UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Assessing, visualizing and improving the utility of synthetic data.

Gillian Raab (Scottish Centre for Administrative Data Research)
gillian.raab@ed.ac.uk

Abstract

The open-source synthpop package for R (www.synthpop.org.uk) provides tools to allow custodians of confidential microdata to create synthetic data based on the original. The synthesis can be customised to ensure that relations evident in the real data are reproduced in the synthetic data. A number of measures have been proposed to assess this aspect, commonly known as the utility of the synthetic data. These include measures based on distances and others based on predictive scores. The methods will be reviewed and compared, and some surprising relations between them illustrated. These measures are incorporated into an easy-to-use utility module in the synthpop package that incorporates methods to visualise the results and thus provide immediate feedback for the person creating the synthetic data. The utility functions can be used to assess synthetic data created by methods other than synthpop.

Assessing, visualizing and improving the utility of synthetic data.

Gillian M Raab*, Beata Nowok*, Chris Dibben*

* Scottish Centre for Administrative Data Research, School of Geosciences,
University of Edinburgh, Edinburgh, EH8 9YL, Scotland, UK.
E-mail: {gillian.raab, beata.nowok,chris.dibben}@ed.ac.uk
URL: https://www.synthpop.org.uk/

Abstract. The synthpop package for R (www.synthpop.org.uk) provides tools to allow data custodians to create synthetic versions of confidential microdata that can be distributed with fewer restrictions than the original. The synthesis can be customized to ensure that relationships evident in the real data are reproduced in the synthetic data. A number of measures have been proposed to assess this aspect, commonly known as the utility of the synthetic data. These include measures based on distances between the two distributions and others based on predictive scores. We show that all these measures can be derived from a propensity score model. The methods will be reviewed and compared, and relations between them illustrated. These measures are incorporated into an easy-to-use utility module in the `synthpop` package that incorporates methods to visualize the results and thus provide immediate feedback to allow the person creating the synthetic data to improve its quality. The utility functions were originally designed to be used for synthetic data objects of class `synds`, created by `synthpop`, but they can now be used to compare synthetic data created by other methods with the original records.

1 Overview

The utility of synthetic data will ultimately be measured by how results from analyses of synthetic data, and the conclusions following from them, will differ from those derived from the real data. Comparisons of specific analyses for the original and synthetic data are often termed ”narrow utility measures”. It is not advisable to tune synthesis methods to make the results of a specific analysis agree with those from the original. Details of the final analyses are seldom known and, even if they were, creating the synthesis to give agreement for an analysis model will give answers that will agree, but the residuals from the model fitted to the synthetic data will not give any evidence of model inadequacy that might have been found with the original. There is a need for measures that compare wider aspects of the differences between the synthetic and original data to give feed backon the utility of the synthesis when it is being created. Such measures are termed ”broad”, ”global” or ”general” utility measures, as opposed to ”narrow” or ”specific” measures that focus on the results of particular analyses.

Here we present details of the general utility measures that can be calculated by functions in the R package `synthpop` and show how they can be used in practice. There are two main reasons we might wish to evaluate the utility of synthetic data:

1. To compare different synthesis methods for the same data set
2. To diagnose where the original and synthetic data distributions differ and thus tune the synthesis methods to improve the utility of the synthetic data.

For both of these reasons we recommend the propensity mean squared error ($pMSE$) as a utility measure. The default printed output from all the utility functions therefore presents only the $pMSE$ measure and its standardized ratio (S_pMSE). All the other utility measures discussed here are available as outputs from the functions. The $pMSE$ is calculated from propensity score that predicts whether the synthetic data can be distinguished from the original. Different methods can be used for this prediction. For the first reason we advise fitting the propensity score model by a classification and regression tree (CART) model. We illustrate this here using the synthesis of a data set consisting of 5,000 records for 10 variables, selected from SD2011, the survey data that are part of `synthpop`. [1] We compare a synthesis that uses a parametric model[2] while the second synthesizes from a CART model. Both the utility evaluations used a propensity score model fitted by CART and the results, below, show that the utilty score is 3 times higher (worse) for the parametric synthesis compared to the CART synthesis.

```
Parametric pMSE  0.03826426
CART       pMSE  0.008475425
Utility ratio parametric to cart  4.51473
```

[1] The variables consist of 6 categorical variables (sex, edu, socprof, trust, smoke and region) and 4 numeric variables (age, income, weight amnd height).

[2] The method used for numeric variables uses a transformation to the expected Normal ranks so as to preserve the univariate distriutions for skew variables.

For the second reason, diagnosing problems and tuning the synthesis, we recommend visualizing the relationships between subsets of the variables ,e.g. all oneway, twoway or threeway combinations. As an example Figure 1 visualizes the utility for all twoway tables from the parametric synthesis, with five groups created from the non-missing values of continuous variables. It is clear that the variable with the most problems is weight. Section 4 will illustrate further syntheses of these data.

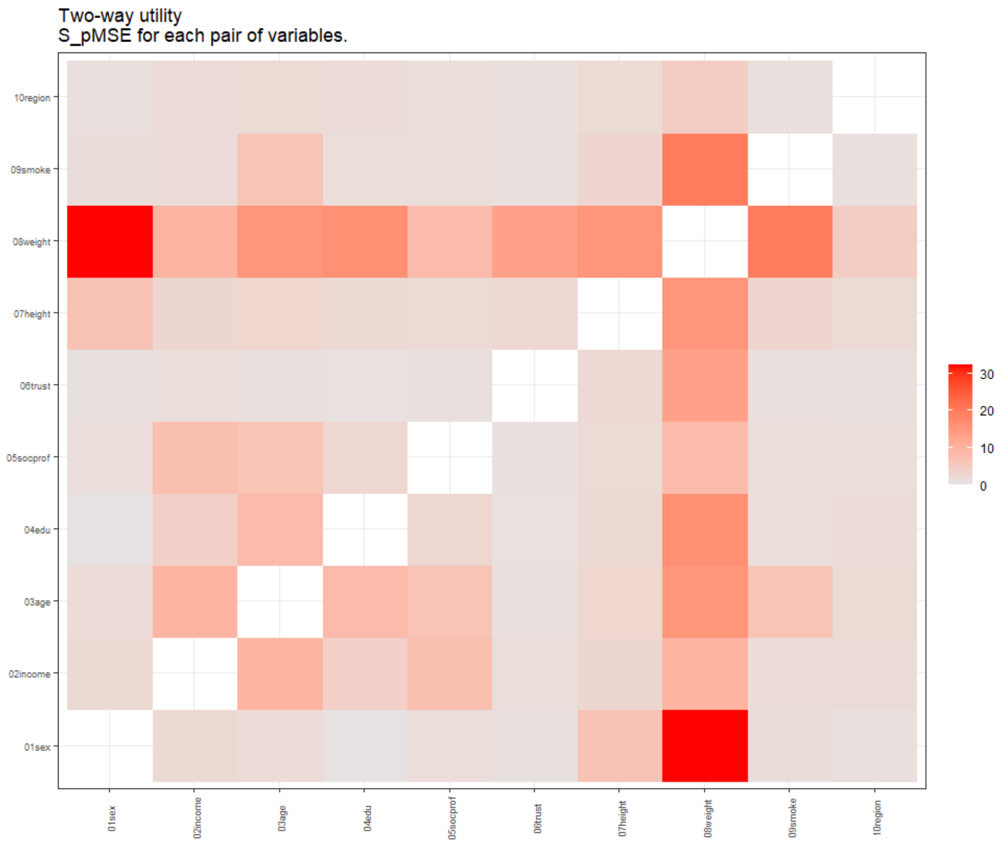

Figure 1: Plot produced from synthesis of `ods` by parametric synthesis.

These recommendations are based on our practical experiences and on empirical evaluations that are detailed in the rest of this paper. In section 2 we present details of all the utility measures, their performance in evaluating syntheses and the relationships between the measures. We show that two sets of seemingly unrelated utility measures (one pair and one set of three) are identical to each other. Section 3 evaluates models for computing the propensity score. Section 4 provides examples of using the utility functions to diagnose problems and tune the synthesis methods to improve utility. The final section summarizes the paper and makes suggestions as to possible future enhancements.

2 Choice of utility measures

One approach to general utility measures involves combining the original and synthetic records and measuring how well the data values can predict the source of the records as real or synthetic (Karr et al. 2006). This method uses the propensity score, $\hat{p}$, the predicted probability that a record comes from the synthetic data. If the synthesis has been carried out from a model that is compatible with the original data distribution then the expected mean of $\hat{p}$ will be $c = n_2/N$, where there are n_1 records from the original data and n_2 from the synthetic data and $N = n_1 + n_2$. We refer to the distributions of utility measures in this case as their Null distributions. The most commonly suggested utility measure is known as the propensity score mean square error ($pMSE$). The Null distribution of the $pMSE$, for prediction models with a fixed number of parameters has been derived by Snoke etal., 2018 and its expectation is $dfc(1-c)^2/N$, where df is the number of degrees of freedom constrained by fitting the propensity score model. Other utility measures can also be derived from the propensity score, e.g. the percentage above 50% of records correctly predicted (PO50) and the Kolmogorov-Smirnov statistic (KS) which is the maximum distance between the cumulative distributions functions (CDFs) of the propensity score for the synthetic and original distributions, Bowen et al., 2021. Further measures that compare $\hat{p}$ values between the original and synthetic data could be considered. One such, the Wilcoxon signed-rank statistic (U).

An alternative approach to utility measures is to group the original and synthetic data, usually by constructing tables based on their values, and to compute measures of difference between the tables. Voas and Williamson, 2001, investigated measures based on the family of goodness-of-fit measures discussed by Read and Cressie, 1988. They note that the usual Pearson χ^2 statistic needs to be adjusted because synthetic data may be generated in cells where the count from the original data is zero. They propose replacing the denominator in the formula with the average of the original and synthetic counts. This statistic and its generalization when $n_1 \neq n_2$ are designated as VW. Other goodness-of-fit measures that can be calculated from tables include the the Freeman-Tukey statistic (FT)[3], the Jensen-Shannon divergence (JSD) and the likelihood ratio χ^2 statistic (G). The likelihood ratio has no contributions from

[3] This measure is proportional to the discrete Hellinger distance between two distributions

cells where the original counts are zero. It would be desirable for these cells to contribute to utility measures since they may be a substantial proportion of all cells, especially for sparse tables. Another possible measure derived from tables is the sum of the absolute differences between the proportions of original and synthetic counts, designated as $MabsD$[4]. A related quantity is $WMabsD$, where the absolute differences are weighted in proportion to the inverse of the standard deviation of their Null expectations, so that this measure has a known Null expectation.

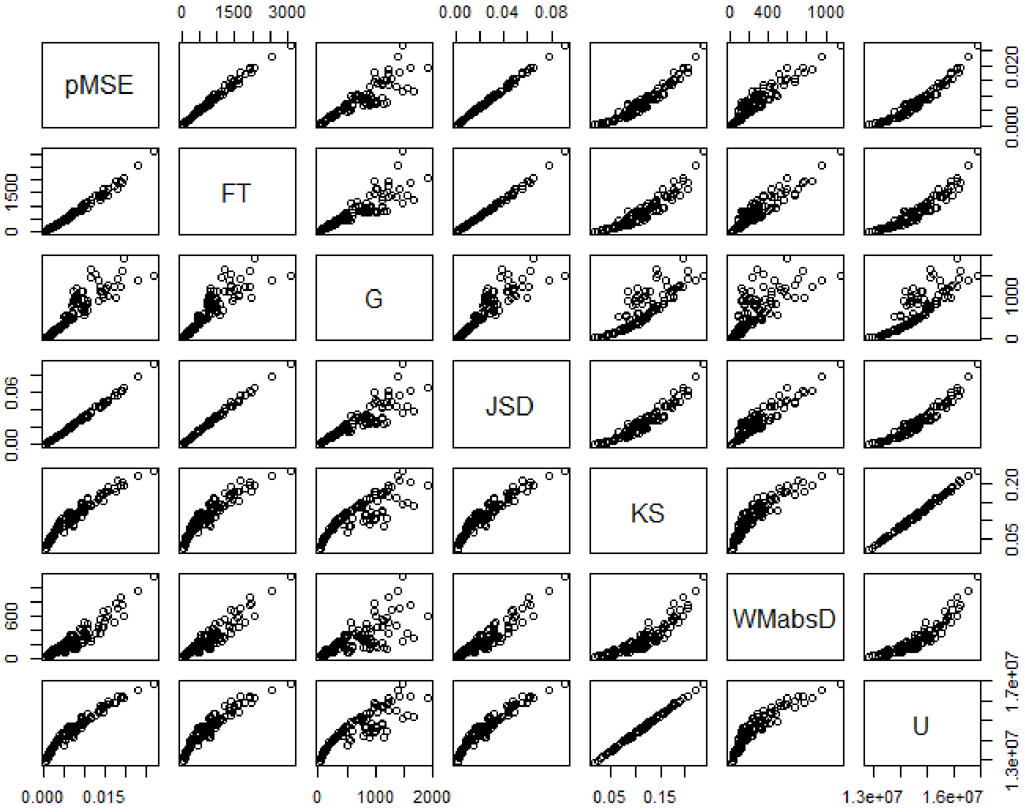

Figure 2: Pairs plot of utility measures for all 84 threeway tables from the synthesis of `ods` by parametric synthesis.

Corresponding to the two approaches, `synthpop` provides two functions to calculate utility measures, `utility.gen` and `utility.tab`. Comparing tables of original

[4] Suggested by Christine Task,as used to evaluate the NIST challenges, see here

and synthetic data can be framed as a prediction model, with the propensity score for records in each cell is the ratio of the synthetic counts to the sum of the original and synthetic counts. For synthetic data where all variables are categorical, a comparison of $n$way tables is equivalent to fitting a propensity score model by logistic regression including all interactions up to order n. Thus any measure defined from the propensity score can also be computed for tables, but some tabular utility measures do not correspond to measures from the propensity score approach.

We have shown [5] that there are fewer utility measures than the paragraph above would suggest. The measures $pMSE$ and VW are linearly related, as are the three measures KS, $PO50$ and $MabsD$. When all utility measures are compared [6] for different syntheses some are very highly correlated; see e.g. Figure 2. The three measures $pMSE$, FT and JSD are one correlated set and the pair KS and U is another.

Number of variables	mean pMSE	mean FT	mean G	mean JSD	mean KS	mean U	mean WMabsD	median df*
2	40.3	40.5	40.9	40.4	17.1	16.3	16.4	13
3	124.0	131.8	151.8	129.4	40.0	43.6	38.9	69
4	142.5	142.5	169.7	147.0	61.5	64.5	57.8	292
5	255.6	281.9	62.5	274.8	125.0	113.3	164.4	2178
6	235.0	263.1	26.6	253.2	127.1	112.9	183.0	3555

*Effective degrees of freedom is one minus the number of cells in the cross-tabulation of all variables that contain any original or synthetic counts.

Table 1: Power of different utility measures to compare "incorrect" with "correct" syntheses from tables cross-tabulating different numbers of variables.

Table 1 summarises a simulation of the power of all the statistics that can be calculated by `utility.tab`. As expected $pMSE$, FT and JSD all have similar power. The likelihood ratio statistic, G, has similar power for tables with large expectated counts but loses power for large, sparse tables. The other three measures have lower power for these examples. One desirable feature of utility measures are that they should be available from propensity score methods not based on tables and another is that they should have a known Null expectation to allow standardised from a single synthetic data set. Only $pMSE$ has both of these properties. Multiple synthetic data sets are needed to obtain a standardized measure of KS from a replication

[5] for details a longer version of this paper can be accessed at https://arxiv.org/pdf/2109.12717.pdf

[6] Using only one from each colinear set

method. The measure U has the poorest power for this example, while the power of G deteriorates for large sparse tables.

3 Models for the propensity score: practical considerations.

As well as chosing a utility measure, the synthesizer must decide on which model is to be used to calculate the propensity score. The two possible classes of model are logistic regression and adaptive classification models such as CART. Within each class, a variety of models can be specified by defining predictors for logistic models and by altering the methods and settings of classification models. The three models now available in the utility modules of `synthpop` are given in Table 2. All can be computed from `utility.gen` but only one also from `utility.tab`. The choices between these models are largely based on practical considerations, as we discuss below.

Model	Description	`utility.gen`	`utility.tab`
(a). Saturated logistic	Logistic regression with all interactions up to the number of variables in data.	x	x
(b) Logistic to order n	Logistic regression with all interactions up to order n	x	
(c) CART models	Classification and regression trees	x	

Table 2: Propensity score models implemented in `synthpop`

Models of type (a) calculated from `utility.tab` are limited by the memory required to hold large tables and by the fact that large tables can become sparse so that their statistical properties may be uncertain. The six variables contributing to the evaluation of the Null models used to calculate resaults for Table 1 defined a table with 14 thousand cells, though only 3,500[7] of the cells contain any counts from either the original or synthetic data. A table of all 10 variables in the data set `ods` would contain over 14 million cells although only 0.04% of them would contain any counts. Memory problems would prevent this method from being used for 7 or more variables from this data set, and the sparsity of the tables would advise against using tables of more than 5 variables. To try to fit model (a) via logistic regression does

[7] Median from 10,000 syntheses

not help either because it is constrained by its large number of parameters. For the first 5 variables from the `ods` data set, including all possible interactions requires a model with 3,500 parameters that failed to converge in several hours of computing time. Thus method (a) can only be used for a few variables at a time.

Other logistic models for data sets with many variables may also be limited by the large number of parameters required to fit the propensity score model. Using method (b) with the default setting of all second-order interactions for the 10 variables in `ods` gives a model with 753 parameters. This model fitted in under 2 minutes [8]. A model that included 3 level interactions of all variables would have defined a model with over 7K parameters. It would only be possible to fit higher-order models for data sets with a small number of variables. Even models with just second order interactions may have problems with large and complex data sets, especially if they contain factors with many levels. The choice of model to fit the propensity score for a data set with many variables is between logistic regression (b), with restrictons on interactions, and a CART model (c). A CART model (c) requires the use of resampling methods if a standardized measure is required, but only a single synthetic data set is required and results seem satisfactory; see Appendix **??**. We have found that CART models can diagnose differences more easily with fewer computational problems than logistic models. To get a single summary measure to compare syntheses we recommend using a CART propensity score model, with $pMSE$ standardised by a permutation method; see the example in Section 1.

The choice of model for diagnosing and fixing problems with a synthesis is different. For this we need a method that will pinpoint the parts of the distribution of synthetic data that differ from that of the original. It is possible to examine the trees that have been used to calculate the propensity score from CART models but it is difficult to decide from such output which of the variables contribute most to the differences between the synthetic and original data. Parametric models for the propensity score can provide more information by highlighting the coefficients of the propensity score model with the largest standardized coefficients. However, for over-parameterized models, certain coefficients may be aliased, because they are too strongly related to others, so they have values set to missing. This is exactly what happened to the coefficient of the indicator that weight was missing, which turned out to be the problem with the synthesis illustrated in Figure 1. Only by restricting the utility evaluation to a few variables , could the problem be diagnosed from coefficients, as we see from the output below.

```
OUTPUT
```

[8] On a Windows laptop with spec to add

```
Utility score calculated by method:  logit

Call:
utility.gen(object = syn_para, data = ods, method = "logit",
maxorder = 1, stats = "pMSE", vars = c("weight",
"sex", "age"), print.zscores = TRUE)

Utility score results
pMSE 0.004101  Ratio to NULL expected, S_pMSE 36.45 degees of freedom 9

z-scores (or mean z-scores if m > 1) greater than the threshold of +/- 1.6
weight_NA age:weight_NA
6.477145     -4.969725
```

This shows that it was coefficients involving the missing value codes for weight that were different for the original and synthetic data, as we will see below. We had to know what we were looking for before finding the coefficients, so this is clearly unsatisfactory as a method of identifying where the synthetic data differs from the original.

An approach that is much more practical and useful is to examine the agreement between the synthetic and original data for low-order margins; starting with oneway marginals, then twoway and perhaps threeway. This approach is illustrated in the next section.

4 Using utility measures to tune the synthesis methods.

The `synthpop` allows the syntheses to be tuned in various ways to adapt to the needs of particular data sets. These include:

- Changing the order in which the conditional distributions are formed
- Stratifying the synthesis by important variables
- Changing the methods for individual variables

- Modifying the predictor matrix to exclude certain variables as predictors of others

The first two are the ones we have found most useful. We have found the need for the third and fourth only in special circumstances, one of which we will describe here. Some survey or administrative data contain very detailed fields that can be grouped into wider classes. Examples are the classification of occupations or diagnostic codes. The detailed variables are nested within the wider one. The detailed variables have too many classes to be used as predictors. To overcome this they need to be synthesised after the wider class and are given the method `nested`. This creates synthetic data for the detailed variable by taking bootstrap samples within the groups. The prediction matrix needs to be modified to remove the detailed variable as a predictor of other variables. Details of this and other possible strategies to improve syntheses are discussed in Raab et al., 2017b.

Use of the first two methods is illustrated in the example below. In all the examples we have specified the default utility value, the standardised $pMSE$ ratio, calculated from its expectation for logit models and by a resampling method for CART models. The target value for this utility model is 1.0, but we do not believe that real world data is ever generated exactly from a model. Thus we do not calculate any significance tests. We have found that a useful rule for practical use is to aim for utility ratios below 10.

The first step of evaluating the utility of any synthesis is to compare the univariate distributions for each variable. This is done using the `compare.synds` that produces histograms for each variable and since Version 1.7 of `synthpop` it function also procuces a table of utility measures for each oneway table, illustrated below for the parametric synthesis of the `ods` data set introduced in Section 1.

```
Utility results for each variable, means  if m > 1.
          pMSE Ratio deg fr
sex         0.5490      1
income      1.0022      6
age         1.0550      4
edu         0.3553      4
socprof     0.4053      9
trust       0.8958      3
height      2.9677      5
weight     52.6684      5
smoke       0.5397      2
```

```
region      1.3915     15

 Comparing counts observed with synthetic for weight

         35 40  45  50  55  60  65  70  75  80  85  90  95 100 105 110 115 120
observed  7 26 165 295 631 535 672 500 680 418 414 188 242  70  46  21  20   4
synthetic 6 15 142 268 586 497 677 475 657 421 428 176 243  71  43  17  18   6
    125 130 135 140 145 miss.NA
observed    7   2   1   2   1      53
synthetic   6   0   1   1   0     246
```

It is immediately clear that it is the missing values of the weight variable that are causing the problem. This can happen with parametric models that are synthesized towards the end of the list of conditional distributions. This is easily fixed by moving the variable weight up towards the start of the visit sequence, giving the following table.

```
 Utility results for each variable, means  if m > 1.
           pMSE Ratio deg fr
 sex         0.5490      1
 income      0.4294      6
 age         0.8788      4
 edu         3.7320      4
 socprof     2.2899      9
 trust       2.0142      3
 height      3.5429      5
 weight      0.3366      5
 smoke       0.0760      2
 region      1.4365     15
```

With this new order it is now time to investigate the twoway relationships between variables for the reordered synthesis. Figure c (a) to (d) shows the plots from the default twoway plots from `utility.tables` from four different syntheses. Note that these plots are all scaled to approximately the same legend as was generated by the range of utilities in the first synthesis: Figure 1 and reproduced as Figure 3(a). Figure 3 (b) shows the twoway plots from the reordered synthesis, clearly much better, although with some high values, notably those for interactions with 'age' where there are some utility values above 10.0. Stratifying the synthesis by dividing into two strata, above and below age 55, brings the maximum utility ratio down to below 7.0 (Figure 3(c)), but note that had we used CART synthesis with the original ordering (Figure 3(d)) the maximum utility ratio would have been below 3.

In this example our preferred CART models did not require any improvemen. Large complex data sets, even synthesized by CART, often require strategies mentioned above to improve their utility. Stratifying the synthesis by variables of interest to the researcher is a good strategy to ensure relationships will be maintained in the synthetic data.

The function `utility.tables` can also calculate utility measues for all threeway tables. Plots like Figure 1 are produced for three way tables holding one of the variables fixed. This third variable can be specified by the user. If this is not done then the program selects the variable with the highest utility score over all tables it contributes to.

5 Conclusion

This paper started life as a simple 'how-to-do-it' explanation of the routines we have written to measure data utility. Documenting them all in detail has led to some unexpected insights into the utility measures. What is more, it provides a firm foundation for our rules as to how to proceed to assess the utility of synthetic data and improve its quality. Briefly:

1. To compare the overall utility of two methods of synthesizing the same original data, you should fit a propensity score model with an adaptive model such as CART and compare the $pMSE$ measures, calculated by `utiity.gen`, for the two methods.

2. To judge and improve the utility of a synthesis method:

   - start by visualising all the oneway tables with `compare.syn` and check the S_pMSE values in the table..
   - Next visualise all twoway ratios with `utility.tables`.
   - If all the S_pMSE ratios are below 10, or better still below 3, it is probably not necessary to do anything more as the utility seems acceptable.

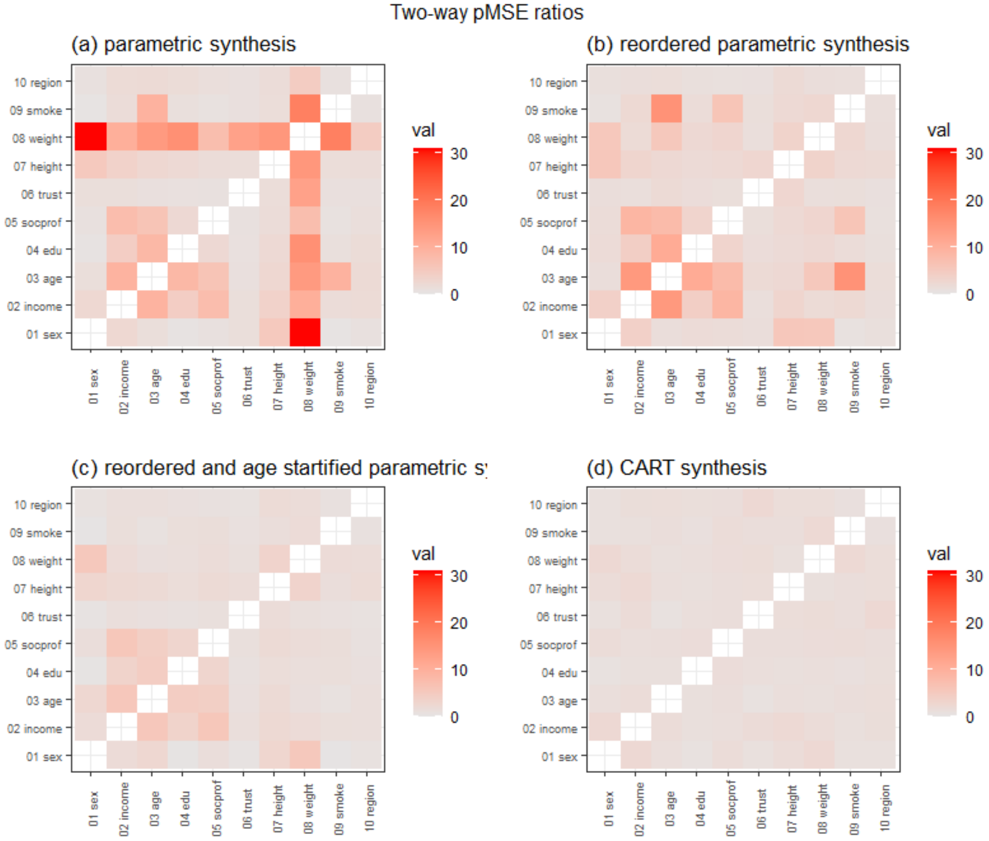

Figure 3: Plots produced from twoway utility measures for synthesis of `ods` by parametric synthesis.

- At each of the steps above you should try to improve the utility by tuning the synthesis with stratification and/or by changing the default parameters of `syn`

These recommendations have been exemplified on just one example, but we have found similar results from other data sets. We hope that other `synthpop` users can try out these functions on their own data and provide feedback on ways we might improve the utility functions and the ways they can be used.

Other measures of differences between the original and synthetic data could also be considered. One such is the discreetised Earth Mover's Distance (EMD) or Wasserstein distance (add ref) that can be defined for tables. It measues the cost of moving the probability mass from one distribution to make it match the second. It requires a cost function for each pair of cells in the table. If the costs for every pair of cells were the same, then the EMD would just be the same as the $PO50$. A measure that gave different costs would clearly be preferable, especially for ordered categories, but would involve a detailed specification. Suggestions on how this might be achieved would be welcome. Further metrics or methods may also be possible and we would welcome suggestions for these.

Another important aspect of utility is feedback from those to whom the synthetic data are supplied. One example of this was a synthesis we carried out of dates when children were excluded from school. By definition these dates need to be weekdays (although this was not true for a few original records). The synthetic data spread the dates over weekends too. To overcome this the data would need to be pre-processed to define the variables differently. This is a different aspect of utility and examples like this are common and as important as the more formal utility measures discussed here.

Acknowledgments

Thanks to users of `synthpop` who have provided feedback on the methods used.

References

Bowen CM, Lui F, Su B (2021). "Differentially private data release via statistical election to partition sequentially." METRON, 79(1), 1–31. URL https://doi.org/10.1007/s40300-021-00201-0.

Karr A, Oganian A, Reiter J, Woo M (2006). "New measures of data utility." Technical report available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.576.5002.pdf

Raab GM, Nowok B (2017) Inference from fitted models in synthpop. R Vignette, URL https://cran.r-project.org/web/packages/synthpop/vignettes/inference.pdf

Nowok B, Raab GM, Dibben C (2015). synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control. R package URL https://cran.r-project.org/web/packages/synthpop/

Raab GM, Nowok B, Dibben C (2017a). "Practical data synthesis for large samples." Journal of Privacy and Confidentiality, 7, 67–97.

Raab GM, Nowok B, Dibben C (2017b). “Guidelines for Producing Useful Synthetic Data.” Available from http://arxiv.org/abs/1712.04078.

Read T, Cressie RC (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data. NAC, Springer, Berlin.

Snoke J, Raab G, Nowok B, Dibben C, Slavkovic A (2018). “General and specific utility measures for synthetic data.” J. R. Statist. Soc. A, 181(3), 663–668.

Voas D, Williamson P (2001). “Evaluating goodness-of-fit measures for synthetic microdata.” Geographical and Environmental Modelling, 5, 177–200.

Woo MJ, Reiter JP, Oganian A, Karr AF (2009). “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation.” Journal of Privacy and Confidentiality, 1, 111–124.