Evolutionary Algorithms for Output Disclosure Risk Assessment

Mark Elliot, Claire Little and Richard Allmendinger

Abstract

This paper demonstrates a new approach for assessing the disclosure risk of analytical outputs from trusted research environments. The method involves embedding the information contained in the output into synthetic data. The embedding is achieved by turning the analytical output into an objective function in an evolutionary algorithm. Once the synthetic data have been produced the data can be assessed using standard disclosure risk metrics for microdata – in this case the correct attribution probability (CAP) measure. The results are promising but further work is needed to understand the capacity of the approach to analyse outputs at the speed required by TRE environments.

1. Introduction

The number of Trusted Research Environments (TREs) and other forms of *safe setting* has grown substantially in the face of an increasingly rich external data environment [1] and increasingly sophisticated machine learning techniques [2] both of which opens a larger range of potential attack vectors. But these same attack vectors present a secondary problem. The rationale for storing data in a safe setting is that the data themselves are not safe to release openly. But most research aims to publish analytical outputs and those outputs will – by definition – be informative about the data that has been analysed. Since the underlying data are not safe to release, we cannot assume a priori that the outputs themselves are safe and the risk that releasing those outputs presents must be assessed.

Systematically assessing the disclosure risk of analytical outputs can be a problem. Most commonly this relies on a set of heuristic rules (Rules-Based Output Statistical Disclosure Control - OSDC) and the expertise of human output checkers, and indeed the researcher whose output is being assessed (principles based OSDC); [3]. These have been developed through developing practice [4] and it is reasonable to say that the approach has been effective to date as we there have been no known attacks on statistical output.

However, laurels should not be rested on – the data environment is becoming increasingly complicated [1, 5] and potential adversaries have increasingly sophisticated tools at their disposal and both the principled and rules-based approaches have weaknesses in assessing complex and multiple outputs (and the possibility of outputs from different researchers is not checked). The principles-based approach suffers from a lack of consistency.

Some attempts have been made recently to automate the output checking process. [6, 7] have developed software to semi automate the output checking process with a resulting improvement to consistency. However, the issue of multiple outputs

remains. Fundamentally, the approach is based on a set of heuristic rules which have never been formally validated and are not directly related to underlying risk measures.

Our research proposes using the analytical output (for example, a set of model coefficients and summary statistics) as the objectives for an Evolutionary Algorithm (EA) synthesiser. That is, the synthetic data is produced without directly interrogating the original (safeguarded) data, using only the output. This will create a synthetic dataset that has the same metadata structure as the original data but with the only data structure captured in the synthesis as that represented in the analytical output. In previous work [8,9] we have examined how this methodology can be used to develop teaching datasets which could also then be utilised by researchers prior to applying to analyse the original data within the TRE. The point being that since the analytical output has been cleared for release the marginal disclosure risk of the synthetic data created from that output ought to be zero.

In the current paper we flip this rationale on its head and posit that we might use the method to more formally assess the risk of the outputs themselves. The key point is by producing microdata that captures the information in the output we can then employ disclosure risk metrics developed for use with synthetic microdata to more systematically assess the risk of that output. An additional potential benefit is that there is no algorithmic restriction on the number of outputs that can be set as objectives and therefore this can – in principle - be used to assess the cumulative risk of multiple outputs.

The remainder of this paper is structured as follows. In section 2 we provide a general description of the two methodological tools we employ here: the use of evolutionary algorithms for data synthesis and the CAP measure for assessing attribution risk (in synthetic data). Section 3 describes the case study experiment design that we have conducted to demonstrate how the approach would work in practice. Section 4 give the results of those experiments and section 5 discussed the implications of the study.

2. Background

Using evolutionary algorithms for Data Synthesis

Data synthesis for structured data [10,11] generates an artificial dataset that has the same structure and statistical properties as a real data set (usually referred to as the *original data*) but (in the case of full synthesis) not containing any of the records of the original data. Synthetic data may be used where access to the original data is not possible or restricted due to privacy constraints, or to augment (add more records to) existing datasets. In general, synthetic data is usually generated by modelling of the original dataset, however it is possible to generate synthetic data without this, using analytical output from the original dataset instead and in previous work [9,13] we have demonstrated how it is possible to use evolutionary algorithms to do this effectively and it is this approach that we utilise here.

Genetic Algorithms (GAs) [13,14] perform iterative optimisation and are a type of EA that use biologically inspired operators. There are three main operators: selection, crossover, and mutation. Broadly speaking, an initial population of candidate solutions is specified (in this case, a candidate solution is a synthetic dataset), and the fitness of the candidates is calculated. The parental selection operator is used to select candidates (parents) to reproduce for a new population, with fitter candidates more likely to be selected. A crossover operator combines some of the parents (there are a variety of methods for this) to produce new candidate solutions (children). A mutation operator then mutates some of the candidates (i.e., randomly changes some of the features). The children or a combination of children and parents form the population of the next generation (this step is called environmental selection). This process is repeated multiple times (generations), using the fitness to guide it, with ideally fitter solutions produced with each generation. Commonly, the process terminates when a specified number of generations has been produced, or a particular fitness level has been reached. GAs are flexible in that there are many parameters that can be changed or set, and the fitness function can be designed for the specific purpose. Initial Work by Chen et al. [15,16] has shown the feasibility of using GAs to generate synthetic microdata. Elliot et al. [12] used an EA to generate synthetic teaching datasets without any access to the original dataset, using only analytical output from the original dataset as an objective within the EA.

Measuring disclosure risk in synthetic data

In theory synthetic data should have low disclosure risk; the risk of re-identification is not meaningful since the synthesis process breaks the link between the data subjects and the data (i.e., synthetic data does not contain "real" records), however there is still a risk of attribution. Attribution occurs when some attribute can be associated with a data subject (such as learning from a synthetic dataset that all men aged over 85 in a particular geographical area have cancer).

The CAP methodology can be used to measure the risk of synthetic data. The Differential Correct Attribution Probability (DCAP), introduced by Elliot [17], measures the attribution risk as the probability of making a correct attribution (on a target variable), given knowledge of a set of key variables. An extension to this is the Targeted Correct Attribution Probability (TCAP) [18] which considers the case where an adversary will focus only on records in equivalence classes with an l-diversity on the target variable which is no lower than the value of the parameter Tau (which is usually set to 1). The TCAP metric is therefore the probability that those matched records yield a correct value for the target variable (i.e. that the adversary is able to make a correct attribution inference). A full description of the CAP methodology is given in Appendix A.

3. Case study design

This study presented here simulates the situation where model output is requested by the user of a TRE. Specifically, we will be assessing a logistic regression model of a

subset of the Samples of Anonymised Records from the 1991 UK census (SARS)[1]. to drive the creation of a synthetic dataset using an Evolutionary Algorithm (EA). The resulting synthetic dataset will have the same metadata structure as the original data but with the only data structure maintained in the synthesis being that represented in the regression model. That is, the synthetic dataset should be able to recreate the analytical output (the coefficients of a logistic regression model).

Data

The 1991 UK SARS Census dataset was downloaded from the UK data archive and then subset on seven areas in the West Midlands region (using the AREAP variable) and using seven explanatory variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, SEX, TENURE). TENURE (housing tenure) was used as the response variable; having been converted to a binary variable (representing whether an individual lives in an owned or rented house). 940 records with missing values for TENURE were removed leaving 50357 records. See Appendix B for a data dictionary.

The data were randomly split into training and holdout disjoint subsamples using simple random sampling.

Experiments were performed on various splits but here we report on just three.

1. Synthesising whole dataset
2. 2-folds (one training dataset which is a ~50% sample containing 25178 records and one holdout dataset also a ~50% sample containing 25179 records)
3. A two-fold-fold experiment where the one of the folds was itself split into two.

The rationale for this series of experiments was as follows. The holdout dataset serves as a baseline. If the inferences that can be made about individuals in the original dataset (from the synthetic data/output) are no better than those than inferences made from the holdout dataset, then any inferences can be determined as normal statistical outcomes rather than statistical disclosures. The problem with this insight is that the TRE owner will have no holdout dataset to make this comparison. But we can envisage that they could split the data in two and then run the analysis on one of the halves and use the other as a hold out (and then vice versa). This in principle enables them to assess the risk of an output of that sort using the original-holdout comparison averaged across the two halves. The secondary issue with that approach is that it assumes that the relationship between the half dataset original and half dataset holdout is the same (in terms of risk) as that between the full dataset and an (imaginary) full holdout. By splitting the data again in experiment three we can examine whether the relationship holds across different sizes of data (at least for this case study).

Henceforth, we refer to the training dataset as the original dataset for the rest of this document. The EA was run using the regression run on the original dataset (the hold-

[1] https://doi.org/10.5255/UKDA-SN-7210-1

out dataset was not involved and used only to provide a comparison of the results and a baseline for the TCAP score).

The analytical output of the original dataset was a binary logistic regression performed on the TENURE variable, with all remaining variables used as explanatory variables. The variables were pre-processed by one-hot encoding the multi-categorical variables. See Appendix C for the reference categories.

EA strategy

The EA was initialised using a population (size = 32) of synthetic datasets (candidates) derived from the uniform distribution of the original (training) data. The fitness of these datasets (candidates) was measured by performing the same logistic regression (using the same data pre-processing steps as performed on the original data) and comparing the resulting regression coefficients to the original coefficients.[2] The fitness was calculated at each generation. Fitness was calculated using the mean squared error (MSE) between synthetic and original coefficients.[3] The fitness score should tend towards zero, i.e., a lower fitness score is optimal (a score of zero would indicate that the logistic regression model has been recreated identically).

Preliminary exploratory experiments determined that crossover with stepped mutation produces good solutions efficiently. (Stepped mutation is where the mutation rate is divided by three every 250 generations). Parameter settings for the EA are detailed in Table 1.

Parameter	Value	Details
Number of generations	2000	
Initial population distribution	Uniform	
Population size	32	
Mutation rate	Stepped	Starting at 0.005, divide by 3 every 250 generations
Crossover	On	Probability of whether to crossover is 0.8 Crossover probability for each record is 0.1

Table 1: Parameter settings for each run of the EA

For each run the EA used information from the training dataset (regression coefficients, table values and distribution information) and did not access the holdout in any way.

For the whole dataset experiment 10 runs of the EA were performed, as there was no holdout dataset. For the 2-fold split experiments the holdout was each of the folds in turn (with the other fold forming the original dataset) and a run of the EA was

[2] The EA is programmed in Python and the package used for regression was statsmodels, details here: https://www.statsmodels.org/

[3] See Taub et al 2020 and Little et al 2025 for discussion of different utility metrics that might be used to drive the synthesiser. Where a table is the analytical output, the synthetic and original tables are compared cell-wise.

performed for each combination of folds, therefore there were two runs. Each of the runs of the EA were randomly initialised, this means that each time the process was started (initial synthetic datasets were generated) a different random seed was used, to ensure that the starting place for each run was different.

4. Results

Experiment 1: Synthesising whole original dataset

There was no holdout dataset, and the regression and univariate distributions were gathered from the whole original dataset. Ten randomly initialised runs were used. The regression coefficients for the original and the mean of ten runs of the EA are shown in Table 2.

Feature	Regression Coefficients		Absolute difference
	Original	Synthetic	
Constant	0.2498	0.2515	0.0017
AGE	0.0012	0.0013	0.0002
AREAP_58	-0.3754	-0.3753	0.0001
AREAP_59	-0.3389	-0.3389	0.0000
AREAP_60	0.3404	0.3405	0.0001
AREAP_61	-0.7305	-0.7302	0.0003
AREAP_62	0.0805	0.0807	0.0002
AREAP_63	0.1464	0.1467	0.0003
ETHGROUP_10	0.7427	0.7429	0.0002
ETHGROUP_2	0.5044	0.5045	0.0000
ETHGROUP_3	1.0247	1.0248	0.0001
ETHGROUP_4	0.9189	0.9191	0.0002
ETHGROUP_5	-1.6371	-1.6368	0.0003
ETHGROUP_6	-0.7816	-0.7815	0.0001
ETHGROUP_7	0.0191	0.0193	0.0003
ETHGROUP_8	0.5156	0.5160	0.0004
ETHGROUP_9	0.3519	0.3520	0.0001
LTILL_2	-0.6933	-0.6913	0.0020
MSTATUS_2	-0.8421	-0.8413	0.0007
MSTATUS_3	-0.3890	-0.3887	0.0003
MSTATUS_4	0.3676	0.3676	0.0001
MSTATUS_5	-0.0176	-0.0174	0.0002
SEX_2	0.0709	0.0723	0.0014
		MAE	0.0004

Table 2: Absolute differences between the regression coefficients for the original and the mean of ten runs of the EA

Below plots the MSE fitness (mean squared error of the logistic regression coefficients) for the EA for each of the ten runs:

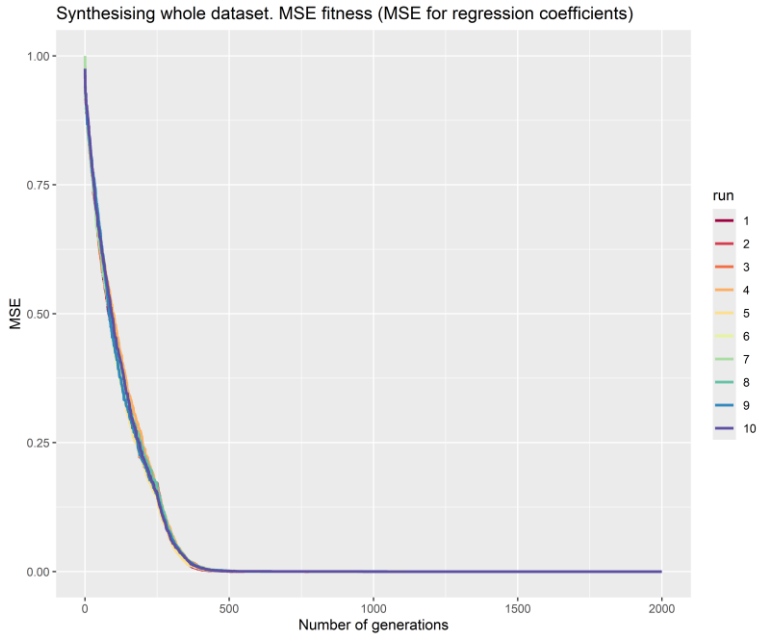

Figure 1: the MSE fitness (mean squared error of the logistic regression coefficients) for the EA for each of the ten runs.

The mean TCAP score across the ten runs was 0.4962 (using TENURE as target and all combinations of 3-6 keys) and 0.6317 (when using a single calculation of TENURE as target and all six remaining variables as the keys).

Experiment 2: a two-fold split

The data was split into 2 folds with a training dataset and a holdout. This was rotated so that each fold was a holdout once. There were therefore two runs of the EA. The MSE fitness (error of the regression coefficients) of the EA is plotted below for each of the two runs:

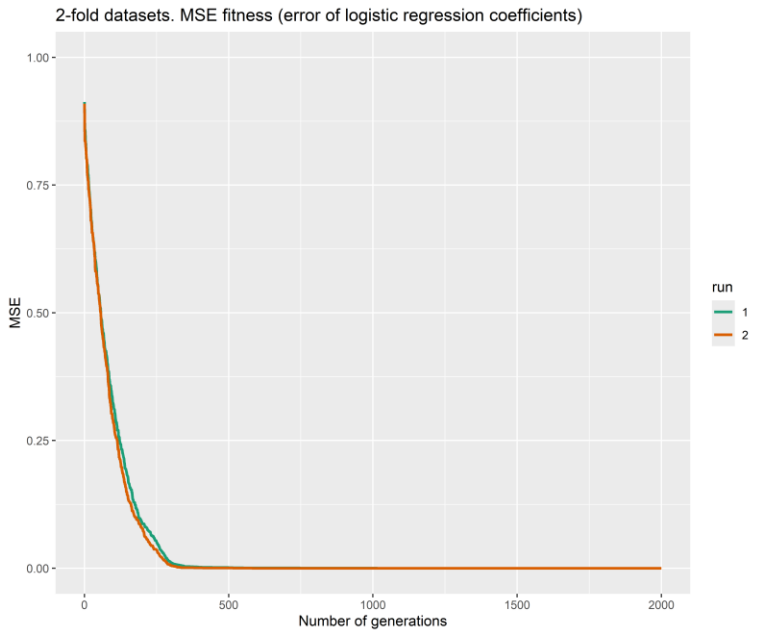

Figure 2: the MSE fitness (mean squared error of the logistic regression coefficients) for the EA for each of the two runs in the twofold split.

The regression coefficients for the original and the synthetic dataset produced by fold 1 are shown in Table 3. Similar figures were obtained for fold 2.

Feature	Regression Coefficients		Absolute difference
	Original	Synthetic	
Constant	0.2677	0.2707	0.0030
AGE	0.0016	0.0009	0.0007
AREAP_58	-0.3797	-0.3794	0.0003
AREAP_59	-0.3700	-0.3697	0.0003
AREAP_60	0.2810	0.2816	0.0006
AREAP_61	-0.7700	-0.7701	0.0001
AREAP_62	0.0248	0.0245	0.0003
AREAP_63	0.1416	0.1414	0.0002
ETHGROUP_10	0.7738	0.7742	0.0004
ETHGROUP_2	0.5627	0.5628	0.0001
ETHGROUP_3	1.2134	1.2140	0.0006
ETHGROUP_4	0.8100	0.8107	0.0007
ETHGROUP_5	-1.6291	-1.6293	0.0002
ETHGROUP_6	-0.8158	-0.8149	0.0009
ETHGROUP_7	0.1094	0.1104	0.0010
ETHGROUP_8	0.2285	0.2286	0.0001
ETHGROUP_9	0.5768	0.5760	0.0008
LTILL_2	-0.6865	-0.6840	0.0025
MSTATUS_2	-0.8830	-0.8824	0.0006
MSTATUS_3	-0.3673	-0.3672	0.0001
MSTATUS_4	0.4125	0.4130	0.0005
MSTATUS_5	0.0214	0.0219	0.0005
SEX_2	0.0462	0.0482	0.0020
MAE			0.0007

Table 3: Absolute differences between the regression coefficients for the original and the synthetic for 2-fold experiment.

The regression coefficients for the original and the synthetic dataset produced by fold 1 are shown in Table 4.

Feature	Regression Coefficients		Absolute difference
	Original	Synthetic	
Constant	0.3346	0.3347	0.0001
AGE	0.0003	0.0011	0.0008
AREAP_58	-0.4424	-0.4423	0.0001
AREAP_59	-0.3913	-0.3913	0.0000
AREAP_60	0.3211	0.3213	0.0002
AREAP_61	-0.7996	-0.7996	0.0000
AREAP_62	0.0586	0.0583	0.0003
AREAP_63	0.1922	0.1922	0.0000
ETHGROUP_10	1.1453	1.1459	0.0006
ETHGROUP_2	0.6494	0.6489	0.0005
ETHGROUP_3	1.0444	1.0445	0.0001
ETHGROUP_4	0.6424	0.6423	0.0001
ETHGROUP_5	-1.7081	-1.7079	0.0002
ETHGROUP_6	-0.9576	-0.9576	0.0000
ETHGROUP_7	-0.1042	-0.1037	0.0005
ETHGROUP_8	0.1549	0.1548	0.0001
ETHGROUP_9	1.1065	1.1061	0.0004
LTILL_2	-0.7122	-0.7119	0.0003
MSTATUS_2	-0.8900	-0.8901	0.0001
MSTATUS_3	-0.3501	-0.3503	0.0002
MSTATUS_4	0.4877	0.4881	0.0004
MSTATUS_5	0.0044	0.0044	0.0000
SEX_2	0.0591	0.0591	0.0000
MAE			0.0002

Table 4: Absolute differences between the regression coefficients for the original and the synthetic, for 2-fold-fold experiment

TCAP comparison for different fold combinations

Table 5 compares the TCAP of the synthetic to the original (TCAP O) and the synthetic to the holdout (TCAP H), for each of the different folds. (The TCAP used is the one calculated across all combinations of 3-6 keys).

	TCAP O	TCAP H	Difference
Fold 1	0.523784	0.529662	-0.00588
Fold 2	0.489204	0.466176	0.023028
Mean	0.506494	0.497919	0.008575
Standard Deviation	0.01729	0.031743	0.014453

	TCAP O	TCAP H	Difference
Fold-Fold 1.1	0.504837	0.535716	-0.03088
Fold-Fold 1.2	0.52773	0.522121	0.005609
Mean	0.516284	0.528919	-0.01264
Standard deviation	0.011447	0.006798	0.018244

Table 5: TCAP measurements for the 2-fold and 2-fold-fold experiments

The basic observation is that on average the results of the second level folds is very similar to the first level fold – both have slightly negative marginal risk scores – indicating the marginal risk is effectively zero and that the release of this output would be safe.

5. Discussion

The forgoing experiments demonstrate how the method might be used by a trusted research environment (TRE) to assess one type of statistical output. In this example, our imaginary TRE – using this method - would be happy to release the regression coefficients as the model is unable to discriminate between individuals who contributed to the underlying dataset and those from another sample drawn from the same population.

There are several issues which mean that at this stage we must be a cautious about what we claim. Firstly, we have not in any sense proved that the method works. This is a quantitative demonstration. As a minimum in future work would need to test this on a wider range of data and outputs. Secondly the synthetic data that we have produced although incredibly close to the original in the model coefficients it has

produced is not identical. Arguably therefore not all the information in the model has transferred to the synthetic data.

Figure 3 shows how TCAP evolves as the synthetic data evolve (moving from right to left - i.e. decreasing MAE). On the left-hand side of this graph are datasets from the final generations of the EA. Visually they are so close the vertical axis that their token overlaps it but they are just to the right of the axis. We can extrapolate down to MAE=0 using the regression equation and if we do that then that has a negligible effect on the risk. However, this would still be an extrapolation with all the perils that implies. We do know that there is one dataset on the vertical axis – the original data - and that has a TCAP score of 1.

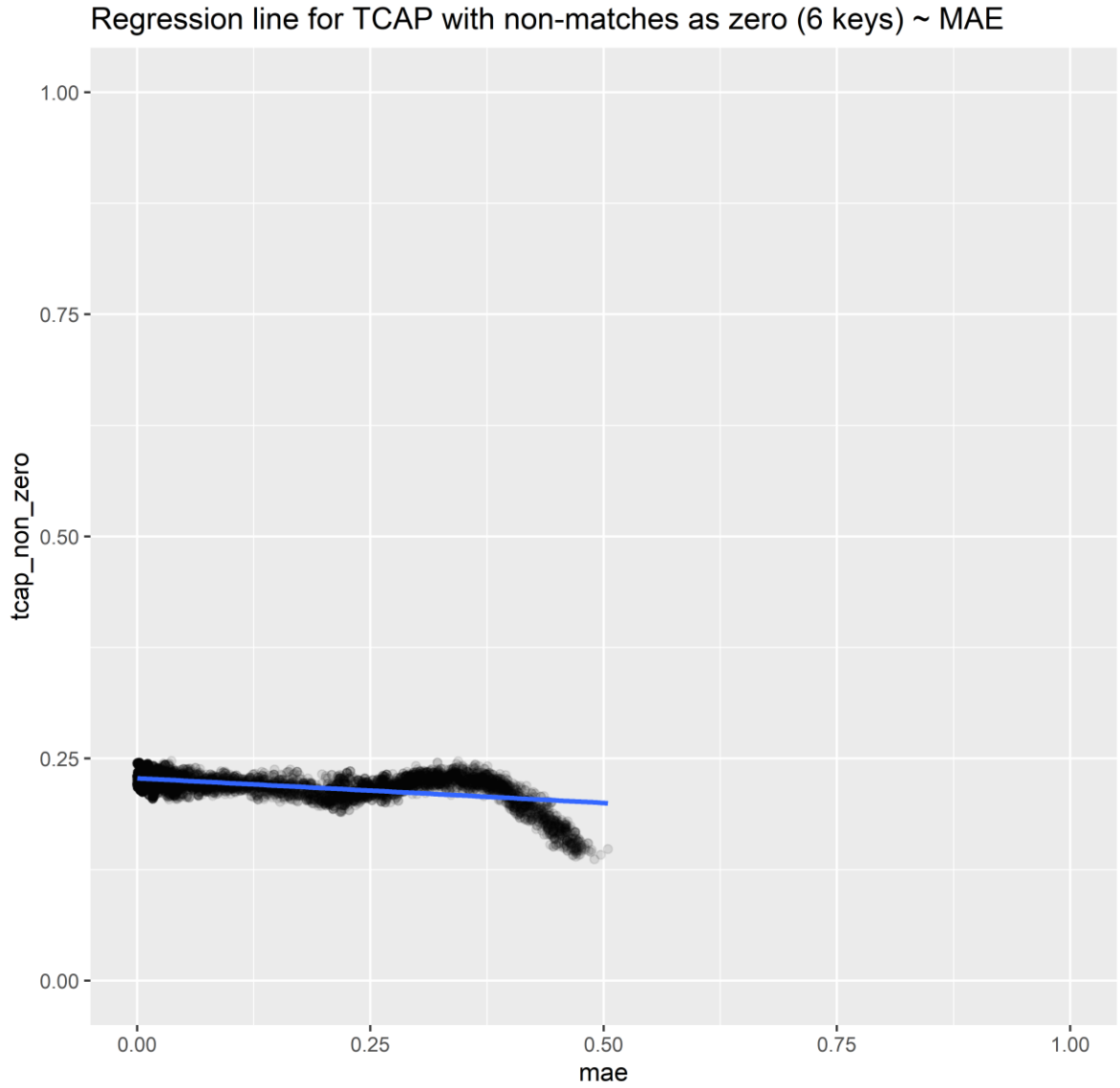

Figure 3: Plot of regression line and points for TCAP (single calculation using 6 keys) against MAE. This was performed using all points from all ten randomly initialised runs. Correlation is -0.9374 Regression line: ***TCAP = 0.2279 – 0.0556 MAE***

A key point here is the level of granularity that we are measuring the regression coefficients at an arbitrary fine level of granularity (say 100 decimal places) even the slightest change in the dataset (say of one value on one record) will be detectable in

non-zero errors. So, if we want absolute zero to be our standard then there will be only one dataset that meets that criteria – the original data. Given that we have virtually hit the vertical axis (and this does happen with all examples we have tested) any improvement in the fitness can only happen if the trajectory becomes vertical (again assuming super-fine granularity).

On the surface it might seem that we have uncovered a new peril here – surely this approach is open to an intruder as a form of reconstruction attack (see [21] for a discussion about the efficacy of these). Specifically, if an adversary runs the algorithm long enough and with enough precision on the objective function, will they find the original data? Well yes and no. Yes, in principle it's a form of reconstruction attack and in essence it is the risk of this that we are measuring. But to be clear, even for very modest sized datasets the size of the search space is unimaginably large, and the algorithm is essentially blind. So, if an adversary had an infinite amount of time, they might eventually find the original data using this method

ology, but for practical purposes with currently imaginable technology they could not.

The above applies for single outputs (that would be cleared by a TRE using traditional output checking). For these, the method is a good way to measure disclosure risk but a poor way for an adversary to attack those outputs. Or to put this another way if the method itself says the output is safe then an adversary should not be able to use the method to attack the data.

The situation is more complicated when we consider multiple outputs. To the extent that these intersect but provide different information about the data we can imagine that emergent disclosure of unreleased data structures as a potential risk. Our exploratory work on this suggests that the algorithm becomes slower to converge as the number of outputs increases and converges at a lower level of overall fitness so that emergent structure may not be computable using this type of algorithm – this remains to be explored.

Limitations and future work

One obvious limitation of the work presented is the extrapolation across different sample sizes with the fold mechanism. We have experimented with various fold splits, in particular shrinking the size of the holdout dataset. This has the advantage of moving the training data closer to the distribution of the original data at the costs of increasing measurement error in the hold out comparison. Realistically though this will only be a significant problem where the size of the original dataset is small. However, this needs extensive testing with datasets of different sizes and shapes and different types of analytical output.

The work here has focused on categorical data – but data in other forms would need to be looked at for wider adoption – particularly numerical data for which the mutation operator might need to operate differently.

There is also a further question of whether synthesisers other than evolutionary algorithms might be used. In principle any algorithm that can use analytical outputs

as an objective function for an optimiser could be used and this merits further exploration.

Conclusion

This paper has presented a methodology for formalising the output checking in Trusted Research Environments using an EA algorithm which uses the analytical output in question as an objective function. We have demonstrated that the method has promise – further work is needed to understand its capacities and limitations.

6. References

[1] Mackey, E., & Elliot, M. (2013). Understanding the data environment. *XRDS: Crossroads, The ACM Magazine for Students*, *20*(1), 36-39.

[2] Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, *10*(1), 3069.

[3] Ritchie, F., & Elliot, M. (2015). Principles-versus rules-based output statistical disclosure control in remote access environments. *IaSSIST Quarterly*, *39*(2), 5-5.

[4] Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A. and Woods, C., 2019. Handbook on statistical disclosure control for outputs. *Safe Data Access Professionals Working Group*. https://ukdataservice.ac.uk/app/uploads/sdc-handbook-v2.0.pdf

[5] Elliot, M., Mackey, E., & O'Hara, K. (2020). The anonymisation decision-making framework 2nd Edition: European practitioners' guide. Manchester, UKAN publications. https://ukanon.net/wp-content/uploads/2024/01/adf-2nd-edition-european-practitioners-guide-final-version-cover-2024-version-2.pdf

[6] Green, E., F. Ritchie, and J. Smith (2021, October). Automatic checking of research outputs (ACRO): A tool for dynamic disclosure checks. *ESS Statistical Working Papers* 2021, 1–27. doi: 10.2785/75954.

[7] Smith, J., Preen, R., Albashir, M., Ritchie, F., Green, E., Davy, S., Stokes, P. and Bacon, S., 2023. SACRO: semi-automated checking of research outputs. In *UNECE Expert Meeting on Statistical Data Confidentiality*.

[8] Elliot, M., Little, C., & Allmendinger, R. (2024). The production of bespoke synthetic teaching datasets without access to the original data. In *International Conference on Privacy in Statistical Databases* (pp. 144-157). Cham: Springer Nature Switzerland.

[9] Little, C., Elliot, M., & Allmendinger, R. (2025, October). Producing synthetic teaching datasets using evolutionary algorithms. In *UNECE Expert Meeting on Statistical Data Confidentiality: Expert meeting on Statistical Data Confidentiality*. https://unece.org/sites/default/files/2025-12/SDC2025_Sd_UnivManchester_SyntheticTeachingDatasets_Elliot_D.pdf

[10] Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

[11] Little, R. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics 9(2), 407–426. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/ statistical-analysis-of-masked-data.pdf

[12] Taub, J., Elliot, M., Pampaka, M., & Smith, D. (2018). Differential correct attribution probability for synthetic data: an exploration. In *International conference on privacy in statistical databases* (pp. 122-137). Cham: Springer International Publishing.

[13] Holland, J. H. (1992). Genetic algorithms. *Scientific American*, *267*(1), 66-73.

[14] Reeves, C. R., & Rowe, J. E. (2002). *Genetic algorithms—principles and perspectives: a guide to GA theory*. Boston, MA: Springer US.

[15] Chen, Y., M. J. Elliot, and J. W. Sakshaug (2017). Genetic algorithms in matrix representation and its application in synthetic data. In UNECE Worksession on Statistical Confidentiality. https://unece.org/fileadmin/

[16] Chen, Y., M. Elliot, and D. Smith (2018). The application of genetic algorithms to data synthesis: a comparison of three crossover methods. In Privacy in Statistical Databases. PSD 2018, pp. 160–171. Springer. DOI: 10.1007/978-3-319-99771-1_11.

[17] Elliot, M. (2023). Final report on the disclosure risk associated with synthetic data produced by the SYLLS team (2014). *https://www.cmi.manchester.ac.uk/research/publications/reports/[accessed 2020-06-27].*

[18] Garfinkel, S., Abowd, J. M., & Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Communications of the ACM*, *62*(3), 46-53.

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

[20] Hu, J., Reiter, J. P., & Wang, Q. (2014, September). Disclosure risk evaluation for fully synthetic categorical data. In *International conference on privacy in statistical databases* (pp. 185-199). Cham: Springer International Publishing.

[21] Garfinkel, S., Abowd, J. M., & Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Communications of the ACM*, *62*(3), 46-53.

Appendix

A. Measuring Disclosure Risk using TCAP

For the experiments within this paper, we refer to the attribute disclosure risk simply as the disclosure risk, since this is the main form of disclosure risk associated with synthetic data. [17] and [12] introduced a measure for the disclosure risk of synthetic data called the Correct Attribution Probability (CAP) score. The disclosure risk is calculated using an adaptation used by [19] called the *Targeted Correct Attribution Probability* (TCAP). TCAP is based on a scenario whereby we assume that an intruder has partial knowledge about a particular individual, and they wish to infer the value of a sensitive variable (the target) for that individual. Specifically, we assume that they know:

- the individual is in the original dataset (that was used to generate the synthetic data), and
- the individuals' values for some of the variables in the dataset (the keys).

These are strong assumptions, which has the benefit of then dominating most other scenarios, with the one possible exception being a membership inference attack. 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).

The TCAP measure is used because it is easily computable across all methods (since it simply compares matching records from the original and synthetic datasets), which is a benefit over other measures such as that proposed by [20] which could be computationally intractable for larger datasets and relies on the very strong assumption that the intruder knows every case but one.

Following [18], TCAP is calculated as follows: define d_o as the original data, and K_o and T_o as vectors for the key and target information, respectively $d_o = \{K_o, T_o\}$.

Likewise, $d__{s}$ is the synthetic dataset $d_s = \{K_s, T_s\}$.

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

$$WEAP_{s,j} = \Pr\left(T_{s,j} \middle| K_{s,j}\right) = \frac{\sum_{i=1}^{n}[T_{s,i} = T_{s,j}, K_{s,i} = K_{s,j}])}{\sum_{i=1}^{n}[K_{s,i} = K_{s,j}]}$$

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. Using the WEAP score the synthetic dataset is then reduced to records with a score of 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:

$$TCAP_{o,j} = \Pr\left(T_{s,j} \middle| K_{s,j}\right)_o = \frac{\sum_{i=1}^{n}[T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j}])}{\sum_{i=1}^{n}[K_{o,i} = K_{s,j}]}$$

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.

TCAP has a value between 0 and 1; a low value would indicate that the synthetic dataset carries little risk of disclosure, whereas a TCAP score close to 1 indicates a higher risk. With the same rationale as with the baseline datasets, a baseline risk value can be calculated (essentially the probability of the intruder being correct if they drew randomly from the univariate distribution of the target variable).

For each census dataset, three targets and six key variables were used, and the corresponding TCAP scores calculated for sets of 6 keys (based on the standard key variable sets produced by [5]. The overall mean of the TCAP scores was then calculated as the overall disclosure risk score. Where possible, the selected key/target variables were consistent across each country.

B. UK 1991 SARS Census data dictionary

Variable name	Description	Label	Values
AGE	age		
AREAP	Individual SAR area	57	Birmingham
		58	Coventry
		59	Dudley
		60	Sandwell
		61	Solihull
		62	Walsall
		63	Wolverhampton
ETHGROUP	Ethnic group	1	White
		2	Black Caribbean
		3	Black African
		4	Black other
		5	Indian
		6	Pakistani
		7	Bangladeshi
		8	Chinese
		9	Other-Asian
		10	Other-other
LTILL	Limiting long-term illness	1	Yes
		2	No
MSTATUS	Marital status	1	Single
		2	Married
		3	Remarried
		4	Divorced
		5	Widowed
SEX	Sex	1	Male
		2	Female
TENURE	Tenure of household space (derived into a binary variable)	0	Own house
		1	Rent house

C. Regression reference categories

Variable	Reference category	Label
AREAP	57	Birmingham
ETHGROUP	1	White
LTILL	1	Yes
MSTATUS	1	Single
SEX	1	Male