A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records

Jennifer Taub*, Mark Elliot**, and Joseph Sakshaug***

*The University of Manchester, jennifer.taub@postgrad.manchester.ac.uk
**The University of Manchester, mark.elliot@manchester.ac.uk
***The University of Manchester, joe.sakshaug@manchester.ac.uk

Abstract: Synthetic data is an alternative to controlling confidentiality risk through traditional statistical disclosure control (SDC) methods. A barrier to the use of synthetic data for real analyses is uncertainty about its reliability and validity. Surprisingly, there has been a relative dearth of research into the measurement of utility of synthetic data. Utility measures developed to date have been either information theoretic abstractions, such as the Propensity Score Measure mean-squared error, or somewhat arbitrary collations of statistics and there has been no systematic investigation into how synthetic data holds in response with real data analyses.

In this paper, we adopt the methodology used by Purdam and Elliot (2007), in which they reran published analyses on disclosure-controlled microdata and evaluate the impact of the disclosure control on the analytical outcomes. We utilise the same studies as Purdam and Elliot to facilitate comparisons of data utility between synthetic and disclosure controlled versions of the same data.

The results will be of interest to academics and practitioners who wish to know the extent to which synthetic data delivers utility under a variety of analytic objectives.

1 Introduction

With the increasing centrality of data in our lives, societies and economies and the drive for greater government transparency and release of open data there has been a concomitant increase in demand for public release microdata. This demand cannot always be met by using traditional statistical disclosure control (SDC) techniques as SDC techniques can also strongly affect data utility. This is illustrated by Purdam and Elliot (2007) who tested the data utility of two SDC techniques – local suppression, which creates missing values to replace some of the key variables, and Post-Randomization (PRAM) which swaps categories for selected variables based on a pre-defined transition matrix – and found that these SDC techniques had a significant impact on a sample of published analytical outputs.

Rubin (1993) first introduced synthetic data as an alternative to using traditional SDC techniques. Rubin's proposal entailed treating all data as if they were missing values and imputing the data conditional on the observed data. As an alternative, Little (1993) introduced a method that would only replace the sensitive values referred to as partially synthetic data. Since fully synthetic data does not contain any original data the disclosure of sensitive information from these data is much less likely to occur. Likewise for partially synthetic data, since the sensitive values are synthetic, disclosure of sensitive information is also less likely to occur. Nevertheless, to be useful synthetic data must yield valid statistical analyses.

Initially synthetic data was only produced using multiple imputation (MI) techniques. However, recent research has examined non-parametric methods - including machine learning techniques - which are better at capturing non-linear relationships (Drechsler and Reiter, 2011). These methods include CART (Reiter, 2005), random forests (Caiola and Reiter, 2010), bagging (Drechsler and Reiter, 2011), support vector machines (Drechsler, 2010), and genetic algorithms

(Chen et al, 2016). CART, originally developed by Breiman et al (1984) as a non-parametric modelling tool based on decision trees, has become a commonly used method for generating synthetic data. To generate synthetic data from a CART model, Reiter (2005) explains that for each Y_k a tree is fit on (X, Y_{-k}) where Y_{-k} is all variables in Y except for Y_k . Sequential imputations are generated starting with Y_1. Let L_{1w} be the w^{th} leaf in the tree Y_1 and $Y_1^{L_{1w}}$ be the $n_{L_{1w}}$ values of Y_1 in leaf L_{1w}, then in each L_{1w} a new set of values is generated by drawing from $Y_1^{L_{1w}}$ using the Bayesian bootstrap (see Rubin, 1981 for more information on Bayesian bootstrapping). Imputations for Y_2 are made using the same procedure, but for tree Y_2, Y_1 is replaced with the imputed values from the last imputation $Y_{1rep,i}$. According to Reiter (2005), some advantages that CART models have over parametric models are that they (i) are more easily applied especially to non-smooth continuous data and (ii) provide a semi-automatic way to fit the most important relationships in the data (p. 12).

This paper will focus on assessing how well CART generated synthetic data can replicate real world analyses. Previous work by Drechsler and Reiter (2011) found that when comparing different methods of machine learning for synthetic data generation, CART yielded the highest data utility. Likewise, Nowok (2015) compared different tree-based synthetic data against a parametric synthetic dataset. Compared to the other tree-based synthetic data approaches CART performed better, but not as well as the parametric synthetic data. Drechsler and Reiter (2011) tested the utility by comparing distributions of different variables and coefficient estimates from a fitted a logistic regression model. Nowok (2015) compared coefficient estimates from a linear regression, a logistic regression, and poisson regression model to test data utility.

In this paper, following Purdam and Elliot (2007), we will be using 9 different sets of analyses and 28 different tests and models to test the data utility of CART synthetic data. CART derived synthetic data was chosen due to the high utility it has shown in the literature, as well as the fact that it is easier than multiply imputed synthetic data to generate. However, the tests could be equally well be applied to other sorts of synthetic data. Section 2 discusses some alternative definitions for data utility and gives a brief overview of data utility tests. In Section 3 we give an overview of the software we used for synthesizing our data. In section 4 we discuss our results leading to our conclusion in section 5.

2 Data Utility

To describe data utility, Winkler (2005) introduces the terms *analytically valid* and *analytically interesting.* He classifies a dataset as *analytically valid* if the following criteria are approximately preserved:

- Means and covariances on a small set of subdomains
- Marginal values for a few tabulations of data
- At least one distributional characteristic

Winkler (2005) classifies a dataset as *analytically interesting* if it provides at least six variables on important subdomains that can be validly analysed. However, both concepts, *analytically valid* and *analytically interesting,* are very abstract notions on data utility.

In contrast, Purdam and Elliot (2007) define the loss of analytical validity as occurring when “a disclosure control method has changed a dataset to the point at which a user reaches a different conclusion from the same analysis” (p. 1102). Whilst Winkler uses utility metrics to define data utility, which is easier to test than Purdam and Elliot’s definition, it could classify data as high utility but allow analysis that come to erroneous conclusions.

2.1 Common Utility Tests

Drechsler and Reiter (2009) classify two existing types of utility measures for synthetic data:

i. *Comparisons of broad differences between the original and released data (broad measures)*

ii. *Comparisons of differences in specific models between original and released data (narrow measures)* (Drechsler and Reiter, 2009, p.592)

Traditionally, in synthetic data, narrow measures have been used to measure data utility. For example, Drechsler et al (2008a) and Drechsler and Reiter (2009) used confidence interval overlap, a narrow measure developed by Karr et al (2006), to test the utility of fully and partially synthetic data. The confidence interval overlap is calculated as:

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

where $U_{,k}$ and $L_{,k}$, denote the upper and lower bound of the intersection of the confidence intervals from both the original and synthetic data for estimate k, and $U_{org,k}$ and $L_{org,k}$ represent the upper and lower bound of the original data, and $U_{syn,k}$ and $L_{syn,k}$ the synthetic data.

Broad measures tend to quantify some kind of statistical distance between the original and released data, utilising measures such as the Kullback-Leibler divergence or the Hellinger distance. Woo et al (2009) developed the Propensity Score Measure mean-squared error (pMSE), which has been adapted by Snoke et al (2016) to be compatible with synthetic data, since pMSE is not compatible with synthetic data in its original iteration. Snoke et al created two alternatives to the traditional pMSE; the pMSE ratio and the standardised pMSE, calculated by standardizing the statistics by its null expected value and standard deviation in order to define a statistical test.

Both broad and narrow utility measures have their failings. Karr et al (2006) reflect that while narrow measures are really good for certain analyses, they do not give a complete picture of the dataset and that broad measures are “pretty good” for many analyses and “really good” for none. This critique of utility tests reflects the fact that many of the utility tests are theoretic abstractions, which do not deal directly with the uses that data serve in real life. To avoid the problems associated with utility measures, Raab (2016) suggested asking users what kind of data they want to produce so that the synthetic data produced can have high utility for their needs. However, Mateo-Sanz et al (2005) caution that protection of microdata in a data-use specific manner can be difficult since the potential uses of data are so vast and if datasets are created for each use it may increase disclosure risk.

Purdam and Elliot (2007) introduce a more hands-on approach to evaluating data utility by repeatedly using narrow measures across a sample of analyses, so as to not fall into the trap where narrow measures report artificially high or low utility measures based a single set of analyses. Their method could be considered a *systematically sampled narrow measure,* as opposed to a *traditional ad hoc narrow measure.* Using the 1991 Sample of Anonymised Records (SAR) from the British census, they examined a sample of 23 published papers and re-ran the analyses of ten which they found to be replicable (with SDC perturbed data). This method for accessing utility perturbed data will henceforth be referred to as the Purdam-Elliot methodology. Overall, Purdam and Elliot (2007) found that the SDC techniques of local suppression and PRAM had a significant impact on the analytical outcomes. The Purdam-Elliot methodology has the benefit of testing data utility by using analyses that have actually been published and therefore actual uses of data rather than hypothetical ones (such as those used by Nowok (2015)). Using particular sets of analyses to test data utility has been used by previous papers to evaluate synthetic data (Drechsler et al, 2008b; Lee et al, 2013), however the Purdam-Elliot methodology will give us wider range of applicability due to the volume and variety of their analyses.

For this paper, we used the same base data and same analyses as Purdam and Elliot (2007)[1] (See Appendix 1 and 2 for description of the analyses) This is beneficial for two reasons: First, the papers that they selected show-cased a variety of analytic techniques, including cross-tabulations, logistic regression, probit regression, and multi-level modelling; second, it will allow us to directly compare the synthetic data results of this paper with the SDC controlled data results in the original paper.

3 Method

We will be using two datasets: the original 1991 SAR without any perturbation as a control and a CART generated synthetic version of the 1991 SAR.

3.1 Description of the 1991 SAR

As in Purdam and Elliot (2007), we used the 1991 SAR, specifically the Individual SAR, which utilises individual level data. The 1991 Individual SAR is a 2% sample of the population of Great Britain containing 1,116,181 records. The SAR is publically available and contains information on topics such as age, gender, ethnicity, household size, household type, employment, and health.

3.2 Reanalyses of Synthetic data files

This paper will be using partially synthetic data since CART only produces partially synthetic data. While Little (1993) envisioned partially synthetic data as only replacing the sensitive variables, in common nomenclature partially synthetic data refers to any synthetic dataset that still contains original values. In this instance only the geographical variables *area* (278) and *region* (12) will be left unchanged and the remaining 28 variables will all be synthesized (see appendix 3 for synthesizing order and appendix 4 for description of the variables). Due to the size of the dataset and the subsequent computational load, block synthesises was used in which the dataset was divided in 12 datasets by region, synthesized and recombined.

The missing data from the original 1991 SAR was left in the synthesis models. The CART generated synthetic data were generated using the r-package, *synthpop* version 1.3-0 (Nowok et al, 2016). The method for *synthpop* was set to “ctree”, which is derived from the *party* package, this is opposed to the default CART method which is derived from the *rpart* package. The “ctree” method allowed for a larger tree size, which given the volume of variables was preferable in this instance.[2] The minimum number of observations per terminal leaf was set to 5, and the tree size was set to 0, which is read by the *party* package as unlimited size. For comparison a control file, the original SAR is used to ensure that all analyses are correctly replicated.

[1] Of the original ten papers that were reanalyzed in Purdam and Elliot (2007), only nine were used for this paper.

[2] We did experiment briefly with using the default cart method in *synthpop* and found lower utility synthetic data. However, work by Nowok (2015) shows the default CART method as providing higher utility synthetic data than the “ctree” cart synthetic data. This is an enquiry that can be followed up in another paper.

4 Results

The replicated papers contain a combination of different kinds of analyses. For papers using frequency tables we took the ratio of counts for the results. The ratio of counts is calculated by dividing the smaller number by the larger number. So given two corresponding counts or percentages, where y_{orig}^{1} is the count or percentage the first box in a frequency table for the original data and y_{synth}^{1} is the corresponding count or percentage for the synthetic data:

$$\begin{gathered}\text{If } y_{orig}^{1} > y_{synth}^{1} \text{ then:}\\ \frac{y_{synth}^{1}}{y_{orig}^{1}}\\ \text{If } y_{orig}^{1} < y_{synth}^{1} \text{ then:}\\ \frac{y_{orig}^{1}}{y_{synth}^{1}}\\ \text{If } y_{orig}^{1} = y_{synth}^{1} \text{ then:}\\ \text{The ratio of counts is 1}\end{gathered} \tag{2}$$

The ratio of counts provides a decimal between 0 and 1. For each frequency table the ratio of counts were averaged together to give an overall ratio of counts for each synthetic frequency table. The average ratio of counts for the frequency tables are shown in table 1.

For the regression models the 95% confidence interval was calculated for the coefficients in both the original and synthetic models. From these confidence intervals, the confidence interval overlap was calculated using equation 1.The confidence interval overlap for all of the coefficients in the model was then averaged together to give an average for each model. The average confidence interval overlaps are reported in Table 2. We will be referring to the scores in both tables 1 and 2 as utility metric scores, since both the ratio of counts and CIO fit this description.

In contrast to the utility metrics in Tables 1 and 2, Table 4 takes a more holistic approach to assessing data utility. Using the key in Table 3, which is taken from the definitions provided by Purdam and Elliot (2007, p. 1109), each synthetic analysis was evaluated on how severely[3] it affected the analysis, from this point onwards referred to as the severity rating. Table 4 is not based upon the findings of table 1 and 2, but rather in Table 4 the conclusions of each paper were taken into account to see whether or not the authors of the original papers would be able to come to the same conclusions that they had in their original papers. Table 4 also includes the average utility metric scores from Tables 1 and 2 to facilitate comparison. In some cases it was harder to assess the severity rating than others where the effect was more obvious. For example, for the Champion (1996) paper the CART synthetic data follows the relationship that the white ethnic group has more people moving long distances than other ethnic groups, and that black ethnic groups have the lowest numbers moving 200+ km or more, as discussed in the original paper. However, it does not pick up on the finding that more Chinese move 200+ km than other non-white ethnic groups, which was also discussed in the original paper. Since the synthetic data overall has the same findings as the original paper, even with the exception of the Chinese migrant data, we classed it as “no effect.” Whilst there is an argument to be made for a severity rating of “moderate”, we ultimately decided that since the Chinese migrant group was only one of ten ethnic groups and in that case one of the smaller ethnic groups, that the change was unsubstantial enough to warrant a severity rating of “no effect”.

[3] “Severity” is a subjective measure based on the authors judgment.

Likewise, the Eade et al (1996) paper showed similar difficulties in classification, this time being on the boundary between “severe” and “moderate” effect. The Eade et al paper is about the Bangladeshi experience in the UK. However, the frequency tables created from the 1991 SAR include all 10 ethnic groups, as well as a category for those born in Ireland. For many of the ethnic groups, the trends shown by the CART synthetic data is very similar to that of the original data. However, it is not for the Bangladeshi ethnic group, which was the focus of their paper. Ultimately, we rated the synthetic data as having a moderate effect on the Eade paper analyses, since the overall tables were similar, even though they were not for the Bangladeshi ethnic group. This translates to the Eade paper having a high average ratio of counts, but a moderate effect. If the Eade et al analysis only included the Bangladeshi ethnic group the average ratio of counts for table 6.3 and 6.4 would be 0.706 and 0.694, which is lower than the average and includes some very low scores in important parts of the analysis.

Paper	**Table or Figure**	**Ratio of Counts**
Ballard (1996)	Figure 5.1	0.504
	Figure 5.2	0.490
	Figure 5.3	0.817
	Figure 5.4	0.887
	Average	**0.675**
Champion (1996)	**Table 4.7**	**0.836**
Eade et al (1996)	Table 6.3	0.799
	Table 6.4	0.799
	Average	**0.799**
Gardiner and Hill (1997)	Table 1	0.726
	Table 2	0.627
	Table 3	0.496
	Table 4	0.395
	Average	**0.561**
Green (1997)	Table 4.1	0.715
	Table 4.2	0.746
	Table 4.3	0.760
	Table 4.4	0.792
	Table 4.5	0.795
	Table 4.6	0.800
	Average	**0.768**
Leventhal (1994)	Figure 1	0.924
	Figure 2	0.654
	Figure 3	0.791
	Figure 4	0.822
	Figure 5	0.815
	Figure 6	0.958
	Figure 7	0.920
	Average	**0.840**

Table 1: Average Ratio of Counts for Frequency Tables Replicated

Paper	Regression model	Confidence Interval Overlap
Drinkwater and O'Leary (1997)	Table 5 for males	0.466
	Table 5 for females	0.484
	Average	0.475
Gardiner and Hill (1996)	Table 4	0.414
Gould and Jones (2000)	Table 4 Model A	0.583

Table 2: Average 95% CI Overlap for Regression Models Replicated

Severe	The results were sufficiently different
Moderate	Change in emphasis rather than a completely different finding
No effect	Indicates that the figures may have been slightly different but the overall pattern was not, indicating the same conclusion could be consistently drawn

Table 3: Key for paper effect severity (From Purdam and Elliot, 2007, p. 1109)

Paper	Severity Rating	Utility Metric Average score
Ballard (1996)	No effect	0.675
Champion (1996)	No effect	0.836
Drinkwater and O'Leary (1997)	Severe	0.475
Eade et al (1996)	Moderate	0.799
Gardiner and Hill (1996)	Severe	0.414
Gardiner and Hill (1997)	Moderate	0.561
Gould and Jones (2000)	Moderate	0.583
Green (1997)	No effect	0.768
Leventhal (1994)	No effect	0.840

Table 4: Description of the Effect of the Synthetic data

	No Effect	Moderate	Severe
CART	4	3	2
Suppressions	5	5	0
Post-randomisation	2	7	1
Both Suppressions and PRAM	1	5	4

Table 5: Summary of the Effect of perturbations (Information on PRAM and Suppressions from Purdam and Elliot, 2007, p. 1110)

In some cases Table 4 did not align with the percentages shown in Tables 1 and 2. For example while the ratio of counts for the Ballard (1996) analyses are low in Table 1, it still has a severity rating of "no effect". This is in part because Ballard is using count data, which given that the CART synthetic data produced a lower number of Pakistani participants will yield a lower result (comparatively the CART synthetic data had 9115 Pakistani respondents, while the original had 9416 Pakistani participants (see Appendix 2)), despite the fact that the CART synthetic data followed the same trends as in the original Ballard paper and the same conclusions could be drawn.

To determine if there is a relationship between the severity ratings and the utility metrics, a one-way ANOVA was conducted. The analysis of variance showed that the association was significant, F(2,6) = 8.316, $p < 0.05$. The average utility metric score for the severity ratings is; no effect = 0.780, moderate = 0.648, and severe = 0.445. This does indicate that it might be possible to tie metrics to analytical impact, although more work would be needed to ground these.

Table 5 summarises the results of Table 4, and also includes the findings from Purdam and Elliot on suppression and PRAM. The CART synthetic data shows lower utility than the suppression-perturbed data. The CART synthetic data shows similar results to the PRAM perturbed data and better results than when the data was perturbed by both PRAM and suppression.

4.1 Discussion

The F-value showing that there is a correlation between the utility metrics and the severity ratings is interesting because it does prove that the utility metrics do have meaning in the real-world. With further investigation it will be possible to create a scale for determining an analyses severity rating based upon its utility metric. However, these results are meant merely to give the suggestion of the idea of the correlation and not as an official scale. This is because we used two different types of utility metrics for the tables and the models, so they should not be directly comparable and second because the sample size of the number of analyses is still small. Additionally, there were analyses that had utility scores that were not congruent with their severity ratings, most notably Ballard (1996) and Eade et al (1996).

The results also raise the question of why some papers had high utility or low utility and if there are certain kinds of analyses that are more suited to be conducted with synthetic data. For example, many of the papers reanalysed involved subsetting the data and for small subsets the synthetic data tended to be less accurate. Appendix 2 shows the sample sizes for the different analyses. Appendix 2 shows that Gardiner and Hill (1996) had a very small sample size (less than 4,000), as did Gardiner and Hill (1997) for Table 4. Drinkwater and O’Leary (1997) have sample sizes ranging between 10,000 and 14,000, which while no means small, is still smaller than the very large sample sizes utilised by high utility analyses such as Green (1997). Alternatively, Ballard (1996) had a small sample size of 4,500 and maintained high utility, while Gould and Jones (2000) had low utility and a very large sample size (though different dummy variables in their model did refer to small groups).

While for Gardiner and Hill (1997) the sample size for their table 4 was 3,532, it ran into problems where most of that sample consisted of white participants and they were looking at racial variation. Given that in 1991 Great Britain was 94.6% white and Leicester was 72.7% white, Gardiner and Hill (1997) table 4 records that 4.08% of black people in Leicester cycle to work, however that statistic is made up of only two cycling black Leicester residents who make up only 0.000179 % of the entire data set. This micro-sample is unlikely to be replicated in the synthetic dataset, but should not be regarded as problematic because Gardiner and Hill (1997) reporting of it is somewhat misleading. In general, we should not be concerned about replicating findings that appear to be spurious.

A paper like Leventhal (1994) was easier for the synthetic datasets to replicate. He was interested in how many people fit a target demographic (head of household, aged 55+, employed, owns home). This was a large category consisting of 1.46% of the population and then he only ever compared his target variable to one other variable at a time. So when looking at the prevalence of his target demographic in different ethnic groups and so forth the results of the synthetic data stayed true to that of the original.

For comparison, Nowok (2016) has run similar utility tests using CART synthetic data. She used the same method, "*ctree"* of CART synthetic data from *synthpop*, but additionally combined 10 synthetic datasets for CART. She found the mean 95% CI overlap for the logistic regression model she used to be 0.50 for the CART synthetic data. Here, the Gardiner and Hill (1996) logistic regression had a mean 95% CI overlap of 0.41 for CART. While these results are similar, the Nowok results are higher. This could arise for many reasons. First as previously mentioned, Nowok used multiply imputed CART synthetic data by creating multiple CART synthetic datasets. However, part of the difference may also stem from the fact that the Gardiner and Hill (1996) paper heavily subsetted the dataset to only include residents of Sheffield, 50 and older (See appendix 2). Nowok's logistic regression did not involve subsetting the data prior to modelling it and therefore the sample size is the same size as the original data, while the Purdam-Elliot set of analyses had varying sample sizes. Furthermore, Nowok only used three different models and built her data set with these three models in mind (as evidenced by her including dummy variables in her synthetic model). This is a normative practice for evaluating the utility of synthetic data or other types of SDC data, but this does raise the question if papers assessing data utility are producing artificially high utility. This was less prone to happen in this paper since with the sheer number of analyses it is near impossible to create a synthetic data set with all 28 tests in mind.

5 Conclusion

This paper uses the Purdam-Elliot methodology for testing synthetic data utility. The methodology has the advantage of running a broad sample of real analyses; striking a balance between the non-specificity of general metrics and the ad hoc nature of traditional narrow measures (since for a particular set of analyses a synthetic dataset may perform with high utility, it may still falter for another analyses, as demonstrated by the results herein). The Purdam-Elliot methodology takes into account that researchers use the data in unexpected ways, by subsetting the data and/or creating their own compilation variables and that it is important to have synthetic data that fulfils these diverse needs within reason.

This paper has also provided more meaning to general utility metrics; the exercise of mapping these onto the judged impact of replicated analyses proved interesting. This type of mapping exercise hints at the possibility of producing more empirically grounded general metrics.

Finally we have also given an indication of the kinds of analyses that might produce higher utility when using CART derived synthetic data although more work is needed here examining the impact of different parameters for CART and data sets and types. We will also be doing further research on how the Purdam- Elliot methodology holds up with other kinds of synthetic data, including MI synthetic data and whether for other types of synthetic data the utility metrics and severity ratings show the same correlation.

References

Ballard, R. (1996). The Pakistanis: stability and introspection. In: C. Peach, ed., *Ethnicity in the 1991 Census: Volume 2 The ethnic minority populations of Great Britain*, 1st ed. London: HMSO, pp.121-149.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. 1st ed. Belmont, California: Wadsworth, Inc.

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

Chen, Y., Elliot, M. and Sakshaug, J. (2016). A Genetic Algorithm Approach to Synthetic Data Production. *PrAISe '16 Proceedings of the 1st International Workshop on AI for Privacy and Security,* (13).

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

Drechsler, J. and Reiter, J. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic data. *Computational Statistics and Data Analysis*, 55, pp.3232-3243.

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. *Journal of Official Statistics,* 25(4), 589-603.

Drechsler, J., Bender, S., and Rässler, S. (2008a). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions in Data Privacy*, 1, 105-130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S. and Zwick, T. (2008b). A new approach for disclosure control in the IAB establishment panel—multiple imputation for a better data access. *AStA Advances in Statistical Analysis*, 92, pp.439-458.

Drinkwater, S. and O'Leary, N. (1997). Unemployment in Wales: Does Language Matter?. *Regional Studies*, 31(6), pp.583-591.

Eade, J., Vamplew, T. and Peach, C. (1996). The Bangladeshis: the encapsulated community. In: C. Peach, ed., *Ethnicity in the 1991 Census: Volume 2 The ethnic minority populations of Great Britain*, 1st ed. London: HMSO, pp.150-160.

Gardiner, C. and Hill, R. (1997). Cycling on the Journey to Work: Analysis of the Socioeconomic Variables from the UK 1991 Population Census Samples of Anonymised Records. *Planning Practice & Research*, 12(3), pp.251-261.

Gardiner, C. and Hill, R. (1996). Analysis of Access to Cars from the 1991 UK Census Samples of Anonymised Records: A Case Study of Elderly Population of Sheffield. *Urban Studies*, 33(2), pp.269-281.

Green, A. (1997). Patterns of ethnic minority employment in context of industrial and occupational growth and decline. In: V. Karn, ed., *Employment, Education and Housing Among the Ethnic Minority Populations of Britain*, 1st ed. London: The Stationary Office, pp.67-90.

Gould, M. and Jones, K. (1996). Analyzing perceived limiting long-term illness using U.K. census microdata. *Social Science & Medicine*, 42(6), pp.857-869.

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

Lee, J., Kim, I. and O'Keefe, C. (2013). On Regression-Tree-Based Synthetic Data Methods for Business Data. *Journal of Privacy and Confidentiality*, 5(1), pp.107-135.

Leventhal, B. (1994). "Case Study Examples to Demonstrate the use of Samples of Anonymised Records in Marketing Analysis", *Occasional Paper* 5, CMU, University of Manchester, Manchester.

Little, R. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2), pp.407-426.

Mateo-Sanz, J., Domingo-Ferrer, J. and Sebe, F. (2005). Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. *Data Mining and Knowledge Discovery*, 11, pp.181-193.

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

Nowok, B. (2015). Utility of synthetic microdata generated using tree-based methods. In: *UNECE Statistical Data Confidentiality Work Session*.

Office for National Statistics. Census Division, University of Manchester. Cathie Marsh Centre for Census and Survey Research, 2013, *Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs)*, [data collection], UK Data Service, Accessed 29 June 2017. SN: 7210, http://doi.org/10.5255/UKDA-SN-7210-1

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

Raab, G. (2016). *Analysis methods and utility measures*. Presentation at Synthetic Data Workshop at Isaac Newton Institute for Mathematical Sciences (Cambridge, UK, October, 2016)

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

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

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. 1st ed. Hoboken, N.J: John Wiley & Sons.

Rubin, D. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9(1), pp.130-134.

Snoke, J., Raab, G., Nowok, B., Dibben, C. and Slavkovic, A. (2016). General and specific utility measures for synthetic data.

Winkler, W. (2005). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. *Research Report Series*, 2005(09).

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

Appendix 1: Description of analyses used in papers

Paper	Variables Used	Table	Description
Ballard (1996)	AGE, COBIRTH, DCOUNTY, ETHGROUP, MSTATUS, REGIONP, SEX	Figure 5.1	Bar chart of country of origin and age
		Figure 5.2	Bar char of age and marital status
		Figure 5.3	Pie chart of region
		Figure 5.4	Bar chart of region and ethnicity
Champion (1996)	DISTMOVE, ETHGROUP	Table 4.7	Frequency table of distance moved by ethnicity
Drinkwater and O'Leary (1997)	AGE, CESTSTAT, COBIRTH, DCOUNTY, DEPCHILD, ECOPRIM, MSTATUS, REGIONP, SEX, WELSHLAN	Table 5 males	Probit regression of whether employed by age, marital status, long-term limiting illness, qualifications, housing status, area, and welsh fluency
		Table 5 females	
Eade et al (1996)	AGE, COBIRTH, ETHGROUP, SEX, SOCLASS	Table 6.3	Cross-tab ethnicity by social class, gender by table
		Table 6.4	
Gardiner and Hill (1996)	AREAP, CARS, ETHGROUP, LTILL, SEX, TENURE	Table 4	Logistic Regression car ownership by age, gender, house ownership, ethnicity and long-term limiting illness
Gardiner and Hill (1997)	AREAP, ETHGROUP, QUALEVEL, SEX, SOCLASS, TRANWORK	Table 1	Frequency table of cycling rate for area
		Table 2	Frequency table of cycling rate for area by gender
		Table 3	Frequency table of cycling rate for area by ethnicity
		Table 4	Frequency table of cycling rate by ethnicity, for Leicester
Gould and Jones (2000)	AGE, ETHGROUP, LTILL, RESIDSTA, SEX	Table 4 Model A	Multilevel logistic model for long-term limiting illness, level 2: area, by age, gender, ethnicity, and interaction terms
Green (1997)	DISTWORK, ETHGROUP, HOURS, INDUSDIV, OCCMAJOR, OCCSUBMJ, SEX, TRANWORK	Table 4.1	Frequency table of ethnicity by industry growth of employment, gender by table
		Table 4.2	
		Table 4.3	Frequency table of ethnicity by occupation growth
		Table 4.4	
		Table 4.5	Frequency table ethnicity by hours worked per week
		Table 4.6	
Leventhal (1994)	AGE, CARS, CENHEAT, ECPOSFHP, ETHGROUP, RELAT, SEGROUP, SOCLASS, TENURE	Figure 1	Bar chart comparing prevalence of target group by region
		Figure 2	Marital status
		Figure 3	Ethnic group
		Figure 4	Social class
		Figure 5	Socio-economic group
		Figure 6	Availability of home central heating
		Figure 7	Number of cars

Appendix 2: Description of subsetting and sample sizes used in the papers

Paper	**Table/Figure**	**Subset**	**Sample size-original**	**Sample size-cart synthetic data**
Ballard (1996)	Figure 5.1	Pakistani male	4,904	4,501
	Figure 5.2	Pakistani female	4,512	4,614
	Figure 5.3	Pakistani	9,416	9,115
	Figure 5.4	South Asian (Pakistani, Indian, and Bangladeshi)	29,724	29,020
Champion(1996)	Table 4.7	Migrants within Britain	95,177	95,376
Drinkwater and O'Leary (1997)	Table 5 males	Wales, age 16-64, resident of Britain, male	14,016	13,756
	Table 5 females	Wales, age 16-59, resident of Britain, female	9,843	10,060
Eade et al (1996)	Table 6.3	Age 16+, answered question on social class, not armed forces, male	340, 958	335,172
	Table 6.4	Age 16+, answered question on social class, not armed forces, female	292, 852	299,022
Gardiner and Hill (1996)	Table 4	Sheffield, Age 50+	3,532	3,598
Gardiner and Hill (1997)	Table 1 Table 2 Table 3	Answered question on travel to work and from selected areas	26,124	27,023
	Table 4	Answered question on travel to work, and from Leicester	2,048	2,419
Gould and Jones (2000)	Table 4 Model A	Resident of Britain, Aged 30-60	419,550	416,620
Green (1997)	Table 4.1 Table 4.3 Table 4.5	Male	540,967	534,974
	Table 4.2 Table 4.4 Table 4.6	Female	575,241	581,207
Leventhal (1994)	Figure 1 Figure 2 Figure 3 Figure 6	Age 16+	894,115	894,620
	Figure 4 Figure 5	Age 16+, answered social class question	647,323	647,464
	Figure 7	Age 16+, answered car question	866,097	867,058

Appendix 3: the ordering for synthesis

REGIONP, AREAP, ECONPRIM, DISTMOVE, LTILL, TRANWORK, SOCLASS, CARS, OCCMAJOR, OCCSUBMJ, INDUSDIV, COBIRTH, AGE, MSTATUS, SEX, DCOUNTY, DISTWORK, ETHGROUP, HOURS, DEPCHILD, QUALEVEL, WELSHLAN, QUALNUM, TENURE, CESTSTAT, RESIDSTA, RELAT, SEGROUP, CENHEAT, ECPOSFHP

Appendix 4- Description of variables from the codebook

No.	**Variable name**	**Number of categories** (excluding N/A)	**Description**
1	REGIONP	12	Individual SAR region
2	AREAP	278	Individual SAR area
3	ECONPRIM	10	Economic position (primary)
4	DISTMOVE	13	Distance of move-migrants
5	LTILL	2	Limiting long-term illness
6	TRANWORK	9	Mode of transport to work
7	SOCLASS	9	Social class (based on occupation)
8	CARS	3	Number of cars
9	OCCMAJOR	9	Occupation: SOC Major groups
10	OCCSUBMJ	22	Occupation: SOC Sub-major groups
11	INDUDSDIV	10	Industry (SIC Divisions)
12	COBIRTH	42	Country of birth
13	AGE	95	Age
14	MSTATUS	5	Marital status
15	SEX	2	Sex
16	DCOUNTY	63	Counties (Aggregations of SAR areas)
17	DISTWORK	7	Distance to work
18	ETHGROUP	10	Ethnic group
19	HOURS	72	Hours worked weekly
20	DEPCHILD	2	Number of resident dependent children
21	QUALEVEL	3	Level of highest qualification
22	WELSHLAN	5	Welsh language (Wales Only)
23	QUALNUM	3	No. of higher educational qualifications
24	TENURE	10	Tenure of household space
25	CESTTSTAT	3	Status in communal establishments
26	RESIDSTA	3	Resident status
27	RELAT	8	Relationship to household head
28	SEGROUP	20	Socio-economic group
29	CENHEAT	3	Availability of central heating
30	ECPOSFHP	3	Economic position of family head