Challenges in Measuring Utility for Fully Synthetic Data

Jörg Drechsler[1]

Institute for Employment Research, Regensburger Str. 104,
90478 Nuremberg, Germany, joerg.drechsler@iab.de

Abstract. Evaluating the utility of the generated data is a pivotal step in any synthetic data project. Most projects start by exploring various synthesis approaches trying to identify the most suitable synthesis strategy for the data at hand. Utility evaluations are also always necessary to decide whether the data are of sufficient quality to be released. Various utility measures have been proposed for this purpose in the literature. However, as I will show in this paper, some of these measures can be misleading when considered in isolation while others seem to be inappropriate to assess whether the synthetic data are suitable to be released. This illustrates that a detailed validity assessment looking at various dimensions of utility will always be inevitable to find the optimal synthesis strategy.

Keywords: confidence interval overlap, confidentiality, global utility, pMSE, privacy

1 Introduction

The synthetic data approach for disclosure protection gained substantial popularity in recent years. While applications were mostly limited to the U.S. Census Bureau ([1, 13, 12]) a decade ago, more and more statistical agencies and other data collecting organizations are now exploring this idea as a possible strategy to broaden access to their sensitive data ([29, 16, 4, 2]). Recent developments in computer science, most notably the use of Generative Adversarial Networks (GANs, [10]) further stimulated the synthetic data movement and several start-up companies now offer synthetic data as a product, often with high flying promises regarding the unlimited usefulness of the data paired with claims of zero risk of disclosure. However, from an information theoretic stand point it is obvious that we can never have both, preservation of all the information from the original data while offering full protection of any sensitive information (except for the corner case in which the original data can be released without risk making the release of synthetic data pointless). Thus, all we can hope for is to find the optimal trade-off between utility and data protection, that is, we can try to maximize the utility for a desired level of data protection or maximize the level of protection for a level of utility that is still deemed acceptable. Of course, in practice this is easier said than done. To fully utilize this optimization problem

the data disseminating agency would need to know which levels of utility and disclosure protection the different stakeholders consider acceptable. Even more important, the agency needs reliable measures of utility and risk. Various metrics have been proposed in the literature for measuring the risk and utility for datasets that have undergone some procedure for statistical disclosure control ([5, 23]). However, not all of them are suitable for synthetic data. Especially with fully synthetic data measuring the disclosure risk remains an open research question. Most risk measures that have been proposed in the literature try to estimate the risk of re-identification. Given that there is no one-to-one mapping between the original and the fully synthetic data, these measures cannot be meaningfully applied. The few proposals for measuring the risk of disclosure for fully synthetic data either rely on the opaque concept of perceived risk (the synthetic record looks too close to a real record), are computationally infeasible in practical settings ([22]) or make unrealistic assumptions regarding the knowledge of the attacker ([25]) (but see [26] for an interesting approach for measuring the risk of attribute disclosure).

However, even measuring the utility of the generated data can be more difficult than it might seem. A key challenge is that the data disseminating agencies typically only have limited knowledge for which purposes the data will be used (if they had this information they could simply publish all the analyses of interest as protecting the analysis output is typically much easier than protecting the full microdata). Thus, utility is typically measured by running a couple of analyses deemed to be of interest for the users and comparing the results from the synthetic data with those obtained for the original data. Alternatively, utility measures have been proposed that try to directly compare the synthetic data with the original data. In this paper, I will demonstrate that some of the measures that have been proposed in the literature can be misleading when considered in isolation while others seem to be inappropriate to assess whether the synthetic data are suitable to be released. The main conclusion based on this small assessment is that users of the synthetic data approach should always ensure that they evaluate several dimensions of utility before they decide which synthesis method works best for their data and whether the data are ready to be released.

2 Measuring the utility

Utility measures are typically divided into two broad categories: narrow or analysis-specific measures and broad or global measures. The former focus on evaluating the utility by measuring how well the protected data preserve the results for a specific analysis of interest, while the letter try to directly compare the original and synthetic data providing a measure of similarity between the two datasets. Examples of analysis-specific utility measures are the confidence interval overlap measure proposed by [11] or the ratio of estimates (ROE) proposed by [27] for tabular data. The global utility is commonly assessed using distance measures such as the Kullback-Leibler divergence [11] or the propen-

sity score mean squared error (pMSE) proposed in [30] and further developed in [24]. As pointed out by various authors ([11, 17, 9]), both types of measures have important drawbacks. While narrow measures provide useful information regarding the specific analysis considered, high utility based on these measures does not automatically imply high utility for other types of analyses. Since the data providers typically do not know which purposes the data will be used for later, it will be impossible to fully assess the utility of the synthetic data based on these measures. The global utility measures on the other hand are so broad that they might miss important weaknesses of the synthetic data. Furthermore, the measures are typically difficult to interpret, that is, it is difficult to decide whether the level of utility is acceptable or not. In practice, these measures are therefore mostly used to compare different synthesis approaches and not to decide whether the synthetic data offer enough utility to be released.

A final class of measures–termed *fit-for-purpose* measures here–can be considered to lie between the previous two. These measures typically only focus on specific aspects of the data, that is, they cannot be considered as global measures but also do not necessarily reflect statistics users might be interested in directly. Examples include plausibility checks such as ensuring only positive age values in the synthetic data, but also visual comparisons of univariate and bivariate distributions. Goodness-of-fit measures such as the χ^2-statistic for various cross-tabulations of the data or Kolmogoroff-Smirnov tests for continuous variables also belong to this group. As illustrated by [17], the pMSE can also be used for this purpose by simply including only the variables to be evaluated as predictors in the propensity model. Fit-for-purpose measures are typically the first step when assessing the utility of the generated data and we will illustrate the importance of these measures in the next section demonstrating that both, the global and the analysis specific measures of utility can be misleading when considered in isolation. Before we discuss the empirical evaluations, we provide further details regarding the utility measures used.

2.1 A global utility measure: the pMSE

As mentioned above the pMSE has become a popular measure in recent years to assess the utility of the generated data. The procedure consists of the following steps:

1. Stack the n_{org} original records and the n_{syn} synthetic records adding an indicator, which is one if the record is from the synthetic data and zero otherwise.
2. Fit a model to predict the data source (original/synthetic) using the information contained in the data. Let p_i, $i = 1, \ldots, N$ with $N = n_{org} + n_{syn}$ denote the predicted value for record i obtained from the model.
3. Calculate the pMSE as $1/N \sum_N (p_i - c)^2$, with $c = n_{syn}/N$.

The smaller the pMSE the higher the analytical validity of the synthetic data. A downside of the pMSE is that it increases with the number of predictors included

in the propensity model even if the model is correctly specified. To overcome this problem, [24] derived the expected value and standard deviation of the pMSE under the hypothesis that both, the original and the synthetic data are generated from the same distribution, that is, the synthesis model is correctly specified. Based on these derivations, the authors propose two utility measures: The pMSE ratio, which is the empirical pMSE divided by its expected value under the null and the standardized pMSE (S_pMSE), which is the empirical pMSE minus its expectation under the null divided by its standard deviation under the null.

2.2 Two Outcome-Specific Measure: the confidence interval overlap and the mean absolute standardized coefficient difference

The confidence interval overlap measure was first proposed by [30]. Paraphrasing from [6], its computation can be summarized as follows: For any estimand, we first compute the 95% confidence intervals for the estimand from the synthetic data, (L_s, U_s), and from the original data, (L_o, U_o). Then, we compute the intersection of these two intervals, (L_i, U_i). The utility measure is

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)}. \tag{1}$$

When the intervals are nearly identical, corresponding to high utility, $I \approx 1$. When the intervals do not overlap, corresponding to low utility, $I = 0$. The second term in (1) is included to differentiate between intervals with $(U_i - L_i)/(U_o - L_o) = 1$ but different lengths.

The mean absolute standardized coefficient difference (MASD) is implemented in the *synthpop* package as a utility measure for regression models. It computes the standardized difference for each regression coefficient as $z_j = (\bar{q}_m - \hat{Q})/(\sqrt{(}v_{org}/m)$, where $\bar{q}_m$ and $\hat{Q}$ denote the estimated coefficient from the synthetic and original data, respectively, v_{org} is the estimated variance of $\hat{Q}$ and m is the number of synthetic datasets. The MASD is then computed as $\sum_{j=1}^{p} |z_j|/p$, where p is the total number of regression coefficients in the model.

3 Misleading utility measures: an illustration

For this small illustration, I use a subset of variables and records from the public use file of the March 2000 U.S. Current Population Survey (CPS). The data comprise eight variables measured on $N = 5,000$ heads of households (see Table 1 for details). Similar data are used in [19, 20, 7, 8] to illustrate and evaluate various aspects of synthetic data. To simplify the modeling task I have removed some variables, subsampled the data, excluded some records, and recoded some variables compared to previous applications.

3.1 Synthesis strategies

Overall we use four different synthesis strategies, three based on fully parametric models and one using a CART approach. We use the R package *synthpop* ([15])

Table 1. Description of variables used in the empirical studies

Variable	Label	Range
Sex	*sex*	male, female
Race	*race*	white, other
Marital status	*marital*	5 categories
Highest attained education level	*educ*	4 categories
Age (years)	*age*	15 – 90
Social security payments ($)	*ss*	0, 1 – 50,000
Household property taxes ($)	*tax*	0, 1 – 98,366
Household income ($)	*income*	1 – 582,896

to generate the synthetic data leaving most of the parameters at their default values. Specifically, we always synthesize all variables, keeping the size of the synthetic data the same as the size of the original data. We also keep the hyperparameters for the CART models at their default values and use standard options for the parametric variables: All continuous variables are synthesized using a linear regression model, while *sex* and *race* are synthesized using a logit model, and *marital* and *educ* are synthesized using multinomial regression. We always use the same synthesis order relying on the order in which the variables appear in the dataset with the minor adjustment that synthesis always starts with the variable *sex*. This adjustment was necessary as *synthpop* currently forces the synthesis for the first variable to be based on sampling when generating fully synthetic data (according to the maintainers of *synthpop* this issue will be fixed in future versions of the package). Since simply sampling from the marginal distribution arguably can be risky for continuous variables as exact values from the original data will be revealed, we decided to start the synthesis with a binary variable for which sampling original values does not pose any risks. Based on the same concerns–releasing exact values for continuous variables–we also use the smoothing option for the CART models. This option fits a kernel density estimator to the original values in any leaf of the tree and samples synthetic values from this estimator instead of sampling original values directly. We always generate $m = 5$ synthetic datasets.

The three parametric synthesizers differ in the way they account for distributional aspects of the original data. The first synthesizer (which we label the *naive* synthesizer below) does not consider these aspects at all, running the synthesis models without preprocessing the data. The second synthesizer (*transform*) tries to address the skewness of the continuous variables by taking the cubic root of all continuous variables before the imputation. Figure 1 shows the distribution of income and age before and after the transformation. The transformations make the distribution more symmetric, which can help to make the assumptions of the linear model more plausible. The final synthesis model (*two-stage*) additionally accounts for the fact that *ss* and *tax* have large spikes at zero as illustrated in

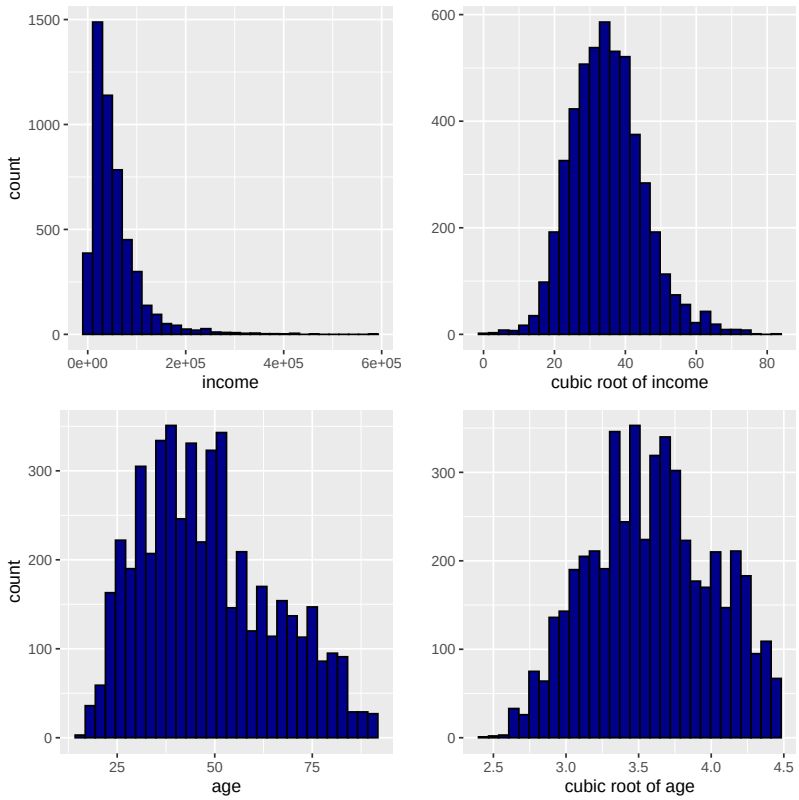

Fig. 1. Histogram of the variables *income* and *age* on the original scale and after transformation by taking the cubic root.

Figure 2. To account for these spikes, we use the *semicont* option in *synthpop* which implements the two-stage imputation approach described in [18].

Looking at these synthesis strategies we would expect that the utility of the synthetic data would improve with each model, as the synthesis models better account for the properties of the original data. The only question should be, how these strategies perform relative to the CART synthesis model. We note that the synthesis models could certainly be further improved. The goal of this small exercise is not to find the best way of synthesizing the CPS data. We only use these synthetic datasets to highlight some caveats when relying on commonly used utility metrics.

3.2 Results for the fit-for-purpose measures

Figures 3 and 4 provide visual comparisons of the distribution of the original and synthetic data for the variables *tax* and *income* generated using the *compare*

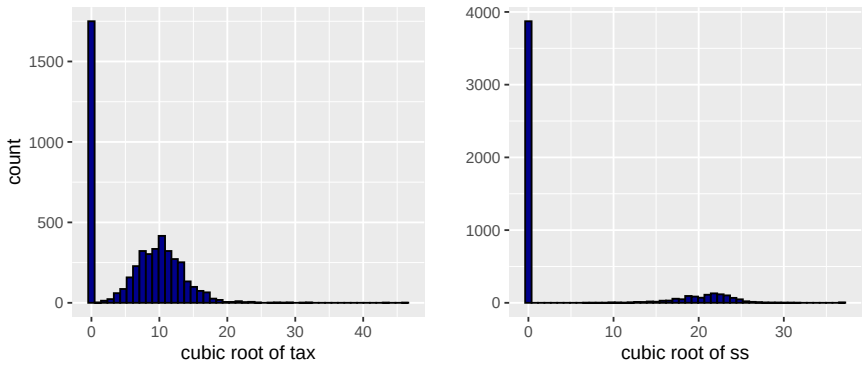

Fig. 2. Histogram of the variables *tax* and *ss* after transformation by taking the cubic root. Both variables have large spikes at zero.

function in *synthpop*. The findings for *age* and *ss* are comparable to the findings for *income* and *tax*, respectively and thus are omitted for brevity. We also do not report the results for the categorical variables as all methods performed similarly well for those. We transform the variables in Figure 3 and 4 by taking the cubic root as the skewness of the variables would make visual comparisons difficult on the original scale (more precisely, we transform the variables using $f(x) = sign(x)|x|^{1/3}$ to also allow for x to be negative). The numbers included in the title of each figure are the standardized pMSEs computed by using only the depicted variables as predictors in the model to estimate the propensity scores.

Several points are noteworthy: The *naive* approach that neither transforms the data nor considers the spikes at zero performs very poorly, as expected. Both variables have a considerable amount of negative values in the synthetic data despite the fact that they all only contain positive values in the original data. The spread of the synthetic data is also considerably larger than for the original data. This is especially true for *tax* due to its large spike at zero. Moving to the second synthesis approach (*transform*), which transformed the continuous variables to bring them closer to normality, we see that this approach helped to improve the quality of the synthetic data. Especially for *income* the distribution is much better preserved. This is also reflected in the substantial reduction in the standardized pMSE from over 260 to 7.66. However, the problems from not modeling the spike at zero are still obvious for *tax*.

This problem is also taken into account in the *two-stage* synthesis strategy where we see the positive impacts of separately modeling the spike. With this approach the spike is well preserved and the distributions in the synthetic data never differ substantially from the distributions in the original data. The results for the CART synthesizer are compatible with the results for the *two-stage* synthesis. We see some minor deviations in the distributions between the original and the synthetic data, but overall the distributions are well preserved. In terms

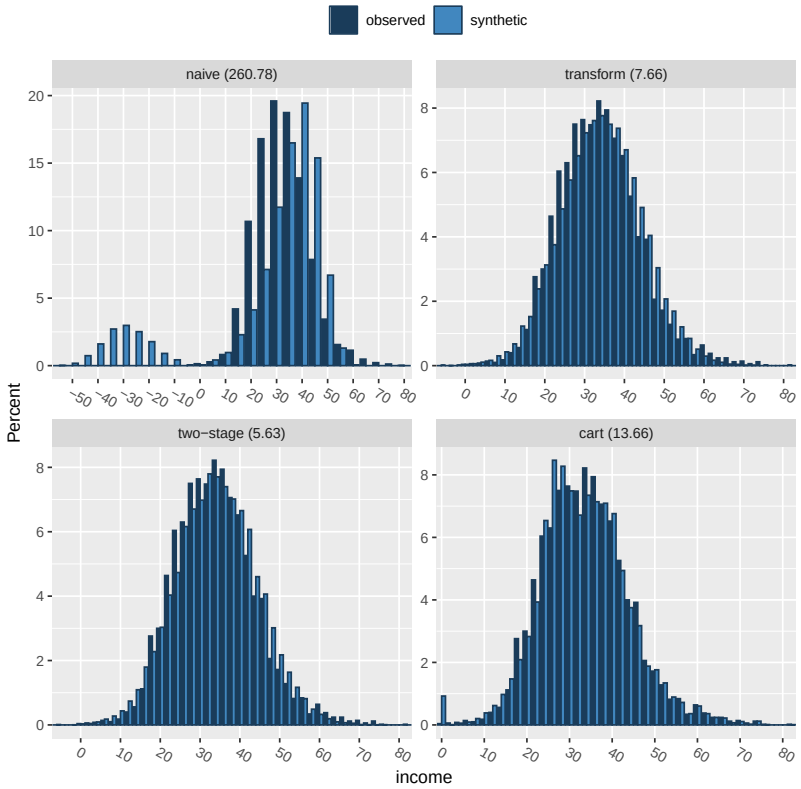

Fig. 3. Histogram of the variable *income* based on the original and synthetic data for various synthesis methods. The numbers in parentheses are the standardized pMSEs.

of the standardized pMSE the *two-stage* approach outperforms the CART synthesis for *income*. Interestingly, the pMSE for *tax* is much small for the CART approach, although the CART synthesis does not preserve the spike at zero as well as the *two-stage* approach. We speculate that this is due to *synthpop* not using the variable directly to compute the measure. Instead, a categorical variable is derived by grouping the units into five equal sized bins using quantiles. The *S_pMSE* is computed using class membership as a predictor. This is obviously a crude measure as it ignores heterogeneity within the bins.

3.3 Results for the outcome specific measures

We focus on one linear regression example to illustrate that weaker performance regarding the preservation of the marginal distributions does not necessarily imply worse results for a specific analysis task. We assume the analyses of interest is a linear regression of log(income) on the other variables contained in

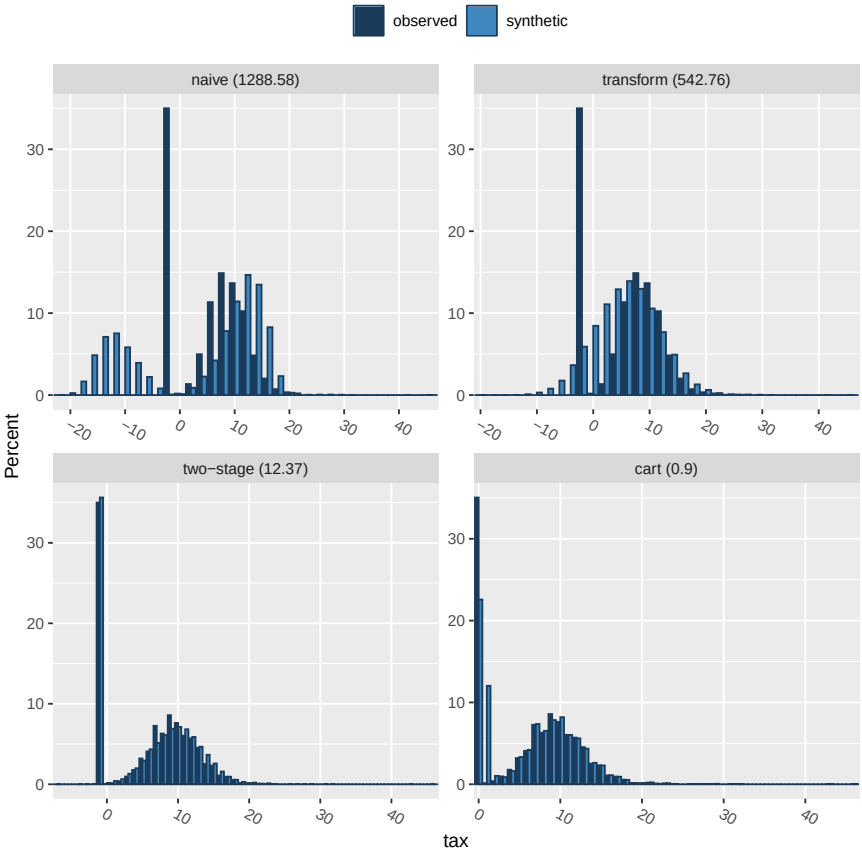

Fig. 4. Histogram of the variable *tax* based on the original and synthetic data for various synthesis methods. The numbers in parentheses are the standardized pMSEs.

the dataset. This model is only for illustrative purposes and we do not claim that the model specification is appropriate. Results for the different synthesis methods are depicted in Figure 5. The numbers in parentheses are the average confidence interval overlaps (CIO) computed as the average across all regression coefficients from the model and the mean absolute standardized coefficient difference (MASD). We note that the CART synthesis model offers the lowest utility for both measures (note that larger values are better for CIO, while smaller values indicate higher utility for MASD). The CART model did not capture the relationship between marital status and income correctly. Even more problematically, the *sex* variable has the wrong sign in the synthetic data. For the CIO measure the utility order matches the order from the previous section, that is, *two-stage* offers higher utility than *transform*, which has higher utility

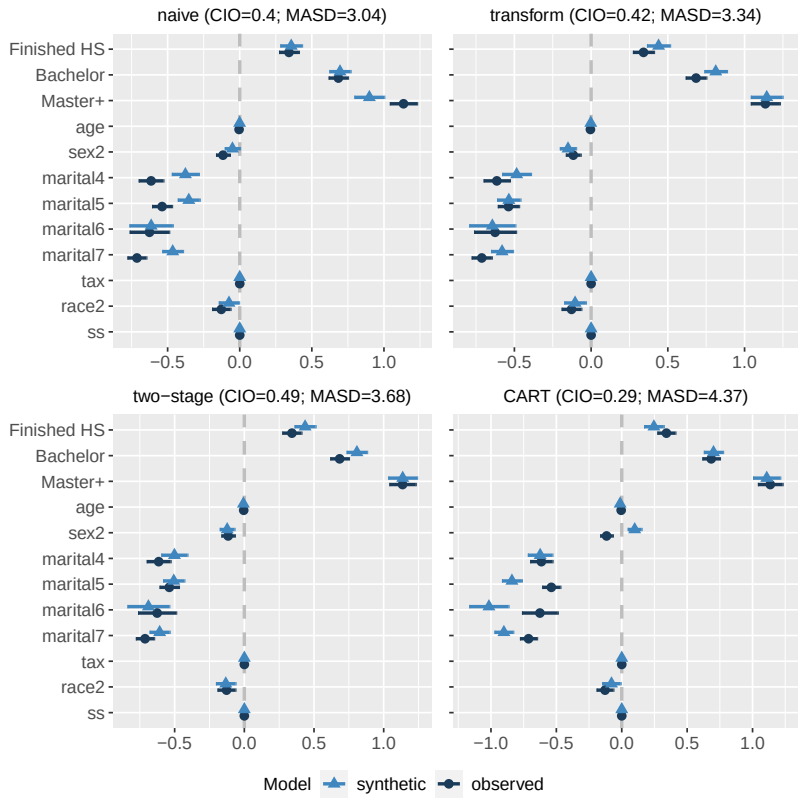

Fig. 5. Comparison of results of a regression of log(income) on the other variables included in the CPS data for various synthesis methods. The lines indicate the length of the 95% confidence intervals. The numbers in parentheses are the average confidence interval overlap (CIO) and the mean absolute standardized coefficient difference (MASD). Both are computed by averaging across all regression coefficients.

than *naive*. Interestingly, the MASD measure indicates a reversed order with the *naive* approach offering the highest utility.

The results illustrate that utility can be high for certain analysis goals even if some aspects of the data are poorly captured but also that aggregated measures which average results across various estimates can potentially be misleading and visualizations such as those shown in Figure 5 might be better suited to identify strengths and weaknesses of the synthetic data.

3.4 Results for the global utility measures

To estimate the standardized pMSE for the entire dataset, we need to specify a model for estimating the individual membership propensities p_i. The most com-

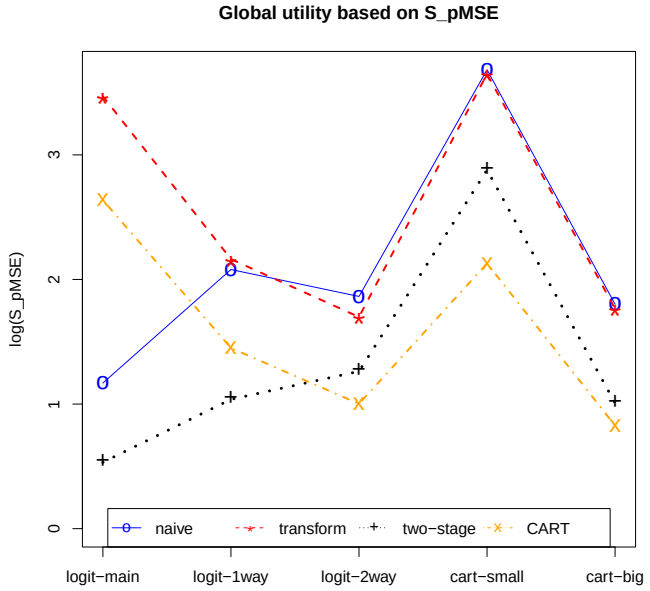

Fig. 6. Standardized pMSE (on the log-scale) for various combinations of synthesis strategies and propensity score models.

mon choice in the propensity score literature is the logit model, but any model can be used as long is it returns predicted probabilities of class membership. In our application, we use the two models available in *synthpop*: the logit model and CART models. We always include all variables, when fitting the different models. However, for the logit model we vary the exact specification of the model evaluating three different settings: The first, labeled *main* in the figure below, only includes all main effects. The second (*1-way*) additionally includes all one-way interactions between the variables. The final model (*2-way*) also includes all two-way interactions. For the CART models we do not need to specify the model, as the method will automatically identify the best splits of the data. However, an important tuning parameter with CART models as implemented in *rpart* ([28]), the library used in this application, is the complexity parameter *cp*. Any split that does not decrease the overall lack of fit by a factor of *cp* is not attempted. Thus, smaller values of *cp* generally imply that larger trees are grown. We evaluate two settings: In the first setting (*CART_small*), we use the default settings of *synthpop*, which presumably use the default values from the *rpart* package, which is $cp = 0.01$. In the second setting (*CART_big*), we use a very small value of $cp = 10^{-7}$.

Results are presented in Figure 6. A couple of things are noteworthy: First, the results confirm that like other global utility measures, the *S_pMSE* cannot be used to assess whether the data are ready to be released. It can vary substantially depending on which model is used to estimate the propensity score. For example, for the *naive* synthesis approach, the *S_pMSE* changes from 39.76 to 6.06 when switching from *CART_small* to *CART_big*. Second, the measure is unable to detect the substantial improvements in the synthetic data when switching from the *naive* approach to the *transform* approach. In fact, the 2-way logit model is the only model that suggests that transforming the variables before the synthesis improves the results. All other models indicate that the *naive* synthesis offers at least similar utility, with the *main* model suggesting substantial quality improvements without transformation. Finally, while two of the logit models (*main* and *1-way*) suggest that a careful parametric synthesis approach (the *two-stage* approach) should be preferred over CART synthesis, the two CART based propensity score models always prefer the CART synthesis strategy.

The results indicate that the global utility measure is not a reliable indicator for deciding which synthesis strategy should be preferred. Results are highly dependent on the model used to estimate the propensity scores and the approach sometimes seems incapable of detecting major differences in the quality of the synthesis models.

4 Conclusions

Given the large variety of synthesis strategies that have been proposed in recent years, picking the most suitable synthesis method is a difficult task. In this situation it seems tempting to rely on global utility measures that return only one number and just pick the strategy that achieves the highest utility according to this score. As I have shown in this paper, things unfortunately are not that easy. The small illustration included in this paper demonstrates fundamental weaknesses of one popular global utility metric: the standardized pMSE. I showed that results for this metric are highly dependent on the model used to estimate the propensity score. Maybe even more worrying, the metric was unable to detect important differences in the utility for most of the model specifications.

On the other hand, I also showed (perhaps unsurprisingly) that utility can still be relatively high for certain types of analyses even if some distributional features of the data are poorly preserved. This implies that a thorough assessment of the utility is inevitable when deciding which synthesis method to pick and whether the data are ready to be released. This assessment should start by evaluating if the data are fit for purpose in various dimensions. If the data disseminating agency already has some information for which types of analyses the data will be used later, it will also be useful to compute outcome specific utility measures for a large variety of analyses to better understand, which analysis models still cause problems and use this information to refine the synthesis

models. The findings from this paper seem to indicate that decisions based on global utility measures should better be avoided.

From a user's perspective, verification servers can also be an important alternative tool to increase confidence in the results obtained from the synthetic data. These servers hold both the synthetic and the original data. Researchers can submit their analysis of interest to the server, it runs the analysis on both datasets, and reports back some fidelity measure how close the results from the synthetic data are to the results based on the original data. However, some care must be taken, as even fidelity measures might spill sensitive information. Developing such measures is currently an active area of research ([21, 14, 3, 31]).

References

1. Abowd, J.M., Stinson, M., Benedetto, G.: Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Tech. rep., Longitudinal Employer–Household Dynamics Program, U.S. Bureau of the Census, Washington, DC (2006)
2. Australian Bureau of Statistics: Methodological news, dec 2021 (2021), `https://www.abs.gov.au/statistics/research/methodological-news-dec-2021`, Last accessed on 2022-05-17
3. Barrientos, A.F., Bolton, A., Balmat, T., Reiter, J.P., de Figueiredo, J.M., Machanavajjhala, A., Chen, Y., Kneifel, C., DeLong, M.: Providing access to confidential research data through synthesis and verification: An application to data on employees of the us federal government. The Annals of Applied Statistics 12(2), 1124–1156 (2018)
4. Bowen, C.M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., Ueyama, K., Williams, A.R.: A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. In: International Conference on Privacy in Statistical Databases. pp. 257–270. Springer (2020)
5. Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K., de Wolf, P., Hundepool, A.: Statistical Disclosure Control. Wiley Series in Survey Methodology, Wiley (2012)
6. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics 201. New York: Springer (2011)
7. Drechsler, J., Reiter, J.P.: Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In: Domingo-Ferrer, J., Saygin, Y. (eds.) Privacy in Statistical Databases, pp. 227–238. New York: Springer (2008)
8. Drechsler, J., Reiter, J.P.: Sampling with synthesis: A new approach for releasing public use census microdata. Journal of the American Statistical Association 105, 1347—1357 (2010)
9. Drechsler, J., Hu, J.: Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data. Journal of Survey Statistics and Methodology 9(3), 523–548 (2021)
10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. arXiv:1406.2661 [cs, stat] (2014), `http://arxiv.org/abs/1406.2661`, arXiv: 1406.2661

11. 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, 224–232 (2006)
12. Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., Abowd, J.M.: Towards unrestricted public use business microdata: The synthetic longitudinal business database. International Statistical Review 79(3), 362–384 (2011)
13. Machanavajjhala, A., Kifer, D., Abowd, J.M., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: IEEE 24th International Conference on Data Engineering. pp. 277–286 (2008)
14. McClure, D.R., Reiter, J.P.: Towards providing automated feedback on the quality of inferences from synthetic datasets. Journal of Privacy and Confidentiality 4(1) (2012)
15. Nowok, B., Raab, G.M., Dibben, C.: synthpop: Bespoke creation of synthetic data in r. Journal of statistical software 74, 1–26 (2016)
16. 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 1. Statistical Journal of the IAOS 33(3), 785–796 (2017)
17. Raab, G.M., Nowok, B., Dibben, C.: Guidelines for producing useful synthetic data. arXiv preprint arXiv:1712.04078 (2017)
18. Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a series of regression models. Survey Methodology 27, 85–96 (2001)
19. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A 168, 185–205 (2005)
20. Reiter, J.P.: Using CART to generate partially synthetic, public use microdata. Journal of Official Statistics 21, 441–462 (2005)
21. Reiter, J.P., Oganian, A., Karr, A.F.: Verification servers: Enabling analysts to assess the quality of inferences from public use data. Computational Statistics & Data Analysis 53(4), 1475–1482 (2009)
22. Reiter, J.P., Wang, Q., Zhang, B.: Bayesian estimation of disclosure risks for multiply imputed, synthetic data. Journal of Privacy and Confidentiality 6(1) (2014)
23. Shlomo, N., Skinner, C.: Measuring risk of re-identification in microdata: State-of-the art and new directions. Journal of the Royal Statistical Society, Series A p. (forthcoming) (2022)
24. 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)
25. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data–a privacy mirage. arXiv e-prints pp. arXiv–2011 (2020)
26. Taub, J., Elliot, M.: The synthetic data challenge. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, The Netherlands (2019)
27. 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)
28. Therneau, T., Atkinson, B., Ripley, B.: rpart: Recursive Partitioning and Regression Trees (2015), `https://CRAN.R-project.org/package=rpart`, r package version 4.1-10
29. de Wolf, P.P.: Public use files of eu-silc and eu-lfs data. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Helsinki, Finland pp. 1–10 (2015)

30. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1, 111–124 (2009)
31. Yu, H., Reiter, J.P.: Differentially private verification of regression predictions from synthetic data. Trans. Data Priv. 11(3), 279–297 (2018)