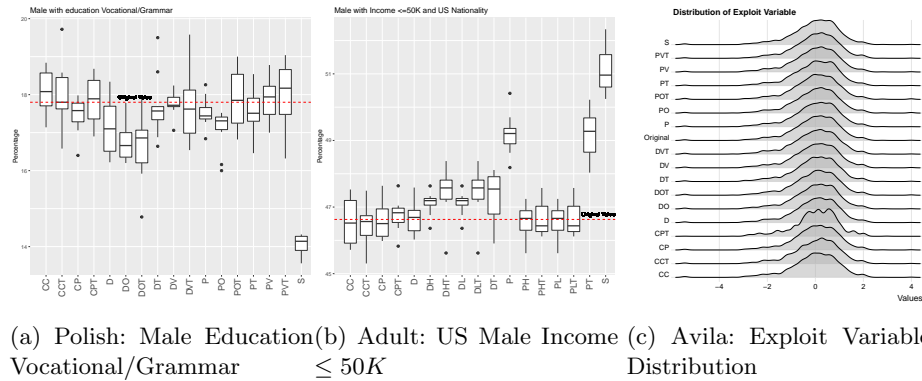


Fig. 5: Ad-hoc Analysis on the Datasets

Ad hoc Analysis. There is no easy way to determine before the release of a dataset what analysis will be performed on it once it is released. Thus, to be useful, synthetic data should provide desirable utility close to original data for any ad hoc analysis. To determine the effectiveness of synthetic data on such analysis, we look at the multivariate relationship and univariate distributions of the variables present in the datasets.

Figure 5 shows the results of ad hoc analysis on all three dataset. For Polish dataset (Fig. 5a) we look at the relationship between the variables gender and education, for Adult (Fig. 5b) it is between gender, income, and nationality, and for Avila (Fig. 5c) we look at the univariate distribution of exploit variable.

In the case of Polish (Fig. 5a), we see that CC and CP performs well. For D and P the imputation order V improved the performance. For Adult (Fig. 5b), CC, CP and D performed well. In terms of imputation order, for P both the order H and L improved the performance significantly. However, in the case of D it was the opposite. Finally, for Avila (Fig. 5c), except CP with proper synthesis (CPT) all synthesizers have similar distributions as the original dataset. In terms of synthesis method, in all three scenarios we see that non-proper synthesis performs better overall than proper.

Table 4: Correlation between Utility Metrics and Analysis Accuracy

Relation	Polish Dataset		Relation	Adult Dataset		Relation	Avila Dataset	
	P Value	Correlation Co-efficient		P Value	Correlation Co-efficient		P Value	Correlation Co-efficient
Accuracy of Income Classification & CIO Raab	1.35E-05	-0.8530482	Accuracy of Income Classification & CIO Raab	7.48E-05	-0.8122135	Accuracy of Author Classification & CIO Raab	5.47E-07	-0.9061566
Accuracy of Income Classification & KL Divergence	0.5338	-0.1622655	Accuracy of Income Classification & KL Divergence	0.6213	-0.1291404	Accuracy of Author Classification & KL Divergence	2.43E-02	0.5430744
Accuracy of Income Classification & APO 90%	1.54E-03	-0.7059498	Accuracy of Income Classification & APO 90%	2.15E-02	-0.552303	Accuracy of Author Classification & APO 90%	3.54E-06	-0.8782808
CIO Raab & APO above 90%	4.14E-07	0.9097029	CIO Raab & APO above 90%	5.54E-06	0.8703936	CIO Raab & APO above 90%	3.04E-11	0.9752702
CIO Raab and KL Divergence	3.97E-01	0.219777	CIO Raab and KL Divergence	1.69E-01	-0.3499996	CIO Raab and KL Divergence	9.80E-02	-0.4145976
APO 90% and KL Divergence	8.43E-01	-0.0520225	APO 90% and KL Divergence	2.28E-01	-0.308567	APO 90% and KL Divergence	4.66E-02	-0.4886557
APO above 90% & Female Inc 1K to 2K	0.06245	-0.4611436	APO above 90% & Male <=50K US	8.42E-03	-0.6163045	APO above 90% & Units Class A	0.2718	-0.2825843
APO above 90% & Male Edu Voc/Grm	4.80E-04	-0.7534067	APO above 90% & Units Income <=50K	0.4505	-0.1961783	CIO Raab & Units Class A	0.3071	-0.2633795
APO above 90% & Units No Income	0.5931	-0.139596	APO above 90% & White below 50K	0.0776	-0.439407			
CIO Raab & Female Inc 1K to 2K	0.04712	-0.4875589	CIO Raab & Male <=50K US	1.12E-04	-0.8008562			
CIO Raab & Male Edu Voc/Grm	1.14E-06	-0.8960993	CIO Raab & Units Income <=50K	0.4568	-0.1935205			
CIO Raab & Married alcause Yes	0.8505	-0.04943915	CIO Raab & White above 50K	0.08458	-0.4304355			
CIO Raab & Units No Income	0.7308	-0.09016015	CIO Raab & White below 50K	0.02481	-0.5413657			

4.4 Correlation between Utility Metrics and Analysis Accuracy

To determine whether there is any relationship between the general utility metrics and analysis accuracy, we conduct several pearson correlation tests between them. For the tests, we measure the correlation of different analysis accuracy results with CIO and APO individually and in some cases with KL divergence as well. We also measure the correlation between the utility metrics CIO, APO, and KL divergence. For the classification accuracy and ad hoc analysis, the relationship is tested between the deviation from the original for each score and the corresponding utility metrics. Table 4 shows the results of the correlation tests.

For the classification accuracy, we see that there exists a strong correlation between accuracy score and CIO for all three datasets. The correlation coefficient for all of them is above 0.8 in the negative direction. There is also a moderate correlation between classification accuracy and APO above 90% with the lowest

coefficient being Adult with a value of -0.55. However, when we look at the correlation between classification accuracy and KL divergence score, there is no such consistent correlation.

In the case of ad hoc analysis, we see a mixed batch of results. For univariate distributions and CIO relationship (i.e., Polish: CIO Raab & Units No Income, Adult: CIO Raab & Units Income $\leq 50K$, and Avila: CIO Raab & Units Class A), we see that there is no strong correlation for all three datasets. However, for multivariate relationships where two or more variables are involved, we see mixed results. For instance, Male Edu Voc/Grm and Male $\leq 50K$ US for Polish and Adult respectively show strong correlation with CIO with a correlation coefficient above -0.80. We also see moderate (e.g., Polish: CIO Raab & Female Inc 1K to 2K, Adult: CIO Raab & Female Inc 1K to 2K) and no correlation (e.g., Polish: CIO Raab & Married alcabuse Yes) between the multivariate analysis and CIO. Similar tests with APO above 90% show that APO either has a similar or lower correlation coefficient than CIO.

The correlation test between the general utility metrics shows that there is a strong correlation between CIO and APO above 90%. Nonetheless, there seems to be no such correlation between CIO or APO above 90% with KL divergence. From the correlation test, we find that CIO is a better utility metric than APO above 90% and KL divergence. However, CIO also has limitations and it is not possible to guarantee that a higher CIO score will ensure higher accuracy in all types of analysis performed on the synthetic data.

Table 5: APO above 90% Comparison between Simple and Selective with $m = 10$

Dataset	metric	P	PSe	D	DSe	PO	POSe	DO	DOSe	PV	PVSe	DV	DVSe	CP	CPSe	CC	CCSe
Polish	Average	0.38	0.22	0.50	0.30	0.33	0.16	0.21	0.16	0.39	0.21	0.45	0.29	0.49	0.40	0.58	0.48
	Ratio	0.78	0.22	0.85	0.15	0.83	0.17	0.63	0.37	0.76	0.24	0.90	0.10	0.65	0.35	0.75	0.25
		P	PSe	D	DSe	PH	PHSe	DH	DHSe	PL	PLSe	DL	DLSe	CP	CPSe	CC	CCSe
Adult	Average	0.29	0.20	0.27	0.20	0.31	0.20	0.24	0.21	0.26	0.39	0.31	0.23	0.51	0.55	0.51	0.55
	Ratio	0.81	0.19	0.79	0.21	0.88	0.12	0.70	0.30	0.29	0.71	0.75	0.25	0.30	0.70	0.46	0.54
		P	PSe	D	DSe	PO	POSe	DO	DOSe	PV	PVSe	DV	DVSe	CP	CPSe	CC	CCSe
Avila	Average	0.10	0.08	0.14	0.04	0.06	0.08	0.18	0.14	0.11	0.10	0.24	0.07	0.23	0.10	0.24	0.16
	Ratio	0.68	0.32	0.91	0.09	0.37	0.63	0.61	0.39	0.58	0.42	0.96	0.04	0.88	0.12	0.74	0.26

Comparison between Simple and Selective Imputation. For the comparison between Simple and Selective imputation methods, we looked at the APO above 90% using both the techniques with Non-Proper imputation and $m = 10$ (Table 5). The results show that for majority of the synthesizers in all of the three datasets simple works better than selective. For Polish, the simple method was better regardless of the synthesizer combinations. However, for Adult and Avila, we find that few synthesizer combinations performed better with selective imputation. For Adult, CP and CC with Selective performed slightly better than simple. Nonetheless, P with imputation order largest categorical variable last (PL) achieved a significant improvement with selective. For Avila, except for P with opposite imputation order (PO), simple performed better than selective

for all the other synthesizer combinations. Thus, in terms of accuracy simple seems to be a better choice.

4.5 Generation Time

For generation time, we look at APO above 90% with the time it took to generate 100 datasets for each synthesizer combination for both the Simple and the Selective methodology (Figure 6). The result of generation time reveals that CC and D are faster overall and achieve better accuracy. However, CP is much slower (i.e., 40 times slower in some cases) than CC and D though they all achieve similar accuracy. For numerical variables, P performs much faster than other synthesizers, however, the accuracy it achieves is much worse than best performing synthesizers. While Selective is faster for most, it is not significantly faster than Simple overall. The choice of imputation order can have an huge impact on the generation time in some scenarios (e.g., for P in Adult the slowest order takes 10 times more than the fastest). We do not see any huge difference between proper and non-proper in terms of generation time.

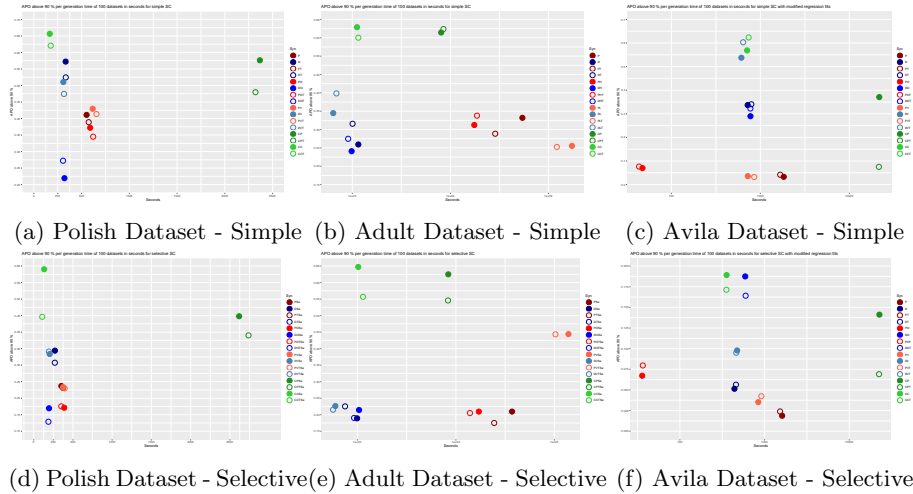


Fig. 6: APO above 90% and Generation Time for 100 Datasets

5 Discussion & Conclusion

In this work, we look at the impact of different parameters and synthesizers on the utility of synthetic data. We also perform a thorough investigation of the correlation between the utility metrics and the analysis accuracy of synthetic data. Our investigation reveals that the choice of synthetic data generation method,

the number of datasets to release, and sometimes the imputation order can impact the utility of synthetic data. We find that Kulback-Leibler divergence, a broad utility metric, does not correlate with analysis accuracy and some narrow metrics such as confidence-interval overlaps (CIO) show varying correlations for specific univariate and multivariate analyses. Simply put, these metrics compare data similarity but cannot guarantee that for any given analysis the results are similar enough. Nevertheless, the synthesizers with the best CIOs also performed best in terms of analysis accuracy. We found another more promising effect when comparing the results of classification tasks by machine-learning models trained on the synthetic versus on the original data, both tested on original data. There, CIO correlates with how similarly the machine-models perform the classification task, both in terms of accuracy and which records they classified correctly or incorrectly, respectively. Machine-learning models have been shown to be susceptible to privacy threats such as membership-inference and model-inversion attacks. In future work, we will therefore investigate whether our results generalize in terms of utility and to what extent training on synthetic data can mitigate privacy threats to machine-learning models.

Acknowledgment

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Czapinski, J., Panek, T.: Social Diagnosis 2011. Objective and Subjective Quality of Life in Poland. *Contemporary Economics* **5**(3) (2011)
2. Dandekar, A., Zen, R.A.M., Bressan, S.: A Comparative Study of Synthetic Dataset Generation Techniques. In: Hartmann, S., Ma, H., Hameurlain, A., Pernul, G., Wagner, R.R. (eds.) *Database and Expert Systems Applications*. pp. 387–395. Springer International Publishing, Cham (2018)
3. Dankar, F.K., Ibrahim, M.K., Ismail, L.: A Multi-Dimensional Evaluation of Synthetic Data Generators. *IEEE Access* (2022)
4. De Stefano, C., Maniaci, M., Fontanella, F., di Freca, A.S.: Reliable Writer Identification in Medieval Manuscripts through Page Layout Features: The “Avila” Bible Case. *Engineering Applications of Artificial Intelligence* **72**, 99–110 (2018)
5. Drechsler, J.: Background on Multiply Imputed Synthetic Datasets, pp. 7–11. Springer New York, New York, NY (2011). https://doi.org/10.1007/978-1-4614-0326-5_2, https://doi.org/10.1007/978-1-4614-0326-5_2
6. Drechsler, J.: Some Clarifications Regarding Fully Synthetic Data. In: Domingo-Ferrer, J., Montes, F. (eds.) *Privacy in Statistical Databases*. pp. 109–121. Springer International Publishing, Cham (2018)
7. Drechsler, J., Bender, S., Rässler, S.: Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. *Trans. Data Privacy* **1**(3), 105–130 (dec 2008)

8. Drechsler, J., Dundler, A., Bender, S., Rässler, S., Zwick, T.: A New Approach for Disclosure Control in the IAB Establishment Panel — Multiple Imputation for a Better Data Access. *ASTA Advances in Statistical Analysis* **92**(4), 439–458 (2008)
9. Drechsler, J., Reiter, J.: Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey. *Journal of Official Statistics* **25**(4), 589 (2009)
10. Dua, D., Graff, C.: UCI Machine Learning Repository (2017), <http://archive.ics.uci.edu/ml>
11. Dwork, C.: Differential Privacy: A Survey of Results. In: International conference on theory and applications of models of computation. pp. 1–19. Springer (2008)
12. El Emam, K., Mosquera, L., Jonker, E., Sood, H.: Evaluating the Utility of Synthetic COVID-19 Case Data. *JAMIA open* **4**(1), ooab012 (2021)
13. Graham, P., Young, J., Penny, R.: Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models. *Journal of Official Statistics* **25**(2), 245 (2009)
14. Hittmeir, M., Ekelhart, A., Mayer, R.: On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In: Proceedings of the 14th International Conference on Availability, Reliability and Security. pp. 1–6 (2019)
15. Hittmeir, M., Ekelhart, A., Mayer, R.: Utility and Privacy Assessments of Synthetic Data for Regression Tasks. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 5763–5772. IEEE (2019)
16. 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. *The American Statistician* **60**(3), 224–232 (2006). <https://doi.org/10.1198/000313006X124640>, <https://doi.org/10.1198/000313006X124640>
17. Kohavi, R., Becker, B.: Adult Dataset. UCI machine learning repository **5**, 2093 (1996)
18. Lee, A.: Generating Synthetic Microdata from Published Marginal Tables and Confidentialised Files. *Statistics New Zealand* (2009)
19. Lee, J.H., Kim, I.Y., O’Keefe, C.M.: On Regression-tree-based Synthetic Data Methods for Business Data. *Journal of Privacy and Confidentiality* **5**(1) (2013)
20. Nowok, B.: Utility of Synthetic Microdata Generated Using Tree-based Methods. UNECE Statistical Data Confidentiality Work Session (2015)
21. Nowok, B., Raab, G., Dibben, C.: synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software, Articles* **74**(11), 1–26 (2016). <https://doi.org/10.18637/jss.v074.i11>, <https://www.jstatsoft.org/v074/i11>
22. Nowok, B., Raab, G.M., Dibben, C.: Providing Bespoke Synthetic Data for the UK Longitudinal Studies and Other Sensitive Data with the synthpop Package for R. *Statistical Journal of the IAOS* **33**(3), 785–796 (2017)
23. Nowok, B., Raab, G.M., Dibben, C.: synthpop: Catall (2019), <https://cran.r-project.org/web/packages/synthpop/synthpop.pdf#nameddest=syn.catall>
24. Purdam, K., Elliot, M.: 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**(5), 1101–1118 (2007)
25. Raab, G.M., Nowok, B., Dibben, C.: Practical Data Synthesis for Large Samples. *Journal of Privacy and Confidentiality* **7**(3), 67–97 (2016)
26. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple Imputation for Statistical Disclosure Limitation. *Journal of official statistics* **19**(1), 1 (2003)
27. Reiter, J.P.: Using CART to Generate Partially Synthetic Public Use Microdata. *Journal of Official Statistics* **21**, 441–462 (2005)

28. Reje, N.: Synthetic Data Generation for Anonymization. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS) (2020)
29. Rubin, D.B.: Multiple Imputation for Survey Nonresponse (1987)
30. Rubin, D.B.: Statistical Disclosure Limitation. *Journal of official Statistics* **9**(2), 461–468 (1993)
31. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **181**(3), 663–688 (2018)
32. Sweeney, L.: k-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(05), 557–570 (2002)
33. Taub, J., Elliot, M., Sakshaug, J.W.: The Impact of Synthetic Data Generation on Data Utility with Application to the 1991 UK Samples of Anonymised Records. *Transactions on Data Privacy* **13**(1), 1–23 (2020)
34. Wang, Z., Myles, P., Tucker, A.: Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility & Patient Privacy. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). pp. 126–131. IEEE (2019)
35. Woo, M.J., Reiter, J., Oganian, A., Karr, A.: Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *Journal of Privacy and Confidentiality* **1**, 111–124 (2009)