ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)
181, *Part* 3, *pp.* 663–688

General and specific utility measures for synthetic data

Joshua Snoke,

Pennsylvania State University, University Park, USA

Gillian M. Raab, Beata Nowok and Chris Dibben

University of Edinburgh, UK

and Aleksandra Slavkovic

Pennsylvania State University, University Park, USA

[Received April 2016. Final revision January 2018]

Summary. Data holders can produce synthetic versions of data sets when concerns about potential disclosure restrict the availability of the original records. The paper is concerned with methods to judge whether such synthetic data have a distribution that is comparable with that of the original data: what we term general utility. We consider how general utility compares with specific utility: the similarity of results of analyses from the synthetic data and the original data. We adapt a previous general measure of data utility, the propensity score mean-squared error pMSE, to the specific case of synthetic data and derive its distribution for the case when the correct synthesis model is used to create the synthetic data. Our asymptotic results are confirmed by a simulation study. We also consider two specific utility measures, confidence interval overlap and standardized difference in summary statistics, which we compare with the general utility results. We present two contrasting examples of data syntheses: one illustrating synthetic data that is evaluated as being useful by both general and specific measures and the second where neither is the case. For the second case we show how the general utility measures can identify the deficiencies of the synthetic data and suggest how this can inform possible improvements to the synthesis method.

Keywords: Classification and regression trees; Disclosure control; Privacy; Propensity score; Synthetic data; Utility

1. Introduction

Dissemination of data to external researchers is an important goal for statistical agencies. With sensitive data, the agencies may be constrained in their ability to allow access to raw records, except perhaps to approved users in restricted locations, such as data safe havens (e.g. US Census Research Data Centers). To make their data more available agencies have developed methods of statistical disclosure control, which is also known as statistical disclosure limitation. Statistical disclosure control methods alter the data to reduce the risk of disclosure for sensitive information, i.e. to protect privacy, while maintaining the utility of the data as judged by the validity of inference carried out using the altered data. Traditional methods include micro-aggregation, top or bottom coding, perturbation by adding random noise and the swapping of

Address for correspondence: Joshua Snoke, Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802-1503, USA.
E-mail: snoke@psu.edu

values (for example, for more details see Fienberg and Slavković (2011) and Hundepool *et al.* (2012)).

An alternative statistical disclosure control method involves the generation of synthetic data sets where some or all of the observed data have been replaced by synthetic values generated from models based on the original data. The risk of disclosure is reduced by replacing the original sensitive values. There is an extensive literature on methods for generating synthetic data and making inferences from them, e.g. Raghunathan *et al.* (2003), Reiter (2003), Kinney and Reiter (2010), Slavković and Lee (2010), Drechsler (2011) and Raab *et al.* (2017b), to cite only a few key references. The US Census Bureau, in partnership with academics, has made significant advances in practical applications and has released several synthetic data products, including the Survey of Income and Program Participation synthetic beta data (Benedetto *et al.*, 2013), the synthetic longitudinal business database (Kinney *et al.*, 2011) and Machanavajjhala *et al.* (2015), which is a Web-based interface to a partially synthetic version of the longitudinal employer–household dynamics data set. Synthetic data are now becoming more widely accepted and are being developed by other institutions world wide. For example, bespoke synthetic data are provided to individual users of the Scottish longitudinal study (SLS), and the `synthpop` package for R (R Core Team, 2017) has been developed by Nowok *et al.* (2016) to facilitate the generation of synthetic data extracts. Synthetic microdata are, however, still experimental and for the examples that were mentioned above they are supplied to users to carry out exploratory analyses, but the final results for publication are almost always obtained from the original data. This final analysis is referred to as the gold standard analysis.

It is well understood that inferences from synthetic data will only be valid if the models that are used to synthesize the data correspond to those that can be considered as having generated the original data. It is important for staff synthesizing the data to assess how well this condition is fulfilled by their synthetic data set, and this can be done by so-called *general* and *specific* measures of utility. The former are summaries of differences between the distributions of the original and the altered data whereas the latter compare the differences between results from particular analyses.

Synthetic data utility has most often been assessed by *analysis-specific measures* which compare data summaries and/or the coefficients of models fitted to synthetic data with those from the original data. If inferences from original and synthetic data agree, the synthetic data are said to have high utility. Published evaluations of synthetic data using specific utility measures, usually for just a few selected analyses, have highlighted differences in the quality of syntheses (Reiter, 2005a; Drechsler and Reiter, 2009; Kinney *et al.*, 2011; Miranda and Vilhuber, 2016; Nowok, 2015). However, when an agency prepares synthetic data for a user they will not know, except in very general terms, what analyses will be carried out. In practice, a user usually carries out a number of exploratory analyses to decide which models to fit and present. When the synthesizer does have some knowledge of the models that the analyst has in mind and bases the synthesis on these models, this may falsely reassure the analyst that their model is the correct model. When the generative model that informs the synthesis adheres too closely to the proposed utility model, the validity checks such as the existence of other interactions will not be apparent in the synthesized data; see Nowok *et al.* (2017) and Raab *et al.* (2017a) for examples. Thus *general measures* of utility could be more helpful in allowing an assessment of how well the final inference might agree with what would have been obtained if the user had access to the unchanged data for all of the analyses, rather than just at the final stage of a gold standard analysis. Global measures of utility that can be used for any type of altered data have been proposed by several researchers, such as Karr *et al.* (2006) and Woo *et al.* (2009), and Drechsler (2011) has illustrated their use for a real example of synthetic data.

The disclosure risk that is associated with releasing any data from a statistical agency is clearly important. Agencies most commonly release data in the form of cross-tabulations or other summaries and there is an extensive literature on methods for assessing disclosure risk for such data; see Willenborg and De Waal (2001) for a review. Disclosure risk measures for microdata releases, such as synthetic data, are less well developed. Methods have often been tailored to individual data products, e.g. Elliot (2015), US Census Bureau (2006), Drechsler and Reiter (2009) and Loong *et al.* (2013). More recent research with synthesized categorical data has proposed methods that can be used to identify individual records with disclosure potential (e.g. Hu *et al.* (2014), Reiter *et al.* (2014) and McClure and Reiter (2016)) but at present does not provide measures that can be used with the complexity of real data sets. This is clearly an area where further research is required but we do not address it here where our focus is on utility measures.

In this paper we evaluate and recommend extensions to existing global and specific measures of utility for the special case of synthetic data, and we compare general utility results with specific utility for data generated by different methods of synthesis. In Section 2 we review methods for generating and making inference from synthetic data and introduce our notation. In Section 3 we review previous work on general utility measures. In Section 4 we extend previous work for a propensity-score-based general utility measure by proposing two statistics that have been specifically designed for synthetic data. In Section 5 we cover specific utility measures that are typically used for synthetic data. In Section 6 we give two data examples comparing outcomes based on general and specific measures and highlighting differences in their evaluation of different syntheses. In Section 7 we offer concluding remarks and recommendations.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Brief review of synthetic data methodology

2.1. Data synthesis

Synthesis is performed by a researcher with access to the original data which we denote as (X, Y), where X denotes the data that will be released with their original values and Y are the sensitive data that will be replaced with synthetic values. It is possible for all the data to be synthesized, in which case X is empty. The synthesizing process assumes that the data come from an underlying joint generative distribution $f(Y|X, \theta)$.

Here we consider the situation when new values of Y are generated by fitting the observed data to $f(Y|X, \theta)$ to give an estimate $\hat{\theta}$ and by generating a new sample from $f(Y|X, \hat{\theta})$. In practice this is typically approximated with a sequence of conditional models. A total of m synthetic data sets are produced, where $m = 1$ gives just a single data set. Some methods of inference from synthetic data require that the synthetic data are generated from the posterior distribution of Y, given the observed data. In the next section we discuss why this requirement does not apply in our case.

2.2. Inference for synthetic data

Inference from synthetic data can be required for a particular model or a set of summary statistics defined by a parameter vector Q that could be estimated from the original data. Inference involves carrying out the same procedure on each of the synthetic data sets and details on notation are given in Table 1.

Table 1. Notation for inference to vector Q

Value	*Description*
$\hat{Q}$	Estimate of Q from the original data
V_{orig}	Estimated variance matrix of $\hat{Q}$ from the original data
v_{orig}	Diagonal elements of V_{orig}
$q_1,\ldots,q_i,\ldots,q_m$	Estimate from the fit to the ith synthesis
$\bar{q}_m$	Mean vector, i.e. $\Sigma_{i=1}^{m} q_i/m$
$v_1,\ldots,v_i,\ldots,v_m$	Estimated variances of each q_i calculated from each synthetic data set as if it were the original
$\bar{v}_m$	Mean vector, i.e. $\Sigma_{i=1}^{m} v_i/m$

We are assuming throughout this paper that methods that are appropriate for simple random sampling are used for inference from both the original and the synthetic data. To make inferences for Q from only the synthetic data we use the average $\bar{q}_m$ as an estimate.

Much of the literature on synthetic data is concerned with using the synthetic data to make inferences for the population parameter Q, allowing for both the variation between Q and $\hat{Q}$ and those between $\hat{Q}$ and $\bar{q}_m$. However, for this paper we focus on the situation where researchers use synthetic data produced for exploratory purposes and then will carry out a gold standard analysis (i.e. using the original data) after models have been chosen. This scenario has been implemented by some researchers and agency staff; see Nowok *et al.* (2017) or Reiter *et al.* (2009) for further examples. When such a gold standard analysis is to be carried out, the user of synthetic data is interested in estimating $\hat{Q}$ and its variance–covariance matrix V_{orig}.

When the original data are generated from the same model as used for synthesis, and when the asymptotic conditions that are specified in Raghunathan *et al.* (2003) and Raab *et al.* (2017a) are met, $\bar{q}_m$ is a consistent estimator of $\hat{Q}$, and the simple plug-in estimator $\bar{v}_m$ is a consistent estimator of V_{orig}. Note that neither multiple syntheses with combining rules nor sampling from the posterior distribution of Y are required to calculate these quantities; see Raab *et al.* (2017a).

To evaluate specific utility, we compare results from the synthetic data sets with what would be obtained from the original data. Thus we need not be concerned with population inference and can compare confidence intervals and standardized coefficients from the original data with the equivalent quantities for synthetic data, calculated from $\bar{q}_m$ and $\bar{v}_m$. This approach uses the same estimator for any type of synthetic data, e.g. whether all the observations or only selected variables or data values are synthesized. If the data-generating model that is used for the synthetic data is the model that generated the original data then the confidence intervals from the synthetic data will be consistently estimated by this approach; see Raab *et al.* (2017a) for more on inference with synthetic data under different situations and Raab and Nowok (2017) for details of their implementation in the `synthpop` package.

3. General utility measures for masked data

Previous work has suggested various general measures of utility for data that have undergone disclosure control. Generally these measures consider the distributional similarity between the original and the masked data sets, with greater utility attributed to masked data that are more similar to the original data. In the broadest sense, measures such as the distance between empirical cumulative distribution functions or the Kullback–Leibler divergence give an estimate of difference.

Karr *et al.* (2006) and the follow-up Woo *et al.* (2009) discussed and implemented various distributional measures such as the Kullback–Leibler divergence, an empirical cumulative distribution function measure, a method based on clustering and a measure that uses propensity scores to estimate general utility. They compared these measures for microaggregation, additive noise, swapping and resampling methods, and they evaluated the propensity score method as the most promising. In this paper we focus on expanding this measure for the specific case of synthetic data. The notation that is used in calculating the propensity scores is given in Table 2.

Propensity scores represent probabilities of group memberships, commonly used in causal inference studies. To use them as a measure of utility, we need to model group membership between the original and the masked data to obtain an estimate of distinguishability. Small distinguishability relates to high distributional similarity between the original and masked data. If we can model the propensity scores well, this general measure should capture relationships between the data that methods such as the empirical cumulative distribution function may miss. The propensity score method, given in Woo *et al.* (2009), described in algorithm 1 (Table 3), proceeds as follows. A set of predictor variables is specified and calculated for the original and masked data. The two data sets are combined with the addition of an indicator variable I giving the source of the data (0 for original data and 1 for altered). A propensity score is estimated for each of the rows of the combined data, as the probability of classification for the indicator variable.

The mean-squared difference between these estimated probabilities and the true proportion of records from the masked data in the combined data (denoted by c; usually 0.5) gives the utility statistic $(1/N)\Sigma(\hat{p}_i - c)^2$: the propensity score mean-squared error $pMSE$. In the case of synthetic data with $m > 1$ pMSE would be calculated for each data set and the mean taken as the overall utility. The method can be thought of as a classification problem where the desired result is poor classification (all records giving a probability of being synthetic close to 0.5), giving better utility for low values of pMSE.

Table 2. Notation for estimation of the propensity scores and following algorithms

Value	*Description*
n_1, n_2, N	Number of records in the original, synthesized and combined data
c	Proportion of synthesized rows in the combined data
k	Number of predictors in the propensity score model, including intercept
Z_{orig}	$n_1 \times k$ matrix of predictors from the original data
Z_{syn}	$n_2 \times k$ matrix of predictors from the synthetic data
Z	$N \times k$ matrix of predictors from the combined data
I	Indicator vector with 0 for original rows and 1 for synthetic rows
$\hat{p}$	Length N vector of predicted probabilities: $P(I = 1\|Z)$

Table 3. Algorithm 1: general utility statistic based on the propensity score mean-squared error

1: stack the original n_1 rows, Y_{real} and the n_2 rows of masked data Y_{syn} to create the $N = n_1 + n_2$ rows of Y_{comb}
2: add an indicator variable I to Y_{comb} such that $I = \{1 : y_i \in Y_{\text{syn}}\}$
3: fit a model to predict I by using predictors Z calculated from Y_{comb}
4: predict propensity scores $\hat{p}_i$ for each row of Z
5: obtain the utility statistic from $(1/N)\Sigma_{i=1}^{N}(\hat{p}_i - c)^2$

4. General utility for synthetic data

We extend the propensity score method for general utility specifically for the case of synthetic data. In particular, when pMSE is calculated from a logistic regression, we derive its large sample expectation and variance under the null case of synthesizing data from the correct generative model of the original data, and we use this to standardize the observed pMSE.

This standardization transforms to a scale that has a clear interpretation for synthetic data. The previous use of the propensity score measure for general utility gave better utility as the value became closer to 0, where a value of 0 would occur when the original and altered data are identical. This is highly unlikely for synthetic data as the goal is not to have identical entries, but to achieve the distributional similarity between the distribution of the observed data and the model used to generate the synthetic data. This condition is required for any inferences from synthetic data to be valid, and we shall refer to it as correct synthesis and when we refer to the null distribution of a statistic this will imply the distribution under correct synthesis. With expressions for the null expectation and standard deviation of pMSE for synthetic data, we can use two standardized statistics: either the ratio to its null expectation, the pMSE-ratio, or the standardized pMSE calculated as its difference from this expectation in units of the estimated null standard deviation. The former has an expected value of 1 and the latter an expectation of 0 and a standard deviation of 1 in the null case. In both cases, increased values of these statistics are expected if correct synthesis does not hold.

We also consider other models that are used to compute the pMSEs, such as non-parametric classification and regression trees (CARTs) which may improve the specification of utility for complex data sets over previously used models such as logistic models, general additive models or polynomial splines. In this case the theoretical results for the null pMSE do not hold, but we show that null values can be approximated by using resampling techniques. CART models were found to be promising for measuring utility in complex data sets and are included in the real data examples.

These general utility measures, with a choice of model for the propensity score, are implemented in the `synthpop` package (Nowok *et al.*, 2016), so data custodians creating synthetic data can compute pMSE, the pMSE ratio or the standardized pMSE as measures of the appropriateness of different synthesis models.

4.1. Null distribution of the mean-squared error calculated from a logistic regression

We first consider the null distribution of pMSE when all the data are synthesized and we derive asymptotic expressions for the expectation and variance of this null pMSE. Using simulated data we show that these expressions are valid and that pMSE-values grow further from their null expectation as the difference between the models generating the original and synthetic data increases.

4.1.1. Theoretical results: the null pMSE-distribution

For the distribution of pMSE, Z_{orig} is a fixed quantity and Z_{syn} is a matrix of random variables generated by the synthesis process. Under correct synthesis Z_{orig} has been generated as a sample from $f(y|\theta)$ and Z_{syn} from $f(y|\hat{\theta})$, where $\hat{\theta}$ is estimated from the original data. We show in Appendix A.1 that the null pMSE is distributed as a multiple of a χ^2-distribution with $k-1$ degrees of freedom and expectation and standard deviation given by

$$E[\text{pMSE}] = (k-1)\left(\frac{n_1}{N}\right)^2\left(\frac{n_2}{N}\right)\Big/N = (k-1)(1-c)^2c/N, \tag{1}$$

$$\text{StDev(pMSE)} = \sqrt{\{2(k-1)\}}\left(\frac{n_1}{N}\right)^2\left(\frac{n_2}{N}\right)\Big/ N = \sqrt{\{2(k-1)\}}(1-c)^2 c/N, \qquad (2)$$

where n_1 is the number of observations in the original data, n_2 the number of observations in the synthetic data, $N = n_1 + n_2$ and $c = n_2/N$. In the most common case when n_1 and n_2 are equal, the expectation becomes $(k-1)/(8N)$ and the standard deviation $\sqrt{\{2(k-1)\}}/(8N)$. The primary assumptions underlying these results are that the estimated propensity scores are not close to 0 or 1 and that the expectations of the means of the synthetic predictors, Z_{syn}, over repeated syntheses will be the means of the original predictors, Z_{orig}. These are discussed further in Appendix A.1.

Appendix A.3 derives the expectation of pMSE calculated from two synthetic data sets, generated from the same original data by the same method as used to compare the synthetic data with the original. We discuss in Section 4.3.1 why a comparison of pairs of syntheses can be useful as a method of estimating the null pMSE-distribution when it cannot be derived theoretically.

4.2. Incompletely synthesized data

When some part of the data are left unchanged this may involve synthesizing only selected variables (incomplete by variables), only selected records (incomplete by rows) or only some variables for some observations (incomplete by observations). When synthesis is incomplete by rows or by observations, the selection is usually restricted to those observations that are expected to pose a high risk of disclosure such as observations with extreme, potentially disclosive, values. When this is so, estimation of the models that are used to create the synthetic data must use records from only those observations that will be replaced (Reiter, 2003). Our theoretical results will not apply because the observations selected will not follow the same distribution as the complete data. This will also be so even for randomly selected rows, unless pMSE is calculated from only the synthesized rows.

The results in Section 4.1.1 easily extended to the case of incomplete variables; see Appendix A.2. In that case, the contribution from predictors depending only on unsynthesized columns is zero, since all values are unchanged. Equations (1) and (2) still hold with k replaced by k^*, the number of variables in the predictor matrix which relate to synthesized variables (including interaction terms between synthesized and unsynthesized variables). The following section presents simulation studies confirming these results, both for complete and for incomplete synthesis, with a multivariate normal example. The simulation also illustrates the behaviour of the pMSE-ratio, or the standardized pMSE under increasingly incorrect synthesis.

4.3. Simulation to validate asymptotic expressions for the expectation and variance of pMSE

We present simulation studies to show that the asymptotic results that are derived in Appendix A.1 and Appendix A.2 hold under correct synthesis and to show how they deviate from the expectations for incorrect synthesis. We ran 1000 simulations, and for each simulation we generated 10 original data sets (which are referred to henceforth as *real* data sets) of size 5000 from a multivariate normal distribution of dimension 10 with means 0, variances 1 and off-diagonal covariances of the ith data set taking values $0, 0.1, \ldots, 0.9$ for $i = 1, \ldots, 10$.

In the first simulation, for each real data set we generated a correct and incorrect complete synthesis. For the correct synthesis we use the variance matrix fitted to the real data to generate synthetic multivariate normal data. For the incorrect synthesis we use the sample means and a variance matrix with its off-diagonal elements set to 0. The incorrect synthesis uses a model

that is progressively further from the true generative model as the real data are generated from a model with covariances that increase from 0 to 0.9. This emulates synthesis that fails to account for correlations between the variables.

We model the propensity scores with a logistic regression model including all main effects and first-order interactions for the variables, but omitting the quadratic terms, giving us $k = 56$ parameters. The expected null mean of pMSE becomes

$$E[\text{pMSE}] = (k-1)(1-c)^2 c/N = 55 \times 0.5^3/10000 = 0.000688 \tag{3}$$

and its standard deviation is

$$\text{StDev(pMSE)} = \sqrt{\{2(k-1)\}}(1-c)^2 c/N = \sqrt{110} \times 0.5^3/10000 = 0.000131. \tag{4}$$

Table 4 gives the means of the simulation results. For correct synthesis the mean pMSE agrees with equation (3) and that of its standard deviation with equation (4) (the data are not shown for equation (4)). Thus the pMSE-ratio (the mean pMSE divided by equation (3)) and the standardized pMSE (the mean pMSE minus equation (3) divided by equation (4)) are close to 1 and 0 respectively, as expected. Values below 1 for the ratio pMSE or 0 for the standardized pMSE are acceptable, simply a result of random variation, and implying correct synthesis.

For the incorrect syntheses models that fail to capture the correlations between the variables, pMSE-values compared with the original data increase as the covariance values increase as does its standard deviation (the standard deviations are not shown). Note that, for the first row of Table 4 when the synthetic data are generated from a model with covariances of 0, it still does not give a value at the expectation, as was the case for synthesis from the correct model. This is because, even though the population covariances are set to 0, the simulated real data do not have exactly zero covariances, so the incorrect synthesis here is not generated from a model that is correctly fitted to the observed data. As the covariances in the original data increase, the pMSE-ratio and the standardized pMSE increase, the latter very steeply. The ratio is an appropriate measure of the discrepancy which the pMSE-model finds between the two distributions. The

Table 4. Results from 1000 simulated syntheses of multivariate normal data using correct and incorrect models with the pMSE calculated from a logistic model including all main effects and first-order interactions†

Population covariance	*Results for correct synthesis pMSE*			*Results for incorrect synthesis pMSE*		
	Mean	*Ratio*	*Standardized score*	*Mean*	*Ratio*	*Standardized score*
0.0	0.000684	0.995	−0.024	0.00124	1.805	4.221
0.1	0.000693	1.007	0.039	0.01428	20.77	103.7
0.2	0.000696	1.013	0.068	0.03158	45.93	235.6
0.3	0.000688	1.000	0.001	0.04696	68.31	353.0
0.4	0.000686	0.998	−0.008	0.06021	87.57	454.0
0.5	0.000686	0.998	−0.010	0.07202	104.8	544.1
0.6	0.000684	0.996	−0.024	0.08248	120.0	623.9
0.7	0.000686	0.998	−0.010	0.09192	133.7	695.9
0.8	0.000688	1.001	0.005	0.10054	146.2	761.7
0.9	0.000691	1.005	0.029	0.10830	157.5	820.9

†Ratios and standardized scores from theoretical expectations.

standardized value gives a measure (like a t-statistic) of its deviation from the null value. Given that we know that correct synthesis can rarely be fully achieved, except for simulated data, the standardized measure may be oversensitive to small differences and the ratio pMSE is likely to be a more useful measure.

In the second simulation, for each real data set we generated a correct and incorrect incomplete synthesis, leaving eight of the 10 original variables unchanged. For the correct synthesis we fitted linear models using all unsynthesized variables as predictors (and the first synthesized variable as a predictor for the second) to generate new synthetic draws. For the incorrect synthesis we took a parametric bootstrap of the two variables by using the sample means and standard deviations. In the same way as the complete synthesis, the incorrect synthesis ignores the correlations between variables and grows progressively further from the true generative model as the real data are generated from a model with covariances that increase from 0 to 0.9.

Equations (5) and (6) give the new expected value and standard deviation of pMSE with only two synthesized variables. Recall that k^* is the dimension of the propensity score predictor matrix that involves synthesized variables. Including main effects and first-order interactions, this reduces from 55 previously to 19. The simulation results that are given in Table 5 confirm this, as well as showing a similar pattern for the ratio and standardized pMSE-values for incorrect synthesis as was seen in Table 4:

$$E[\text{pMSE}] = (k^* - 1)(1 - c)^2 c/N = 19 \times 0.5^3/10\,000 = 0.000\,237\,5 \tag{5}$$

and its standard deviation

$$\text{StDev(pMSE)} = \sqrt{\{2(k^* - 1)\}}(1 - c)^2 c/N = \sqrt{38} \times 0.5^3/10\,000 = 0.000\,077\,055. \tag{6}$$

4.3.1. Using resampling techniques for the distribution of pMSE

We can use the results above when calculating propensity scores by using a fully specified logistic model which provides a value of k for the number of fitted parameters, but we may be interested

Table 5. Results from 1000 simulated syntheses of multivariate normal data with only two of the 10 columns synthesized, using correct and incorrect models with pMSE calculated from a logistic model including all main effects and first-order interactions†

Population covariance	*Results for correct synthesis pMSE*			*Results for incorrect synthesis pMSE*		
	Mean	*Ratio*	*Standardized score*	*Mean*	*Ratio*	*Standardized score*
0.0	0.000244	1.027	0.083	0.00045	1.902	2.781
0.1	0.000239	1.007	0.022	0.00618	26.00	77.05
0.2	0.000239	1.007	0.022	0.01553	65.39	198.5
0.3	0.000237	0.996	−0.013	0.02551	107.4	328.0
0.4	0.000232	0.975	−0.076	0.03563	150.0	459.3
0.5	0.000236	0.994	−0.019	0.04576	192.7	590.8
0.6	0.000233	0.982	−0.055	0.05614	236.4	725.5
0.7	0.000236	0.995	−0.015	0.06697	282.0	866.0
0.8	0.000233	0.981	−0.060	0.07849	330.5	1016
0.9	0.000232	0.978	−0.066	0.09118	383.9	1180

†Ratios and standardized scores from theoretical expectations.

in using data-adaptive models, such as stepwise regressions or CARTs. In these cases, we cannot use the previous results, but we would still like to estimate the null pMSE. We show that the null distribution can be estimated by using resampling techniques. The theoretical derivations in the previous section assumed that the two data sets that were compared were drawn from the same underlying generative model. By resampling, we can compare two data sets which we know were generated from identical distributions, and we can use the resulting pMSE-values to estimate the theoretical null pMSE.

One such resampling method is to calculate the pMSE between pairs of synthetic data sets generated from the same original data. The pairs can be used to obtain an estimate of the expected pMSE even when the synthesizing model is incorrect, since both data sets are drawn from the same distribution. This method requires much additional computation if only one synthetic set is planned. An alternative method in the case of a single synthetic data set is to use a permutation test to obtain null expectations. We describe it here for the case when the synthetic data have the same number of records as the original. The indicator variable that is used with the Z-matrix from the original and a single synthetic data set is permuted and pMSE calculated from each permutation (see algorithm 2 in Table 6). This method can be less computationally burdensome than producing extra syntheses, and it can also produce utility estimates when only a single synthetic data set has been produced. Its disadvantage is that it does not give the correct null pMSE unless all the data are synthesized. This can be understood by considering the contribution to pMSE from columns of Z corresponding to the unsynthesized data X. In calculating pMSE from the original data there will be no contribution from these columns because the difference in means will be 0 (see Appendix A.1). But the contribution will not be nothing with the permutation distribution because the permutation no longer treats X as fixed. An alternative approach would be to omit any Z-variables that depend on X-variables from the calculation of pMSE, but this would be unsatisfactory since it would not evaluate whether the relationships between Y_{syn} and X were maintained.

We discuss the expectation of the pMSE calculated from pairs or permutations in Appendix A.3. For most large complex data sets, synthesized by CART models, the expected pMSE from pairs will be close to, or slightly lower than, the null pMSE. Thus we propose two resampling methods that can be used when methods, such as CARTs, without a known number of parameters are used to calculate the distribution of pMSE and derive the pMSE-ratio and the standardized pMSE utility statistics. To confirm our results, the simulation study that was described above was extended to include our evaluation of the resampling method, and it is included in Appendix B. For logistic models with known k, the resampling methods gave estimates of the null distribution of pMSE that agreed with the theoretical results (the data are not shown). For CART propensity score models, where we do not know k, the expected values under permutation

Table 6. Algorithm 2: permutation test for null mean and standard deviation estimates

1: *if* $m > 1$ synthetic data sets *then*
2: randomly assign a synthetic data set for each permutation
3: *end if*
4: *for* each permutation *do*
5: randomly shuffle the group indicator variable I to produce I_p
6: follow algorithm 1 using I_p in place of I
7: obtain $\text{pMSE}_{\text{Perm}_i}$ from the predicted propensity scores
8: *end for*
9: return $\overline{\text{pMSE}_{\text{Perm}_i}}$ and $\text{sd}(\text{pMSE}_{\text{Perm}_i})$ for null mean and standard deviation values

Table 7. Estimation methods of the null pMSE for various synthesis and propensity score model scenarios

	Propensity score model type	
	Logistic regression	*CART*
Complete synthesis	Theoretical	Permutation (or pairwise)
Incomplete (columns)	Theoretical	Pairwise

stayed constant across different syntheses as expected and the ratio of pMSE to the null expectation increased as the model that was used for synthesis was further from the correct model. We present results for complete synthesis and for incomplete columns. We also investigated the possibility of using resampling methods for the null distribution of pMSE for synthesis with selected rows. Although the pairwise method gave satisfactory results for randomly selected incomplete rows, we have not investigated the important, but more complicated, situation when the data to be replaced are selected according to their perceived disclosure potential.

Table 7 summarizes the applicable methods under various synthesis and propensity score model scenarios. If pMSE is calculated from a method with a known number of parameters, k, then the ratio and standardized measures can be calculated from equations (1) and (2) for both complete and incomplete (by variables) syntheses. For complete synthesis with a model where k is unknown, permutation methods are recommended to obtain the ratio and standardized pMSE. Repeated pairwise syntheses could also be used in this case but have the disadvantage of requiring multiple syntheses. When only some of the variables are synthesized then the only method that is possible for CART models is the paired comparisons of multiple syntheses.

4.4. Choice of model for the propensity score

As Woo *et al.* (2009) have discussed, the choice of model for the propensity score is crucial to the way in which the pMSE measures compare masking methods. Woo *et al.* (2009) evaluated some different logistic regression propensity score models, and they found that it was important to include higher order terms, including cubic terms, for pMSE to discriminate between methods such as incorrect simulation, adding random noise and aggregation. However, their simulated data largely relied on inappropriate marginal distributions for the incorrect model. This type of inadequacy should be readily checked for synthetic data by visual comparisons of the real and synthetic data, as is done in the `synthpop` package (Nowok, 2015). For their real data example, Woo *et al.* (2009) used a model with all main effects and first-order interactions between variables, where generalized additive models were used for the continuous variables. This approach would seem to be a useful starting point, although it might be more helpful to use the transformations that would normally be used in modelling continuous variables, rather than the additive models.

We consider expanding the propensity score models to include CART models (Breiman *et al.*, 1984). These models have proved useful for generating synthetic data (Reiter, 2005b) and have been shown to outperform other machine learning techniques (Drechsler and Reiter, 2011; Nowok, 2015) and parametric models (Nowok *et al.*, 2017) for this purpose. Additionally, boosted tree models have been found to be useful for estimating propensity scores in causal inference applications; see McCaffrey *et al.* (2013).

It is well known that CART models are subject to overfitting and parameters can be set to control the complexity to prevent this. This is not generally a problem for generating synthetic data but it can be when the pMSE-score is calculated, since a substantial proportion of the propensity scores may be close to 0 or 1 even under data that have been generated from a correct synthesis. This leaves little room for the pMSE-value to increase when an incorrect synthesis model is used, since the overfitted model picks up higher differences even when the synthetic data are drawn from the correct model. It is important to check whether drastic overfitting is occurring, by looking at the propensity scores, and if necessary adjusting the tuning parameters.

At the other extreme the parameters should be set to allow adequate discrimination. If the classification tree fails to perform any splits all estimated propensity scores will equal 0.5 and pMSE will be 0. Although you may argue that this indicates good synthesis, it is more likely that the tuning parameters for the decision tree were not set appropriately. The resamples that are used to evaluate the expected null utility can inform the choice of tuning parameter. If any CART models that are used to calculate the null distribution fail to split, then the tuning parameters should be adjusted to produce larger trees. For a simple synthetic data set, logistic models with first-order interactions should be tried first. As the data become more complex, we recommend also fitting parametric models with higher order interactions (if computationally feasible) and CART models for comparison. The utility function in the `synthpop` package currently includes both CART models and logistic models with a maximum order of interactions between variables to be specified.

5. Specific utility measures for synthetic data

In contrast with the general utility approach, we can measure the utility of a synthetic data set by assessing the similarity of results for specific analyses using both the original and the synthetic data. For high utility we expect close similarity between the results for the same analysis calculated from the two different sources of data. Most of the previous literature has used specific utility measures rather than general measures, usually for other types of disclosure-controlled data, e.g. produced by top coding or microaggregation, rather than for synthesized data. Karr *et al.* (2006) and Reiter *et al.* (2009) referred to this type of utility as fidelity measures, since it provides the masked data users with a measure of trustworthiness for the analysis compared with the analysis on the unreleased data.

The most common and understandable examples of analysis-specific measures compare estimated summary statistics or general linear model coefficients obtained from the original and masked data. The percentage overlap of confidence intervals, for each of the coefficients or summary statistics of interest, are calculated from the observed and masked data, e.g. Karr *et al.* (2006), Reiter *et al.* (2009), Drechsler and Reiter (2009), Slavković and Lee (2010) and Woo and Slavkovic (2015). An interval overlap measure, which is given in equation (7), can then be calculated for each statistic of interest and summarized by the average, with a higher IO corresponding to greater utility. Note that this measure is negative when there is no overlap and will decrease as the intervals move further apart:

$$\mathrm{IO} = 0.5\left\{ \frac{\min(u_\mathrm{o}, u_\mathrm{s}) - \max(l_\mathrm{o}, l_\mathrm{s})}{u_\mathrm{o} - l_\mathrm{o}} + \frac{\min(u_\mathrm{o}, u_\mathrm{s}) - \max(l_\mathrm{o}, l_\mathrm{s})}{u_\mathrm{s} - l_\mathrm{s}} \right\}, \tag{7}$$

where the subscripts ‘o’ and ‘s’ denote the confidence interval bounds for the original and synthetic data respectively.

The IO-measure has been extended by Karr *et al.* (2006) to a measure of ellipsoid overlap,

EO, which uses an estimate of the overlap between the joint posterior distribution of all the parameters for the original and synthetic data. EO is a more satisfactory measure because it allows for the correlation between the parameter estimates. However, it is much more onerous to compute, the easiest method involving simulation, and may be less easily understood by those analysing the data.

An alternative summary of the differences in summary statistics is the standardized difference between the original estimate and the synthetic estimate calculated as $|\hat{\beta}_{\text{orig}} - \hat{\beta}_{\text{syn}}|/\text{se}(\hat{\beta}_{\text{orig}})$, where $\hat{\beta}_{\text{orig}}$ and $\hat{\beta}_{\text{syn}}$ are the coefficients of the same model estimated from the real and synthetic data and $\text{se}(\hat{\beta}_{\text{orig}})$ is the estimated standard error of the coefficients from the original data. This measure was used in Woo and Slavkovic (2015) to test data that had undergone the post-randomization method, and it is similar to the standardized bias that was suggested by Loong *et al.* (2013), which differs only by using the estimated standard error from the synthesized data.

For our examples we present both the confidence interval overlap and the standardized difference as measures of specific utility. These two related measures are implemented in the `synthpop` package under the `compare.fit.synds()` function and can be used to compare results from synthetic data with a gold standard analysis once a researcher's code has been run on the original data. For a model with many coefficients IO and the standardized difference can be summarized by their mean or their median and range or, more usually, displayed graphically.

6. Data examples

As discussed in previous sections, specific utility measures the inferential usefulness of a data set for a given model. When the model to be used with synthetic data is well specified, it may be misleading to rely only on specific measures if they are close to the model that is used for the synthesis (Raab *et al.*, 2017a). General measures assess a broader set of differences, including those that might influence the results of exploratory analyses and of model validation. Different models for the propensity score will be sensitive to different aspects of these differences. General utility and specific utility, each with a range of models, should be used along with data visualizations and marginal distribution checks to aid synthetic data producers in determining which synthesis is best for release. We use two real data examples to illustrate the need for this holistic approach.

6.1. Scottish health survey

We use data from the 2013 SHS, focusing specifically on the data that were used for the 2015 report on mental health and wellbeing; see Wilson *et al.* (2015). This report uses a subset of the SHS data set containing 8721 observations on 15 variables covering demographic information, behavioural factors and mental health indicators: Table 8.

The study focused on mental health outcomes for males and females as measured by the two scores, the Warwick–Edinburgh mental wellbeing scale WEMWBS and the general health questionnaire GHQ12, while controlling for demographic and behavioural factors. WEMWBS is derived from 14 questions concerning personal thoughts and feelings with self-reported answers. GHQ12 entails 12 experiential questions, six positively worded and six negatively worded, with self-reported responses of the participants' level of agreement. Specifically the models that were estimated, which we replicate, were four logistic regression models, two for men and two for women with the two mental health indicators as the response variables, detailed in Table 9. Although these responses were originally continuous they were dichotomized to 0–1 variables, with 1 indicating a higher level of mental health problems.

We create three synthetic data sets by different methods: sequential parametric regression

Table 8. Summary of data for the report on mental health and wellbeing

Variable	*Label*	*Range*
Sex	Sex	male ≡ 1; female ≡ 2
Age group	ag16g10	7 categories, minimum = 16
Martial status	maritalg	6 categories
Parental employment type	pnssec5	7 categories
Income quintile	eqv5	6 categories
In 15% most deprived area	SIMD15_12	1 ≡ no; 2 ≡ yes
Economic activity	econac12	6 categories
Provides caregiving	RG17a	5 categories
Physical activity level	adt10gpTW	4 categories
Servings of fruits and vegetables	porftvg3	3 categories
Has alcohol dependence	AUDIT20	1 ≡ no; 2 ≡ no answer; 3 ≡ possibly
Smoker status	cigst3	1 ≡ current; 2 ≡ never; 3 ≡ ex
Chronic obstructive pulmonary disease diagnosis	COPDDoct	1 ≡ yes; 2 ≡ no
WEMWBS mental health score	wemwbs	1 ≡ 'issues'; 0 ≡ 'standard'
GHQ12 mental health score	ghq12scr	1 ≡ 'issues'; 0 ≡ 'standard'

Table 9. Wellbeing fitted models for the SHS data used to calculate specific utility

Model	*Sex*	*Response*	*Covariates*
1	Male	wemwbs	ag16g10, maritalg, SIMD15_12, econac12, eqv5, RG17a, adt10gpTW, AUDIT20, cigst3, porftvg3, COPDDoct
2	Female	wemwbs	ag16g10, maritalg, SIMD15_12, econac12, eqv5, RG17a, adt10gpTW, AUDIT20, cigst3, porftvg3, COPDDoct
3	Male	ghq12scr	ag16g10, maritalg, pnssec5, econac12, eqv5, RG17a, adt10gpTW, AUDIT20, cigst3, COPDDoct
4	Female	ghq12scr	ag16g10, maritalg, pnssec5, econac12, eqv5, RG17a, adt10gpTW, AUDIT20, cigst3, COPDDoct

models, sequential non-parametric CART models and non-parametric bootstrap samples of each variable (sampling). In each case we replace all the original observations by synthesized values. These syntheses are evaluated by three general utility measures, each using a different model for the propensity score: logistic regression with main effects only, with main effects and first-order interactions and a CART model. The logistic models had 44 and 964 degrees of freedom. Results are in Table 10 for pMSE and its ratio and standardized values from the theoretical null distribution for logistic models and from permutations for the CART model. Table 10 also presents the specific utility measures that were given in Section 5 for models 1–4, i.e. confidence interval overlap and standardized differences in coefficient values. For both of these, the median across all covariates in the models is reported.

The parametric regression model is evaluated as satisfactory by all three general utility measures. The CART method is judged less satisfactory by the logistic interactions model. The sampling method is judged inadequate by the logistic interactions and the sampling methods. The propensity score from main effects only fails to identify any of these problems.

The specific utility models assess the synthesis models similarly. Parametric synthesis gives highest interval overlaps and lowest standardized differences for all models. Model 4 has the

Table 10. SHS general and specific utility results; comparing synthesis models with different pMSE-models

Measure	*Model*	*Results for the following methods for synthesis:*		
		Parametric	*CART*	*Sampling*
General utility				
pMSE	pMSE logistic main effects	0.000384	0.000327	0.000420
pMSE	pMSE logistic interactions	0.00726	0.0144	0.1277
pMSE	pMSE CART	0.0512	0.0457	0.1221
pMSE-ratio	pMSE logistic main effects	1.19	1.01	1.30
pMSE-ratio	pMSE logistic interactions	1.05	2.09	18.5
pMSE-ratio	pMSE CART	0.996	0.920	2.09
Standardized pMSE	pMSE logistic main effects	0.908	0.064	1.43
Standardized pMSE	pMSE logistic interactions	1.14	23.8	384
Standardized pMSE	pMSE CART	−0.157	−2.86	45.7
Specific utility				
Median confidence interval overlap	Fitted model 1	0.737	0.670	0.630
Median standardized $\hat{\beta}$ difference	Fitted model 1	1.03	1.29	1.45
Median confidence interval overlap	Fitted model 2	0.906	0.797	0.270
Median standardized $\hat{\beta}$ difference	Fitted model 2	0.369	0.796	2.86
Median confidence interval overlap	Fitted model 3	0.834	0.695	0.587
Median standardized $\hat{\beta}$ difference	Fitted model 3	0.651	1.19	1.62
Median confidence interval overlap	Fitted model 4	0.822	0.675	0.487
Median standardized $\hat{\beta}$ difference	Fitted model 4	0.697	1.27	2.01

largest differences between specific utility for different synthesizing models and this is illustrated in Fig. 1 where confidence intervals overlap and standardized differences for all 39 coefficients are displayed as boxplots. For this example, general and specific utility measures agree that the parametric synthesis method performs best and both evaluate the synthesis model as compatible with the distribution of the original data.

6.2. Historical census data

Our second example uses data from the 1901 Census of Scotland made available by the integrated census microdata (ICM) project (`https://www.essex.ac.uk/history/research/icem/`). These data sets have many features that make them similar to current census data from the UK, such as large sample sizes, mainly categorical data (some with small categories), some highly skewed continuous variables and data organized by household, but they have the advantage that the original data are freely available to disseminate.

To illustrate our methods we use a subset of the data consisting of private households in the historic county of Midlothian and the parish of the City of Edinburgh, leaving 46 110 records. The variables that are shown in Table 11 consist of individual characteristics of the head of the household, plus data on household composition and number of rooms. The variable 'disability' had a large number of categories, many with only a few cases, but only 164 individuals affected. Thus this category is reduced to a binary category indicating any disability. There were also many small categories for country of birth so all individuals who were born outside the UK are coded as 'other'. The variable 'work_status' is derived from the census data on employer status and occupation and has four categories according to whether a head of household was a worker, an employer, had independent means or this was irrelevant (e.g. students or retired

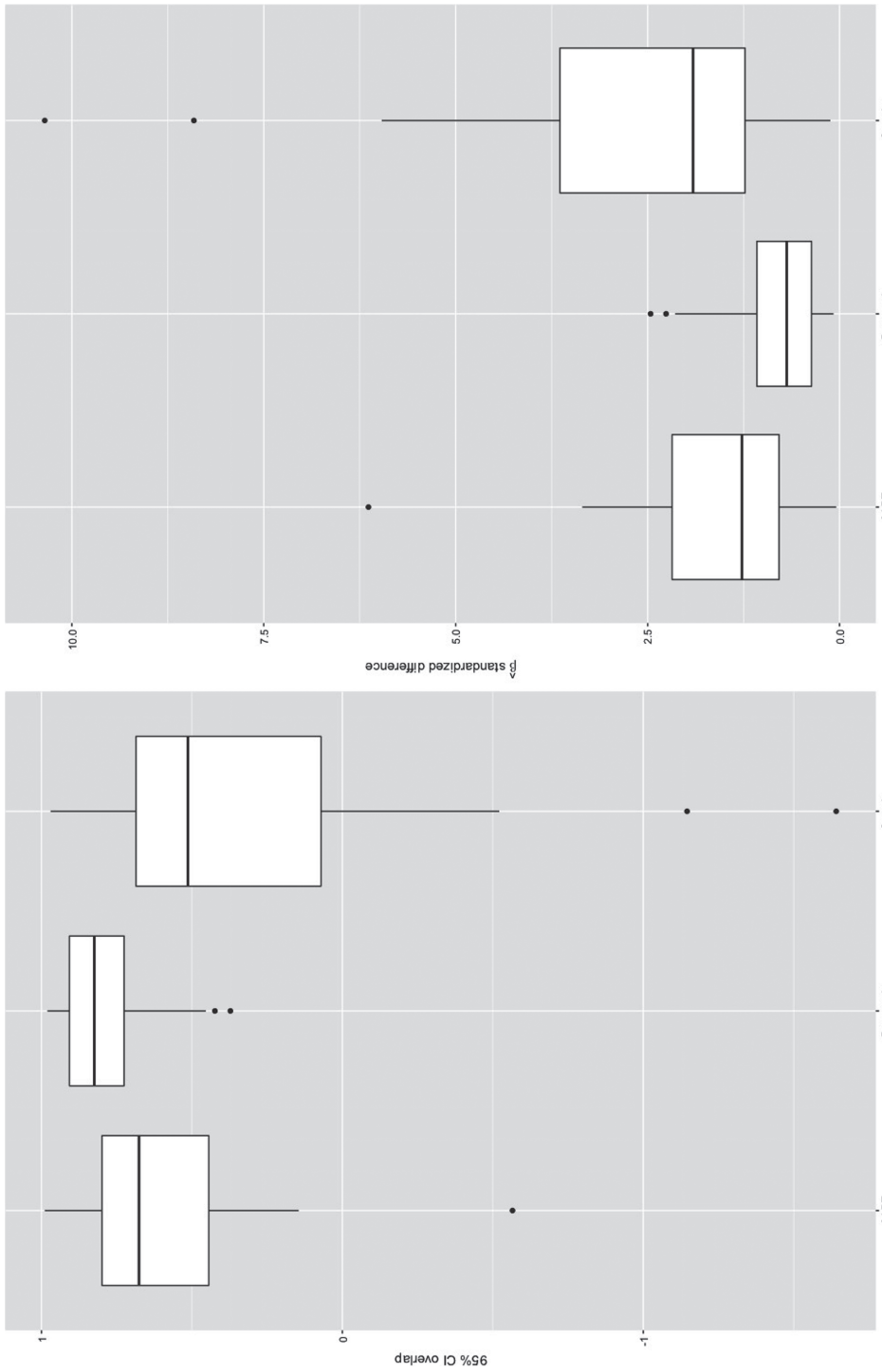

Fig. 1. SHS model 4 specific utility: boxplots, one for each synthesizing method, of (a) the 95% confidence interval overlap and (b) standardized $\hat{\beta}$ difference for each coefficient

Table 11. ICM 1901 historical census data, $N = 46110$

Variable	*Label*	*Range*
Head-of-household characteristics		
Sex	sex	2 categories
Disability	disability	2 categories
Marital status	mar_stat	4 categories
Age (years)	age	10–96 years
Working status	work_status	4 categories
Country of birth	ctry_bth	5 categories
Household characteristics		
Number of related individuals	n_relations	1–26
Number of lodgers and boarders	n_lodgers	0–11
Number of others (servants, visitors or unknown)	n_others	0–12
Total rooms in dwelling	totrooms	1–54

people). This data set and the code that was used to create the synthetic data sets and to evaluate their utility are available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

The variables *sex* and *age* are left unsynthesized. The remaining variables are synthesized by using three different methods: CARTs, parametric models with normal linear regression for numeric variables ('normal') and the same parametric model with normal replaced by linear regression on the expected normal deviates from the ranks (rank). The synthesis was conditional on the values of the unsynthesized variables and the order of the remaining variables was chosen by synthesizing the numerical variables first and the categorical variables second, giving

$$\{\text{totrooms}, \text{n\_relations}, \text{n\_others}, \text{n\_lodgers}, \text{work\_status}, \text{disability}, \text{mar\_stat}, \text{ctry\_bth}\}$$

and 15 synthetic data sets were generated from each method, and the observed pMSE taken as the average from the pMSE that was calculated with each data set. General utility was assessed from a CART model and from a logistic model with first-order interactions. For the CART propensity models the null pMSE was calculated from the 105 pairwise comparisons of the 15 synthetic data sets. The logistic interactions propensity score model had 141 parameters, but the expectation of pMSE was calculated from the 138 parameters that included synthesized variables. Table 12 gives the observed pMSE for each combination of synthesis method and propensity score model, along with the two measures rescaled by the null propensity score estimates.

For specific utility, three models are estimated as shown in Table 13, and measures of confidence interval overlap and standardized $\hat{\beta}$ difference are calculated. Results for each are also given in Table 12.

We can see that the general utility measures judge the two parametric syntheses as quite unsatisfactory. The coefficients of the logistic utility model and the trees that were produced by CARTs (which are not presented here) show that it is the distribution of n_relations, n_lodgers, n_others and totrooms that discriminates between the real and synthetic data. Plots comparing the original and synthetic data, by each of the three methods, for these four variables are available in the on-line supplementary material. The normal method produces negative values in the synthetic data for each of these variables which readily allow the propensity score model to predict that they are not part of the original distribution. The rank method attempts to

Table 12. General and specific utility results for syntheses of the ICM data by three methods, evaluated with two propensity score models

Measure	*Model*	*Results for the following methods for synthesis:*		
		CART	*Parametric*	
			Normal	*Rank*
General utility				
pMSE	pMSE logistic interactions	0.000833	0.00268	0.0157
pMSE	pMSE CART	0.000663	0.118	0.0294
pMSE ratio	pMSE logistic interactions	4.46	14.3	83.9
pMSE ratio	pMSE CART	1.51	18.8	6.73
Standardized pMSE	pMSE logistic interactions	28.7	110	689
Standardized pMSE	pMSE CART	4.76	257	79.9
Specific utility				
Mean confidence interval overlap	Fitted model 1	0.357	0.621	0.560
Mean standardized $\hat{\beta}$ difference	Fitted model 1	2.52	1.48	1.72
Mean confidence interval overlap	Fitted model 2	0.564	0.778	0.844
Mean standardized $\hat{\beta}$ difference	Fitted model 2	1.71	0.869	0.611
Mean confidence interval overlap	Fitted model 3	0.68	0.476	−0.250
Mean standardized $\hat{\beta}$ difference	Fitted model 3	1.24	2.06	4.90

Table 13. ICM fitted models

Model	*Response*	*Type*	*Covariates*
1	work_status	Multinomial	sex, age, mar_stat, ctry_bth, n_relations, n_lodgers, n_others
2	disability	Logistic	sex, age, mar_stat
3	totrooms	Linear	sex, age, mar_stat, n_relations, n_lodgers, n_others

overcome this by fitting models to z-scores. But this is not satisfactory, especially for the most skewed variables (n_lodgers and n_others) giving synthetic data with too high a proportion of 0s. The general utility for the rank method is slightly better than the normal synthesis when evaluated with a CART propensity score model, but much worse when evaluated with the logistic interactions model. The coefficients of the logistic propensity score model show that the rank synthesis is worse than the normal synthesis at modelling the interactions between pairs of variables. The CART synthesis method, in contrast, reproduces the distributions of these four variables very well. However, the general utility results indicate that there are other problems with this synthesis. Examination of the tree for the propensity score CART model and the coefficients of the logistic propensity score models indicate that the problem lies with the interaction between certain pairs of variables.

Clearly none of these methods of synthesis is satisfactory, and this is also reflected in the specific utility measures for the three chosen models, none of which are satisfactory. The parametric models perform best for model 2 which does not involve any of the four problematic variables and worst for model 3 which includes them all. Fig. 2 visualizes the confidence interval overlaps

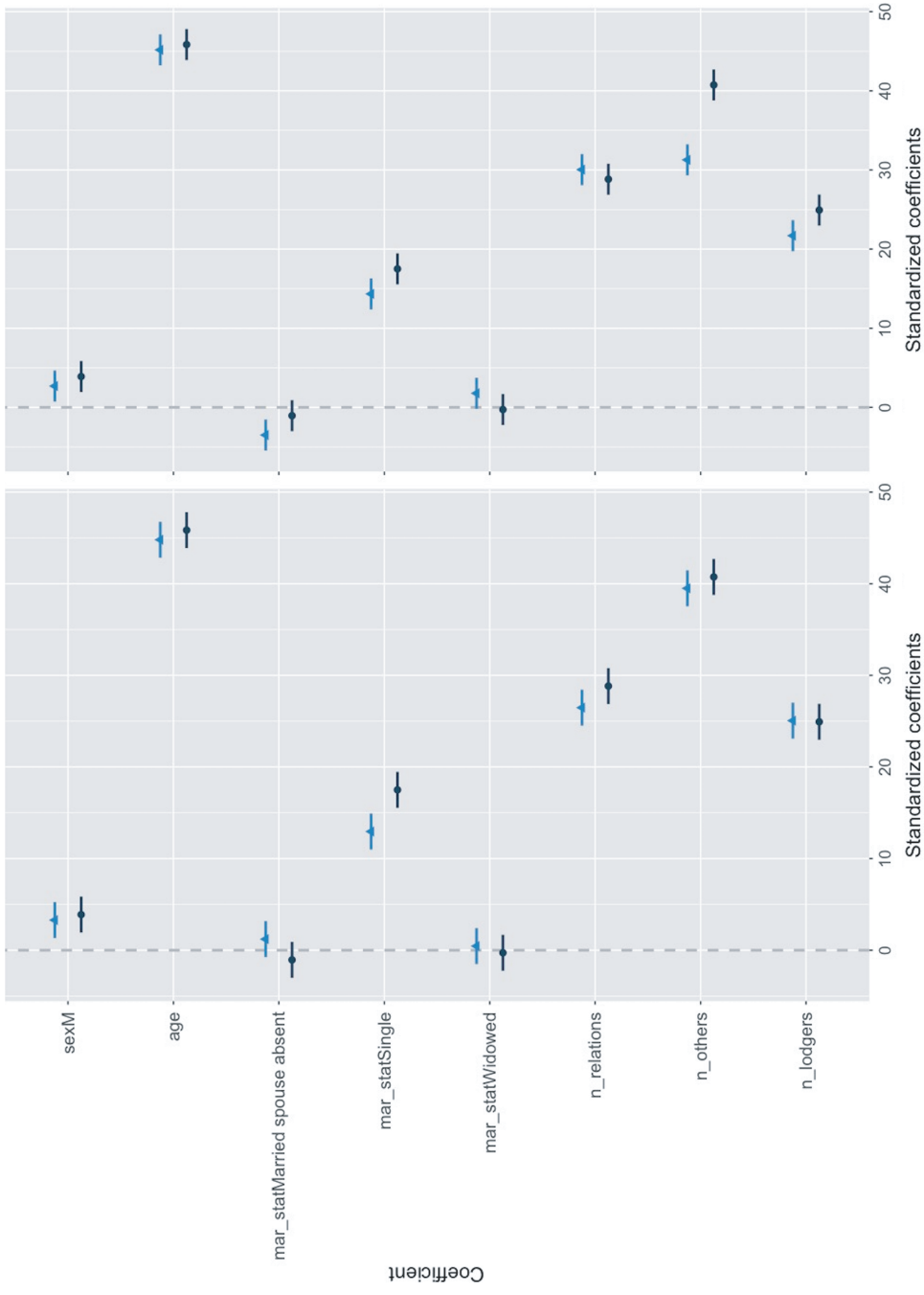

Fig. 2. ICM model 3 confidence interval overlap plots for (a) CART and (b) normal synthesis: synthetic (▲) and observed (●) confidence intervals for each standardized coefficient

for model 3, synthesized by CART and normal models. We see that the CART method performs poorly primarily with the mar_stat variable, whereas the others are fine. The interaction between mar_stat and totrooms was one of the interactions that was identified by the diagnostics from the utility models for the CART synthesis. Marital status comes late in the sequence of conditional models so it will generate a complex prediction tree. Other strong predictors of marital status, like age, will produce many splits that prevent the effect of marital status on the number of rooms from being captured. The normal model has a different problem where it underestimates the strong relationship between the number of rooms and the number of other individuals. (The explanation for this finding is probably related to the fact that many of the others were servants who were predominantly found in households with more rooms.) The rank method (the data are not shown) distorts the relationship between the number of rooms and family members even more than does the normal model.

The synthesis methods that were illustrated here were chosen arbitrarily to illustrate our methodology. The two models for the propensity score along with visualization of the distributions have enabled us to identify two different problems with the syntheses. To produce a satisfactory synthetic extract from data like these will require a customized synthesis where the models selected, the methods used, the choice of predictors and the order in the sequence for each conditional model are specified, as discussed in Raab *et al.* (2017b). The utility measures can inform the staff producing synthetic data in making these choices.

7. Conclusions

In this paper we develop extensions to general utility measures for synthetic data. Our extensions include two new general utility statistics, the pMSE-ratio and standardized pMSE, calculated by standardizing the statistic by its null expected value and standard deviation. Rescaling by the null statistic aids the interpretation of utility, for the specific case of synthetic data. Rather than basing our distance measure on identical data matrices, the standardized distance measures evaluate whether the synthesizing model is the correct model for the original data. Our measures are easier to compare because they do not depend on sample size. The methods that are described here are implemented as functions in the `synthpop` package for R and they are used in the code that is available from `http://wileyonlinelibrary.com/journl/rss-datasets` that produced the analyses of the ICM data that were described in Section 6.2. As Woo *et al.* (2009) have discussed, the choice of model that is used to calculate pMSE is crucial to its performance as a utility measure. We proposed extending the models suggested by including non-parametric CART models to estimate propensity score values. Our examples of both real and simulated data suggest that CART models may be useful, particularly in the case of complex data. Parametric models with higher order interactions deserve further exploration but are often not computationally feasible with many categorical variables. One solution might be to identify a subset of variables and to investigate how well general utility suggests that relationships between them are maintained. Different propensity score models are sensitive to particular aspects of the shortcomings of the synthesizing model. Thus we recommend that a synthesis should be evaluated with more than one such model.

The utility evaluations for our two examples of synthesizing real data are very different from one another. For the SHS example one of the synthesizing methods evaluates well and the synthetic data can reproduce the conclusions of models from a published analysis of these data. For the ICM data neither of our initial choices of synthesizing method is satisfactory, and a customized synthesis will be needed to produce useful synthetic data for this example. In our experience of producing synthetic versions of confidential data we have encountered

many examples that resemble the ICM data in their need to tailor the synthesis methods to the structure of the data and the needs of potential users. All utility measures in this paper are being implemented in the `synthpop` package in R. The general utility measures can be used by staff creating synthetic data extracts to tailor the methods that are used to synthesize a data set to provide users with synthetic data that will be fit for their purposes. The researchers can use the specific utility measures when they carry out a gold standard analysis at the end of their projects to be reassured that the synthetic data that were used for exploratory analyses were not misleading. Thus, our methodological developments will help agency staff to produce useful synthetic data and thus to widen access to the use of confidential data by researchers.

Acknowledgements

This work was supported in part by National Science Foundation grants 'Big data social sciences' IGERT DGE-1144860 to Pennsylvania State University, and BCS-0941553 and SES-1534433 to the Department of Statistics, Pennsylvania State University. Beata Nowok and Gillian Raab are funded by the UK Economic and Social Research Council's Administrative Data Research Centre–Scotland, grant ES/L007487/1.

Appendix A

A.1. Null distribution of pMSE calculated from a logistic model

We assume that the original data have n_1 observations and the synthesized data n_2. To compute pMSE we fit a logistic regression model of the indicator variable I on an $(n_1+n_2)\times k$ matrix of predictors Z where

$$Z=\begin{pmatrix} Z_{\text{orig}} \\ Z_{\text{syn}} \end{pmatrix},$$
$$I=\begin{pmatrix} I_1 \\ I_2 \end{pmatrix}$$

and I_1 is an n_1-vector of 0s, I_2 an n_2-vector of 1s, Z_{orig} is derived from the original data and Z_{syn} from the synthesized data. Note that the usual formulae for the standard errors of logistic regression will not apply here, since I is fixed, and not random. The distribution of any statistic is derived from that of the random variables Z_{syn}, conditionally on the observed values of Z_{orig} and I. Note that the Z-matrix here will include a column of 1s for the intercept and will usually contain the original Y_{orig}- and Y_{syn}-values as well as interaction and product terms or other functions calculated from them.

A logistic regression model can be fitted by updating the current estimate of the coefficient vector β^* by iterative reweighted least squares. The weights are $w^*=p^*(1-p^*)$ where p^* is the current estimate of the fitted proportion and the dependent variable is $Z'\beta^*+W^{-1}(I-p^*)$ where W^* is an $N\times N$ diagonal matrix with elements w^* (McCullagh and Nelder, 1989). Once the fitting has converged to give estimates $\hat{\beta}$ we can write the estimated coefficients of the logistic regression as

$$\hat{\beta}=(Z'WZ)^{-1}Z'W\{\text{logit}(\hat{p})+W^{-1}(I-\hat{p})\} \tag{8}$$

where $\hat{p}$ is the vector of predicted probabilities for each row of Z, i.e. the propensity score, and W is an $N\times N$ diagonal matrix with i element $w_i=\hat{p}_i(1-\hat{p}_i)$. Thus at convergence $W^{-1}(I-\hat{p})$ becomes 0, leading to a set of k equations:

$$[Z'_{\text{orig}}:Z'_{\text{syn}}]\begin{pmatrix} -\hat{p}_1 \\ 1-\hat{p}_2 \end{pmatrix}=0 \tag{9}$$

where $\hat{p}_1$ and $\hat{p}_2$ are vectors of length n_1 and n_2 of the propensity scores for the original and synthetic data respectively. Thus the first equation corresponding to the intercept gives the following expression for the mean of the propensity score:

$$\bar{\hat{p}}=n_2/(n_1+n_2)=n_2/N=c. \tag{10}$$

The assumption that $\hat{p} - c \ll I - c$, for every row of Z, leads to the following results. Since I takes the values 1 and 0 only, it follows that all elements of w can be approximated by $c(1-c)$. This approximation means that we can express the deviation of $\text{logit}(\hat{p})$ from its mean as $w^{-1}(\hat{p} - c)$ since the derivative of $\text{logit}(\hat{p})$ at its mean is w^{-1}. Thus from equation (8) we obtain

$$Z\hat{\beta} - \overline{Z\hat{\beta}} = Z(Z'Z)^{-1}Z'w^{-1}(\hat{p} - c) \tag{11}$$

which can become

$$Z\hat{\beta} - \overline{Z\hat{\beta}} = Z(Z'Z)^{-1}Z'w^{-1}((\hat{p} - I) + (I - \hat{p})), \tag{12}$$

because $W^{-1}(I - \hat{p})$ is 0 at convergence as we saw in equation (8), and then

$$Z\hat{\beta} - \overline{Z\hat{\beta}} = Z(Z'Z)^{-1}Z'w^{-1}(I - c). \tag{13}$$

We can write this in terms of its component matrices as

$$Z\hat{\beta} - \overline{Z\hat{\beta}} = \begin{pmatrix} Z_{\text{orig}} \\ Z_{\text{syn}} \end{pmatrix} (Z'_{\text{orig}}Z_{\text{orig}} + Z'_{\text{syn}}Z_{\text{syn}})^{-1}[Z_{\text{orig}} : Z_{\text{syn}}]w^{-1} \begin{pmatrix} -n_2/N \\ n_1/N \end{pmatrix} \tag{14}$$

where the final column vector consists of a unit vector with n_1 entries equal to $-n_2/N$ and n_2 entries equal to n_1/N. Using the approximation

$$\hat{p} - c = (Z\hat{\beta} - Z\bar{\hat{\beta}}) \frac{\mathrm{d}\hat{p}}{\mathrm{d}(Z\hat{\beta})}\bigg|_{\hat{p}=c} \tag{15}$$

we obtain $\hat{p} - c = (Z\hat{\beta} - Z\bar{\hat{\beta}})w$, since the derivative becomes $c(1-c) = w$. Applying this with equation (14) we obtain

$$\hat{p} - c = \begin{pmatrix} Z_{\text{orig}} \\ Z_{\text{syn}} \end{pmatrix} (Z'_{\text{orig}}Z_{\text{orig}} + Z'_{\text{syn}}Z_{\text{syn}})^{-1}(\bar{Z}_{\text{syn}} - \bar{Z}_{\text{orig}})n_1n_2/N \tag{16}$$

and the mean-squared error from the propensity score becomes

$$\text{pMSE} = (\hat{p} - c)'(\hat{p} - c)/N = (\bar{Z}'_{\text{syn}} - \bar{Z}'_{\text{orig}})(Z'_{\text{orig}}Z_{\text{orig}} + Z'_{\text{syn}}Z_{\text{syn}})^{-1}(\bar{Z}_{\text{syn}} - \bar{Z}_{\text{orig}}) \frac{(n_1n_2/N)^2}{N}. \tag{17}$$

The first element of $(\bar{Z}_{\text{syn}} - \bar{Z}_{\text{orig}})$, from the intercept term of the regression, becomes 0, and the expectation of the matrix

$$(Z'_{\text{orig}}Z_{\text{orig}} + Z'_{\text{syn}}Z_{\text{syn}})^{-1} \tag{18}$$

without its first row and column becomes

$$\{(Z'_{\text{orig}} - \bar{Z}'_{\text{orig}})(Z_{\text{orig}} - \bar{Z}_{\text{orig}}) + (Z'_{\text{syn}} - \bar{Z}'_{\text{syn}})(Z_{\text{syn}} - \bar{Z}_{\text{syn}})\}^{-1}, \tag{19}$$

because the independence of Z_{orig} and Z_{syn} ensures that the contribution of cross-product terms to the inverse will be zero. When the synthetic data are generated from the distribution that generated Z_{orig} the expected value of $\bar{Z}_{\text{syn}} - \bar{Z}_{\text{orig}}$ will converge to 0 for large samples and its variance to V/n_2 where V is the variance of Z_{orig}. Also, the expression

$$(Z'_{\text{orig}} - \bar{Z}'_{\text{orig}})(Z_{\text{orig}} - \bar{Z}_{\text{orig}}) + (Z'_{\text{syn}} - \bar{Z}'_{\text{syn}})(Z_{\text{syn}} - \bar{Z}_{\text{syn}}) \tag{20}$$

will converge to NV for large samples. Thus we can see that equation (17) is a multiple of a quadratic form in $\bar{Z}_{\text{orig}} - \bar{Z}_{\text{syn}}$ of dimension $k - 1$, so it is distributed as

$$\frac{(n_1/N)^2n_2/N}{N}\chi^2_{k-1}.$$

Thus the expected value and standard deviation of pMSE are

$$E[\text{pMSE}] = \frac{(1-c)^2c}{N}(k-1),$$

$$\text{StDev}(\text{pMSE}) = \frac{(1-c)^2c}{N}\sqrt{\{2(k-1)\}},$$

which become $(k-1)/(8N)$ and $\sqrt{\{2(k-1)\}}/(8N)$ when n_1 and n_2 are equal.

The assumption that $\hat{p} - c \ll I - c$ appears to be rather strong. The assumption is required for the null distribution when we would not expect the regression to provide much discrimination, so we would expect all the predicted values to be close to c in that case. We have found through simulation in some cases with very high order interactions and non-linear terms that it is possible to produce scores that are not close to c. In this case the assumption can be violated, but in practice we do not think that these would be advisable propensity score models.

A crucial assumption in this derivation is that the large sample expectation of each column of Z_{syn}, under repeated syntheses from the same original data, will be the mean of the corresponding column of Z_{orig}. This follows trivially, and without the asymptotic assumption, for the columns of Z_{orig} and Z_{syn} that correspond to Y_{orig}. For other columns we note that the expectation of any function of the variables in a distribution can be written as a function of its parameters θ and that any function of consistent estimators is a consistent estimator of the corresponding function of θ. Thus, for large samples, the means of the columns of Z_{orig} will be functions of $\hat{\theta}$. Since the columns of Z_{syn} are combinations of variables generated from $f(y|\hat{\theta})$, their expectation will be given by the same function of $\hat{\theta}$ that defines the mean of the corresponding column of Z_{orig}.

A.2. Distribution of pMSE when some variables are left unchanged

The derivations above require that all elements of each variable are replaced by synthetic values. When synthesis is incomplete because only some variables are synthesized, whereas others remain as in the original data, a variant of this result can be used as follows. Some of the predictors Z that are used in the logistic regression will only use the unsynthesized variables. We can denote this subset by Z^{fix} and the remaining variables by Z^* which will be assumed to have k^* columns, including the intercept term. The values of $(\bar{Z}^{\text{fix}}_{\text{syn}} - \bar{Z}^{\text{fix}}_{\text{orig}})$ will all be identically 0 because their synthesized values are identical to the original values. Thus pMSE can be written in terms of a quadratic form from $(\bar{Z}^*_{\text{syn}} - \bar{Z}^*_{\text{orig}})$. Thus, by arguments paralleling those above, pMSE for this type of incompletely synthesized data will have the distribution given above with k^* replacing k, where k^* is the dimension of the predictor matrix involving only synthesized variables (including interactions between synthesized and unsynthesized variables).

A.3. Distribution of pMSE calculated from two synthetic data sets from the same original data

When pMSE is calculated from two synthetic data sets, each synthesized from the same original data, the expression $(\bar{Z}_{\text{syn}} - \bar{Z}_{\text{orig}})$ is replaced by $(\bar{Z}_{\text{syn1}} - \bar{Z}_{\text{syn2}})$ which can be written as $((\bar{Z}_{\text{syn1}} - \bar{Z}_{\text{orig}}) - (\bar{Z}_{\text{syn2}} - \bar{Z}_{\text{orig}}))$. Conditionally on Z_{orig}, these two terms are independent and each has variance–covariance matrix V. Thus pMSE calculated from two synthesized data sets, each of size n_2, will have expected value and standard deviation $2(k-1)(1-c)^2c/N$, which is twice that of the null expected pMSE for a sample of size n_2. This result was confirmed by simulations from logistic models and applies when the propensity score models use the same Z-matrix for each resynthesis.

When CART methods are used for the propensity score the argument above does not hold because the null pMSE will include a component for model selection that will be the same for a comparison of pairs as for the comparison of synthesized data with the original. This implies that the mean between-pair estimates of pMSE for CART models will have expectation between the expected null pMSE and twice the expected null pMSE. This was confirmed by simulations. For data sets where the propensity score models can generate a large number of potential trees the expectation of pMSE from pairs will be close to the null expectation, as was found in Appendix A.1. When the original data consist of only a few categorical variables, with a small choice of possible trees, the estimate from pairs will be close to twice the expected null pMSE. In practice most data sets will be sufficiently complex to allow the assumption that the estimate from pairs gives the null expectation. These results apply to the pMSE calculated either from pairs or from permutations, since in both cases the data sets being compared will have the same distribution.

Appendix B: Simulations of null pMSE with classification and regression tree propensity score models

Here we show results for simulations using the same set-up as described in Section 4.3, but we switch the logistic propensity score models for non-parametric CART models. In the first case all the data are synthesized, and in the second case only two of the 10 variables are synthesized, as before. First we estimate

Table 14. Results from 1000 simulated complete syntheses of multivariate normal data, using correct and incorrect models with the pMSE calculated from non-parametric CART models†

Population covariance	*Results for correct synthesis pMSE*			*Results for incorrect synthesis pMSE*		
	Mean	*Ratio*	*Standardized score*	*Mean*	*Ratio*	*Standardized score*
0.0	0.02134	0.96584	−0.20497	0.02194	0.99360	−0.04966
0.1	0.02162	0.98190	−0.11288	0.02812	1.27088	1.45820
0.2	0.02134	0.97505	−0.14870	0.03800	1.71642	3.88034
0.3	0.02067	0.95014	−0.27136	0.05137	2.32482	7.11181
0.4	0.02055	0.95326	−0.25137	0.06837	3.08777	11.31911
0.5	0.02051	0.95946	−0.21945	0.08866	4.01128	16.16698
0.6	0.01991	0.94439	−0.28992	0.11269	5.10356	22.04270
0.7	0.01944	0.93813	−0.30129	0.14155	6.40335	29.09392
0.8	0.01893	0.93915	−0.28086	0.17431	7.89067	37.06338
0.9	0.01748	0.92756	−0.26846	0.20365	9.23024	43.92087

†Ratios and standardized scores calculated from the null distribution estimated from 45 pairs formed from 10 multiple syntheses of each simulated data set. The CART method was carried out with the `rpart` function from the `rpart` package for R, with the complexity parameter cp 1×10^{-3}.

Table 15. Results from 1000 simulated syntheses of multivariate normal data with only two of the 10 columns, synthesized by using correct and incorrect models with pMSE calculated from non-parametric CART models†

Population covariance	*Results for correct synthesis pMSE*			*Results for incorrect synthesis pMSE*		
	Mean	*Ratio*	*Standardized score*	*Mean*	*Ratio*	*Standardized score*
0.0	0.016106	0.991741	−0.028395	0.016259	0.987100	0.004145
0.1	0.015956	0.996262	−0.012367	0.019528	1.207032	0.738865
0.2	0.015331	0.993628	−0.019922	0.024894	1.534692	1.912159
0.3	0.015085	1.012724	0.038179	0.031738	1.967164	3.618090
0.4	0.013968	0.998023	−0.005400	0.040502	2.559000	5.430248
0.5	0.012879	0.989164	−0.028306	0.051799	3.290095	7.915648
0.6	0.012119	1.004021	0.009555	0.066470	4.217474	11.584360
0.7	0.010523	1.002788	0.005776	0.085163	5.488950	16.148351
0.8	0.009099	0.989744	−0.019050	0.110642	7.366080	20.769889
0.9	0.007091	1.054471	0.087836	0.148356	10.417990	29.718231

†Ratios and standardized scores calculated from the null distribution estimated from 45 pairs formed from 10 multiple syntheses of each simulated data set.

the null as described in Section 4.3.1 by resampling and comparing pairs of synthetic data sets which we know were drawn from the same generative model.

Table 14 shows the simulation results for the CART propensity score models. Note that the pMSEs that are calculated for CART models are much larger than those from the logistic models, with the estimated null indicating that differences have been introduced via model selection resulting in what we might term an overfitting component of variation. We can see that in this case, as with the simulations using a parametric model, the ratio and standardized pMSE-values stay constant for the correct synthesis across different correlations. As can be seen from the first column, the expected value under different correct synthesis changes, since the number of parameters in the propensity score model is no longer fixed. As expected

from Appendix A.3 pMSE from the simulated data is slightly lower than that between synthesis for pairs, giving ratios just below 1. This ratio is sufficiently close to 1.0 to allow standardization by the between-pair pMSE in practical situations. Another important difference from using the logistic propensity score model is that the ratios for the wrong synthesis increase on a slower scale, which is due to an increase in the pMSE-scores both for the observed and for the null due to the size of the trees. All of this amounts to a trade-off in return for the greater flexibility to assess a wider range of departures from the synthesis model with CART models.

Relating to what was discussed in Section 4.3.1, it is important not to use overly large trees, since it will make it more difficult to discern between worse syntheses. As the number of splits in the tree approaches the number of observations, the expected pMSE-ratio will be limited to 2, no matter how bad the synthesis. This is because the maximum value for pMSE is 0.25, and the maximum expectation under the null is 0.125 ($\frac{1}{8}$) when the original and synthetic data sets have the same number of rows. When using CART propensity score models, it is important to monitor the size of the models being produced. The complexity parameters that control the fitting must allow models to be sufficiently large to provide splits in the null case, from pairs or permutations. However, the models must not be so large, in this null case, that there is little room for improvement and hence low power to detect problems with the synthesis models.

Next, we replicate the incomplete synthesis simulation from Section 4.3 by using the CART propensity score model and null approximation. The results shown in Table 15 are once again what we expect, exhibiting the same patterns as the complete-data simulation, and showing that this resampling method can be used to estimate the null pMSE for CART models when some variables are not synthesized.

We also carried out corresponding simulations using the permutation method (the data are not shown) and obtained results for complete data which were in agreement with those in Table 4 to within the uncertainty of the simulations. As anticipated in Section 4.3.1, for incomplete data like Table 15 the permutation method overestimated the null pMSE.

References

Benedetto, G., Stinson, M. H. and Abowd, J. M. (2013) The creation and use of the SIPP Synthetic Beta. US Census Bureau, Washington DC. (Available from `http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf`.)

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Drechsler, J. (2011) *Synthetic Data Sets for Statistical Disclosure Control*. New York: Springer Science and Business Media.

Drechsler, J. and Reiter, J. P. (2009) Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB Establishment Survey. *J. Off. Statist.*, **25**, 589–603.

Drechsler, J. and Reiter, J. P. (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computnl Statist. Data Anal.*, **55**, 3232–3243.

Elliot, M. (2015) Final report on the disclosure risk associated with the synthetic data produced by the sylls team. *Report 2015-2*. Cathie Marsh Institute for Social Research, University of Manchester, Manchester.

Fienberg, S. E. and Slavković, A. B. (2011) Data privacy and confidentiality. In *International Encyclopedia of Statistical Science* (ed. M. Lovric), pp. 342–345. New York: Springer.

Hu, J., Reiter, J. P. and Wang, Q. (2014) Disclosure risk evaluation for fully synthetic data. In *Privacy in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 185–199. New York: Springer.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and De Wolf, P.-P. (2012) *Statistical Disclosure Control*. New York: Wiley.

Karr, A., Kohnen, C. N., Organian, A., Reiter, J. P. and Sanil, A. P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *Am. Statistn*, **60**, 224–232.

Karr, A., Oganian, A., Reiter, J. and Woo, M.-J. (2006) New measures of data utility. *Wrkshp Data Confidentiality, a Working Group in National Defense and Homeland Security*. (Available from `http://sisla06.samsi.info/ndhs/dc/Papers/NewDataUtility-01-10-06.pdf`.)

Kinney, S. K. and Reiter, J. P. (2010) Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. *J. Off. Statist.*, **26**, 301–315.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S. and Abowd, J. M. (2011) Towards unrestricted public use business microdata: the Synthetic Longitudinal Business Database. *Int. Statist. Rev.*, **79**, 362–384.

Loong, B., Zaslavsky, A. M., He, Y. and Harrington, D. P. (2013) Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Statist. Med.*, **32**, 4139–4169.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008) Privacy: theory meets practice on the map. In *Proc. 24th Int. Conf. Data Engineering*, pp. 277–286. New York: Institute of Electrical and Electronics Engineers Computer Society.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013) A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statist. Med.*, **32**, 3388–3414.
McClure, D. and Reiter, J. P. (2016) Assessing disclosure risks for synthetic data with arbitrary intruder knowledge. *Statist. J. Int. Ass. Off. Statist.*, **32**, 109–126.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Miranda, J. and Vilhuber, L. (2016) Using partially synthetic microdata to protect sensitive cells in business statistics. *Statist. J. Int. Ass. Off. Statist.*, **32**, 69–80.
Nowok, B. (2015) Utility of synthetic microdata generated using tree-based methods. *Wrkshp Statistical Data Confidentiality, Oct 7th*. (Available from `http://www1.unece.org/stat/platform/display/SDCWS15/Statistical+Data+Confidentiality+Work+Session+Oct+2015+Home`.)
Nowok, B., Raab, G. M. and Dibben, C. (2016) synthpop: bespoke creation of synthetic data in R. *J. Statist. Softwr.*, **74**, 1–26.
Nowok, B., Raab, G. M. and Dibben, C. (2017) Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statist. J. Int. Ass. Off. Statist.*, **33**, 785–796.
Raab, G. and Nowok, B. (2017) Inference from synthetic data: package vignette for the synthpop package. (Available from `https://cran.r-project.org/web/packages/synthpop/vignettes/inference.pdf`.) University of Edinburgh, Edinburgh.
Raab, G. M., Nowok, B. and Dibben, C. (2017a) Practical data synthesis for large samples. *J. Priv. Confident.*, **7**, 67–97.
Raab, G. M., Nowok, B. and Dibben, C. (2017b) Guidelines for producing useful synthetic data. *Preprint*. University of Edinburgh, Edinburgh. (Available from `https://arxiv.org/abs/1712.04078`.)
Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003) Multiple imputation for statistical disclosure limitation. *J. Off. Statist.*, **19**, 1–17.
R Core Team (2017) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Reiter, J. P. (2003) Inference for partially synthetic, public use microdata sets. *Surv. Methodol.*, **29**, 181–188.
Reiter, J. P. (2005a) Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J. R. Statist. Soc.* A, **168**, 185–205.
Reiter, J. P. (2005b) Using CART to generate partially synthetic, public use microdata. *J. Off. Statist.*, **21**, 441–462.
Reiter, J., Oganian, A. and Karr, A. (2009) Verification servers: enabling analysts to assess the quality of inferences from public use data. *Computnl Statist. Data Anal.*, **53**, 1475–1482.
Reiter, J. P., Wang, Q. and Zhang, B. E. (2014) Bayesian estimation of disclosure risk for multiply imputed, synthetic data. *J. Priv. Confident.*, **6**, 17–33.
Slavković, A. and Lee, J. (2010) Synthetic two-way contingency tables that preserve conditional frequencies. *Statist. Methodol.*, **7**, 225–239.
Therneau, T., Atkinson, B. and Ripley, B. (2015) rpart: recursive partitioning and regression trees. *R Package Version 4.1–10*.
US Census Bureau (2006) DRB memo on disclosure testing the SIPP Synthetic Beta. *Technical Report*. US Census Bureau, Washington DC.
Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control in Practice*, 2nd edn. New York: Springer.
Wilson, M., Kellock, C., Adams, D. and Landsberg, J. (2015) *The Scottish Health Survey Topic Report: Mental Health and Wellbeing; 2015*. Edinburgh: Scottish Government.
Woo, M.-J., Reiter, J. P., Oganian, A. and Karr, A. F. (2009) Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confident.*, **1**, 111–124.
Woo, Y. M. J. and Slavkovic, A. (2015) Generalised linear models with variables subject to post randomization method. *Statist. Appl.*, **24**, 29–56.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material histograms: I-CeM data example'.