CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 3

29-31 October 2019, the Hague, the Netherlands 21 October 2019

Trade-off between Information Utility and Disclosure Risk in GA Synthetic Data Generator

Yingrui Chen, Jennifer Taub, Mark Elliot (University of Manchester, United Kingdom)

mark.elliot@manchester.ac.uk, yingrui.chen@postgrad.manchester.ac.uk, jennifer.taub@manchester.ac.uk

Abstract and Paper

The trade-off between risk and utility is the essence of the statistical confidentiality problem and although tools such R-U maps have been useful in augmenting our thinking how to operationalise such concepts is not fully understood. Genetic algorithms (GAs) are specifically designed to deal with optimising trade-offs between conflicting objectives and would be well suited to this task. This paper uses a GA system to provide such operationalisation using the example of data synthesis.

To deliver this the paper introduces a new risk measure called the targeted correct attribute probability (TCAP), which measures attribute disclosure risks for synthetic data, as the risk objective to and uses divergence from the full contingency table as the utility objective. Two types of mutliobjective GAs are tested using UK census data.

The results show that the GA approach is effective in optimising the data synthesis. In the discussion, we provide some speculations about how the method could be extended to capture standard SDC and synthesis within a single framework.

Trade-off between Information Utility and Disclosure Risk in GA Synthetic Data Generator

Yingrui Chen*, Jennifer Taub*, Mark Elliot*

* Cathie Marsh Institute, The University of Manchester, Manchester, UK,
{yingrui.chen, jennifer.taub, mark.elliot}@manchester.ac.uk

Abstract. Data synthesis is a data confidentiality method which is applied to microdata to prevent leakage of sensitive information about respondents. Instead of publishing real data, data synthesis produces an artificial dataset that does not contain the original records of respondents. This, in particular, offers significant protection against re-identification attacks. Previous work has shown that Genetic algorithms (GAs) are capable of producing good synthetic data using the full contingency table as the only objective, because it completely specifies the frequencies of equivalence class structure; using this objective could in theory lead to the original data as the solution; although the search space for any non trivial dataset appears be fractal and therefore convergence to the original data is improbable in polynomial time [3].

As a fundamental property for categorical data, the full table necessarily retains all statistical properties present in the original data. However, this implies higher disclosure risks (which are by definition one such statistical property).

This technique therefore presents precisely the same issue of managing the trade-off between disclosure risks and data utility as orthodox SDC. However, GAs have a big advantage. A GA is an iterative learning algorithm that used to solve complicated problems and critically supports multiple contrary objectives by allowing the dynamically monitoring of the relationship between them.

This paper uses the differential correct attribute probability (DCAP), which measures attribute disclosure risks for synthetic data, as the risk objective in GA synthesiser to contest with the full contingency table as the only utility objective. .

1 Introduction

Disclosure control for microdata attempts to protect sensitive information in an original dataset whilst retaining that datasets statistical properties for analysts.

Data synthesis can be regarded a statistical disclosure control (SDC) technique that produces a synthetic dataset that is designed to preserve the statistical properties of the original data and provide sufficient variables to allow proper multivariate analyses (Abowd and Lane, 2004). Therefore, data synthesis can be treated as an optimisation problem with two opposite constraints: data utility and disclosure risks.

In this paper, we use a genetic algorithm (GA) to generate synthetic data. GA's form a branch of evolutionary computing which aims to solve optimisation problems. There are three main operators within GAs: *selection*, *crossover* and *mutation*. Starting with group of candidate solutions (the initial population). The fitness of these candidates are evaluated. Fitter candidates are randomly paired using a crossover operator to generate new candidate solutions. These candidates are then passed to mutation operator and given a probability to mutate. The offspring form the new population and the process is repeated until a desired optimality achieved.

2 Objectives

Objectives within GAs are user defined values for a set of (usually latent) attributes which will hold for any optimal solution and so define a standard against which the fitness of a candidate should be measured. When designing a GA for a real-world problem, there is invariably more than one objective to be considered. These objectives are often opposed to one another, and this situation holds for data utility and disclosure risk in the synthetic data problem [7]. in general, as utility decreases, disclosure risk rises. When dealing with multi-objective problems, we can adopt one of two solutions; the first is to simultaneously optimise all objectives using a method such as *Pareto optimisation*. However, Chen's work suggests that it is very difficult to arrive at a stable optimal solution using this method[3]. This leads to the second alternative which is to simply compose the objectives into a single function with determined weights (referred as SOGA in this paper). The objectives must be normalised to avoid an objective with a larger numerical variance dominating the less varied one, otherwise the selection of weights is easy to get wrong and even a small difference in weighting can lead to different solutions.

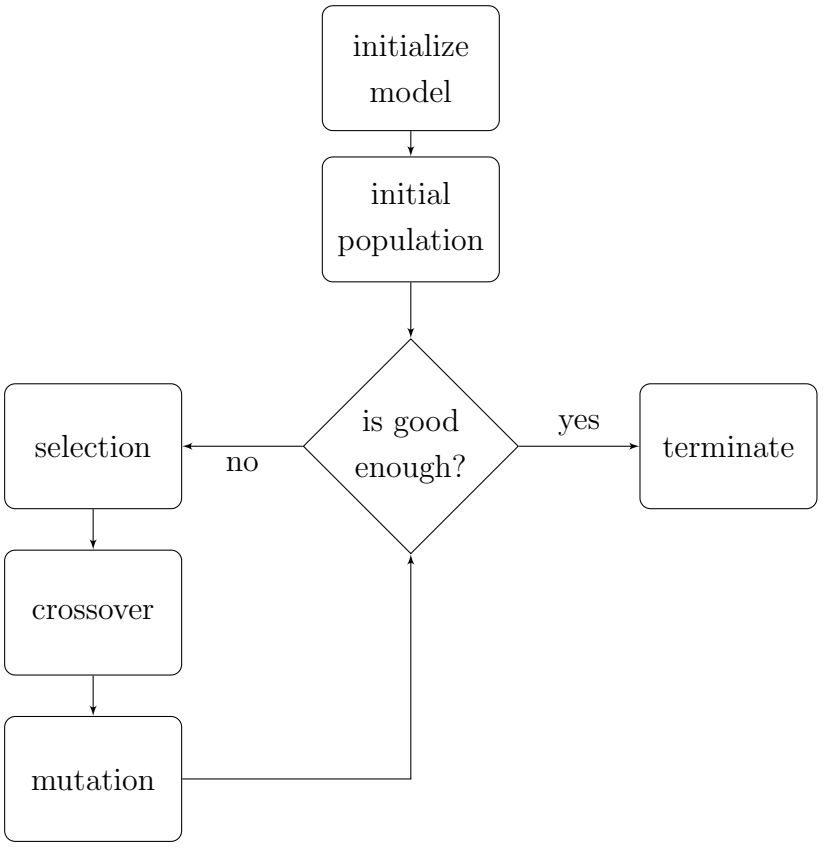

Figure 1: Flowchart of Genetic Algorithms

2.1 The Utility Measure

The between-variate structure in a categorical data can be captured by contingency table. Assume $I = \times_{j\in[1,,m]} I$ denotes the possible configurations of the variables that take values from finite sets, a contingency table is an m-dimensional table containing a count for each member of I. Since many statistical analysis methods are based on contingency tables, it is undoubted that a synthetic data that is close enough to the original data retains most - if not all - of its statistical properties.

Jensen-Shannon distance D_{JS} is a method to measure the level of divergence between two probability distributions, and specifically can measure the divergence between a pair of contingency tables. Suppose P and Q are two discrete probability distribution, $D_{JS}(P||Q)$ is defined by:

$$D_{JS}(P||Q) = (\frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M))^{\frac{1}{2}} \tag{1}$$

, where $M = \frac{1}{2}(P + Q)$ and D_{KL} is Kullback-Leibler divergence, which cannot be used directly because of the requirement for absolute continuity.

The objective is to minimise $U(X, Y)$, which is the divergence of full contingency tables CT_{FULL} from the synthetic and original data.

$$U(X, Y) = D_{JS}(CT_{FULL}(X)||CT_{FULL}(Y)) \tag{2}$$

Compared with other utility measurements, the full contingency table gives a more straightforward view of the divergence between original data and its synthetic version in a case-level. The disadvantage is that the synthesiser may overfit to the original data.This may not be immediately obvious, but there maybe structure within the data that is not useful. That is although the fitter output carries more sufficient information, it has highly chance of exposing sensitive information.

2.2 The disclosure risk measure

Most of the functions that are used to evaluate disclosure risks from released datasets are based on post-hoc measurements. Although there exists an inverse relationship between utility and disclosure risks in data privacy [4], involving risk in fitness evaluation the early stage of process may eliminate candidates with good properties. Therefore, we replace the starting point of the generator from a set of inferior candidates by a set of near optimal candidates (on the utility function) generated by

modest mutation if the original data, which makes the bi-objectives model more reasonable. We note in passing here that this makes the process more like a traditional SDC approach than an traditional data synthesis, but as we will discuss later within an optimisation framework, we consider the distinction between SDC and synthesis to be entirely arbitrary.

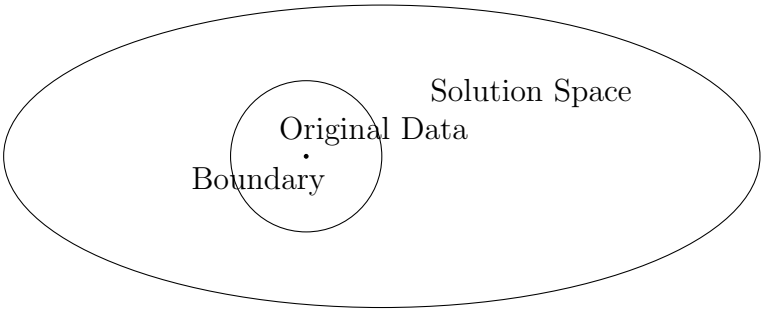

Figure 2: Illustration of Solution Space in Data Synthesis

Elliot (2014) and Taub et al (2018) introduced a measure for disclosure risk of synthetic data called the *Differential Correct Attribution Probability* (DCAP), which consists of a Correct Attribution Probability (CAP) score. DCAP was originally used as a post-hoc test to assess attribution risk [5, 14].

It is considered differential in that the score is to be compared to that of both the original dataset and that of a baseline (based on the univariate distribution of a given sensitive variable). DCAP works on the assumption that the intruder knows the values of a set of key variables for a given unit and is seeking to learn the specific value of a target variable. The Correct Attribution Probability (CAP) for the record indexed j is the empirical probability of its target variables given its key variables,

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

where the square brackets are Iverson brackets, n is the number of records, and d_o is the original data and K_o and T_o as vectors for the key and target information. Likewise, d_s is the synthetic dataset, with the vectors K_s and T_s.

The CAP for record j based on a corresponding synthetic dataset d_s is the same empirical, conditional probability but derived from d_s,

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

However, a test run on 9,000 datasets mutated from the original dataset in with different levels mutation showed an unexpected near perfect correlation between DCAP and $U(X, Y)$ (-0.9920). A linear model with multiple $R^2 = 0.9839$ confirmed the linear relationship between the two objectives. In another words, using this measure there will not be an ideal solution with high utility but low disclosure risk. In effect this mean that the pure DCAP measure is a measure of utility rather than risk[1]. This actually makes sense since this full DCAP measure captures the capability of user to make inferences about based on the conditional distribution of the target variable. On the other hand drawing from the posterior distribution for any arbitrary record regardless of it really makes non sense for an intruder.

As one method of circumnavigating this, Taub et al (2018) introduce a scenario in which rather than using the whole dataset, only the statistical uniques of the original dataset are used in calculating the CAP score [14] (this method corresponds to a common focus for National Statistical Institutes). Correlation between DCAP-Statistical Uniques and $U(X, Y)$ dropped to -0.58. The non-matches (records on the original dataset which do not match any records in synthetic dataset on the key) in this instance of DCAP were scored as 0, this allowed for candidates with more non-matches to have lower scores on the risk measure, which is intuitive.

3 Experiments

3.1 Model Design in GA Synthesiser

The GA synthesiser in our previous work [2] was equipped with Deterministic tournament selection operator with tournament size $t = 2$, Candidates are randomly selected into tournaments of size t (with replacement). The probability that a candidate wins the tournament and enters crossover is given by $p(1 - p)r$ where p is a parameter (such that $1/t < p1$) and r is the rank of the candidates fitness within the tournament. In deterministic tournament selection p is set to 1. And the same mechanism applies to this problem by combining the two objectives using the Euclidean distance from the origin $(0, 0)$:

[1]We are thankful to Gillian Raab [10]for drawing attention to this general problem which led to us running the above experiment

$$
\begin{aligned}
&\text{Minimise } F(X,Y) \\
&\qquad F(X,Y) = \frac{1}{\sqrt{2}}\sqrt{U(X,Y)^2 + R(X,Y)^2} \\
&\text{s.j.t } U(X,Y) \in [0,1] \\
&\qquad R(X,Y) \in [0,1]
\end{aligned}
$$

Crossover and mutation are the main operators used in GA to provide variation within the next generation. We used whole-case crossover and uniform mutation, which has been shown to be efficient in generating synthetic data. The following figures illustrate how the crossover and mutation operators work.

$$
\begin{pmatrix}
\boxed{x^1_{11}\ x^1_{12}\ \ldots\ x^1_{1m}} \\
x^1_{21}\ x^1_{22}\ \ldots\ x^1_{2m} \\
\boxed{x^1_{31}\ x^1_{32}\ \ldots\ x^1_{3m}} \\
x^1_{41}\ x^1_{42}\ \ldots\ x^1_{4m} \\
\boxed{x^1_{51}\ x^1_{52}\ \ldots\ x^1_{5m}} \\
x^1_{61}\ x^1_{62}\ \ldots\ x^1_{6m} \\
\vdots\ \ \vdots\ \ \vdots\ \ \vdots \\
x^1_{n1}\ x^1_{n2}\ \ldots\ x^1_{nm}
\end{pmatrix}
\begin{pmatrix}
\boxed{x^2_{11}\ x^2_{12}\ \ldots\ x^2_{1m}} \\
x^2_{21}\ x^2_{22}\ \ldots\ x^2_{2m} \\
\boxed{x^2_{31}\ x^2_{32}\ \ldots\ x^2_{3m}} \\
x^2_{41}\ x^2_{42}\ \ldots\ x^2_{4m} \\
\boxed{x^2_{51}\ x^2_{52}\ \ldots\ x^2_{5m}} \\
x^2_{61}\ x^2_{62}\ \ldots\ x^2_{6m} \\
\vdots\ \ \vdots\ \ \vdots\ \ \vdots \\
x^2_{n1}\ x^2_{n2}\ \ldots\ x^2_{nm}
\end{pmatrix}
$$

Figure 3: X^1 and X^2 before whole-case crossover

$$
\begin{pmatrix}
x^2_{11}\ x^2_{12}\ \ldots\ x^2_{1m} \\
x^1_{21}\ x^1_{22}\ \ldots\ x^1_{2m} \\
x^2_{31}\ x^2_{32}\ \ldots\ x^2_{3m} \\
x^1_{41}\ x^1_{42}\ \ldots\ x^1_{4m} \\
x^2_{51}\ x^2_{52}\ \ldots\ x^2_{5m} \\
x^1_{61}\ x^1_{62}\ \ldots\ x^1_{6m} \\
\vdots\ \ \vdots\ \ \vdots\ \ \vdots \\
x^1_{n1}\ x^1_{n2}\ \ldots\ x^1_{nm}
\end{pmatrix}
\begin{pmatrix}
x^1_{11}\ x^1_{12}\ \ldots\ x^1_{1m} \\
x^2_{21}\ x^2_{22}\ \ldots\ x^2_{2m} \\
x^1_{31}\ x^1_{32}\ \ldots\ x^1_{3m} \\
x^2_{41}\ x^2_{42}\ \ldots\ x^2_{4m} \\
x^1_{51}\ x^1_{52}\ \ldots\ x^1_{5m} \\
x^2_{61}\ x^2_{62}\ \ldots\ x^2_{6m} \\
\vdots\ \ \vdots\ \ \vdots\ \ \vdots \\
x^2_{n1}\ x^2_{n2}\ \ldots\ x^2_{nm}
\end{pmatrix}
$$

Figure 4: X^1 and X^2 after whole-case crossover

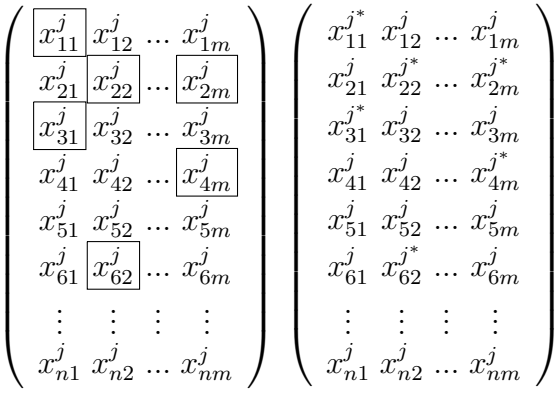

Figure 5: X^j before and after uniform mutation

3.2 Experiment design

The dataset is from the 1901 Scottish Census[9] and consists of 82,851 records. It was subsetted to of 5 variables; parish, sex, marital status, age(recoded into agegroup), and employment status. The employment variable was used as the target variable, with the other variables serving as as the key. We are comparing solutions from a GA with different level of elitism[2], followed by testing if it is possible to find a synthetic data with both of the opposite objectives acceptable. The initial population was formed by 100 datasets that are mutated from the original data, thus they are sufficiently high in utility at the beginning. The generator takes 0.1 and 0.001 as crossover and mutation rates respectively.

3.3 Experiment Results and Discussion

As an output we synthesised a single synthetic dataset which has CAP-U= 0.3964 and $U = 0.1257$. The process took 57 generations to converge to a solution.

[2]Elitism is a concept which indicates the degree to which an GA focuses on the best candidates - it is equivalent to the term "greedy" used in other algorithmic contexts

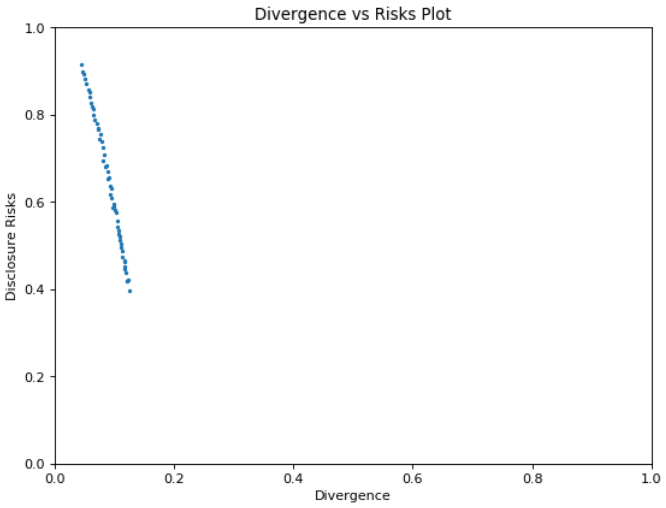

Figure 6: Changing of risk and utility from the best candidates in every generation during optimising process: 2D

3.4 Comparisons to Other Synthesis Methods

As well as the GA synthesised dataset, we also generated a CART and parametric synthetic datasets[3]. Figure 7 shows histograms comparing the counts for the five variables the the original and the synthetic datasets.

The CART synthetic dataset has DCAP= 0.4186 and the parametric synthetic dataset has DCAP= 0.3278. Given that the baseline DCAP score for the univariate is 0.415[4], the GA and parametric dataset would be considered no risk since they are below the baseline and the CART synthetic dataset would have minimal risk since it is very close to the baseline. While visually Figure 7 shows that all three synthetic datasets closely follow the univariate distributions of the original we tested this with a series of chi square tables where:

$$\begin{aligned} \mathrm{H}_0&: \text{ original=synthetic} \\ \mathrm{H}_A&: \text{ original}\neq \text{ synthetic} \end{aligned}$$

The results shown in Table 1 show that for all variables for all synthetic datasets that the null hypothesis is not rejected, implying that the synthetic dataset have the same univariate distributions for their variables.

[3]We used *synthpop* using the generic precepts with synthesising order *parish, sex, marstat, agegroup, employ.*

[4]Wherein the baseline CAP for record j is the marginal probability of its target variables estimated from the original dataset,

$$CAP_{b,j} = Pr(T_{o,j}) = \frac{1}{n}\sum_{i=1}^{n}[T_{o,i} = T_{o,j}] \tag{5}$$

Synthetic Dataset	Variable	Marital Status	Sex	Age Group	Parish	Employment
GA	Chi^2	0.2251	0.0604	1.083	1.6138	0.3425
	P-Value	0.9411	0.8059	0.9992	1.00	0.8426
CART	Chi^2	2.0766	1.4065	5.2701	14.621	2.343
	P-Value	0.7217	0.2356	0.8102	0.9822	0.3099
Parametric	Chi^2	2.4702	0.2878	7.8011	8.7484	1.0358
	P-Value	0.65	0.5916	0.5543	0.9998	0.5958
	DF	4	1	9	28	2

Table 1: Chi Square Comparison Between Original and Synthetic Datasets

We also ran two sets of alternative utility tests. Table 2 shows the propensity mean square error (pMSE) score (Woo et al, 2009) for the three synthetic datasets[5]. Table 2 also includes the standard pMSE and the pMSE ratio. Both introduced by Snoke et al (2018)[12]. The closer the pMSE is to 0 the better the data performs. In this instance all synthetic datasets have quite low pMSE scores, however the GA performs slightly better than the CART and parametric datasets. The GA does not perform as well on the standardised pMSE and pMSE ratio. According to Snoke et al, the standardised pMSE has an expectation of 0 and the pMSE ratio of 1. Hence the CART and the parametric synthetic datasets do perform better in this respect.

[5]To calculate the propensity scores we used a logistic regression consisting of $k = 45$ parameters with no interaction variables.

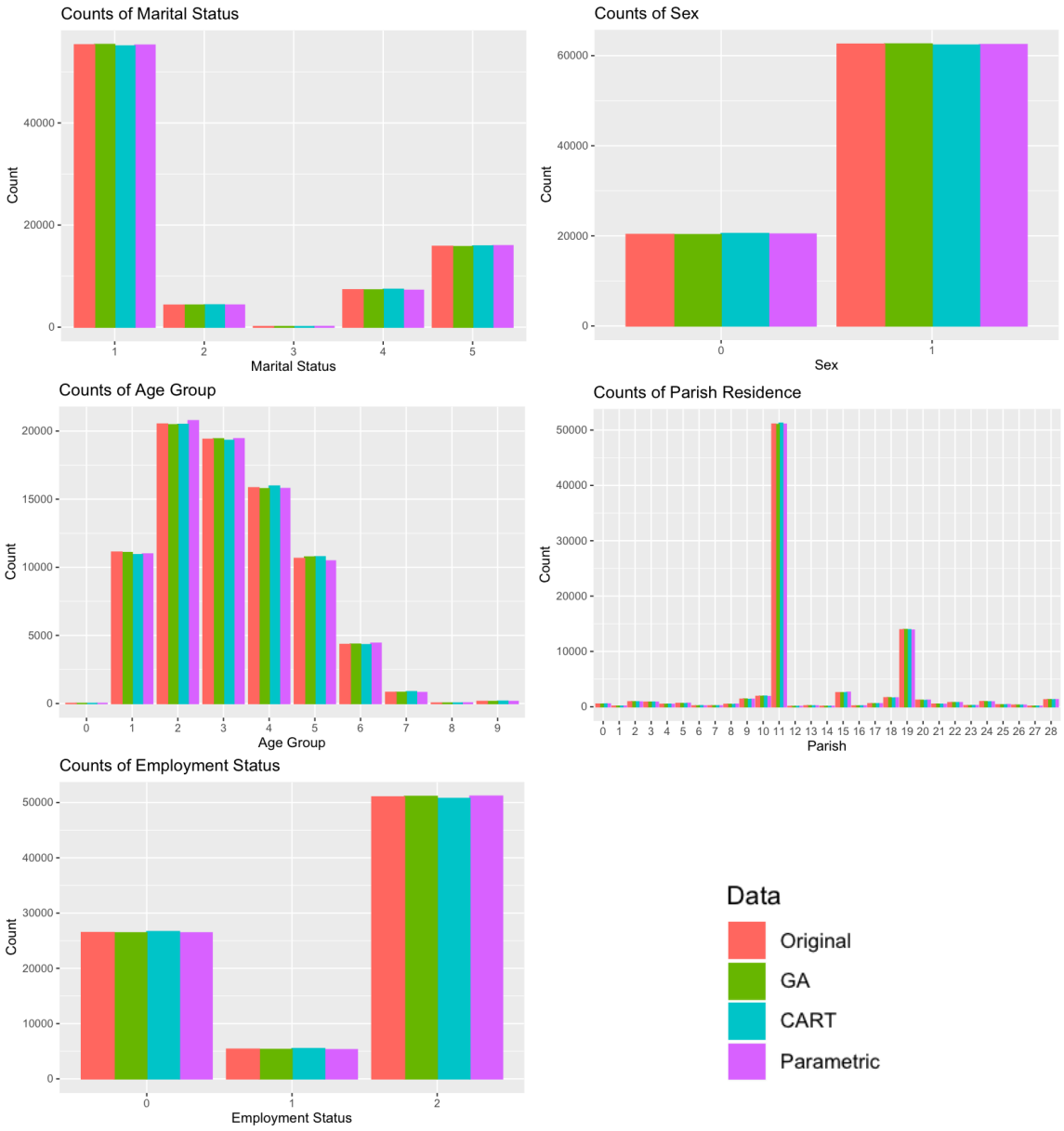

Figure 7: Histogram Comparing Original Data to GA, CART, and Parametric Synthetic Data

Synthetic Dataset	pMSE	Standardized pMSE	pMSE ratio
GA	5.44E-06	-3.9221	0.1638
CART	3.397e-05	0.1106	1.0236
Parametric	3.17e-05	7.077e-06	0.9553

Table 2: Propensity Scores for Synthetic Datasets

While the pMSE is a good broad measure of data utility, we also wanted to test the synthetic datasets in terms of narrow measures. To that end, we also ran a ratio

of estimates (ROE) (see Taub et al, 2016) over a series of bi-variate cross-tabulations, shown in Table 3. Table 3 shows that all three synthetic datasets average high ROE scores, wherein a score of 1 would be a replica of the original datasets. The GA average is slightly less than that of the CART and parametric synthetic datasets.

Variable 1	Variable 2	GA	CART	Parametric
Marital Status[6]	Sex	0.7891	0.9486	0.9454
Marital Status	Age Group	0.8582	0.9218	0.8742
Marital Status	Parish	0.9051	0.8864	0.8487
Marital Status	Employment	0.8858	0.9426	0.9469
Sex	Age Group	0.8988	0.9661	0.9031
Sex	Parish	0.9529	0.9215	0.9462
Sex	Employment	0.8931	0.9886	0.9818
Age Group	Parish	0.9262	0.8548	0.8798
Age Group	Employment	0.8665	0.8693	0.9324
Parish	Employment	0.9485	0.8985	0.9106
	Average	0.8924	0.9198	0.9169

Table 3: ROE Scores for Two Variables Cross-Tabulations

4 Conclusion

In this paper, we have reported on the use of GAs to produce synthetic data and in particular of embodying the risk utility trade-off within a single algorithm.

Our experiments indicate that GAs are viable alternative to standard synthesising methods. The GA produced synthetic data was below the baseline CAP score indicating that it had low disclosure risk. The GA synthetic data performed similarly to the CART and parametric synthetic datasets (two established forms of data synthesis) on the utility tests.

Further experiments testing different parameter settings and then implementations with larger (more realistic) datasets are needed. GA unlike previous methods for data synthesis could prove a very useful tool in that it's disclosure risk level can be pre-set, instead of being left as a post-hoc question.

References

[1] Abowd, J. M., and Lane, J. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centres, In *Privacy in statistical databases*, Springer, Berlin Heidelberg, 282-289

[2] Chen,Y., Elliot M., and Sakshaug, J. (2017). Genetic Algorithms in Matrix Representation and Its Application in Synthetic Data, UN-ECE Work Session on Statistical Data Confidentiality. Ljubljana, October 2018. `https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/2_Genetic_algorithms.pdf.Lastaccess20/12/2017`.

[3] Chen, Y.(in preparation). Genetic Algorithms and their Application to Synthetic Data Generation. PhD Thesis to be submitted to the University of Manchester. Expected submission date December 2019.

[4] Duncan, G. T., Keller-McNulty, S. A. and Stokes, S. L. (2004). Database security and confidentiality: Examining disclosure risk vs. data utility through the R-U confidentiality map, Technical report. Downloaded from https://tinyurl.com/Duncanetal04 [accessed 12/09/2019]

[5] Elliot, M. (2014). Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. [online] CMIST. Available at: https://tinyurl.com/syllsDR

[6] Konak, A., Coit, D. W., and Smith, A. E., (2006). Multi-objective optimization using genetic algorithms: A tutorial, *Reliability Engineering and System Safety*, 91(9), 992-1007.

[7] Lu, H., and Yen, G., (2002), Rank-Density-Based Multiobjective Genetic Algorithm, *Proceedings of the 2002 Congress on Evolutionary Computation 2002*, 944-949

[8] Navarro-Arribas, G. and Torra, V. (2015). Data Privacy: A Survey of Results, Advanced Research in Data Privacy, *Studies in Computational Intelligence*. Vol. 567. Springer Switzerland, 27-37

[9] National Records of Scotland, (1901), 1901 Scottish Census.

[10] Raab, G. (2018). Personal correspondence.

[11] Shlomo, N., (2010). Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility', *Journal of Privacy and Confidentiality*, vol. 2, no. 1, 73-91.

[12] Snoke, J., Raab, G., Nowok, B., Dibben, C. and Slavkovic, A. (2018). General and specific utility measures for synthetic data. Journal of the Royal Statistical Society Series A (Statistics in Society).

[13] Taub, J., Elliot, M., and Saukshaug, J. (2017) A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records. In proceedings of UNECE Statistical Data Confidentiality Work Session.

[14] Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126. pp. 122-137