UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Generative adversarial networks for synthetic data generation: A comparative study.

Claire Little (University of Manchester)
claire.little@manchester.ac.uk

Abstract

Generative Adversarial Networks (GANs) are gaining a lot of attention as means for synthesising data. So far much of this work has been applied to use cases outside of the data confidentiality domain with a common application being the production of artificial images. Here we consider the potential application of GANs for the purposes of generating synthetic census microdata. We employ a battery of utility metrics and a disclosure risk metric (the Targeted Correct Attribution Probability) to compare the data produced by a customised GAN with those produced using orthodox data synthesis methods (using the Synthpop package).

Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study

Claire Little*, Mark Elliot* Richard Allmendinger** Sahel Shariati Samani*

* School of Social Sciences, University of Manchester, Manchester, M13 9PL, UK

** Alliance Manchester Business School, University of Manchester, Manchester, M13 9PL, UK

Abstract. Generative Adversarial Networks (GANs) are gaining increasing attention as a means for synthesising data. So far much of this work has been applied to use cases outside of the data confidentiality domain with a common application being the production of artificial images. Here we consider the potential application of GANs for the purpose of generating synthetic census microdata. We employ a battery of utility metrics and a disclosure risk metric (the Targeted Correct Attribution Probability) to compare the data produced by tabular GANs with those produced using orthodox data synthesis methods.

1 Introduction

Machine Learning (ML) methods are showing increasing promise as an approach to synthetic data generation. Generative Adversarial Networks (GANs), first proposed by Goodfellow et al. (2014), are the focus of much of the research literature. GANs are a generative deep learning technique that use artificial neural networks. The generative element of a GAN does not access the original data whilst training, and can therefore produce synthetic data without directly interacting with the original data; this is an interesting feature that might in theory reduce disclosure risk.

Whilst research into using GANs for mixed-type, or tabular, microdata has so far been limited, GANs are generating much research interest and have been used for various applications, although as detailed by Wang et al. (2020) these are predominantly in the image domain. Research has focussed on purposes such as: synthetic image generation (e.g. Karras et al. (2019) and Radford et al. (2016)); image colorization (e.g. Nazeri et al. (2018)); image-to-image translation (e.g. Huang et al. (2018), Isola et al. (2017), and Zhu et al. (2017)); image inpainting (e.g. Demir and Unal (2018)); super-resolution (e.g. Ledig et al. (2017) and Menon et al. (2020)); and synthetic medical image generation (e.g. Frid-Adar et al. (2018), Iqbal and Ali (2018), Piacentino et al. (2021), and Sandfort et al. (2019)).

In this paper, we provide a framework for assessing the relative merits of GANs compared to traditional statistical methods for producing synthetic data in terms of

the utility and residual risk of the data produced. We first give a brief introduction to the data synthesis problem and deep learning approaches in Section 2. In section 3 we outline the design of study. Section 4 provides the results of comparing two GANS with CART using *Synthpop* and a Bayesian approach using *DataSynthesizer*.

2 Background

2.1 Data Synthesis

Rubin (1993) introduced the idea of synthetic data, proposing using multiple imputation on all variables such that none of the original data was released. Little (1993) proposed an alternative that simulated only sensitive variables, thereby producing partially synthetic data. Rubin's idea was slow to be adopted, as noted by Raghunathan et al. (2003), who along with J. P. Reiter (2002, 2003a, 2003b), formalised the synthetic data problem. Further work has involved using non-parametric methods such as classification and regression trees (CART) and random forests (e.g. Drechsler and Reiter (2010, 2011) and J. Reiter (2005))

There are two competing objectives when producing synthetic data: high data utility (i.e., ensuring that the synthetic data is useful, with a distribution close to the original) and low disclosure risk. Balancing this trade-off can be difficult, as, in general, reducing disclosure risk comes at a cost in utility. This trade-off can be visualised by considering the R-U confidentiality map developed by Duncan et al. (2004). Whilst there are multiple measures of utility, ranging from comparing summary statistics, correlations and cross-tabulations, to considering data performance using predictive algorithms, there are fewer that focus on disclosure risk for synthetic data. As noted by Taub et al. (2018), much of the statistical disclosure control (SDC) literature focusses on re-identification risk, which is not meaningful for synthetic data, rather than the risk of attribution, which is more likely. The Targeted Correct Attribution Probability (TCAP) developed by Elliot (2014) and Taub et al. (2018) can be used to assess attribution risk.

2.2 Deep Learning and GANs

Deep learning (Lecun et al., 2015), a subset of the broader field of machine learning, uses artificial neural networks to learn models from data. Neural networks (NNs) are made up of a series of stacked layers of neurons joined by weighted connections (the term 'deep' refers to the number of hidden layers; a 'shallow' NN may contain only 1 or 2 layers). In general, a NN is trained and learns iteratively by backpropagating the loss, or error, through the network and adjusting the weights to reach an optimal solution. Deep learning methods can discover the underlying structure in complex, high-dimensional data and have been responsible for dramatically improved performance in areas such as speech recognition, image recognition, object detection, natural language understanding and genomics (Lecun et al., 2015).

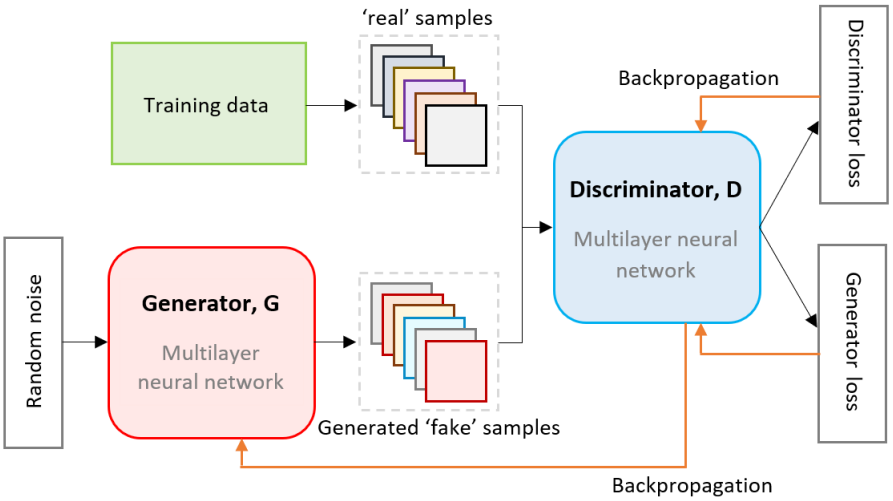

Figure 1: Example of GAN Architecture

GANs (Goodfellow et al., 2014), simultaneously train two NN models: a generative model which captures the data distribution, and a discriminative model that aims to determine whether a sample is from the model distribution or the data distribution. The process corresponds to a minimax two-player game. The generative model starts off with noise as inputs (it does not access the training, or original, dataset at all) and relies on feedback from the discriminative model to generate a data sample.

As described by Goodfellow et al. (2014), the discriminative and generative models are typically both multilayer NNs that are trained using the backpropagation or dropout algorithms. GANs perform alternate training, whereby the discriminator trains whilst the generator is held constant, and vice versa. The discriminator can be thought of as a supervised classification model. It receives batches of labelled real and generated data examples and outputs a single value for each example, the probability that it came from the real distribution, rather than the generator. If this value is close to 1, then it would be considered real; closer to zero would be classified as fake. The discriminator is penalized for misclassifying fake/real instances, and the weights adjusted accordingly. During generator training the weights are updated based on how well the generated samples fool the discriminator (ideally, when a generated image is fed into the discriminator the output will be close to 1). Figure 1 contains a basic example of GAN architecture.

GAN training can be challenging to optimize, as it can be difficult to balance the training of both models (generator and discriminator). If they do not learn at a similar rate then the feedback may not be useful. GANs can also be susceptible to issues such as vanishing gradients (where the discriminator does not feedback enough information for the generator to learn), mode collapse (e.g., the generator finds a small number of samples that fool the discriminator and only produces those, leading to the gradient of the loss function to collapse to near 0) and failure to converge. And, as noted by Lucic et al. (2018), since there is no consistent and generally accepted evaluation metric, it can be difficult to objectively evaluate or compare the

performance of different models.

Microdata tends to contain mixed data (i.e., a combination of numerical and categorical variables), which is likely to be heterogeneous, containing imbalanced categorical variables, and skewed or multimodal numerical distributions. GANs for image generation tend to deal with numerical, homogeneous data; in general, they must be adapted to deal with mixed data. Several studies have done this by adapting the GAN architecture, these are often referred to as tabular GANs. medGAN, developed by Choi et al. (2017) combined an autoencoder with a GAN to produce synthetic electronic health record (EHR) data containing binary (but not multi categorical) and continuous data. Camino et al. (2018) extended this work to include categorical data, however experiments by Goncalves et al. (2020) found the model failed to generate realistic patient data. Chen et al. (2019) proposed ITS-GAN which used a convolutional GAN architecture (normally used for images) and autoencoders to encode the "functional dependencies" within the data.

TableGAN, developed by Park et al. (2018) is based on the convolutional DCGAN (Radford et al., 2016) architecture, but contains three NNs (generator, discriminator and classifier) as opposed to the standard two NNs. The classifier NN is used to learn the "semantics" or rules from the original data and incorporate them into the training process. CTAB-GAN by Zhao et al. (2021) is based on a conditional GAN and also incorporates a classifier which is designed to learn the semantics of the data. TGAN, proposed by Xu and Veeramachaneni (2018) uses a Recurrent Neural Network (RNN) architecture with LSTM (long short-term memory) cells for the generator. RNNs are often used to process sequences of data (such as speech), and using this architecture, TGAN produces data column by column, predicting the value for the next column based on the previous ones. CTGAN, developed by Xu et al. (2019) is by the authors of TGAN, but does not use the same RNN GAN architecture. CTGAN uses "mode-specific normalization" to overcome non-Gaussian and multimodal distribution problems, and employs oversampling methods to handle class imbalance in the categorical variables. Much as with the GANs designed for numeric data, it is notable across much of the tabular GAN research, that there is inconsistency in terms of how they are evaluated, with no dominant method employed (other than assessing machine learning utility).

3 Study design

This study compares the performance of tabular GANs with more orthodox data synthesis methods.

3.1 Data

The 1991 Individual Sample of Anonymised Records (SAR) for the British Census (Office for National Statistics, Census Division, University of Manchester, Cathie

Marsh Centre for Census and Survey Research, 2013) was used. This contains a 2% sample (1,116,181 records) of the population of Great Britain (excluding Northern Ireland), including adults and children. There are 67 variables containing information such as age, gender, ethnicity, employment and housing. For the purposes of this experiment a subset of the overall dataset was selected; the geographical region of the West Midlands (containing 104,267 records, 9.34% of the total dataset). Twelve variables were selected, 1 numerical and 11 categorical, described in Appendix A. Minimal pre-processing was applied, and missing values were retained.

3.2 System selection

The statistical methods used were Synthpop, developed by Nowok et al. (2016) and DataSynthesizer, developed by Ping et al. (2017). The GAN methods were CTGAN, developed by Xu et al. (2019) and TableGAN, developed by Park et al. (2018). All methods were used with default parameters. It is recognised that the default parameters may not always produce the optimal performance (particularly with GANs) but the defaults are those most commonly applied and are used to provide a fair comparison across all methods.

Synthpop, an open source package written in R, by default uses methods based on classification and regression trees (CART, developed by Breiman et al. (1984)), which can handle mixed data types and is non-parametric. Synthpop synthesises the data sequentially, one variable at a time; the first is sampled, then the following are predicted using CART (in the default mode) with the previous variables used as predictors. This means that the order of variables is important (and can be set by the user). As suggested by Raab et al. (2017), variables with many categories may be moved to the end of the sequence, therefore the ordering was set by the least to maximum number of categories, with age first. Raab et al. (2017) also suggest that data rules should be included if they exist; since the SARs data contained four rules (e.g. if age ≤ 15, then marital status is single) these were included.

DataSynthesizer, a Python package, implements a version of the PrivBayes (Zhang et al., 2017) algorithm. DataSynthesizer learns a differentially private Bayesian Network which captures the correlation structure between attributes and then draws samples. A settable parameter, ϵ, controls differential privacy; 1 was used as the default. For the selected GAN based methods, CTGAN (described in the previous section) is a Python package developed to deal with mixed data; default parameters were used. TableGAN, implemented in Python, has a low, medium and high privacy setting; low was used as the default.

3.3 Measuring Disclosure Risk using TCAP

Taub et al. (2018) introduced a measure for disclosure risk of synthetic data called the Correct Attribution Probability (CAP). We will be using an adaptation used in Taub and Elliot (2019) called the *Targeted Correct Attribution Probability* (TCAP).

The TCAP method is based on a strong intruder scenario in which two data owners produce a linked dataset (using a trusted third party), which is then synthesised and the synthetic data published. The adversary is one of the data owners who attempts to use the synthetic data to make inferences about the others' dataset.

More modestly, at the individual record level, the adversary is somebody who has partial knowledge about a particular population unit (including the values for some of the variables in the dataset – the keys – and knowledge that the population unit is in the original dataset) and wishes to infer the value of a sensitive variable (the target) for that population unit.

Following Taub and Elliot (2019), we assume that the adversary will focus on records which are in equivalence classes with corresponding l-diversity of 1 on the target and attempts to match them to their data. The TCAP metric is then the probability that those matched records yield a correct value for the target variable (i.e. that the adversary makes a correct attribution inference).

Following Taub and Elliot (2019), TCAP is calculated as follows: We define d_o as the original data, and K_o and T_o as vectors for the key and target information, respectively

$$d_o = \{K_o, T_o\} \quad . \tag{1}$$

Likewise, d_s is the synthetic dataset

$$d_s = \{K_s, T_s\} \quad . \tag{2}$$

We then calculate the Within Equivalence Class Attribution Probability (WEAP) score for the synthetic dataset. The WEAP score for the record indexed j is the empirical probability of its target variables given its key variables

$$WEAP_{s,j} = Pr(T_{s,j}|K_{s,j}) = \frac{\sum_{i=1}^{n}[T_{s,i} = T_{s,j}, K_{s,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{s,i} = K_{s,j}]} \quad , \tag{3}$$

where the square brackets are Iverson brackets, n is the number of records, and K and T are vectors for the key and target information, respectively. Then using the WEAP score the synthetic dataset will be reduced to records with a WEAP score that is 1.

The TCAP for record j based on a corresponding original dataset d_o is the same empirical, conditional probability but derived from d_o,

$$TCAP_{o,j} = Pr(T_{s,j}|K_{s,j})_o = \frac{\sum_{i=1}^{n}[T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{o,i} = K_{s,j}]} \quad . \tag{4}$$

For any record in the synthetic dataset for which there is no corresponding record in the original dataset with the same key variable values, the denominator in Equation 4 will be zero and the TCAP is therefore undefined. If the TCAP score is 0 then the synthetic dataset carries little risk of disclosure; if the dataset has a TCAP score close to 1, then for most of the riskier records disclosure is possible.

3.4 Evaluating utility

Following Taub et al. (2020), we assess the utility of the synthetic data using three measures: confidence interval overlaps (CIO), the ratio of estimates (ROE) and the propensity mean squared error (pMSE). To calculate the CIO we use 95% confidence intervals, and use the coefficients from regression models built on the original and synthetic datasets. The CIO, proposed by Karr et al. (2006), is defined as:

$$J_k = \frac{1}{2}\left(\frac{U_{,k} - L_{,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{,k} - L_{,k}}{U_{syn,k} - L_{syn,k}}\right) \quad , \tag{5}$$

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

ROE is calculated by taking the ratio of the synthetic and original data estimates, where the smaller of these two estimates is divided by the larger one. Thus, given two corresponding estimates (e.g. totals, proportions), where y^1_{orig} is the estimate from the original data and y^1_{synth} is the corresponding estimate from the synthetic data, the ROE is calculated as:

$$ROE = \frac{\min(y^1_{orig}, y^1_{synth})}{\max(y^1_{orig}, y^1_{synth})} \tag{6}$$

If $y^1_{orig} = y^1_{synth}$ then the ROE = 1. The ROE will be calculated over bivariate and univariate data, and takes a value between 0 and 1. For each categorical variable the ratio of estimates are averaged across categories to give an overall ratio of estimates.

The propensity score, or pMSE, developed in Woo et al. (2009) and Snoke et al. (2018) is a measure of data utility designed to determine how easy it is to discern between two datasets based upon a classifier. It is calculated by merging the original and synthetic datasets and creating a variable T, where $T = 1$ for the synthetic dataset and $T = 0$ for the original dataset. For each record in the combined dataset the probability of being in the synthetic dataset is computed; this is the propensity score. The propensity score can be computed via logistic regression. The distributions of the propensity scores for the original and synthetic data are compared; where these are similar data utility should be high. In summary:

$$pMSE = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2 \quad , \tag{7}$$

where N is the number of records in the merged dataset, $\hat{p}_i$ is the estimated propensity score for record i, and c is the proportion of data in the merged dataset that is synthetic (which is often 1/2). A pMSE score close to 0 would indicate high utility (a score of 0 indicates the original and synthetic data are identical).

4 Results

Synthetic datasets the same size as the original ($n = 104,267$) were generated (aside from TableGAN which generated $n = 104,000$ records). No post-processing of the data was performed. Figure 3 in Appendix B contains a comparison of age distributions for each dataset. Figure 4 in Appendix B shows indicative bar graphs for four of the variables. These plots illustrate that whilst the data produced by Synthpop and DataSynthesizer (and to a lesser extent CTGAN) show similar counts to the original dataset, the data produced by TableGAN in many cases does not. TableGAN did not identify some categories (for example, classifying all individuals as having zero higher educational qualifications, or no individual as being divorced) or over/under estimated in other cases (e.g. classifying a large proportion of individuals with missing Tenure).

4.1 Disclosure risk

Table 1 shows the TCAP score for each key, with LTILL (long-term illness), FAMTYPE (family type) and ECONPRIM (primary economic position) as the targets. Four sets of keys, ranging from three to six variables were used (see Appendix A).

4.2 Utility

Table 2 presents all utility metrics. Following Taub and Elliot (2019), $1-4pMSE$ is used to scale the pMSE between 0 and 1. The mean of the ROE scores for the univariate and bivariate cross-tabulations (applied across all 66 possible combinations of two variables) is also presented. The CIO and standardized difference are presented as the mean across multiple regression models using each variable in the dataset as a target. The CIO could not be calculated for TableGAN because the variables in the dataset had insufficient variation to provide comparable logistic regression models.

The overall utility score is calculated by taking the mean of the ROE scores, the CI overlap and the value for $1 - 4pMSE$. The overall utility score for TableGAN takes the mean excluding the CIO, which could not be calculated. The R-U (Risk-Utility) confidentiality map shown in Figure 2 plots the overall utility score against the average TCAP (risk) score. Included for reference is the original data (which has an overall utility and risk score of 1) and the average TCAP baseline.

Target	Key	Synthpop	DataSynthesizer	CTGAN	TableGAN	Baseline
LTILL	6	0.935	0.929	0.912	0.911	
	5	0.897	0.898	0.891	0.907	0.774
	4	0.894	0.899	0.889	0.907	
	3	0.936	0.951	0.931	0.901	
FAMTYPE	6	0.709	0.23	0.598	0.301	
	5	0.725	0.658	0.639	0.384	0.223
	4	0.736	0.654	0.651	0.416	
	3	0.809	0.608	0.648	0.420	
TENURE	6	0.596	0.429	0.490	0.217	
	5	0.504	0.376	0.453	0.336	0.329
	4	0.500	0.350	0.447	0.341	
	3	0.496	0.353	0.482	0.279	
Average		0.728	0.644	0.669	0.527	0.442

Table 1: TCAP scores.

5 Discussion

The results show that the orthodox statistical method – CART using Synthpop – shows the highest utility and highest risk of the four methods tested (for the data and the parameter settings used). Table GAN produced the lowest level of risk but at levels of data quality that appeared to be unacceptably low. The other two methods were between the two.

It should be noted that this study is a proof of concept for the methodology and we make no assertions at this stage about generalisability. The limitations are:

- We have only tested the methods on a single dataset.
- We have only tested here a sub-sample of records. Ideally we would want these to be applied to a population file.
- We have used the default settings for each of the systems. Different settings would produce different outcomes.
- We have only considered a small selection of GANs and no other machine learning methods.
- At present, the methods we are employing here are only attempting to optimize the closeness of the synthetic data to the original (i.e. analog utility). Ideally we would want a system that optimises risk and utility.

Notwithstanding the above limitations, Figure 2 does indicate that the trade off is happening even with synthetic data. The four data points (excluding the original data) that we have appear to fall on a straight line, and with the summary measure we have used the utility is traded quite heavily for reductions in risk.

Metric	Synthpop	DataSynthesizer	CTGAN	TableGAN
pMSE	**0.00015**	0.01438	0.03162	0.17529
log(pMSE ratio)	**1.058**	5.655	6.442	8.153
1-(4 x pMSE)	**0.9994**	0.9425	0.8735	0.2988
ROE univariate	**0.981**	0.821	0.743	0.499
ROE bivariate	**0.847**	0.616	0.587	0.255
CI Overlap	**0.506**	0.365	0.410	-
Standardized difference	**3.721**	4.297	4.034	-
Overall utility	**0.833**	0.686	0.653	0.351

Table 2: pMSE, ROE and CIO scores with best results for each measure in **bold**.

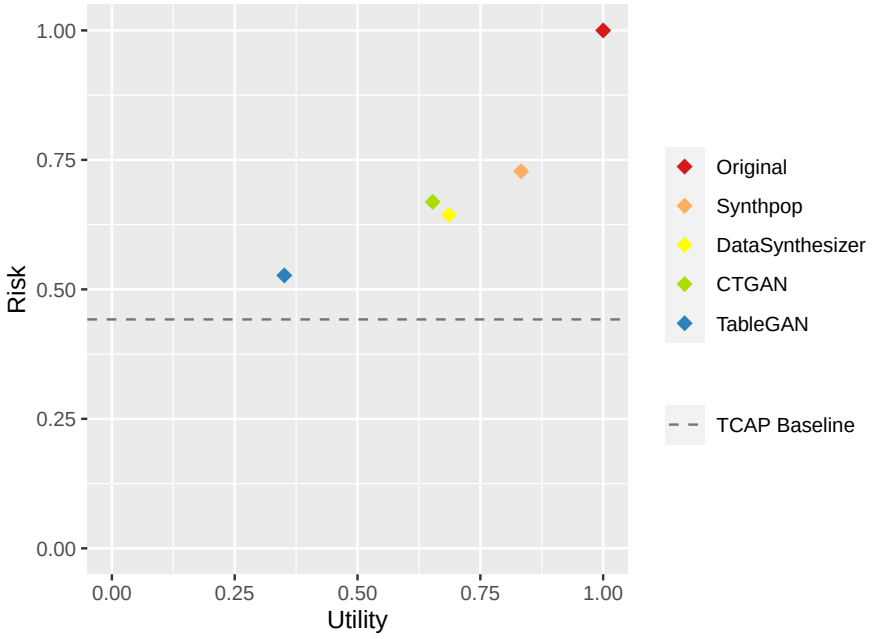

Figure 2: The Risk-Utility map of the synthetic datasets.

As a general observation on the methodology, the TCAP measure is itself relatively new and has never been calibrated. TCAP is a natural evolution of DCAP (Taub et al., 2018) (which can be regarded as an absolute but unrealistic measure of inferential disclosure risk). TCAP poses a more realistic intruder scenario but ideally we would want to pen test this scenario using a methodology such as that of Elliott et al. (2016). In future work we will:

1. Run a much wider range of tests examining the effects of changes in parameter settings on each methods position on the RU map.
2. Investigate other GAN architectures with a view to developing the most appropriate one for the data synthesis task.
3. Test the above on larger datasets in terms of the number of variables and cases and investigate the extent to which it is possible to synthesise a whole (UK) census.

References

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Camino, R. D., Hammerschmidt, C. A., & State, R. (2018). Generating multi-categorical samples with generative adversarial networks. *ICML 2018 workshop on Theoretical Foundations and Applications ofDeep Generative Models*.

Chen, H., Jajodia, S., Liu, J., Park, N., Sokolov, V., & Subrahmanian, V. S. (2019). Faketables: Using GANs to generate functional dependency preserving tables with bounded real data. *IJCAI International Joint Conference on Artificial Intelligence*, 2074–2080. https://doi.org/10.24963/ijcai.2019/287

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. *Proceedings of Machine Learning for Healthcare 2017*, *68*, 1–20. https://github.com/mp2893/medgan

Demir, U., & Unal, G. (2018). Patch-Based Image Inpainting with Generative Adversarial Networks. *arXiv preprint arXiv:1803.07422*. http://arxiv.org/abs/1803.07422

Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, *105*(492), 1347–1357. https://doi.org/10.1198/jasa.2010.ap09480

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, *55*(12), 3232–3243. https://doi.org/10.1016/j.csda.2011.06.006

Duncan, G. T., Keller-McNulty, S. A., & Stokes, S. L. (2004). *Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map* (tech. rep.). National Institute of Statistical Sciences.

Elliot, M. (2014). *Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team* (tech. rep. October). https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf

Elliott, M., Mackey, E., O'Shea, S., Tudor, C., & Spicer, K. (2016). End user licence to open government data? a simulated penetration attack on two social survey datasets. *Journal of official statistics*, *32*(2), 329–348.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, *321*, 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1), 1–40. https://doi.org/10.1186/s12874-020-00977-1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *arXiv preprint arXiv:1406.2661*. https://doi.org/10.1145/3422622

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal Unsupervised Image-to-Image Translation. *Proceedings of the European conference on computer vision (ECCV)*, 172–189. https://doi.org/10.1007/978-3-030-01219-9{_}11

Iqbal, T., & Ali, H. (2018). Generative adversarial network for medical images (MI-GAN). *Journal of Medical Systems*, *42*(231).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *American Statistician*, *60*(3), 224–232. https://doi.org/10.1198/000313006X124640

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 105–114. https://doi.org/10.1109/CVPR.2017.19

Little, R. J. A. (1993). Statistical Analysis of Masked Data.

Lucic, M., Kurach, K., Michalski, M., Bousquet, O., & Gelly, S. (2018). Are Gans created equal? A large-scale study. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 698–707.

Menon, S., Damian, A., Hu, S., Ravi, N., & Rudin, C. (2020). PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.00251

Nazeri, K., Ng, E., & Ebrahimi, M. (2018). Image colorization using generative adversarial networks. *International conference on articulated motion and de-*

formable objects, *10945 LNCS*, 85–94. https://doi.org/10.1007/978-3-319-94544-6{_}9

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11). https://doi.org/10.18637/jss.v074.i11

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). https://doi.org/http://doi.org/10.5255/UKDA-SN-7210-1

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, *11*(10), 1071–1083. https://doi.org/10.14778/3231751.3231757

Piacentino, E., Guarner, A., & Angulo, C. (2021). Generating synthetic ecgs using gans for anonymizing healthcare data. *Electronics (Switzerland)*, *10*(4), 1–21. https://doi.org/10.3390/electronics10040389

Ping, H., Stoyanovich, J., & Howe, B. (2017). DataSynthesizer: Privacy-Preserving Synthetic Datasets. *Proceedings of SSDBM '17*, *Part F1286*, 1–5. https://doi.org/10.1145/3085504.3091117

Raab, G. M., Nowok, B., & Dibben, C. (2017). Guidelines for producing useful synthetic data. *arXiv*.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016*.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple Imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*(1), 1–16.

Reiter, J. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics-Stockholm-*, *18*(4), 1–19. http://www.stat.duke.edu/~jerry/Papers/jos02.pdf

Reiter, J. P. (2003a). Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology*, *29*(2), 181–188.

Reiter, J. P. (2003b). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *168*(1), 185–205. https://doi.org/10.1111/j.1467-985X.2004.00343.x

Rubin, D. B. (1993). Statistical Disclosure Limitation. *Journal of Official Statistics*, *9*(2), 461–468. https://doi.org/10.1007/978-0-387-39940-9{_}3686

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve gen-

eralizability in CT segmentation tasks. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-52737-x

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *181*(3), 663–688. https://doi.org/10.1111/rssa.12358

Taub, J., & Elliot, M. (2019). The synthetic data challenge. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf

Taub, J., Elliot, M., Pampaka, M., & Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. *Privacy in Statistical Databases*, 122–137. https://doi.org/10.1007/978-3-319-99771-1

Taub, J., Elliot, M., & Sakshaug, J. W. (2020). The impact of synthetic data generation on data utility with application to the 1991 UK samples of anonymised records. *Transactions on Data Privacy*, *13*(1), 1–23.

Wang, L., Chen, W., Yang, W., Bi, F., & Yu, F. R. (2020). A State-of-the-Art Review on Image Synthesis with Generative Adversarial Networks. *IEEE Access*, *8*, 63514–63537. https://doi.org/10.1109/ACCESS.2020.2982224

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

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *arXiv:1811.11264*.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems*, *42*(4). https://doi.org/10.1145/2588555.2588573

Zhao, Z., Kunar, A., Van der Scheer, H., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective Table Data Synthesizing. *arXiv preprint arXiv:2102.08369*. http://arxiv.org/abs/2102.08369

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

A The SARS Dataset variables

Twelve variables from the 1991 Great Britain Individual Sample of Anonymised Records (Office for National Statistics, Census Division, University of Manchester, Cathie Marsh Centre for Census and Survey Research, 2013), or SARs dataset were used:

- AREAP: individual SAR area, 21 categories
- AGE: age, integer range 0-95
- COBIRTH: country of birth, 13 categories
- ECONPRIM: primary economic position, 10 categories
- ETHGROUP: ethnic group, 10 categories
- FAMTYPE: family type, 9 categories
- LTILL: limiting long-term illness, 2 categories
- MSTAUS: marital status, 5 categories
- QUALNUM: number of higher educational qualifications, 3 categories
- SEX: sex, 2 categories
- SOCLASS: social class, 9 categories
- TENURE: tenure of household space, 7 categories

Note, the COBIRTH variable was aggregated to 13 categories (countries of the UK, Ireland and continents) as Synthpop can struggle with datasets containing many categorical attributes with many categories.

A.1 Key variables used for TCAP

For each of the target variables (LTILL, FAMTYPE and TENURE) the following were used as key variables:

- 6 keys: AREAP, AGE, SEX, MSTATUS, ETHGROUP, ECONPRIM
- 5 keys: AREAP, AGE, SEX, MSTATUS, ETHGROUP
- 4 keys: AREAP, AGE, SEX, MSTATUS
- 3 keys: AREAP, AGE, SEX

B Indicative plots of univariates

Figure 3: Histograms comparing original data with synthetic data for age

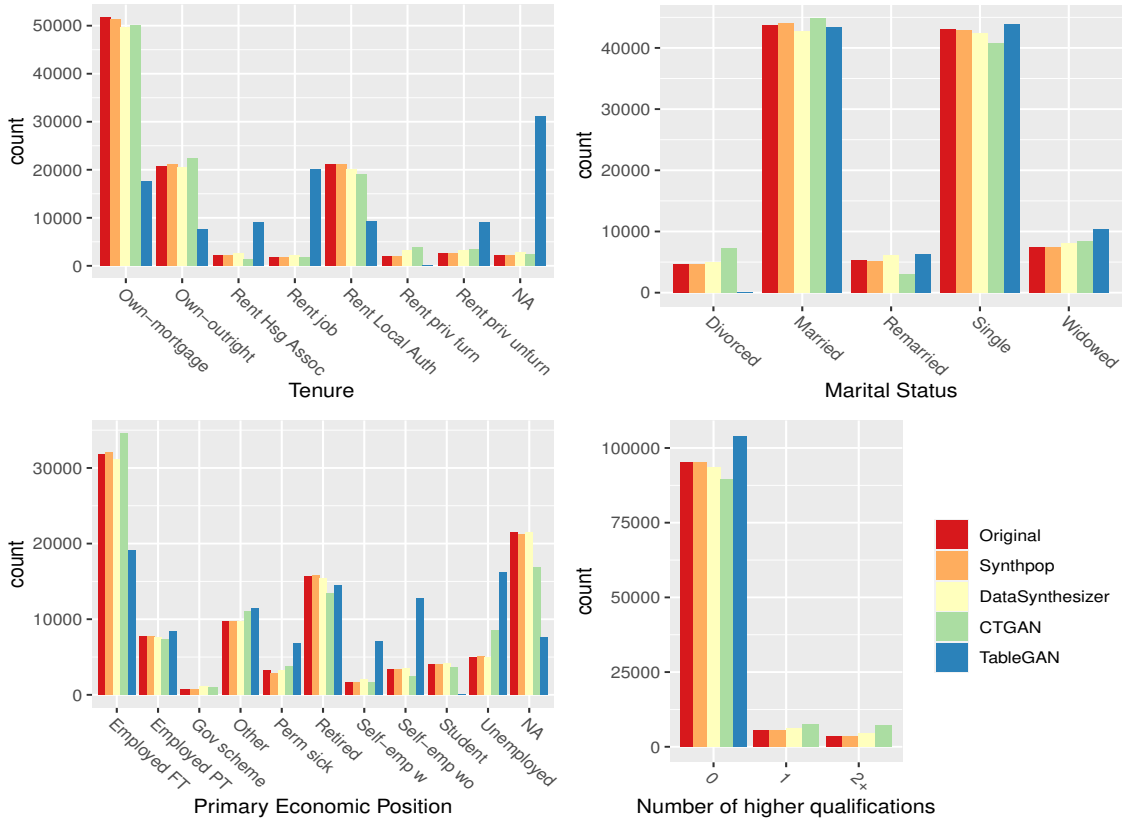

Figure 4: Bar graphs comparing original data to synthetic data